Methods in Molecular Biology 1702

Springer Protocols

Mariano Bizzarri
Editor

Systems Biology

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Systems Biology

Edited by

Mariano Bizzarri

Department of Experimental Medicine, Systems Biology Group Lab
Sapienza University of Rome, Rome, Italy

Humana Press

Editor
Mariano Bizzarri
Department of Experimental Medicine
Systems Biology Group Lab
Sapienza University of Rome
Rome, Italy

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7455-9 ISBN 978-1-4939-7456-6 (eBook)
https://doi.org/10.1007/978-1-4939-7456-6

Library of Congress Control Number: 2017957710

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media, LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Challenges and Promises of Systems Biology: Methodological Issues

There are many definitions striving to capture the hidden meaning of "Systems Biology." Indeed, despite many efforts, the concept of Systems Biology remains quite uncertain. As already highlighted [1], currently two primary streams can be recognized within Systems Biology: (1) pragmatic Systems Biology, which emphasizes the use of large-scale molecular interactions ("omic" approach), aimed at building huge signaling networks by applying mathematical modeling, and thus showing how cells make decisions based on the "information" flowing through their networks. (2) Theoretic Systems Biology which posits that the theoretical (and consequently the methodological) basis of biological study should be deeply modified [2].

Molecular Biology tries to explain the mysteries of the living being by exclusively considering it a consequence of a linear translation of the "DNA code." As originally formulated, the "central dogma" posits that "information" flows from DNA to proteins, and not the other way around [3].

However, environmental factors do change the genome, by both genetic and epigenetic mechanisms, and a number of both molecular and biophysical factors participate in shaping gene activity and cell functions [4]. Moreover, genomic functions are inherently interactive and biological processes flow along complex circuits, involving RNA, proteins, and context-dependent factors (extracellular matrix, stroma, chemical gradients, and biophysical forces [5]) within which vital processes occur [6]. As a result, no simple, one to-one correspondence between genes and phenotypes can be made [7].

Reassessment of the fundamental concepts of biological science is therefore necessary. This is happening in all fields, from genetics to cancer [8–10]. Thus, what once were heresies seem to be creeping back into mainstream biology.

Ultimately, the authors contributing to this volume do not believe Systems Biology should be considered a "simple" "gradual" extension of Molecular Biology [11], despite efforts leaning in such direction [12].

At first glance, Systems Biology can be definitely deemed as a way to rethink biology. Systems Biology is indeed more than just a "sum up" of different sciences, given that Systems Biology deals with "systems," and it is concerned with the complex, emergent properties that arise from the relationship between molecules, cells, and tissues. Functional properties are not yet in the "molecules," instead they "emerge" from a self-organized process, which shape geometrically the living structure into a system, characterized by hierarchical levels. The interaction among them leads to both top- and downward causation [13].

Systems Biology is currently committed to promoting an integration of different kinds of knowledge, not a simple collation of disciplines, but a true multidisciplinary synergy. There is no doubt that this challenging task needs a new epistemology and scientific methodology for the third millennium [14].

Therefore, we have to address some critical methodological issues, just to mention a few.

1. What kind of relationships exists among the lower levels (i.e., molecular) and the highest ones (cell, tissues, organs)? This question can be more precisely reframed as follows: how the intrinsic stochastic activity occurring at the genome level could ultimately end up into a deterministic behavior, as such we observe at the cell or tissue level [15]? Indeed, both stochastic gene expression and protein conformational noise [16] contribute in generating phenotypic heterogeneity from which the most suitable configuration can be explored by the cell to "make" appropriate decisions conferring it with remarkable phenotypic plasticity. There is no doubt that gene expression and enzymatic pathways are finely tuned by gene regulatory networks (GRN) according to nonequilibrium dynamics [17]. We can now appreciate and understand more deeply such complex behavior than a few years ago. Yet, gene and molecular activity regulation only partially rely on driving cues acting at the molecular, local level, while they are strongly modulated by cues and constraints dependent on higher levels and tightly embedded within the specific biological field [18]. Compelling evidence claims that order in living systems is mostly imposed by high-level, general constraints and forces (including electromagnetic, gravity, and cell-tissue-dependent mechano-transduced forces) [19, 20]. Definitely, phenotypic switching can result from stochastic (genetic and nongenetic) rather than by deterministic events alone (genetic), while higher order constraints altogether with the activation of specific regulatory network configuration will help in stabilizing the cell fate commitment [21]. Systems Biology tries to identify these factors by investigating the levels upon which these kinds of interactions are likely to happen, that is to say the mesoscopic level, according to the definition provided by Laughlin [22]. The search for parameters that can help in (quantitatively) describing biological process implies a double effort: (a) identification of the minimum number of "observables" required for a proper description of the system's behavior, (b) assessment of the (quantitative) relationships among variables in order to reconstruct a reliable mathematical model. That approach will likely enable us to infer previsions from data as well as to detect critical transition points.
2. Molecular biology, through a "classical" reductionist approach, taught us how some selected and "compartmentalized" biochemical processes are mechanistically linked to each other and how biochemical cascades operate within the cell. Yet, how the "parts" are integrated is still an open question. Moreover, we still do not know how those parts contribute in shaping the whole and, in turn, how the whole drives and "canalizes" biochemical pathways. This observation has a huge body of consequences and implies that we have to rethink not only the theoretical basis of biology but also our experimental methodology. The natural world consists of hierarchical levels of organization that range from subatomic particles to molecules, ecosystems, and beyond. Each level is both characterized and governed by emergent laws that do not appear at the lower levels of organization [23]. This implies that, in order to explain the features and behavior of a whole system, we require a *theory* that operates at the corresponding hierarchical level. For instance, emergent phenomena that occur at the level of the organism cannot be fully explained by theories that describe events at the level of cells or macromolecules. As forecasted by Paul Weiss: "There is no phenomenon in a living system that is not molecular, but there is none that is only molecular, either. [The molecule-based approach] ignores the relevance of

morphological form and morphogenesis, which are of salient importance for investigations on cancer, which after all is also a phenomenon of development at the tissue level" [24].

The present volume precisely focuses on methodological aspects of Systems Biology in order to provide this new theoretical approach with a robust and tailored experimental support.

As Editor, I have collected an eclectic assortment of articles. This is not a "one view fits all" approach. It is rather one to "let a hundred flowers bloom," specifically aimed to identify key methodological issues that actually challenge the reliability of "true" Systems Biology studies.

The present volume addresses many of these questions: First, by introducing a few key conceptual hallmarks required for proper modeling. Indeed, the universe of potential models for any complex system like the function of a cell has very large dimensions and, in the absence of any theory of the system, there is no guide to constrain the choice of model. The authors A. Paldi, M. Montevil, and M. Bertolaso extensively discuss the theoretical principles on which the "architecture of the model is conceived. The conceptual framework proposed is alleged to capture the insights made at different levels of cellular organization and considered previously as contradictory. It also provides a formal strategy for further experimental studies. This is a preliminary, fundamental task as "a typical research project in biology usually follows a naive inductive logic and the role of the underlying theory is usually underestimated. Concepts are usually taken for granted and rarely questioned directly. As a consequence, biology has a tendency to see methodological or technical problems even when the difficulty is conceptual" (A. Paldi). M. Bertolaso and E. Ratti aptly examine this difficulty. Accordingly, "a relational ontology is a necessary tool to ground both the conceptual and explanatory aspects of Systems Biology" (M. Bertolaso and E. Ratti). A relational ontology emphasizes the fact that even if properties that seem to be "internal" are actually relational. This is because a relational ontology assumes that the identity of the objects depends strictly on the existence of the web of relations an object is embedded within it. Therefore, "in order to understand what certain biological entities (e.g., genes, proteins) do, we need to recreate the web of relations they are usually part of (M. Bertolaso and E. Ratti)".

Sonnenschein and Soto discussed the need of a theory of organisms on which a reliable Systems Biology approach (both from the theoretical and the methodological point of view) needs to be established. As far as "theory" is needed, basic premises on which any theoretical attempt should be based are also required. The authors identified three of such very fundamental principles: the default state of the cell, a principle of variation, and one of organization. Accordingly, development as well as pathological developments (like cancer) are argued to be explained as a relational problem whereby release of the constraints created by cell interactions and the physical forces generated by cellular agency lead cells within a tissue to regain their default state of proliferation with variation and motility.

M. Montevil outlines limits and possibility of (mathematical) modeling. Indeed, it is not sufficient for a model to reproduce a process (in both its qualitative or quantitative aspects) for this model to be correct. "The validation of a model is based on the validation of a process and of the way this process takes place. As a result, it is necessary to explore the predictions of the model to verify them experimentally" (M. Montevil). Second, modeling usually entails only a specific, limited part of a complex behavior that—obviously—occurs in a tissue and in an organism. Thereby, the biological "meaning" should be mandatorily

investigated by putting that process into the whole context. A true biological understanding cannot benefit from studying parts in isolation. It is time to fill the gap we create long time ago by adopting a reductionist stance. Some hints are provided and, among others suggestions, Montevil proposes a different kind of cooperation between biologists, physicists, and mathematical modelers in order to merge the respective skills and knowledge thereof.

A proper selection among the overwhelming body of biological variables so far provided by high-throughput analytical technologies constitutes a preliminary, essential task. To recall a paradigmatic example given by the history of Physics, dynamics took off as a science only when acceleration was identified as the appropriate parameter instead of velocity. Something analogous is happening in the realm of Systems Biology where the "golden" parameter (s) has still to be unveiled.

For the search of proper parameters is at the core of biological modeling, A. Giuliani sketched a compelling survey on the field. Namely, Giuliani warns us adopting "standardized" procedures, already vindicated in nonbiological contexts. Precisely, as "system's parameter estimation in biology asks for a continuous feedback between biological and procedural information, the data analysis by no way can be considered as a 'separate optimized' set of procedures to be applied to a set of experimental results" (A. Giuliani). We have therefore to look at those networks linking the different players of the system at hand. That network can be deemed as the only relevant "causative agent" with the experimental observables acting as probes of the coordinated motion of the underlying network. However, this approach requires a completely different style of reasoning with respect to the classical approach of biologists used to a neat dependent/independent variables discrimination and considering the observables as autonomous players in the game. To "extract" such observables from the intricacy of the system, the "most fruitful way is letting the network to suggest us (e.g., by the application of unsupervised techniques like PCA) where to look avoiding the overfitting/irrelevance traps" (A. Giuliani).

Selection of proper parameters for identification of the systems is a central tenet, especially if we consider that in most cases the parameters introduced into the set of equations are completely unknown and/or only rough estimates of their values are available. The so-called parameter estimation problem is then formulated as an optimization problem where the objective is finding the parameter set so as to minimize a given cost function that relates model predictions and experimental data. R. Guzzi et al., striving to conceptualize a reliable approach to the so-called "inverse problem," acutely address this question. The inverse problem is indeed a strategy in identifying a minimal set of parameters that can describe the system under examination or to extract from the models the information embedded in the system. Since parameter estimation in dynamic models of biochemical systems is characterized by limited observability, large number of parameters and a limited amount of noisy data, the solution of the problem is in general challenging and, even when using robust and efficient optimization methods, computationally expensive. Yet, the solution of this problem is mandatory, as the parameter recognition is required to "identify" the system. As a result, the "parameterization of a subclass of dynamic systems will be called identifiable if, for any finite but sufficiently long time series of observed input-output trajectories, there exists an unique element in the subclass of systems which represents those observations" (R. Guzzi et al.). The inverse problem-based approach could greatly help in solving that conundrum.

C. Simeoni et al. as well as J.M. Nieto-Villar et al. discuss two worth noting examples of parameter identification and modeling here. According to Nieto-Villar, the phase space of cell differentiation is reconstructed mainly by adopting the formalism borrowed from the

thermodynamics of irreversible processes. Accordingly, the cancer "phenotype" is conceptualized as a self-organized nonlinear dynamical system far from thermodynamic equilibrium, as an "emergent," specific phenomenon belonging to the cluster of cell-phase transitions. Therefore, the cancer landscape is rebuilt by considering three main parameters: the production of entropy per unit time, the fractal dimension, and the tumor growth rate. The consequent mathematical model shows that cancer can self-organize in time and space far from thermodynamic equilibrium, acquiring high robustness, complexity, and adaptability. The former study from C. Simeoni focuses instead on epithelial-mesenchymal transition (EMT), a key process of cell fate specification as well of cell reprogramming. It is shown that epithelial to mesenchymal (as well as the opposite occurring during mesenchymal-epithelial transition, MET) involves an intermediate step, in which the system displays the classical feature of a "metastable state." The metastable state is instrumental for enacting EMT and for identifying the parameters that are interwoven into a framework of fast-slow dynamics for Ordinary Differential Equations. Noticeably, those parameters are "captured" by looking at the mesoscopic level, i.e., "the realm comprised between the nanometer and the micrometer, where wonderful things start to occur that severely challenge our understanding" (C. Simeoni). That is to say, at the mesoscopic level, nonlinear effects, as well as nonequilibrium processes, are more likely to be appreciated and identified. Again, as in the previous paper from Nieto-Villar, differentiating processes are described as dynamical phase transitions. Yet, a special attention is paid to evidencing the role of global cues and constraints that can be properly assessed at levels higher than the molecular one. This statement "emphasizes the intrinsic limits of studying biological phenomena on the basis of purely microscopic experiments [...] and, therefore, a multi-scale model (with some parameters derived from the microscopic analysis) is better suited from a methodological point of view."

Limits and opportunities for a general "modeling strategy" are widely discussed in the chapter by K. Selvarajoo. This chapter reports that even complex response of living cells may be described by "simple" biochemical models, based on linear and nonlinear differential equations. For linear models, the reaction topology rather than kinetics plays crucial and sensitive roles, while a more complicated picture emerges when nonlinear dynamics is considered. In fact, "for nonlinear dynamics, the parameters need to be precise or the response cannot be accurately determined due to the stability issue" (K. Selvarajoo). This means that the influence of the so-called "initial conditions" in shaping a nonlinear dynamics in no way can be neglected.

A specific application of those concepts to cancer studies is reported by S. Filippi and P. Ao. The former discusses two principal theories on carcinogenesis—the Somatic Mutation Theory (SMT) and the Tissue Organization Field Theory (TOFT)—by providing a simulation of a brain cancer cells growth in a realistic NMR imported geometry. The paper from P. Ao strives to "incorporate" both genetic and epigenetic factors to describe liver cancer development. Such endeavor is summarized in the "endogenous network hypothesis," where a core working network of hepatocellular carcinoma is depicted by means of a nonlinear dynamical approach. That model allows one to recognize two stable states within the "hepatic landscape," and those states reproduce "the main known features of normal liver and hepatocellular carcinoma at both modular and molecular levels" (P. Ao). It is worth noting that the model highlights that specific, multiscale positive feedback loops are responsible for the maintenance of normal liver and cancer, respectively. Namely, the model evidences that by inhibiting proliferation and inflammation related positive feedback

loops, and simultaneously inducing liver-specific positive feedback loops, liver cancer can be successfully antagonized, and even "reverted."

The complexity of these "loops" is extensively investigated by O. Wolkenhauer and his team, by proposing an integrative workflow to study large-scale biochemical disease networks by combining techniques from bioinformatics and systems biology. Integrating experimental and clinical data with the workflow allows vindicating specific hypotheses, namely by aiming at identifying smaller modules/molecular signatures for tumor-specific disease phenotypes. The workflow discussed herein can be applied to any large-scale biochemical network to unravel the mechanisms underlying complex biological traits and diseases.

The appraisal of specific biological processes through a Systems Biology approach needs, therefore, to capture their dynamics by considering a number of kinetic and thermodynamic parameters over wide spatial and temporal scales in order to integrate in the model the influence of nonlocal factors belonging to higher level of organization. This issue is specifically addressed by the contribution from F. Cardarelli, in which a method to probe the "diffusion law" of molecules directly from imaging is described. The method principally refers to a fluorescence fluctuation-based approach. Of note, the presented approach does not require extraction of the molecular trajectories nor the use of bright fluorophores.

S.A. Ramsay, A. Colosimo, and L. Casadei also discuss specific methodological issues. In his contribution, S.A. Ramsay describes a computational workflow for cross-species visualization and comparison of mRNA-sequence transcriptome profiling data. The workflow is based on gene set variation analysis (GSVA) and is illustrated using commands in the R programming language. In addition, a complete step-by-step procedure for the workflow using mRNA-sequence data sets is provided.

The contribution from A. Colosimo addresses a very intriguing aspect, i.e., the modulation of the collective behavior—with the emergence and spreading of synchronous activities—exerted by environmental force fields. This is a wide diffused phenomenon, even if rarely investigated. The chapter discusses the collective behavior of different kinds of populations, ranging from shape-changing cells in a *Petri dish* to functionally correlated brain areas in vivo, by means of a fruitful, unifying methodological approach, based upon a Multi-Agent Simulation (MAS) paradigm as incorporated in the NETLOGO™ interpreter.

L. Casadei et al. furnish a compelling example of the application of Systems Biology principles in planning the study of the metabolic, dynamic profile (the so-called metabolomic fingerprint) of both cell cultures and individuals. This approach is of particular interest in evidencing how a disease may influence the overall metabolic response of patients, but also can modulate key factors/pathway that could be exploited by specific treatments.

Finally, D'Avenio et al. analyze how a Systems Biology approach can benefit from quantitative morphological studies in which several shape parameters (namely the Fractal Dimension) can provide useful information about a system's behavior.

Systems Biology taught us that physiological and pathological processes are complex context-dependent entities to which our genes make a necessary but only partial contribution [25]. Yet, Systems Biology is not a simple collection of "new theoretical" principle. Rather, the conceptual premises imply a profound reconsideration of the methodological framework. We have to rethink how an experiment is planned, what kind of parameters are worthy of investigation, and how their mutual relationship should be described by means of a different mathematical modeling spanning through different space and temporal scales. That task requires an open-minded attitude and a true multidisciplinary approach. I mean

that biologists, physicists, and mathematicians should learn to work together, in a true cooperative effort for establishing a new experimental method of research.

Taken as a whole, this set of articles not only challenges some of the current methodological paradigms but also lays the groundwork for alternative approaches and in many cases takes those approaches further towards the goal of understanding living systems as complex processes, governed by both local and general control factors operating at different space and temporal scales.

References

1. O'Malley MA, Dupré J (2005) Fundamental issues in systems biology. BioEssays 27:1270–1276
2. Noble D (2008) Genes and causation. Phil Trans R Soc A 366:3001–3015
3. Crick F (1970) Central dogma of molecular biology. Nature 227:561–563
4. Goldenfeld N, Woese C (2007) Biology's next revolution. Nature 445:369
5. Bizzarri M, Pasqualato A, Cucina A, Pasta V (2013) Physical forces and non-linear dynamics mould fractal cell shape. Histol Histopathol 28(2):155–174
6. Keller EF (2000) The century of the gene. Harvard University Press, Cambridge, MA
7. Noble D (2006) The music of life. Oxford University Press, Oxford
8. Beurton PJ, Falk R, Rheinberger H-J (eds) (2008) The concept of the gene in development and evolution: historical and epistemological perspectives. Cambridge University Press, Cambridge
9. Gissis SB, Jablonka E (eds) (2011) From subtle fluids to molecular biology. MIT Press, Cambridge, MA
10. Soto AM, Sonnenschein C, Miquel PA (2008) On physicalism and downward causation in developmental and cancer biology. Acta Biotheoretica 56:257–274
11. Medina M (2013) Systems biology for molecular life sciences and its impact in biomedicine. Cell Mol Life Sci 70(6):1035–1053
12. De Backer P, De Waele D, Van Speybroeck L (2010) Ins and out of systems biology vis-à-vis molecular biology: continuation or clear cut? Acta Biotheoretica 58:15–49
13. Bizzarri M, Palombo A, Cucina A (2013) Theoretical aspects of systems biology. Prog Biophys Mol Biol 112(1–2):33–43
14. Conti F, Valerio MC, Zbilut JP, Giuliani A (2007) Will systems biology offer new holistic paradigms to life sciences? Syst Synth Biol 1:161–165
15. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467(7312):167–173
16. Xue B, Oldfield CJ, Van Y-Y, Dunker AK, Uversky VN (2012) Protein intrinsic disorder and induced pluripotent stem cells. Mol Biosyst 8(1):134–150
17. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature 453(7194):544–547
18. Tsuchiya M, Piras V, Choi S, Akira S, Tomita M, Giuliani A, Selvarajoo K (2009) Emergent genome-wide control in wild-type and genetically mutated lipopolysaccharides-stimulated macrophages. PLoS One 4(3): e4905
19. Bailly F, Longo G (2009) Biological organization and anti-entropy. J Biol Syst 17:63–96
20. Bizzarri M, Monici M, van Loon JJ (2015) How microgravity affects the biology of living systems. Biomed Res Int 2015:863075

21. Mahmoudabadi G, Rajagopalan K, Getzenberg RH, Hannenhalli S, Rangarajan G, Kulkarni P (2013) Intrinsically disordered proteins and conformational noise: implications in cancer. Cell Cycle Georget Tex 12(1):26–31
22. Laughlin RB, Pines D, Schmalian J, Stojkovic BP, Wolynes P (2000) The middle way. Proc Natl Acad Sci USA 97:32–37
23. Mazzocchi F (2008) Complexity in biology. EMBO Rep 9:10–14
24. Drack M, Wolkenhauser O (2011). System approaches of Weiss and Bertalanffy and their relevance for Systems Biology today. Semin Cancer Biol 21:150–155
25. Rossenbloich B (2011) Outline of a concept for organismic systems biology. Sem Cancer Biol 21:156–164

Contents

Contributors

PING AO • *Ministry of Education Key Laboratory of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China; Shanghai Center for Quantitative Life Sciences and Physics Department , Shanghai University, Shanghai, China; State Key Laboratory for Oncogenes and Related Genes, Shanghai Cancer Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China*

MARTA BERTOLASO • *University Campus Biomedico, Rome, Italy*

MARIANO BIZZARRI • *Department of Experimental Medicine, Systems Biology Group Lab, Sapienza University of Rome, Rome, Italy*

FRANCESCO CARDARELLI • *NEST, Scuola Normale Superiore and Istituto Nanoscienze—CNR, Pisa, Italy*

LUCA CASADEI • *Department of Chemistry, "Sapienza" University of Rome, Rome, Italy*

CHRISTIAN CHERUBINI • *Unit of Nonlinear Physics and Mathematical Modeling, Departmental Faculty of Engineering, University Campus Bio-Medico of Rome, Rome, Italy; International Center for Relativistic Astrophysics—I.C.R.A, University Campus Bio-Medico of Rome, Rome, Italy*

GERMINAL COCHO • *Instituto de Física de la UNAM, México, Mexico*

TERESA COLOMBO • *Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council, Rome, Italy*

ALFREDO COLOSIMO • *Department of Scienze Anatomiche, Istologiche, Medico Legali e dell'Apparato Locomotore, Sapienza, Università di Roma, Rome, Italy*

ALESSANDRA CUCINA • *Department of Surgery "Pietro Valdoni", Sapienza University of Rome, Rome, Italy*

GIUSEPPE D'AVENIO • *National Center of Innovative Technologies in Public Health, Istituto Superiore di Sanità, Rome, Italy*

SIMONA DINICOLA • *Department of Clinical and Molecular Medicine, Sapienza University of Rome, Rome, Italy; Department of Surgery "Pietro Valdoni", Sapienza University of Rome, Rome, Italy*

SIMONETTA FILIPPI • *Unit of Nonlinear Physics and Mathematical Modeling, Departmental Faculty of Engineering, University Campus Bio-Medico of Rome, Rome, Italy; International Center for Relativistic Astrophysics—I.C.R.A, University Campus Bio-Medico of Rome, Rome, Italy*

ALESSANDRO GIULIANI • *Department of Environment and Health, Istituto Superiore di Sanità, Rome, Italy*

MAURO GRIGIONI • *National Center of Innovative Technologies in Public Health, Istituto Superiore di Sanità, Rome, Italy*

SHAILENDRA K. GUPTA • *Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany; Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India*

RODOLFO GUZZI • *Systems Biology Group Lab, University La Sapienza, Rome, Italy*

FAIZ M. KHAN • *Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany*

Alessandro Loppini • *Unit of Nonlinear Physics and Mathematical Modeling, Departmental Faculty of Engineering, University Campus Bio-Medico of Rome, Rome, Italy*
Cesare Manetti • *Department of Environmental Biology, "Sapienza" University of Rome, Rome, Italy*
Ricardo Mansilla • *Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades, UNAM, México, Mexico*
Reynaldo Martin • *Department of Chemical-Physics, A. Alzola Group of Thermodynamics of Complex Systems M.V. Lomonosov Chemistry Chair, Faculty of Chemistry, University of Havana, Havana, Cuba*
Corrado Mascia • *Department of Mathematics, Sapienza University of Rome, Rome, Italy*
Morena La Monaca • *Project Consulting s.r.l., Rome, Italy*
Maël Montévil • *Laboratoire "Matière et Systèmes Complexes" (MSC), UMR 7057 CNRS, Université Paris, 7 Diderot, Paris Cedex 13, France; Institut d'Histoire et de Philosophie des Sciences et des Techniques (IHPST), UMR 8590, Paris, France*
Sheyla Montero • *Department of Basics Science, University of Medical Science of Havana, Havana, Cuba*
José Manuel Nieto-Villar • *Department of Chemical-Physics, A. Alzola Group of Thermodynamics of Complex Systems M.V. Lomonosov Chemistry Chair, Faculty of Chemistry, University of Havana, Havana, Cuba*
Paola Paci • *Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council, Rome, Italy*
Andras Paldi • *Ecole Pratique des Hautes Etudes, PSL Research University, UMRS_951, INSERM, Univ-Evry, Genethon, Evry, France*
Alessandro Palombo • *Department of Experimental Medicine, System Biology Group, University La Sapienza, Rome, Italy*
Stephen A. Ramsey • *Oregon State University, Corvallis, OR, USA*
Emanuele Ratti • *Center for Theology, Science and Human Flourishing, University of Notre Dame, Notre Dame, IN, USA*
Mehdi Sadeghi • *Research Institute for Fundamental Sciences (RIFS), University of Tabriz, Tabriz, Iran*
Kumar Selvarajoo • *BioTrans, Agency for Science, Technology and Research (A*STAR), Singapore, Singapore*
Chiara Simeoni • *Department of Mathematics, University of Nice Sophia Antipolis, Nice, France*
Carlos Sonnenschein • *Department of Integrative Physiology and Pathobiology, Tufts University School of Medicine, Boston, MA, USA*
Ana M. Soto • *Department of Integrative Physiology and Pathobiology, Tufts University School of Medicine, Boston, MA, USA*
Mariacristina Valerio • *Department of Chemistry, "Sapienza" University of Rome, Rome, Italy*
Garima Verma • *Department of Experimental Medicine, System Biology Group, University La Sapienza, Rome, Italy*
Gaowei Wang • *Ministry of Education Key Laboratory of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China; Department of Pathology, University of California, San Diego, La Jolla, CA, USA*
Olaf Wolkenhauer • *Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany; Chhattisgarh Swami Vivekanand Technical University, Bhilai,*

Chhattisgarh, India; Stellenbosch Institute of Advanced Study (STIAS), Wallenberg Research Centre, Stellenbosch University, Stellenbosch, South Africa

Ruoshi Yuan • *Ministry of Education Key Laboratory of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China; Department of Systems Biology, Harvard University, Boston, MA, USA*

Xiaomei Zhu • *Ministry of Education Key Laboratory of Systems Biomedicine, Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China; Shanghai Center for Quantitative Life Sciences, Shanghai University, Shanghai, China; Department of Physics, Shanghai University, Shanghai, China*

Chapter 1

Conceptual Challenges in the Theoretical Foundations of Systems Biology

Marta Bertolaso and Emanuele Ratti

Abstract

In the last decade, Systems Biology has emerged as a conceptual and explanatory alternative to reductionist-based approaches in molecular biology. However, the foundations of this new discipline need to be fleshed out more carefully. In this paper, we claim that a relational ontology is a necessary tool to ground both the conceptual and explanatory aspects of Systems Biology. A relational ontology holds that relations are prior—both conceptually and explanatory—to entities, and that in the biological realm entities are defined primarily by the context they are embedded within—and hence by the web of relations they are part of.

Key words Systems biology, Relational ontology, Relational properties, Ontological dependence

1 Introduction

Systems Biology is an approach to the understanding and conceptualization of the biological realm that emphasizes systemic and holistic aspects rather than reductionist and mereological features. However, the conceptualization of the aspects system biologists emphasize has been elusive. Here, we want to introduce some general notions that can account for the scope of Systems Biology. In particular, we will ground our understanding of Systems Biology within a *relational ontology*. A relational ontology is not simply an ontology of relations. Systems thinking, well before the rise of Systems Biology, had proposed an ontology of systems that was aimed at displacing and replacing an ontology of things and objects; these were seen as making sense only in their interconnections as parts of larger systems. Relational ontology, in our acceptation, does not consist in a further replacement of systems with relations. A relational ontology holds that relations are what maintains and dynamically transforms all we can observe; as such, relations are the way to access our observables. In this way, relational ontology involves epistemology; indeed, it is a way to combine ontology and epistemology in the same view of the world and of living

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_1,

entities. A relational ontology is not a trivial alternative to other ontologies, in fact it does not dismiss the existence of systems, parts, and wholes; rather—and very importantly—it dynamizes them, as, for example, parts can be lost, generated, acquired, and wholes can emerge, disappear, be displaced. Something very peculiar can be said for the biological realm (or, when the questions are biological), as here the very properties of objects, parts, wholes, systems—even those that classical ontology defines as "intrinsic"—are themselves relational.

The structure of the chapter is as follows. In Subheading 2, we introduce the dominant view in molecular biology, in order to understand—by contrast—the issues that should be emphasized in a relational ontology. In Subheading 3, we introduce some abstract notions forming a relational ontology. In Subheading 4, we show how evidence from the biological sciences needs to be interpreted in light of the general categories of a relational ontology.

2 Mechanisms and Biological Atomism

A good way to introduce a "relational ontology" is to start with a very different thesis. The success of molecular biology in the second half of the twentieth century [1–3] is associated with a cognitive approach toward biological systems. This cognitive approach, explicitly advocated by several prominent molecular biologists, has been emphasized by so-called *mechanical philosophers* [4, 5]. In order to understand biological systems, we should decompose them in discrete entities, analyze each part in isolation, and then compute the contribution of each part in order to obtain the behavior of the whole system; to understand the whole, we need to *see it as* the sum of its parts. What once was a methodological attitude, became later—at least in molecular biology—a metaphysical thesis, namely that the whole *is* the sum of its parts. The ontological assumption behind this view is that biological entities (genes, proteins, etc.) behave in the specific way we observe because of some intrinsic properties they possess, meaning that an object is itself because of properties that pertain to itself only, and neither because of something wholly distinct from it, nor because of some operation performed by the scientist through a specific technology. Accordingly, some biological entities will be fundamental because their contributions are necessary to make sense of bigger systems. In philosophy of biology the latter assumption is sometimes called *biological atomism*, i.e., the thesis that there are elementary units the sum of which can in principle explain life [6]. Therefore, the process of (material) abstraction—namely studying in vitro the behavior of specific entities, out of their context—is not just a compromise we have to accept to discover something out of

incomputable complexity, it is actually a *manoeuver* that allows biologists to unveil the properties of specific objects that are essential in order to understand the behavior of bigger systems. This ontological assumption also implies that there is a privileged level of organization that is causally responsible and explanatorily relevant for all the other levels. If such level of organization is responsible for the others, the direction of causation will be from this level to the others; therefore, offering an explanation at that level should, in principle, answer all questions arising at the other different levels. Sometimes, this is labeled the thesis of *bottom-up causation*, meaning that lower-levels cause the existence of higher-levels. In molecular biology, the privileged level is the one of macromolecules, mainly genes or their direct functional products (i.e., proteins).

Let us consider a concrete example. In cancer studies, *the somatic mutation theory* (SMT) exemplifies what we have mentioned above. In a nutshell, the SMT is the view that cancer is a disease related to some specific genes. Whatever we observe about high-level dynamics in cancer, it is caused by mutations of nucleotides at the genomic level. The origin of the view that the genome sequence is fundamental to explain mostly everything we observe about biology is controversial and we do not want to discuss it here. It suffices to say that, in the SMT view, somatic mutations provoke some sort of advantage to a cell, such that it proliferates faster or becomes virtually immortal. Therefore, cancer is a complex phenomenon caused by mutations of nucleotides. The level of organization where to look to explain cancer is *the genome*. For this reason, well-funded research projects today are focused on the *identification of mutations*—and in general structural variations—in the genome (*see* refs. 7, 8), because mutations are what causes cancer. Any (mechanistic) explanation of cancer will then be focused on the study of how mutated genes do what they do.

3 A Relational Ontology for Systems Biology

Several studies, especially cancer studies, show that the derived SMT intrinsic properties-based, causally-linear ontology is problematic [9]. The stochastic evolution of cancer, by definition, makes it impossible to establish direct causal relationships with specific genetic or epigenetic features. Reconstructing discrete stages is difficult, and attributing the origin of cancer to a unique intracellular molecular component or specific exogenous factor seems impossible. During the neoplastic process the molecular components are mainly unvaried, but their functional activity changes, due to internal and external factors that eventually involve multiple DNA-damaging events as well. Such change is considered dis-functional as far as it does not respond to the normal regulative factors properly (e.g., aberrant differentiation) and brings about a

change of the subsystems as well (e.g., genetic instability). From this point of view cancer can be considered a disease of the on-going systemic organization of an organism, of its natural dynamism. Parts lose their integrated functional properties and become more rigid falling into apparently functional states that mainly require a lower level of energy to be maintained.

On the one hand, the mechanistic approach has been proven useful, and it will be extremely useful in the future as well to understand how some biological processes are realized: we do not have to throw the baby out with the bathwater. On the other hand, the ontological assumptions are problematic, especially if we take them to be an ontological thesis rather than just an epistemic approach to the study of complexity [10]. The mechanistic approach actually works for some specific scientific questions related to linear pathways and interactions, or punctual events and molecular changes. But scientific shortcomings and anomalies, as well as philosophical reasons, point toward paying more attention to the causal relevance of long-range interactions between entities.

For these reasons, we now introduce a different picture, which is the picture of Systems Biology [11]. We try to be as explicit as possible in depicting the fundamentals of this view, since it is not rare to appeal to buzzword such as "systems," "holism" with no explicit definitions. We further specify that we are interested in "system-level understanding," as opposed to "system-centered view," because the former is the right term combined with a proper relational ontology, while the latter is tightly related to an ontology of systems.

The starting point of Systems Biology is based on the idea that there are no biological entities with their respective properties if not as parts of specific systems. Accordingly, trying to define what biological entities can or cannot do on the basis of their internal properties is rather hopeless. In other words, the process of abstraction of entities from their contexts in order to identify their fundamental properties can be problematic because entities could not have the properties that they have if they were not placed in a specific context. Such context dependency has been addressed and discussed in terms of "system level understanding." The idea is that a biological entity shows certain properties only if part of a specific context suggests the turn from a mechanistic ontology to a systemic one. A *relational ontology* incorporates such an assumption [12]. But what exactly does this mean? Which are the fundamental concepts of relational ontology *in abstracto*?

Appealing to a relational ontology means that the understanding of the biological realm should be characterized as a cognitive process where relations are somehow prior to entities in both explanatory and conceptual terms. By *entities* we mean a specific class of "things" (specific in the sense of "biological"), where things are subjects of the predication of properties. Here comes the

distinction between intrinsic and extrinsic properties. Without going too much into metaphysical details, we define a property *F* of *x* as an intrinsic property if *x* has *F* only in virtue of what *x* is. For example, having a specific mass is an intrinsic property, while weight does not seem so. Having a certain weight is an extrinsic property, i.e. it is a property we have because of the way we interact with the world. Relations might be considered properties [13], and in particular extrinsic properties (though this is not necessarily the case, such as in the case of biological derivation; genealogy is fundamental to the predicate of other intrinsic and extrinsic properties). But since the distinction between intrinsic and extrinsic properties does not always reflect the idea of intrinsic versus relational, we just restrict our focus to the distinction between properties that objects have in virtue of what they are (which we call *intrinsic*) and properties that entities have because of the way they interact with other entities (which we call *relational*). A relational ontology will be an ontology emphasizing the fact that even properties that seem to be *internal* are actually *relational*. A relational ontology in biology will be the recognition that the understanding and the conceptualization of biological entities relies upon the recognition and causal relevance of some relationship (genealogical, ecological, functional, etc.).

Take for instance the notion of gene [14]. A specific gene may be defined in terms of properties that seem to be *internal* (i.e., the gene *x* is a specific sequence of nucleotides), while in other contexts such as in networks biology a specific gene is defined as a node within a network [15], i.e., a gene *x* is a node within a network of interactions, defined by certain value about its connectivity (e.g., degree, clustering coefficient, etc.). However, the fact that a gene has a specific sequence, and the fact that this sequence has a certain *causal role* (i.e., being transcribed as a blueprint for a specific protein) strictly depend on the context where the gene happens to be. Therefore, even properties that seem prominently internal are somehow relational, i.e., they depend on the context.

At this point, it should be noticed that the importance of the organic and functional context—and hence of the related relational properties that any biological entity possesses—are implicitly emphasized also by traditional molecular biology. For instance, the fact that in order to study the action of a gene an in vivo validation is taken as the gold standard means that, in order to observe the real behavior of a certain biological entity, we need to re-create a context that is similar to the *wild-type*. Therefore, even old-fashioned molecular biology implicitly thought that the specific web of relations that a biological entity has in a specific context is actually very important in defining what this biological entity is. The epistemic move of molecular biologists is to put in brackets the characterization of the context, and then to identify the explanatory relevance of certain entities over time and across a number of

different contexts. Through this, they can actually infer, for instance, a causal relationship between a mutation and a cellular behavior, given a specific class of contexts. For this reason, sometimes biologists have thought that there are properties that certain biological entities have regardless of the context. However, much difficulty in cancer research has arisen at this point. There are different cases in which, for example, transplanted cells finally lack the mutated gene, while still retaining the neoplastic phenotype.

In order to better explain these aspects, let us introduce another notion that can account for the fact that even seemingly internal properties are in fact relational properties—i.e., the context determines and constrains the behavior of single biological entities. This is the notion of *ontological dependency*. As Wolff [16] rightly points out, dependency has some advantages with respect to similar notions such as supervenience and reduction. Unlike supervenience, ontological dependency is an *explanatory relation*—i.e., we explain the existence of an object x in terms of another object y instead of vaguely saying that anytime x occurs y occurs too. Unlike reduction, dependence is *not eliminative*—i.e., to say that x depends on y for its existence does not eliminate x like saying that x is *nothing but* y. In other words, "[t]o say that A ontologically depends on B is to say that both A and B exist, but that B is in some sense ontologically and explanatorily prior to A (...) A exists (at least in part) *because* B exists" (p. 618). So the claim we defend here is that the fact that biological entities only have relational properties is to be ascribed to the fact that they depend ontologically on the context they happen to be in, and that their identification and the explanation of their persistence through change are mediated by reciprocal relationships. The illusionary view that some properties of biological entities are strictly internal is based on structural similarities between contexts, implying that biological entities will behave in a similar way across several contexts.

4 Evidence from the Biological Sciences

4.1 Intrinsic Properties as Relational Properties

In Systems Biology it is now common to recognize that properties of a single gene or a single molecule emerge from the properties of huge networks and from the position of the single gene or molecule within those huge networks [17]. Palumbo et al. [18, 19] put it very explicitly: "A gene is defined as 'essential' if its deletion has lethal effects for the organism under a given experimental condition," but essentiality, while being a property of the gene, "is an emergent property of metabolic network wiring." The authors name this the "essentiality-by-location" principle. Palumbo et al. [19] demonstrated that a double mutation involving two enzymes in yeast, not essential per se, causes death of the organism *if* the double knockout provokes a "lack of alternative path" condition in

the metabolic network as a whole. Here, two concurrent non-lethal events acquire an essential property, lethality, from the existence of a global metabolism architecture, not by some deep internal "nature" of the two enzymes. In other words, their lethality would be a collective emergent property of the network system [20, 21] (*see* **Note 1**).

4.2 Relational Properties Constraining the Behavior of Parts

In cancer research, we find many examples and experimental evidence that the most important properties of a cancer cell emerge from properties that can be attributed to more inclusive wholes, such as tissues and "cells + surroundings" ensembles. The maintenance of a *status quo* in adult tissues requires that newly generated cells "adopt the appropriate fate" and contribute to the structure and function of the organ to which they belong. Such dynamic stability takes place by "dynamic and reciprocal" exchanges of information between cells and their surroundings [22, 23]. According to this model, tissues and organs are embedded in an extracellular matrix (ECM)/basement membrane (BM) that provides them structural support and contextual information along with soluble factors. In the same way, tumors exist in intimate relationship with the surrounding microenvironment, and "it is the dynamics of this heterogeneous and ever changing ecosystem that provides additional but crucial information for mutated genes to exert their function" [24]. This view of the properties of cancer cells has revolutionized conventional assumptions based on single-cell studies. Take, for example, the property of "drug resistance." Drug-resistant cells were assumed to emerge as winners in the competition after prolonged exposure to cytotoxic agents; they were thought to be the bearers of multiple mutations that fueled both tumor growth and clinical multidrug resistance. Now it is clear, from new epistemological assumptions and from empirical evidence, that the solid tumor microenvironment/architecture may in fact significantly contribute to the emergence of therapeutic resistance [12, 24]. Many scientific and philosophical articles [25] emphasize that properties of components (e.g., molecules) may be lost when such components wander away *in isolation or in other wholes.*

Another interesting example from cancer biology is the Tissue Organization Field Theory (TOFT) that considers some properties of the tissue as more fundamental than some properties of the component cells. For more than 15 years now, cancer researchers Carlos Sonnenschein and Ana Soto have been pushing their Tissue Organization Field Theory (TOFT) of carcingenesis, according to which neoplasia arises from a problem of three-dimensional organization of a tissue rather than from a normal cell gone awry by mutation or by other mechanisms [26]. For the TOFT, carcinogenesis takes place at the tissue level of biological organization: the appearance of a tumor (*see* **Note 2**) is due to chronic abnormal interactions between the mesenchyme/stroma and the parenchyma

of a given morphogenic. When the structure of a tissue is affected, cells are "disoriented" and no longer constrained, they cannot differentiate properly and revert to the default state of all cells which is proliferation (and migration). Conversely, carcinogenesis is a reversible process, whereby normal tissues (or their components) in contact with neoplastic tissues may normalize the neoplasy. The modes in which cells are organized in a tissue are thus causally and explanatorily relevant, so that in the crucial phases of cancer onset, aberrant stimuli affecting the coordination and structure of the hierarchical organization of cellular systems are sufficient and more explanatory than genetic mutations. Tissue level properties such as Fields (*see* **Note 3**), for example, are attributed causal priority over parts and held accountable for carcinogenesis and for tumor heterogeneity.

4.3 Relating (with) Developing Biological Contexts

Since the genetic turn in the 1970s, the firm goal of cancer research was the search for key mechanisms and elements (e.g., genes) that, being specific, could become the target of treatment. In cancer research, in vitro cultures were long considered sufficiently homogeneous to assume that all the units they contained were causally efficient and equivalent. In this predominance of specificity, in vitro cultures remained the privileged experimental system, to some extent favored by the long-lasting impossibility to study single cells and by the difficulties to deal with the whole organism. Then, evidence accumulated showed that cell lines, established in vitro, do not offer a suitable experimental model, as they reduce the complexity of the phenomena observed in vivo. The equivalence between cell culture results and those obtained in growing animals, which are ultimately a reiteration of the phenomenon to be understood, can often be regarded as incorrect. Also, the *context dependence* of the tumor cells phenotype forced a consideration of the relevance of some established dynamics that take over the control of the tumor cells' behavior.

It became clear that the reconstruction of the functional context of the tissue microenvironment provides a key condition for causal specificity to be studied. Progressively, contextual factors—which include long-range interactions and topological factors—were acknowledged in their role of stabilizing the structural and functional properties of molecular parts. Robustness of networks, reversibility of the effects linked to epigenetic regulation, tissue architecture, and genomic analysis gained importance; the relational conditions under which disorder in the morphostatic gradients generates the precursors of epithelial cancers in the stroma, in the absence of genetic mutations, were described also by computer simulations. The modeled organization of normal tissue and the progression of morphogenetic change linked to diffusion phenomena, show how the destruction of morphogenetic gradients is sufficient to provide the aberrant cell phenotype. The cell is freed from

gradient-based control, irrespective of the presence, or absence, of genetic mutations in cancer cells, during the initial neoplastic process. Basically, we can say that the architecture of normal tissue is a 3-D organizing system that, like morphogenetic fields, carries positional and historical information. Both association patterns and cell types change as tissues and organs are formed. In addition, the immune system was shown to play more important roles than identified genetic alterations [27–33].

The stochastic evolution of cancer, by definition, makes it impossible to establish direct causal relationships with *specific* genetic or epigenetic features. Reconstructing discrete stages is difficult, and attributing the origin of cancer to a unique intracellular molecular component or specific exogenous factor seems impossible. During the neoplastic process the molecular components are mainly unvaried, but *their functional activity* changes, due to internal and external factors that eventually involve multiple DNA-damaging events as well. Such change is considered dis-functional as far as it does not respond to the normal regulative factors properly (e.g., aberrant differentiation) and brings about a change of the subsystems as well (e.g., genetic instability). From this point of view cancer can be considered a disease of the on-going systemic organization of an organism, of its natural dynamism. Parts lose their integrated functional properties and become more rigid falling into apparently functional states that mainly require a lower level of energy to be maintained. This perspective is also interpretive of the numerous studies showing that cancer cells can return to normality when placed in a normal microenvironment and maintain their ability to undergo apparently correct differentiation, despite genetic defects [34–37]. The changes in the genome would then be causally specific only in the context of global destabilization of gene expression (*see* **Note 4**).

Ageing and cancer appear as deeply related. Some data on the role of stem cells in ageing suggest that stem cells age as a result of the alteration of processes that, over the course of life, work to prevent the onset of the neoplastic phenotype. Not only cellular factors that are inheritable through cell duplication (e.g., damage of the DNA), but also alterations in the niches that support stem cells, can contribute to the processes of ageing in mammals [38]. The results of embryonic stem cell research have deepened our understanding of the mechanisms involved in the generation and assembly of tissues and organisms, including those related to ageing and tumorigenesis (*see* **Note 5**). Some authors even started to envisage a unified theory of development, ageing, and cancer [40, 41].

Similarities between carcinogenesis and ontogenesis have paved the way for unified studies of their pathways and protein patterns. A simple example comes from studies of the WNT family proteins whose members—secreted glyco-proteins modified by covalent bonds to lipids—are involved in embryogenesis, adult tissue homeostasis, and carcinogenesis. But if tumorigenesis and

embryogenesis are similar under certain respects, important differences must also be acknowledged, as demonstrated by experiments on the differential effects of the same mutation during embryonic differentiation and neo-plastic transformation [42]. Such context-dependence of the effects of genetic mutations leads to a consideration that will be fundamental in this book: the pathologic character of tumor cells goes beyond any genetic or biochemical alteration [43]. The most noticeable difference between normal and tumor tissue lies in the imbalance between the processes of cell differentiation and proliferation, allowing tumors to produce an accumulation of aberrant undifferentiated, or partially differentiated, mitotically active cells. During embryogenesis there is, in fact, a fine balance between cell proliferation and differentiation essential for the normal development of the fetus, whereas in cancer it is precisely the balance between the two processes that is compromised as it is not brought to a successful completion [44]. Recent research on the early development of prostate cancer supports this idea [45].

5 Conclusion

In this chapter, we tried to specify the fundamental concepts grounding the perspective of Systems Biology. In particular, we have defended the view that a system-level understanding should come through a relational ontology.

In Subheading 3, we defined the main tenets of a relational ontology, namely the relational properties and ontological dependence. Relational properties are those properties that an object has only in light of the relation it has with other objects. A relational ontology emphasizes the fact that even if properties that seem to be "internal" are actually relational. This is because a relational ontology assumes that the identity of the objects depends strictly on the existence of the web of relations an object is embedded in.

In Subheading 4, we showed how evidence from biological sciences supports the adoption of the relational ontology for Systems Biology. For instance, there are specific properties of genes that, while they seem to be internal, they are eminently relational, e.g., synthetic lethality. Other examples showing that "parts" are constrained by the context they are embedded in come from cancer studies. In particular, studies in vitro about the behavior of specific genes or proteins happened to be problematic because the in vitro experimental setting simplified too much the context—and hence the web of relations—of biological entities under investigation. Therefore, in order to understand what certain biological entities (e.g., genes, proteins, etc.) do, we need to recreate the web of relations they are usually part of.

6 Notes

1. In this paper we do not commit to any specific notion of emergence. This is a topic that would deserve a paper alone.
2. Initially and fundamentally, the proponents of the TOFT aimed to explain the origin of *sporadic* cancers, i.e., those cancers that seem unrelated to specific inherited genetic mutations. Yet, the TOFT is also proposed as a unifying theory for sporadic and hereditary cancers, since TOFT authors argue that inherited genetic lesions can be explanatorily relevant *as far as* they are related to tissue organization
3. Actually, Soto and Sonnenschein, citing Gilbert, define a morphogenic field as "the *collection of cells* by whose interactions a particular organ or structure forms in the embryo" ([46] **fn. 2, emphasis added**). They thus seem to be locating a system more than defining a tissue property. Rubin [47] goes more in the direction of defining field in terms of tissue-level properties such as "increased saturation density." Still, fields are "grossly invisible, broad regions." This geometrical/geographical definition meets therapeutical aims, since identifying fields is eventually aimed at excising them along with the tumor in surgery, thereby reducing the possibility of recurrence at the site.
4. On this point of particular interest is the work done by [48]. Special attention has been also devoted to the demonstration that genetic instability itself (therefore, the accumulation of mutations) follows the onset of an abnormal microenvironment, as studies seem to demonstrate the genetic instability of stem cells, when grown without control of the microenvironment [49]. The same could happen in pre-malignant cells, after the loss of the stabilizing effects from the organization of surrounding tissue. The subsequent deregulation of the DNA maintenance pathways, generated by alteration of the microenvironment, would be sufficient to generate the defects observed in cancer cells, so mutations that inactivate specific genes involved in cell differentiation may be, more generally, a consequence of the other non-mutational mechanisms. While here I focus on the context-dependence of the *effects* of mutations (i.e., specificity) once they had occurred, the *origin* of mutations can also be considered context-dependent, as epitomized in the remark that, "It may be more correct to say that cancers beget mutations than it is to say that mutations beget cancers" [50].
5. Examples of regulatory mechanisms are micro-RNA dependent post-transcriptional regulation [39] and the epigenetic control of gene expression. Micro-RNAs (miRNAs) are small non-protein-coding RNAs that negatively regulate gene expression

at the post-transcriptional level. Stem cells express specific profiles of miRNAs that, in turn, can alter the cells' differentiation potential. As for epigenetics, the stemness state of a cell appears to be correlated with its chromatin organization state and epigenetic modifications.

References

1. Morange M (1998) A history of molecular biology. Harvard University Press, Cambridge, MA
2. Rheinberger H-J (2007) What happened to molecular biology? BIF Futura 22:218–223
3. Weinberg RA (2014) Coming full circle-from endless complexity to simplicity and back again. Cell 157(1):267–271. https://doi.org/10.1016/j.cell.2014.03.004
4. Darden L (2006) Reasoning in biological discoveries. Cambridge University Press, Cambridge, UK
5. Bechtel W, Richardson R (2010) Discovering complexity - decomposition and localization as strategies in scientific research. The MIT Press, Cambridge, MA
6. Nicholson DJ (2010) Biological atomism and cell theory. Stud Hist Philos Biol Biomed Sci 41(3):202–211. https://doi.org/10.1016/j.shpsc.2010.07.009
7. Garraway LA, Lander ES (2013) Lessons from the cancer genome. Cell 153(1):17–37. https://doi.org/10.1016/j.cell.2013.03.002
8. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW (2013) Cancer genome landscapes. Science (New York, NY) 339(6127):1546–1558. https://doi.org/10.1126/science.1235122
9. Bizzarri M, Cucina A (2016) SMT and TOFT: why and how they are opposite and incompatible paradigms. Acta Biotheor 64(3):221–239. https://doi.org/10.1007/s10441-016-9281-4
10. Nicholson DJ (2012) The concept of mechanism in biology. Stud Hist Philos Biol Biomed Sci 43(1):152–163. https://doi.org/10.1016/j.shpsc.2011.05.014
11. Bizzarri M, Palombo A, Cucina A (2013) Theoretical aspects of systems biology. Prog Biophys Mol Biol 112(1–2):33–43. https://doi.org/10.1016/j.pbiomolbio.2013.03.019
12. Bertolaso M (2016) Philosophy of cancer: a dynamic and relational view. Springer, New York
13. Esfeld M (2003) Do relations require underlying intrinsic properties?—a physical argument for a metaphysics of relations. Metaphysica 4 (1):5–25
14. Boem F, Ratti E, Andreoletti M, Boniolo G (2016) Why genes are like lemons. Stud Hist Philos Biol Biomed Sci 57:88–95. https://doi.org/10.1016/j.shpsc.2016.04.005
15. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genetics 5(2):101–113. https://doi.org/10.1038/nrg1272
16. Wolff J (2011) Do objects depend on structures? Br J Philos Sci 63(3):607–625. https://doi.org/10.1093/bjps/axr041
17. Craver CF (2016) The explanatory power of network models. Philos Sci 83(5):698–709
18. Palumbo MC et al (2005) Functional essentiality from topology features in metabolic networks: a case study in yeast. FEBS Lett 579 (21):4642–4646
19. Palumbo MC et al (2007) Essentiality is an emergent property of metabolic network wiring. FEBS Lett 581(13):2485–2489
20. Bertolaso M, Giuliani A, Filippi S (2013) The mesoscopic level and its epistemological relevance in systems biology, Recent advances in systems biology. Nova Science Publishers, Inc., New York, pp 19–36
21. Giuliani A (2010) Collective motions and specific effectors: a statistical mechanics perspective on biological regulation. BMC Genomics 11(Suppl 1):S2
22. Bissell MJ, Hall HG, Parry G (1982) How does the extracellular matrix direct gene expression? J Theor Biol 99(1):31–68
23. Bissell MJ et al (2002) The organizing principle: microenvironmental influences in the normal and malignant breast. Differentiation 70 (9–10):537–546
24. Correia AL, Bissell MJ (2012) The tumor microenvironment is a dominant force in multidrug resistance. Drug Resist Updat 15 (1–2):39–49
25. Boogerd FC et al (2005) Emergence and its place in nature: a case study of biochemical networks. Synthese 145(1):131–164

26. Sonnenschein C, Soto AM (1999) The society of cells: cancer and control of cell proliferation. Springer, New York
27. Soto AM, Sonnenschein C (2004) The somatic mutation theory of cancer: growing problems with the paradigm? BioEssays 26 (10):1097–1107
28. Jaffe L (2005) Response to paper by Henry Harris. BioEssays 27(11):1206
29. Feinberg AP, Ohlsson R, Henikoff S (2006) The epigenetic progenitor origin of human cancer. Nat Rev Genet 7(1):21–33
30. Heng HHQ et al (2006) Cancer progression by non-clonal chromosome aberrations. J Cell Biochem 98(6):1424–1435
31. Harris DP et al (2005) Regulation of IFN-gamma production by B effector 1 cells: essential roles for T-bet and the IFN-gamma receptor. J Immunol 174(11):6781–6790
32. Huang A et al (2002) Serum tryptophan decrease correlates with immune activation and impaired quality of life in colorectal cancer. Br J Cancer 86(11):1691–1696
33. Martien S, Abbadie C (2007) Acquisition of oxidative DNA damage during senescence: the first step toward carcinogenesis? Ann N Y Acad Sci 1119:51–63
34. Mintz B, Illmensee K (1975) Normal genetically mosaic mice produced from malignant tera-tocarcinoma cells. Proc Natl Acad Sci U S A 72:3585–3589
35. Hochedlinger K, Blelloch R, Brennan C, Yamada Y, Kim M, Chin L, Jaenisch R (2004) Reprogramming of a melanoma genome by nuclear transplantation. Genes Dev 18:1875–1885
36. Kenny PA, Bissell MJ (2003) Tumor reversion: correction of malignant behaviour by microenvironmental cues. Int J Cancer 107:688–695
37. Lotem J, Sachs L (2002) Epigenetics wins over genetics: induction of differentiation in tumor cells. Semin Cancer Biol 12:339–346
38. Sharpless NE, De Pinho RA (2007) How stem cells age and why this makes us grow old. Nat Rev Mol Cell Biol 8:703–713
39. Oakley EJ, Van Zant G (2007) Unraveling the complex regulation of stem cells: implications for aging and cancer. Leukemia 21:612–621
40. Finkel T, Serrano M, Blasco MA (2007) The common biology of cancer and ageing. Nature 448:767–774
41. Soto AM, Maffini MV, Sonnenschein C (2008) Neoplasia as development gone awry: the role of endocrine disruptors. Int J Androl 31 (2):288–293
42. Biava, P. M. (1999). Complexity and cancer. Leadership Medica 1. Vedi (accesso di marzo 2008)
43. Biava PM (2002) Complessità e biologia. Il cancro come patologia della comunicazione. Mondadori, Milano
44. Abbs S, Bussoli T, Kavalier F (2004) Nature encyclopaedia of the human genome. BJM 328:172
45. Marker PC (2008) Does prostate cancer co-opt the developmental program? Differentiation 76:736–744
46. Soto AM, Sonnenschein C (2011) The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory. BioEssays 33(5):332–340
47. Rubin H (2011) Fields and field cancerization: the preneoplastic origins of cancer: asymptomatic hyperplastic fields are precursors of neoplasia, and their progression to tumors can be tracked by saturation density in culture. BioEssays 33(3):224–231
48. Capp J-P (2005) Stochastic gene expression, disruption of tissue averaging effects and cancer as a disease of development. BioEssays 27 (12):1277–1285
49. Maitra A et al (2005) Genomic alterations in cultured human embryonic stem cells. Nat Genet 37(10):1099–1103
50. Prehn RT (1994) Cancers beget mutations versus mutations beget cancers. Cancer Res 54 (20):5296–5300

Chapter 2

An Integrative Approach Toward Biology, Organisms, and Cancer

Carlos Sonnenschein and Ana M. Soto

Abstract

Over the last two decades, we have challenged the hegemony of the somatic mutation theory of carcinogenesis (SMT) based on the lack of theoretical coherence of the premises adopted by its followers. We offered instead a theoretical alternative, the tissue organization field theory (TOFT), that is based on the premises that cancer is a tissue-based disease and that proliferation and motility is the default state of all cells. We went on to use a theory-neutral experimental protocol that simultaneously tested the TOFT and the SMT. The results of this test favored adopting the TOFT and rejecting the SMT. Recently, an analysis of the differences between the Physics of the inanimate and that of the living matter has led us to propose principles for the construction of a much needed theory of organisms. The three biological principles are (a) a default state, (b) a principle of variation, and (c) one of organization. The TOFT, defined as "development gone awry," fits well within the principles that we propose for a theory of organisms. This radical conceptual change opened up the possibility of anchoring mathematical modeling on genuine biological principles. By identifying constraints to the default state, multilevel biomechanical explanations become as legitimate as the molecular ones on which other modelers that adopt the SMT rely. Expanding research based on the premises of our theory of organisms will enrich a comprehensive understanding of normal development and of the one that goes awry.

Key words Developmental biology, Cancer, Somatic mutation theory, Tissue organization field theory, Constraints, Variation, Theory of organisms

1 Introduction

We scientists are part of the "world" we intend to observe, describe, and understand. This obvious fact poses a problem about the objectivity of our observations. Thus, objectivity is not a given, it has to be constructed. At this effect, scientific theories provide organizing principles and objectivity by framing observations and experiments. Among all the sciences, Physics has followed this general strategy and hence successfully managed to construct a rich set of general theories. Biology, instead, has only one general theory, that of evolution according to Charles Darwin. This is a theory about the relentless change that gave rise to a widely diverse

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_2,

variety of organisms. In this context, just two principles, namely, the generation of variation through reproduction and natural selection, provide the framework to deal with phylogeny.

Biologists, however, have yet to develop a theory of the organism that would deal with the timescale of the life cycle. They have been more adept to propose theoretical constructions during the nineteenth and in the first half of the twentieth century. In the last decades, their prevalent attitude has been to think, instead, that data are devoid of theoretical content, and that theory is unnecessary (as in "data speak by themselves"). Oftentimes, this view is accompanied by the belief that theoretical ideas borrowed from the mathematical theories of information are factual: for example, that development is a "program," that molecules contain "information," and that cells emit and receive signals. As a consequence of this distorted *Zeitgeist*, theories about particular biological phenomena, for example cancer, are kept as independent of the concept of organism. Staying with the particular subject of cancer, since the inception of the somatic mutation theory of carcinogenesis (SMT) at the beginning of the twentieth century, the growing number of lack of fit between this theory and experimental results has been met by a ceaseless list of only temporary, ineffective ad hoc fixes. We conclude that the lack of progress in areas of great biological complexity is a consequence of this theoretical paucity. To remedy this situation, we here address first some key points that illustrate the difference between the inert and the alive, then elaborate on our theory of organisms and later explain carcinogenesis from this theoretical framework.

In PART I of his chapter we offer a brief assessment of the differences between the inert and the alive and a short description of the principles for a theory of organisms, while in PART II, we address carcinogenesis within this context.

2 PART I. From the Inert to the Alive

Physical theories are grounded on stable mathematical structures that, in turn, are based on regularities such as theoretical symmetries. A physical object is both defined and understood by its mathematical transformations. These operations permit a stable description of space, a space that is objectivized as the space providing theoretical determination and which specifies the trajectory of the object (usually done by optimization principles). In sum, from this condensed analysis it can be concluded that physical objects are generic and their trajectories are specific [1, 2]. In Biology, instead, there is instability of theoretical symmetries, which are likely to change when the object is transformed with the passage of time, such as when a zygote develops into an adult animal. Thus, biological objects, i.e., organisms, are specific and hence they are

not interchangeable. Their trajectories are generic and are not specified by the phase space [2].

Moreover, organisms are the result of a history (ontogeny and phylogeny). During such a history, a cascade of changes occur as a result of which organisms acquire variability and show contextuality depending on the environment in which they live. Unlike inert objects, organisms are agents, that is, they can and will initiate actions such as proliferation and movement. Additionally, organisms not only are able to create their own rules, they also have the capacity to change them [3].

We have recently proposed a theory of organisms that deals with ontogenesis and thus complement Darwin's theory of evolution that addresses phylogenesis [4]. Our theory of organisms is based on three principles: namely, (a) the default state of all cells is proliferation with variation and motility, (b) a principle of variation, and (c) the principle of organization. These principles provide a rather comprehensive understanding of the organism's ability to create novelty and stability and to coordinate these apparent counterparts. By profoundly changing both biological observables and their determination with respect to the theoretical framework of physical theories, these principles open up the possibility of anchoring mathematical modeling in Biology. We will next expand on the background under which those principles have been proposed.

2.1 The Root of a Theory of Organisms: The Cell Theory

The cell theory developed from contributions mainly made by Dumortier, Schleiden, Schwann, Virchow, and Remak [5]. In brief, this theory claims that all organisms are made up by cells (one or many), that each cell derives from another cell, and that cells are the fundamental unit of structure and function of organisms. Georges Canguilhem recognized two main components in the cell theory, each of them dealing with a fundamental question: (1) the composition of organisms, in which the cell is the element "bearing all the characteristics of life," and (2) the genesis of organisms [6]. The role of the cell in the genesis of organisms applies to both unicellular and multicellular organisms. In the latter instance, the egg from which sexed organisms are generated is a cell, and the development of such an organism can be explained by the division of the egg into daughter cells by their proliferation. In this regard, Claude Bernard considered the cell as "a vital atom." Bernard stated "In all in-depth analysis of a physiological phenomenon, one always arrives at the same point, the same elementary irreducible agent, the organized element, the cell" [Claude Bernard Revue Scientifique, Sept 26, 1874-cited in [6]].

When considering unicellular organisms, a cell and an organism are the same entity and remain as an individual. However, individuality cannot be attributed to both the cells of the multicellular organism and the organism that contains them. In this instance, the concept of level entanglement provides a useful perspective of the

relationship between the organism and cells. This means that a zygote is both a cell and an organism, and with each cell division by the zygote and its progeny, these two levels of individuation become more obvious. In other words, following Gilbert Simondon, individuation in multicellular organisms becomes a process rather than a thing [7]. All along, the cell theory plays a unifying role between evolutionary and organismal biology because it provides a link between the uni- or the multi-cellular individual and its progeny in which the cell itself is a vehicle of inheritance. Within this theoretical perspective, the cell remains as the irreducible locus of agency.

2.2 The Founding Principles

Which is the lineage of the principles of the theory of organisms? Again, those principles are (a) the default state, (b) the principle of variation, and (c) the principle of organization**.** Each of these principles has its own history. As a consequence of work that we began in the early 1970s while studying the role of estrogens on the proliferation of their target cells, we proposed the default state in order to explain the data we were then collecting [8]. The default state is firmly rooted in the cell theory and in the strict materiality of life. Additionally, the default state is anchored on the notion that *the cell* (the original cell derived from LUCA) was an organism and is the origin of all organisms.

The joint work of Longo, Montévil, Sonnenschein, and Soto resulted in the integration of variation into the default state of proliferation and motility on the grounds that variation is generated at each cell division. In addition to this default state, a supracellular source of variation has been identified, namely, the "framing principle of non-identical iterations of morphogenetic processes in organogenesis" [9]. This type of variation accounts for the generation of mostly regular patterns of non-identical structures typically observed during organogenesis [9]. The work of Miquel, Soto, and Sonnenschein also addressed the generation of new observables, while examining the concepts of emergence, downward causation, and level entanglement [10]. In turn, the principle of variation can be traced back to Bailly and Longo's analysis of the differences between the physical and biological objects, the concept of extended criticality [11], and of course, Darwin's original idea of descent with modification. The relentless change inherent to the principle of variation points to the crucial difference between the theories of the inert and those of the living. The complementary principle of stability requires to be addressed as a main component of biological organization.

Historically, the *principle of organization* can be traced back to the concepts of autopoiesis [12], of closure [13] and of work-constraints cycles [14]. These concepts have been further elaborated by Montévil and Mossio [15]. This principle of organization is the fundamental source of biological stability. The notion of

closure of constraints as the means to achieve and maintain stability was traditionally applied to intracellular processes. Mossio et al. also explored the concept that constraints are conserved at the timescale of the process that is being constrained [15, 16]. Objectively, this concept of constraints opens a point of entry for the mathematization of biology. In fact, we modeled mammary gland morphogenesis using the notion of default state and its constraints [17].

2.3 Articulating These Principles into a Set

Our three principles are firmly anchored in the biotic world. Following Darwin' example, we consider unnecessary to delve into the transition from the prebiotic to the biotic world. By this we mean that we are agnostic about whether or not the principles that we propose to study organisms are relevant to the abiotic world; this is because even a hypothetical biochemical structure capable of instantiating closure is not an organism, and also because a self-replicating molecule is not equivalent to an organism undergoing multiplication. Our theoretical work narrowly addresses both unicellular and multicellular organisms.

In the current analysis about how the three principles we propose for our theory of organisms are related, we posit that they are irreducible to one another and none of them could be construed as the "condition of possibility" for the other two.

2.4 What Is the Role of the Default State?

Our proposal on the biological default state (proliferation with variation and motility) represents a fundamental biological postulate comparable to that of inertia in Physics. Hence, it does not require an explanation and it is implicit in the Darwinian view of evolution. What does require an explanation is the identification and mode of action of the constraints that limit the instantiation of the default state both in unicellular and in multicellular organisms. In other words, what requires an explanation is the departure from the default state, namely, proliferative quiescence, lack of motility, and restrained variation [17].

2.5 What Is the Role of Constraints?

Biological constraints and their effects are crucial targets of research in the framework of a theory of organisms. Constraints force cells out of the default state, or modify them by reducing, hindering, or canalizing their ability to proliferate and/or to move. Such an inhibitory constraint eliminates the need to use the metaphoric and anthropocentric notion of "signal" because it acknowledges the agency of cells. In other words, cells cease to be passive, inanimate things on which one has to act upon (stimulate) in order for them to proliferate or to move.

The principle of organization aims at identifying specific constraints in an organism, and thus to verify whether a given constraint is functional, namely that, together with other constraints it establishes closure. In an organism, constraints are maintained by other constraints and in turn they maintain other constraints. Given

the interdependence of the parts in an organism, it is insufficient to analyze a single constraint or a given set of constraints in isolation. Nonetheless, we obtained an insightful explanation of glandular morphogenesis by analyzing constraints on the default state in a 3D model of the breast [17]. Admittedly, additional constraints at the tissue level and organismal regulation acting via hormones should be studied for an increasingly comprehensive biological analysis.

Given that each cell division generates two similar but not identical cells, and by virtue of the default state together with the Darwinian notion of descent with modification, the principle of variation manifests itself in the default state. The principle of variation also applies at supra-cellular levels of biological organization as in the framing principle of non-identical iterations of morphogenetic processes [9]. According to the principle of variation, constraints should not be considered phylogenetic invariants. To the contrary, they are also subject to variation. For instance, a morphogenetic process that is described as a set of constraints is not necessarily conserved in a lineage. Instead, this process will be altered both for some individuals and at the level of groups of individuals, for example in a particular species. Thus, constraints are subject to change.

2.6 How Does Mathematical Modeling Fits Within the Theory of Organisms?

Symmetries and conservation laws are strictly linked and are basic principles in both Mathematics and Physics. To the contrary, in Biology, variation is crucial to both the theory of evolution and the theory of organisms that we are proposing. Mathematicians have yet to be inspired to create structures that would open the possibility of formalizing biological concepts because of the hindrance posed by the principle of variation in Biology. Highlighting the differences between inert and live objects, however, opens the way to facilitate the understanding of what would take to arrive at the development of a "mathematical biology" that would play a comparable role to that it has played in Physics. Of note, such an approach is very different from the applied mathematics transplanted directly from Physics that is routinely used to model biological phenomena [9, 18]. We favor, instead, to model biological phenomena using biological principles ([17] Montévil, this book).

3 PART II. A New Theory of Cancer, the Tissue Organization Field Theory

The tissue organization field theory (TOFT) adopts two main premises, namely, (a) cancer is a tissue-based disease akin to the process of morphogenesis during development (cancer is development gone awry) [8], and (b) proliferation with variation and motility is the default state of all cells [9, 19]. In PART I, we elaborated about (1) the premises adopted to propose a theory of

organisms, (2) the epistemological basis of the exact sciences (Physics and Mathematics), and (3) the conceptual nuances in the biological sciences dealing with the interpretation of evidence related to unicellular and multicellular organisms. We insist in considering that although theoretical principles do not require experimental observation for their formulation, they frame experimental conditions under which empirical data can reproducibly show patterns consistent with the premises adopted to frame a theory, in this case the theory of organisms [20, 21].

3.1 The Theory of Organisms, the TOFT, Organogenesis, and Modeling from Biological Principles

How, when, and where does carcinogenesis fit within the theory of organisms? The TOFT proposes that carcinogenesis, like morphogenesis, is a relational, contextual process. That is to say, teeth, hair follicles, feathers, mammary glands, lungs are formed due to reciprocal interactions between the mesenchyme and the epithelium. The relational interactions among different components of an organ cannot be reduced to discrete subcellular events [22]. In fact, morphogenesis, i.e., the generation of shape and form, is intimately dependent on physical forces generated by these cell-cell and cell-tissue interactions [23].

We proposed a model of mammary gland morphogenesis resulting from the principles outlined in PART I. Briefly, it consists of two basic components, a cellular one (epithelial cells), and a physical component (collagen-I matrix consisting of collagen fibers). As mentioned in PART I, cells are agents that move, proliferate, and generate mechanical forces that act on both the collagen fibers and their neighboring cells. As the collagen fibers get organized by the cells, they also constrain the ability of cells to move and to proliferate. We interpret this circularity in terms of a closure of constraints. Implementing this mathematical model revealed that constraints to the default state are sufficient to explain the formation of the two main components of the gland: namely, spherical structures called acini and elongated structures that branch (a ductal system). The results of this modeling effort suggest that cells also produce new constraints such as inhibitors of cell proliferation and motility. We posit that alterations of these constraints are at the root of carcinogenesis. This is consistent with reports that excess rigidity of the matrix gives rise to irregular structures unable to form a lumen, which are reminiscent of carcinoma in situ [24]. In the same vein, mammographic density, which is due to enhanced tissue rigidity, is an acknowledged risk factor for breast cancer. We next posit that the relaxation of any of these constraints in the mammary gland morphogenesis model may lead to abnormal tissue organization, which if persistent may lead to carcinogenesis.

3.2 How Does the Above Narrative Relate to Empirical Evidence?

A widely used model in carcinogenesis consists of the treatment of normal, young female rats from susceptible strains with a chemical carcinogen or a physical one, like radiation. In the following few months, all or nearly all these animals develop mammary gland adenocarcinomas. Where does the chemical or the radiation ultimately act to induce cancer? Or, in other words, which is the target of the carcinogen? In order to test whether the target of the carcinogen was either (a) any of the cells in the epithelium, as proposed by the SMT, or (b) relational, namely, the interactions between stroma and epithelium and their cells as posited by the TOFT, we adopted a theory neutral approach. Namely, we separately exposed the stroma and the epithelium of rat mammary glands to N-methylnitrosourea (NMU), a carcinogen that has a short half-life (~20 min). Once the carcinogen was "cleared" from the "exposed" group (that is, 5 days after carcinogen exposure), a series of recombinants between epithelium and stroma were performed. The recombination of exposed stroma with normal non-exposed epithelial cells resulted in adenocarcinomas, which originated in epithelial ducts. The reverse combination did not generate tumors in their hosts [25]. Subsequently, we reported the normalization of epithelial tumor cells isolated from the NMU-induced mammary carcinomas which organized as normal mammary gland ducts when injected into normal mammary gland stroma [26]. Similar outcomes were obtained from recombining a quasi-normal, non-tumorigenic mammary epithelial cell line and irradiated stroma [27], and a non-tumorigenic prostate cell line and prostate cancer-derived fibroblasts [28]. Altogether this empirical evidence was consistent with explanations of carcinogenesis advanced by the TOFT and inconsistent with those of the SMT. Moreover, these experiments invalidate the SMT and contradict the idea that cancer is irreversible as implied by the dictum "once a cancer cell always a cancer cell" [29].

3.3 Using the TOFT to Explain "Cancer Puzzles"

Next, we examined published evidence collected in the field of cancer that has been perceived as representing quirks or "cancer puzzles." This characterization was based on the difficulty in interpreting outcomes of experimental protocols that followed the genocentric approach of the SMT. Among those puzzles, it is worth recalling instances where, on the one hand, normal tissues transplanted into the "wrong" locations resulted in neoplasia while, on the other, genuine cancer tissues and their cells became normalized when placed in the midst of normal tissues (normal niches). One of the most spectacular of those puzzles is exemplified by experiments spanning 8 years, whereby Leroy Stevens, at the Jackson Laboratories in Bar Harbor, Maine, transplanted early mouse embryos into the testis of congenic mice. These embryos generated local teratocarcinomas that were subsequently transplanted for almost 200 generations from mouse to mouse. A group of

researchers under the leadership of Beatrice Mintz verified the normalization of these teratocarcinoma cells when they were placed in early blastocysts of syngeneic mice; moreover, viable offspring showed a mosaic phenotype combining tissues derived from both the host's normal cells and the grafted teratocarcinoma cells [30–32]. Also, some of these teratocarcinoma cells in mosaic male mice that ended up randomly in their testis contributed to their germ-line and formed sperm that carried the genes of these formerly teratocarcinoma cells into their own progeny. The conclusions drawn from these and comparable experiments are that a cell from a neoplasm can behave as a normal cell does, both regarding its proliferative capability (both a normal and a cell belonging to a neoplasia generate two, and only, two daughter cells) and in its ability to carry a genome that responds to cues from distant or neighboring cells and extracellular matrix as a normal cell does [33]. Thus, genuine neoplastic tissues and cells are able to generate normal cells and tissues when grafted among normal cells.

A parsimonious argument can be offered when explaining the occurrence of cancers in offspring resulting from the fusion of mutated parental gametes (sperm and/or oocytes). These neoplasms are what we have described as *inherited inborn errors of development* (Inherited IED) [34, 35]. In such offspring, all the cells in their respective morphogenetic fields carry those genomic mutations. Those mutations may occur in genes whose protein products participate in the establishment of normal morphogenetic fields, and thus, morphogenesis will be impaired and this "development gone awry" may end up forming a neoplasm that would manifest postnatally as an organ malformation or a tumor or both. Examples of these rare Inherited IED are the Li-Fraumeni syndrome, retinoblastoma, BRCA 1 and 2-linked breast and ovarian tumors, the Lynch syndrome, and other syndromes that represent less than 2% of all clinical tumors. Obviously, carriers of these germ-line mutations have all their cells mutated and thus the morphogenetic field as a whole reflects the underlying defect in these syndromes. In these instances, mutations become "proximate" causes of the malformations and/or tumors.

Separately, the other subgroup of *induced inborn errors of development* can be generated when carcinogens (such as environmental endocrine disruptors, viral or radiation exposure, etc.) affect embryos during organogenesis [34]. The evidence already collected in this field is consistent with the notion that in addition to the above-referred documented instances of *inherited* and *induced inborn errors of development* (Induced IED), a percentage of *sporadic* cancers (about 98% of all clinical cancers) may have been initiated in the womb [36, 37]. Altogether, regardless of whether these neoplasms are due to germ-line mutations or the deleterious effects of carcinogens in utero can be attributed to the underlying process of "development gone awry" [19].

Phenotypes of epithelial cells are susceptible of being manipulated experimentally by changing the niche (epithelium/stroma) in which they originally land or are placed. In addition to the examples cited above from B. Mintz and her group, those of Barcellos-Hoff and Ravani [27] and ours [26], others have strengthened this concept. For instance, when mouse mammary tumor virus (MMTV)-"neu-induced" tumor cells mixed with normal mammary mouse epithelial cells were inoculated into cleared mammary fat pads (stroma), these cells became normalized and formed normal ducts together with normal epithelial cells [38]. In addition, these "tumor cells" became normal luminal, myoepithelial, and secretory mammary epithelial cells. Thus, a normal mammary gland microenvironment, comprised of stromal, epithelial, and host-mediated constraints, may combine to suppress the cancer phenotype during glandular tissue regeneration.

4 Conclusions

Twenty years ago we were puzzled by the lack of theoretical coherence in the fields of control of cell proliferation and cancer. After having proposed that proliferation and motility is the default state of all cells, including those in metazoans, we extended our theoretical exploration to carcinogenesis that led us to propose the TOFT. We used a theory-neutral experimental protocol that simultaneously tested the TOFT and the SMT. The results of this test favored adopting the TOFT and rejecting the SMT.

When using the three principles we proposed, namely, (a) a default state, (b) a principle of variation, and (c) one of organization, we have argued that carcinogenesis can be explained as a relational problem; that means that the release of the constraints created by cell interactions and the physical forces generated by cellular agency lead cells within a tissue to regain their default state of proliferation with variation and motility. Ultimately, carcinogenesis, defined as "development gone awry," now fits well with the principles we propose for a theory of organisms.

This radical conceptual change opened up the possibility of anchoring mathematical modeling on genuinely biological principles. Turing identified an epistemological gap between modelization and imitation [39, 40]. While the former is based on a theory about the object being modeled, the latter is not. Thus, an analysis of the differences between the Physics of inanimate and that of the living matter has led us to propose principles for the construction of a much needed theory of organisms. In addition to this theoretical purpose, these founding principles have been useful for framing experiments and mathematical modeling. Finally, biological principles are needed to move beyond imitation. In this regard, the model of ductal morphogenesis referred to above is based on the

basic principles of default state and the intrinsic constraints generated by the epithelial cells. By identifying constraints to the default state, multilevel biomechanical explanations become as legitimate as the molecular ones on which other modelers rely. Expanding research based on the premises of our theory of organisms will enrich a comprehensive understanding of normal development and of the one that goes awry.

Acknowledgments

This work was conducted as part of the research project "Addressing biological organization in the post-genomic era" which is supported by the International Blaise Pascal Chairs, Region Ile de France (AMS: Pascal Chair 2013). Additional support was provided by Award Number R01ES08314 (P.I. AMS) from the U. S. National Institute of Environmental Health Sciences. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors are grateful to Cheryl Schaeberle for her critical input. The authors have no competing financial interests to declare.

References

1. Bailly F, Longo G (2011) Mathematics and natural sciences: the physical singularity of life. Imperial College Press, London
2. Longo G, Montévil M (2014) Perspectives on organisms: biological time, symmetries and singularities. Springer, Berlin
3. Canguilhem G (1991) The normal and the pathological. Zone Books, New York
4. Darwin C (1859) On the origin of species. Clowes and Sons, London
5. Harris H (1995) The cells of the body: a history of somatic cell genetics. Cold Spring Harbor laboratory press. Plainview, NY
6. Canguilhem G (2008) Knowledge of life. Fordham University Press, New York
7. Miquel PA, Hwang SY (2016) From physical to biological individuation. Prog Biophys Mol Biol 122:51–57
8. Sonnenschein C, Soto AM (1999) The society of cells: cancer and control of cell proliferation. Springer Verlag, New York
9. Longo G, Montévil M, Sonnenschein C, Soto AM (2015) In search of principles for a theory of organisms. J Biosci 40:955–968
10. Soto AM, Sonnenschein C, Miquel P-A (2008) On physicalism and downward causation in developmental and cancer biology. Acta Biotheor 56:257–274
11. Longo G, Montévil M (2011) From physics to biology by extending criticality and symmetry breakings. Prog Biophys Mol Biol 106:340–347
12. Varela FG, Maturana HR, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. Biosystems 5:187–196
13. Rosen R (1991) Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life. Columbia University Press, New York
14. Kauffman S (2002) Investigations. Oxford University Press, New York
15. Montévil M, Mossio M (2015) Biological organisation as closure of constraints. J Theor Biol 372:179–191
16. Mossio M, Montévil M, Longo G (2016) Theoretical principles for biology: organization. Prog Biophys Mol Biol 122:24–35
17. Montévil M, Speroni L, Sonnenschein C, Soto AM (2016) Modeling mammary organogenesis from biological first principles: cells and their physical constraints. Prog Biophys Mol Biol 122:58–69
18. Soto AM, Longo G, Miquel PA, Montévil M, Mossio M, Perret N, Pocheville A,

Sonnenschein C (2016) Toward a theory of organisms: three founding principles in search of a useful integration. Prog Biophys Mol Biol 122:77–82

19. Soto AM, Sonnenschein C (2011) The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory. BioEssays 33:332–340
20. Soto AM, Sonnenschein C (1984) Mechanism of estrogen action on cellular proliferation: evidence for indirect and negative control on cloned breast tumor cells. Biochem Biophys Res Commun 122:1097–1103
21. Soto AM, Longo G, Montévil M, Sonnenschein C (2016) The biological default state of cell proliferation with variation and motility, a fundamental principle for a theory of organisms. Prog Biophys Mol Biol 122:16–23
22. Gilbert SF, Sarkar S (2000) Embracing complexity: Organicism for the 21st century. Dev Dyn 219:1–9
23. Shyer AE, Tallinen T, Nerukar NL, Wei Z, Gil ES, Kaplan DL, Tabin CJ, Mahadevan L (2013) Villification: how the gut gets its villi. Science 342:212–218
24. Paszek MJ, Weaver VM (2004) The tension mounts: mechanics meets morphogenesis and malignancy. J Mammary Gland Biol Neoplasia 9:325–342
25. Maffini MV, Soto AM, Calabro JM, Ucci AA, Sonnenschein C (2004) The stroma as a crucial target in rat mammary gland carcinogenesis. J Cell Sci 117:1495–1502
26. Maffini MV, Calabro JM, Soto AM, Sonnenschein C (2005) Stromal regulation of neoplastic development: age-dependent normalization of neoplastic mammary cells by mammary stroma. Am J Pathol 167:1405–1410
27. Barcellos-Hoff MH, Ravani SA (2000) Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells. Cancer Res 60: 1254–1260
28. Barclay WW, Woodruff RD, Hall MC, Cramer SD (2005) A system for studying epithelial-stromal interactions reveals distinct inductive abilities of stromal cells from benign prostatic hyperplasia and prostate cancer. Endocrinology 146:13–18
29. Pierce GB, Shikes R, Fink LM (1978) Cancer: a problem of developmental biology. Prentice-Hall, Englewoods Cliffs, NJ
30. Mintz B, Ilmensee K (1975) Normal genetically mosaic mice produced from malignant teratocarcinoma cells. Proc Natl Acad Sci U S A 72:3585–3589
31. Illmensee K, Mintz B (1976) Totipotency and normal differentiation of single teratocarcinoma cell cloned by injection into blastocysts. Proc Natl Acad Sci U S A 73:549–553
32. Stewart TA, Mintz B (1981) Successful generations of mice produced from an established culture line of euploid teratocarcinoma cells. Proc Natl Acad Sci U S A 78:6314–6318
33. Sonnenschein C, Soto AM (2011) The death of the cancer cell. Cancer Res 71:4334–4337
34. Sonnenschein C, Davis B, Soto AM (2014) A novel pathogenic classification of cancers. Cancer Cell Int 14:113–117
35. Sonnenschein C, Soto AM (2015) Cancer metastasis: so close and so far. J Natl Cancer Inst 107. https://doi.org/10.1093/jnci/djv236
36. Soto AM, Sonnenschein C (2010) Environmental causes of cancer: endocrine disruptors as carcinogens. Nat Rev Endocrinol 6:363–370
37. Paulose T, Speroni L, Sonnenschein C, Soto AM (2015) Estrogens in the wrong place at the wrong time: fetal BPA exposure and mammary cancer. Reprod Toxicol 54:58–65
38. Booth BW, Boulanger CA, Anderson LH, Smith GH (2011) The normal mammary microenvironment suppresses the tumorigenic phenotype of mouse mammary tumor virus-neu-transformed mammary tumor cells. Oncogene 30:679–689
39. Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460
40. Turing AM (1952) The chemical basis of morphogenesis. Philos Trans R Soc Lond B Biol Sci 237:37–72

Chapter 3

Conceptual Challenges of the Systemic Approach in Understanding Cell Differentiation

Andras Paldi

Abstract

The cells of a multicellular organism are derived from a single zygote and genetically identical. Yet, they are phenotypically very different. This difference is the result of a process commonly called cell differentiation. How the phenotypic diversity emerges during ontogenesis or regeneration is a central and intensely studied but still unresolved issue in biology. Cell biology is facing conceptual challenges that are frequently confused with methodological difficulties. How to define a cell type? What stability or change means in the context of cell differentiation and how to deal with the ubiquitous molecular variations seen in the living cells? What are the driving forces of the change? We propose to reframe the problem of cell differentiation in a systemic way by incorporating different theoretical approaches. The new conceptual framework is able to capture the insights made at different levels of cellular organization and considered previously as contradictory. It also provides a formal strategy for further experimental studies.

Key words Conceptual framework, Cell differentiation, Stochastic fluctuations, Metabolism, Epigenetic mechanisms, Chromatin

1 Introduction

Biology is an empirical science; nothing makes sense in the eyes of a biologist unless it is derived from experimental observations. A typical research project in biology usually follows a naive inductive logic and the role of the underlying theory is usually underestimated. Concepts are usually taken for granted and rarely questioned directly. As a consequence, biology has a tendency to see methodological or technical problems even when the difficulty is conceptual. History of science taught us that paradigm shifts and breakthroughs in a field usually require a theoretical re-foundation or conceptual reframing of the major issues. Concepts in biology, as well as in any other scientific discipline, must be revised periodically, upgraded, or replaced if necessary. The last years witnessed emerging discussions on some of the fundamental concepts in cell biology such as "cell identity" or "cell fate." These discussions were

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_3,

made necessary, among others, by the rapid evolution of techniques with single-cell resolution. Perhaps this is the reason why very often they are "hidden" behind methodological issues. In addition to their high resolution, these methods are very efficient in collecting astronomic amounts of morphological, physiological, or molecular data. Contrary to the expectation of many, accumulation of data alone did not provide us with the understanding of why and how cells become phenotypically different (i.e., differentiate) and what are the driving forces of this process. There is a growing feeling that the difficulties to address cell differentiation result from the inappropriate conceptual framing of the problem. Where these difficulties come from? What are the possible ways to resolve them? The present reflection aims to show that it is possible to define a new theoretical approach and to rethink cell differentiation on the basis of our present knowledge.

2 The Quest for Classification

Traditionally, cell differentiation is defined as the phenotypic transformation of the cells from one type into another. Therefore, the first challenge is the definition of the concept of "cell type." Textbooks and reviews usually claim that there are about 200 different cell types in the human organism. They are classified according to their morphological similarity, tissue location, function or patterns of gene expression, etc. This is easy to do when very different cells are compared; there is no difficulty to assign neurons, lymphocytes, or epithelial cells for example to different categories. But how to classify closely related cells such as those separated only by a few cell divisions? Such cells usually resemble to each other, yet they may display a broad spectrum of gene expression levels, physiological or morphological traits, etc. How to decide if they belong to the same or to a different category? The following quote taken from a recent paper illustrates the difficulties: "... should these subcategories be declared distinct cell types? What differences, be they functional, regulatory, or morphological, are sufficient to define an organism's cellular taxonomy?" [1]. The same problem is sometimes formulated as an issue of "cell identity" and cell differentiation as a process of change of identity [2]. Others are debating whether various cell types such as stem cells represent a state or entity [3, 4]. "Identity" or "entity" are concepts borrowed from philosophical ontology. The debates on them are as old as our systematic thinking about the world and can be tracked back to Plato and Aristotle. The cell biological re-formulation of the fundamental ontological question is: do cell types exist as independent entities? If so, what are the essential distinguishing features of cell types? Unfortunately, the issue of cell identity is usually treated in a simple intuitive way. Although experimental biology is not expected to

provide a solution for fundamental philosophical issues, but understanding the origin and true significance of the concepts directly imported from philosophy would stimulate their constructive conceptual framing of the cell type issue.

The intensive quest for a better classification has been triggered by the rapid development of single-cell resolution techniques, hence the illusion of a technological difficulty [5]. Earlier, biochemical or molecular methods used to characterize gene expression, protein levels, or other features needed hundreds, thousands, or even more cells and were able to provide us with population averages only. Single-cell resolution techniques are able to extract similar information from a large number of individual cells. For the first time, in addition to the average we have also a reliable measure of the variability in the population. It is not surprising that the number of recognized cell types increases steadily with the resolution of these techniques. In a recent study for example, using single-cell RNA sequencing 17 different categories of CD34+ hematopoietic cells were identified on the basis of their gene expression patterns versus only two categories when a less sensitive cytometry analysis was used [6]. There are numerous similar examples [1, 7, 8]. In general, highly sensitive single-cell resolution techniques show that even very closely related cells are different to some extent with respect to their gene expression patterns. Although two different populations of cells are easy to discriminate on the basis of the average expression level of some distinctive marker genes, it is usually difficult to assign an individual cell picked up randomly to one of these defined cell types on the basis of the single-cell gene expression profile. Although counterintuitive at a first glance, "cell type" appears as a concept that describes groups rather than individual cells. Then, how to set the limit between "irrelevant" and "important" differences between two cells? As we could learn from philosophy, there is no simple solution to this problem and perhaps the best way is to get rid definitively of these controversial concepts. The existing pragmatic solutions measure the extent of the differences without discriminating what is relevant or irrelevant. The most frequently employed strategy is based on the collection of a large number of parameters on individual cells using single-cell RT-PCR, RNA sequencing, mass cytometry, or high-throughput image analysis of the cell morphology [1, 6, 7, 9–14]. The data obtained are analyzed using multiparametric classification algorithms that group cells in categories on the basis of their phenotypic "similarity." In this context, "similarity" between the cells is calculated as a function of the distance between the cells in a multidimensional space defined by the measured parameters. The cell phenotype is represented as a location in a multidimensional parameter space. If the measured parameters are the gene expression levels, as it is frequently the case, the number of dimensions is equal to the number of genes in the genome, and their

expression levels are the coordinates that determine the exact position. Statistical analysis of the distances between these positions representing the cell phenotypes gives an estimate of the probability (p value) that the groups identified by the classification algorithm can be obtained by chance. If the probability for this is sufficiently low, one can accept that the cells assigned to different phenotypic groups are indeed different. The advantage of these methods is that they provide an accurate measure of the phenotypic differences between the cells on the basis of clearly defined criteria (mRNA level, protein abundance, etc.) that can be used to classify the cells. This is an important methodological step, yet it does not give answer to the original question. Assigning a cell to a given cell type remains a decision of the observer, who sets the list of parameters to be considered, the threshold p-value, sample size, etc. This makes the classification relative, highly dependent on the experimental context, choice of the statistical methods, and, importantly, on the subjective opinion of the investigator. Clustering algorithms are incorrectly assumed to provide objective judgment on phenotypic classification and became a standard procedure for the analysis of single-cell data. Nevertheless, while subjective, the use of the chosen classification method makes the different experiments quantitatively comparable. Therefore, they open the way to testing hypotheses on the mechanisms of cell differentiation [15] without clearly defining what a cell type is.

The single-cell data confirm that every gene expression combination is not equiprobable. Some gene expression profiles and the corresponding cellular phenotypes are more frequent, hence probably more stable than others. Genes interact with each other and this can be described as a complex network where edges represent interactions between the vertices formed by the genes. The network representation of gene-gene interactions led to the proposition that frequently observed gene expression profiles or the corresponding cellular phenotypes reveal states of the gene interaction network that are close to an attractor in the multidimensional parameter space [16, 17]. These attractors emerge as a result of mutually stabilizing interactions between a set of genes making their co-expression more frequent among the possible combinations. In the attractor interpretation therefore, a cell phenotype is a state of the gene interaction network that is more or less close to an attractor and cell differentiation is a process of transition between the attractors [17, 18]. This representation makes direct reference to the now classical “epigenetic landscape” metaphor proposed by Conrad Waddington almost 70 years ago [17, 19]. The attractor concept of cell phenotype circumvents the “continuous versus discrete” dilemma of cell classification and focuses on the temporal dynamics of the phenotypic change.

3 Temporal Dynamics, Stability, and Change

Single-cell studies uncovered another important aspect of the cellular phenotypes: the expression of the genes in a cell and, consequently, the phenotypes are fluctuating continuously. As a result, even cells in a clonal population exhibit a broad distribution of various traits. Stochasticity of gene expression was suggested and experimentally detected long time ago [20, 21], but the phenomenon gained a significant interest only after the publication of a landmark paper in 2002 [22]. Variation of gene expression is the direct consequence of the stochasticity of biochemical reactions involving molecules present in small copy-numbers in the cell. For example, in a typical eukaryotic cell there are only two copies of each gene. Transcription factors, RNA polymerase molecules, and other components of the gene expression machinery are also present in very low concentration. Under these conditions, biochemical reactions are limited by the diffusion of the molecules and occur only when the participating molecules meet by chance. In the case of gene expression involving many different partners this leads to strong fluctuations at a time scale of minutes to hours comparable with the life cycle of the cells. These fluctuations, frequently called "noise," ubiquitous and are unavoidable because they are caused by the very nature of the biochemical reactions [23]. Therefore, stochastic fluctuations rather than stability should be considered the default state of gene expression.

This transforms radically the way we have to consider the problem of gene expression changes during differentiation. Traditionally, gene expression is supposed to be stable. Changes during differentiation are supposed to be strictly controlled, inducing regulated transition of the cell between phenotypic states. Spontaneous gene expression fluctuations have no role in the process, they are just "noise." However, measured and characterized experimentally, we know now that the extent of the gene expression "noise" in individual cells is comparable to the variations supposed to be regulated [24, 25]. Population level measurements provide us only with average values of the gene expression levels; they show population level tendencies and hide the individual variations. Yet, not populations, but individual cells differentiate. How to reconcile then the unstable nature of almost every characteristic of individual cells with their obvious capacity to maintain a stable phenotype? How to explain that these phenotypes can change in an orderly way? Until now the phenotype and the underlying gene expression pattern were supposed to be stable and the *explanandum* was the "change." In the new conceptual frame, stability becomes the *explanandum*. The question to be addressed now is how a naturally fluctuating living cell can be maintained in stability. This is just the

opposite of the traditional deterministic view, which has dominated biology until now.

Obviously, change and stability are a pair of complementary concepts that also raise the question of "continuous versus discrete." A slow "change" can be seen as "stability" depending on the timescale of the observer. Averaging over a sufficiently long interval of time can filter smaller fluctuations and reveal the tendency of an individual cell to conserve or change its phenotype, gene expression levels, protein abundances, etc. The key question is what is a "sufficiently long" time interval? A pragmatic approach to this question is to take the characteristic timescale of the random fluctuations as a starting point. Purely stochastic gene expression changes occur at a characteristic timescale of minutes to hours. If the timescale of the fluctuations is longer than the cell's lifecycle, the phenotype is usually considered stable because the daughter cells remain phenotypically close to the mother cell [26]. Slow fluctuations can therefore be seen as to reflect a kind of "memory", that makes the actual phenotype state of the cell remaining close to the previous one. From this point of view, one extreme is "no change" (full stability), where the past state is identical to the present one. This is only a theoretical possibility. It is opposed to the other extreme, also theoretical, represented by random fluctuations without memory, where no prediction of the present state from the past is possible. Real cells are never fully stable, nor they are fully ergodic. Since the whole problem is further complicated by the fact that the cells divide and usually transmit their phenotype to the daughter cells, the candidate mechanisms for slowing down natural fluctuations and stabilizing cellular phenotypes are also expected to remain active during and after cell divisions.

4 Energy for Stability

We have learned from physics that maintaining order and stability in an open system is principally a matter of energy investment. Indeed, theoretical models and experimental verification have demonstrated that the energetic costs of the noise reduction are very high and of the same order of magnitude as the cell's capacity to produce energy [27]. Consequently, the cell has no capacity to suppress molecular fluctuations such as gene expression noise and their consequences; at best it can reduce them to some extent. The putative mechanisms must be functionally dependent on and limited by the energy-producing cellular processes.

Gene expression is a "birth and death" process. Birth is a multistep process involving transcription and translation and all the steps of maturation of the mRNA-s and proteins. "Death" is also a multistep process involving the degradation of the intermediate or the final gene products. The actual level of the gene

product in the cell is determined by the rate of the synthesis and degradation [28]. Simultaneous high synthesis and degradation rates can produce similar levels as low synthesis/degradation rates. Obviously, fluctuations of the gene product concentration can also be caused by the fluctuations in both the rates of synthesis and degradation. All these processes require ATP or some other form of energy-carrying substrates. Some steps are known to contribute more to the fluctuation/stabilization process than others.

Mechanisms that dissipate chemical energy generated by the metabolism to modulate gene expression fluctuations are now well known. Molecular mechanisms known as "epigenetic modifications" of the chromatin are excellent candidates for the role of the "stabilizer" of phenotype through influencing the fluctuations of the "birth" rate. Chromatin is a macromolecular structure formed by the genomic DNA associated to proteins, essentially histones. When wrapped in the chromatin, DNA is not accessible for transcription. Transcription is only possible if the chromatin dissociates from the DNA. This is a typical stability problem. Each chromatin component carries several covalent modifications, such as acetylation, methylation, phosphorylation, poly-ADP-ribosylation, etc. that determine the overall stability of the structure. The biochemical reactions that introduce or remove these modifications are catalyzed by dedicated enzymes. The reactions form a cooperative network that brings together either a stable repressive chromatin structure (heterochromatin), which makes the DNA inaccessible to the transcriptional machinery, or an open structure (euchromatin) that allows transcription. Thanks to the cooperative nature of the reactions and despite the reversibility and very short half-life of each individual modification, both the structures can stably be maintained for a long period of time. This is a dynamic, steady-state stability resulting from the equilibrium of the permanent action of the modifying and the reverse reactions and the resulting rapid dissociation-association of the corresponding chromatin proteins [29, 30]. As a result, the chromatin around a gene is either open, allowing transcription or repressed, making transcription impossible [31]. The structure is constantly adjusted depending on the dynamic equilibrium of the "on" and "off" reactions. It has been shown that the chromatin behaves as a dynamic bistable system with hysteresis [32]. The transition between the active and repressed states of a gene is switch-like. It depends on the competition of the heterochromatin- or euchromatin-generating reaction networks and on the time spent in the previous state. A heterochromatin structure formed long time ago is more difficult to reverse than a recently generated.

Whether a gene becomes silenced or accessible for transcription is *in fine* determined by the dynamic equilibrium between the processes bringing together the permissive and repressive chromatin and on the pre-existing state of the chromatin. When accessible,

transcription factors can selectively bind to the DNA in a sequence-dependent way and further bias the equilibrium toward the open state usually at the sites of transcription initiation. In this way, transcription factors contribute to the stabilization of the gene expression networks as proposed by the attractor concept of cell phenotypes. However, they cannot specifically activate a repressed gene without a prior transition of the chromatin to an at least partially open state, because the DNA is simply not accessible for binding. This is an essential point, because it means that a silenced gene can be re-activated only after or concomitantly with the change of the chromatin structure. Transcription factors alone are not sufficient to activate a gene; they can only increase the probability of the transcription initiation of the accessible genes.

Another essential distinguishing feature of the chromatin is the capacity to "record" the previous activity of the genes due to the hysteretic dynamical properties. In this way, chromatin becomes a major component of the so-called cellular memory because it can conserve its structure over mitotic and in rare cases even over meiotic divisions. This property confers the cells the capacity to differentiate in an orderly way instead of switching irregularly between the possible phenotypes.

In summary, chromatin is a highly dynamic key player able to slow down the stochastic fluctuations of the transcription, essentially by its reversible repression. The way the chromatin structure is brought together by the epigenetic modification confers to it a memory function. When repressed by the heterochromatin, a gene cannot be transcribed. There is no mRNA production, hence no fluctuations. When accessible for transcription, the RNA synthesis is subject to stochastic effects resulting in a bursting production of mRNA molecules and generating stochastic fluctuations in their number. These fluctuations can be amplified or buffered by the consecutive steps of translation and degradation of the gene products and in this way they contribute to the overall fluctuations of the cellular phenotype.

5 Energy for Change

Stability of a dynamic system requires energy that compensates for the continuous stochastic fluctuations. However, changing a dynamic equilibrium into another, a gene expression profile into another is also energy dependent. Activating repressed genes and repressing active ones is achieved by the cell through changing the chromatin around these genes. The transition between the repressive and permissive configurations depends essentially on the epigenetic modifications. The dynamic nature of these modifications implies that both the maintenance of the chromatin structure and the transition between the different forms are energy-dissipating

processes. Indeed, the substrates used for epigenetic modification are all small molecular intermediates of the core energy metabolic pathways (for comprehensive reviews, *see* ref. 33. I apologize for not citing original papers). For example, acetylation of the histones and many other nuclear proteins is achieved using acetyl-CoA as substrate. Acetyl-CoA is probably one of the most important hubs in the metabolic network of the cell. It is directly generated from pyruvate, the end product of the glycolysis. Acetyl-CoA is either converted into citrate in the first step of the Kebs-cycle or used as a starting point for the biosynthesis of lipids and indirectly of almost any other types of macromolecules in the cell. The levels of Acetyl-CoA fluctuate widely depending on the metabolic flux and directly influence the level of acetylation of the chromatin components in the nucleus. The same is true for all other epigenetic modifications. Methylation is dependent on S-adenosyl-methyonin, a methyl donor synthesized from methionine, an essential amino acid and ATP. Demethylation reactions use a-ketoglutarate, a key Krebs cycle intermediate, poly-ADP ribosylation is dependent on NAD+ as a substrate, phosphorylation requires ATP, etc. The direct substrate level metabolic link between energy production and chromatin structure is more than obvious. In general, the rate of enzymatic reactions is essentially dependent on the substrate concentration. The intracellular concentration of the key metabolic substrates is indeed a major determinant of the epigenetic reaction rates [34].

Energy production depends on a network of red-ox reactions. The concentration of the intermediate metabolites and final high-energy-carrying molecules, in turn, is determined by the flux and activity of the whole metabolic network, which is itself dependent on the nature and availability of electron donors and acceptors. Electron donors are essentially nutriments taken up from the cellular environment and to lesser extent the cell's own reserves. The electron transfer between them is a multistep process and involves intermediate electron transporters (NAD, NADP, FAD). These electron transporters provide electrons to all other electron transfer reactions including biosynthesis. In the presence of a sufficient external carbon source as an electron donor and oxygen, the oxidation into H_2O and CO_2 will be dominant and the ATP production and concentration of reduced electron transporters, as NAD+ and NADP+ will be high. When oxygen is not available, glycolysis will dominate, biosynthesis will eliminate the oxidation by O_2 as a final electron acceptor. The concentration Acetyl-CoA and Krebs-cycle intermediates will be relatively high. Therefore, the nature of metabolic regimes and the transition between them can modulate the concentration of key metabolic substrates for epigenetic reactions. This, in turn, increases or decreases the rate of the corresponding epigenetic reactions and, as a corollary, modulates the frequency and amplitude of the gene expression fluctuations.

The direct dependence of the stabilizing mechanisms on the energy-producing metabolic flux also implies that the cellular environment can impact the rapidity and extent of gene expression fluctuations. In fact, the metabolic flux in the cell is dependent primarily on the external substrates as electron donors. The most efficient terminal electron acceptor O_2 is also provided by the cell's immediate environment. The oxygen concentration usually varies significantly within the tissues as a function of the physical distance to the source (blood vessels) and the local demand. In this way, the cellular microenvironment is of primary importance in determining how a cell can generate energy and impact the transcriptional fluctuations through the epigenetic modifications. Each cell is exposed by a unique microenvironment that is essentially composed by other cells. This may explain why cells in the same tissue are so different and create complementarity and interdependence between neighbors. The cells localized close to the nutrient and oxygen sources use different metabolic pathways than those cells that are located more distantly and exposed to a microenvironment composed by the resources not used by their neighbors and by their secreted metabolites. A tissue or a cell community can be considered analogous to an ecosystem and the interaction between the cells as a Darwinian selective pressure. It has been proposed that cell differentiation is a process analogous to Darwinian evolution [35, 36]. Stochastic fluctuations of gene expression in the cell generate spontaneously phenotypic fluctuations. Interactions between the cells and their microenvironment act as a selective force that can stabilize some phenotypes only. Each cell fluctuates until it can express the characteristics that allow using the available resources and maintaining a metabolic flux that produces the necessary energy in the system of interdependent individual cells placed in a given environment.

6 Conclusion

It is important to keep in mind that living cells are out-of-equilibrium open thermodynamic systems that constantly dissipate energy. The minimal energy flux required to maintain the dynamic equilibrium is a sine qua non-condition for the living state and expected to be the organizing force of the living matter [37]. This theoretical conclusion led to the proposal that the true driving force of cell differentiation is the requirement to continuously dissipate energy produced by the metabolic flux [38, 39]. Chromatin stabilizing/destabilizing epigenetic mechanisms appear as a major evolved molecular mechanism that links the environment to the fluctuations of the genome function [38]. These mechanisms transform metabolic fluctuations into gene expression fluctuations ensuring

the generation of new phenotypic variants until the metabolic adaptation is achieved.

The Darwinian model of cell differentiation conceptualizes the whole process of ontogenesis using the same concepts of variation/selection as in the theory of evolution. Phenotype variations are generated by the stochastic fluctuations of the molecular processes that maintain the continuous fluctuations of gene expression levels [10, 40–42]. The necessity to maintain the permanent energy flux required for the vital cellular processes represents a strong selective pressure continuously acting on the fluctuating phenotype. Suboptimal metabolic flux acts by increasing the fluctuations; return to the steady state decreases them. The metabolic pressure canalizes the cell phenotype through the direct substrate level link between the core energy metabolism and the chromatin modifying epigenetic mechanisms. The same epigenetic mechanisms also ensure the conservation of gene expression profiles after cell divisions.

Redefining the conceptual framework of cell differentiation by considering variation as a central player leads to a unified theory that explains the emergence of different living forms at different time scales without making the distinction between an individual as a unit of evolution and its parts as units of ontogenesis. The two processes are expressions of the same principles [43].

Acknowledgments

I thank my colleagues, Alice Moussy, Daniel Stockholm, and Guillaume Corre, for the helpful discussions and the useful comments on the manuscript.
Financial support: EPHE, Genethon, Stochagene ANR grant n° BSV6 014 02.

References

1. Trapnell C (2015) Defining cell types and states with single-cell genomics. Genome Res 25(10):1491–1498. https://doi.org/10.1101/gr.190595.115
2. Merrell AJ, Stanger BZ (2016) Adult cell plasticity in vivo: de-differentiation and transdifferentiation are back in style. Nat Rev Mol Cell Biol 17(7):413–425. https://doi.org/10.1038/nrm.2016.24
3. Blau HM, Brazelton TR, Weimann JM (2001) The evolving concept of a stem cell: entity or function? Cell 105(7):829–841
4. Zipori D (2004) The nature of stem cells: state rather than entity. Nat Rev Genet 5 (11):873–878. https://doi.org/10.1038/nrg1475
5. Wagner A, Regev A, Yosef N (2016) Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol 34(11):1145–1160. https://doi.org/10.1038/nbt.3711
6. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Lauridsen FK, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, Tanay A, Amit I (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell 163 (7):1663–1677. https://doi.org/10.1016/j.cell.2015.11.013
7. Blakeley P, Fogarty NM, del Valle I, Wamaitha SE, Hu TX, Elder K, Snell P, Christie L,

Robson P, Niakan KK (2015) Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development 142 (18):3151–3165. https://doi.org/10.1242/dev.123547

8. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA (2015) Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12(5):453–457. https://doi.org/10.1038/nmeth.3337
9. Yin Z, Sadok A, Sailem H, McCarthy A, Xia X, Li F, Garcia MA, Evans L, Barr AR, Perrimon N, Marshall CJ, Wong ST, Bakal C (2013) A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol 15 (7):860–871. https://doi.org/10.1038/ncb2764
10. Moussy A, Cosette J, Parmentier R, da Silva C, Corre G, Richard A, Gandrillon O, Stockholm D, Paldi A (2017) Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment. PLoS Biol 15(7):e2001867. https://doi.org/10.1371/journal.pbio.2001867
11. Stockholm D, Edom-Vovard F, Coutant S, Sanatine P, Yamagata Y, Corre G, Le Guillou L, Neildez-Nguyen TM, Paldi A (2010) Bistable cell fate specification as a result of stochastic fluctuations and collective spatial cell behaviour. PLoS One 5(12):e14441. https://doi.org/10.1371/journal.pone.0014441
12. Amir el AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol 31 (6):545–552. https://doi.org/10.1038/nbt.2594
13. Gut G, Tadmor MD, Pe'er D, Pelkmans L, Liberali P (2015) Trajectories of cell-cycle progression from fixed cell populations. Nat Methods 12(10):951–954. https://doi.org/10.1038/nmeth.3545
14. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol 34(6):637–645. https://doi.org/10.1038/nbt.3569
15. Ezer D, Moignard V, Gottgens B, Adryan B (2016) Determining physical mechanisms of gene expression regulation from single cell gene expression data. PLoS Comput Biol 12 (8):e1005072. https://doi.org/10.1371/journal.pcbi.1005072
16. Huang S (2009) Non-genetic heterogeneity of cells in development: more than just noise. Development 136(23):3853–3862. https://doi.org/10.1242/dev.035139
17. Huang S (2012) The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? Bioessays 34(2):149–157. https://doi.org/10.1002/bies.201100031
18. Furusawa C, Kaneko K (2012) A dynamical-systems view of stem cell biology. Science 338 (6104):215–217. https://doi.org/10.1126/science.1224311
19. Waddington CH (ed) (1957) The strategy of the genes. Allen & Unwin, Crows Nest
20. Hume DA (2000) Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. Blood 96(7):2323–2328
21. Ko MS (1991) A stochastic model for gene induction. J Theor Biol 153(2):181–194
22. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297(5584):1183–1186. https://doi.org/10.1126/science.1070919
23. Balazsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. Cell 144 (6):910–925. https://doi.org/10.1016/j.cell.2011.01.030
24. Chen H, Larson DR (2016) What have single-molecule studies taught us about gene expression? Genes Dev 30(16):1796–1810. https://doi.org/10.1101/gad.281725.116
25. Larson DR, Singer RH, Zenklusen D (2009) A single molecule view of gene expression. Trends Cell Biol 19(11):630–637. https://doi.org/10.1016/j.tcb.2009.08.008
26. Corre G, Stockholm D, Arnaud O, Kaneko G, Vinuelas J, Yamagata Y, Neildez-Nguyen TM, Kupiec JJ, Beslon G, Gandrillon O, Paldi A (2014) Stochastic fluctuations and distributed control of gene expression impact cellular memory. PLoS One 9(12):e115574. https://doi.org/10.1371/journal.pone.0115574
27. Lestas I, Vinnicombe G, Paulsson J (2010) Fundamental limits on the suppression of molecular fluctuations. Nature 467 (7312):174–178. https://doi.org/10.1038/nature09333
28. Schwanhausser B, Wolf J, Selbach M, Busse D (2013) Synthesis and degradation jointly determine the responsiveness of the cellular

proteome. Bioessays 35(7):597–601. https://doi.org/10.1002/bies.201300017

29. Misteli T (2001) Protein dynamics: implications for nuclear architecture and gene expression. Science 291(5505):843–847
30. Phair RD, Scaffidi P, Elbi C, Vecerova J, Dey A, Ozato K, Brown DT, Hager G, Bustin M, Misteli T (2004) Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. Mol Cell Biol 24(14):6393–6402. https://doi.org/10.1128/MCB.24.14.6393-6402.2004
31. Turner BM (2012) The adjustable nucleosome: an epigenetic signaling module. Trends Genet 28(9):436–444. https://doi.org/10.1016/j.tig.2012.04.003
32. Dodd IB, Micheelsen MA, Sneppen K, Thon G (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129(4):813–822. https://doi.org/10.1016/j.cell.2007.02.053
33. Cyr AR, Domann FE (2011) The redox basis of epigenetic modifications: from mechanisms to functional consequences. Antioxid Redox Signal 15(2):551–589. https://doi.org/10.1089/ars.2010.3492
34. Lu C, Thompson CB (2012) Metabolic regulation of epigenetics. Cell Metab 16(1):9–17. https://doi.org/10.1016/j.cmet.2012.06.001
35. Kupiec JJ (1996) A chance-selection model for cell differentiation. Cell Death Differ 3(4):385–390
36. Kupiec JJ (1997) A Darwinian theory for the origin of cellular differentiation. Mol Gen Genet 255(2):201–208
37. Lane N, Martin W (2010) The energetics of genome complexity. Nature 467(7318):929–934. https://doi.org/10.1038/nature09486
38. Paldi A (2003) Stochastic gene expression during cell differentiation: order from disorder? Cell Mol Life Sci 60(9):1775–1778. https://doi.org/10.1007/s00018-003-23147-z
39. Paldi A (2012) What makes the cell differentiate? Prog Biophys Mol Biol 110(1):41–43. https://doi.org/10.1016/j.pbiomolbio.2012.04.003
40. Mojtahedi M, Skupin A, Zhou J, Castano IG, Leong-Quong RY, Chang H, Trachana K, Giuliani A, Huang S (2016) Cell fate decision as high-dimensional critical state transition. PLoS Biol 14(12):e2000640. https://doi.org/10.1371/journal.pbio.2000640
41. Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, Guillemin A, Papili Gao N, Gunawan R, Cosette J, Arnaud O, Kupiec J-J, Espinasse T, Gonin-Giraud S, Gandrillon O (2016) Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. PLoS Biol 14:e1002585. https://doi.org/10.1371/journal.pbio.1002585
42. Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, Hirche C, Lutz C, Buss EC, Nowak D, Boch T, Hofmann WK, Ho AD, Huber W, Trumpp A, Essers MA, Steinmetz LM (2017) Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol 19(4):271–281. https://doi.org/10.1038/ncb3493
43. Kupiec J-J (2009) The origin of individuals. World Scientific, Hackensack, NJ

Chapter 4

A Primer on Mathematical Modeling in the Study of Organisms and Their Parts

Maël Montévil

Abstract

Mathematical modeling is a very powerful tool for understanding natural phenomena. Such a tool carries its own assumptions and should always be used critically. In this chapter, we highlight the key ingredients and steps of modeling and focus on their biological interpretation. In particular, we discuss the role of theoretical principles in writing models. We also highlight the meaning and interpretation of equations. The main aim of this chapter is to facilitate the interaction between biologists and mathematical modelers. We focus on the case of cell proliferation and motility in the context of multicellular organisms.

Key words Mathematical modeling, Proliferation, Theory, Equations, Parameters

1 Introduction

Mathematical modeling may serve many purposes such as performing quantitative predictions or making sense of a situation where reciprocal interactions are beyond informal analyses. For example, describing the properties of the different ionic channels of a neuron individually is not sufficient to understand how their combination entails the formation of action potentials. We need a mathematical analysis such as the one performed by the Hodgkin-Huxley model to gain such an understanding [1]. In this sense, mathematical modeling is required at some point in order to understand many biological phenomena. Let us emphasize that the perspective of modelers is usually different than the one of many experimentalists, especially in molecular biology. The latter field tends to emphasize the contribution of individual parts, but traditional reductionism [2] involves both the analysis of parts and the theoretical composition of parts to understand the whole, usually by means of mathematical analysis. Without the latter move, it is never clear whether the parts analyzed individually are sufficient to explain how the phenomenon under study comes to be or whether key processes are missing.

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_4,

We want to emphasize the difference between mathematical models on the one side and theories on the other side. Of course, modelization belongs to the broad category of theoretical work by contrast with experimental work. However, in this text, we will refer to theory in the precise sense of a broad conceptual framework such as evolutionary theory. Evolutionary theory has been initially formulated without explicit mathematics. Evolutionary theory has actually led to different categories of mathematical analyses such as population genetics or phyllogenetic analysis which are very different mathematically. Theoretical frameworks typically guide modelization and contribute to justifying mathematical models.

Mathematical modeling raises several difficulties in the study of organisms.

The first one is that most biologists do not have the mathematical or physical background to assess the meaning and the validity of models. The division of labor in interdisciplinary projects is an efficient way to work, but it should at least be completed by an understanding of the principles at play in every part of the work. Otherwise, the coherence of the knowledge that results from this work is not ensured.

The second difficulty is intrinsic. Living objects have theoretical specificities that make mathematical modeling difficult or at least limit its meaning. These specificities are at least of two kinds.

- Current organisms are the result of an evolutive and developmental history which means that many contingent events are deeply inscribed in the organization of living being. By contrast the aim of mathematical modeling is usually to make explicit the necessity of an outcome. For more on this issue, *see* ref. 3.
- The study of a part X of an organism is not completely meaningful by itself. Instead, the inscription of this part inside the organism and in particular the role that this part plays is a mandatory object of study to assess the biological relevance of the properties of X that are under study. As such, the modelization of X per se is insufficient and requires a supplementary discussion [4].

The third difficulty is that there are no well-established theoretical principles to frame model writing in physiology or developmental biology [5]. In particular, cells are elementary objects since the cell theory states that there are no living things without cells. However, cells have complex organizations themselves. Modeling their behavior (*see* **Note 1**) is therefore challenging and requires appropriate theoretical assumptions to ensure that this modeling has a robust biological meaning.

A theoretical way to organize the mathematical modeling of cell behaviors is to propose a default state, that is to say to make explicit a state of reference that takes place without the need of

particular constraints, input, or signal. We think that proliferation with variation and motility should be used as a default state [6, 7]. Under this assumption, cells spontaneously proliferate. By contrast, quiescence should be explained by constraints explicitly limiting or even preventing cell proliferation. The same reasoning applies *mutadis mutandis* to motility. This assumption has been used to model mammary gland morphogenesis and helps to systematize the mathematical analysis of cellular populations [8].

In this chapter, we will focus on model writing. Our aim is not to emphasize the technical aspects of mathematical analysis. Instead, this text aims to help biologists to understand modelization in order to better interact with modelers. Reciprocally, we also highlight theoretical specificities of biology which may be of help to modelers. Of course, the usual way to divide chapters in this book series is not entirely appropriate for the topic of our chapter. We still kept this structure and follow it in a metaphorical sense. In materials, we are describing key conceptual and mathematical ingredients of models. In methods, we will focus on the writing and analysis of models per se.

2 Materials

2.1 Parameters and States

2.1.1 Parameters

Parameters are quantities that play a role in the system but which are not significantly impacted by the system's behavior at the time scale of the phenomenon under study. From an experimentalist's point of view, there are two kinds of parameters. Some parameters correspond to a quantity that is explicitly set by the experimenter such as the temperature, the size of a plate, or the concentration of a relevant compound in the media. Other parameters correspond to properties of parts under study, such as the speed of a chemical reaction, the elasticity of collagen or the division rate τ of a cell without constraints. Changing the value of these parameters requires changing the part in question, *see* also **Note 2**.

Identifying relevant parameters has actually two different meaning:

- Parameters that will be used explicitly in the model are parameters whose value is required to deduce the behavior of the system. The dynamics of the system depends explicitly on the value of these parameters. *A fortiori*, parameters that correspond to different treatments leading to a response will fall under this category. Note that the importance of some parameters usually appear in other steps of modeling.
- Theoretical parameters correspond to parameters that we know are relevant and even mandatory for the process to take place but that we can keep implicit in our model. For example, the concentration of oxygen in the media is usually not made explicit in

a model of an in vitro experiment even though it is relevant for the very survival of the cells studied. Of course, there is usually a cornucopia of this sort of parameters, for example the many components of the serum.

2.1.2 State Space

The state of an object describes its situation at a given time. The state is composed of one or several quantities, *see* **Note 3**. By contrast with parameters, the notion of state is restricted to those aspects of the system which will change as a result of explicit causes or randomness intrinsic to the system described. The usual approach, inherited from physics, is to propose a set of possible states that does not change during the dynamics. Then the changes of the system will be changes of states while staying among these possible states. For example, we can describe a cell population in a very simple manner by the number of cells $n(t)$. Then, the state space is all the possible values for n, that is to say the positive integers.

Usually, the changes of the state depend on the state of the system which means that the state has a causal power, which can be either direct or indirect. A direct causal power is illustrated by n which is the number of cells that are actively proliferating in the example above and thus trigger the changes in n. An indirect causal power corresponds, for example, to the position of a cell provided that some positions are too crowded for cells to proliferate.

2.1.3 Parameter Versus State

Deciding whether a given quantity should be described as a parameter or as an element of the state space is a theoretical decision that is sometimes difficult (*see* also **Note 4**). The heart of the matter is to analyze the role of this quantity but it also depends on the modeling aims.

- Does this quantity change in a quantitatively significant way at the time scale of the phenomenon of interest? If no it should be a parameter. If yes:
- Are the changes of this quantity required to observe the phenomenon one wants to explain? If yes, it should be a part of the state space. If no:
- Do we want to perform precise quantitative predictions? If yes, then the quantity should be a part of the state space and a parameter otherwise.

In the following, we will call "description space" the combination of the state space and parameters.

2.2 Equations

Equations are often seen as intimidating by experimental biologists. Our aim here and in the following subsection is to help demystify them. In the modeling process, equations are the final explicitation of how changes occur and causes act in a model. As a result

understanding them is of paramount importance to understand the assumptions of a model.

The basic rule of modeling is extremely simple. Parameters do not require equations since they are set externally. However, the value of states is unspecified. As a result, equations are required to describe how states change. More precisely, modelers require an equation for each quantity describing the state. Quantities of the state space are degrees of freedom, and these degrees of freedom have to be "removed" by equations for the model to perform predictions. These equations need to be independent in the sense that they need to capture different aspects of the system: copying twice the same equation obviously does not constrain the states. Equations typically come in two kinds:

- Equations that relate different quantities of the state space. For example, if we have n the total number of cells and two possible cell types with cell counts n_1 and n_2, then we will always have $n = n_1 + n_2$. As a result, it is sufficient to describe how two of these variables change to obtain the third one.
- Equations that describe a change of state as a function of the state. These equations typically take two different forms, depending on the representation of time which may be either continuous or discrete, *see* **Note 5**. In continuous time, modelers use differential equations, for example $dn/dt = n/\tau$. This equation means that the change of n (dn) during a short time (dt) is equal to ndt/τ. This change follows from cell proliferation and we will expand on this equation in the next section. In discrete time, $n(t + \Delta t) - n(t)$ is the change of state which relates to the current state by $n(t + \Delta t) - n(t) = n(t)\Delta t/\tau$. Alternatively and equivalently, the future state can be written as a function of the current state: $n(t + \Delta t) = n(t)\Delta t/\tau + n(t)$. Defining a dynamics requires at least one such equation to bind together the different time points, that is to say to bind causes and their effects.

2.3 Invariants and Symmetries

We have discussed the role of equations, now let us expand on their structure. Let us start with the equation mentioned above: $dn/dt = n/\tau$. What is the meaning of such an equation? This equation states that the change of n, dn/dt, is proportional to n. (1) In conformity with the cell theory, there is no spontaneous generation. There is no migration from outside the system described, which is an assumption proper to a given situation. The only source of cells is then cell proliferation. (2) Every cell divides at a given rate, independently. As a conclusion, the appearance of new cells is proportional to the number of cells which are dividing unconstrained, that is to say n. A cell needs a duration of τ to generate two cells (that is to say increase the cell count by one) which is exemplified by the fact that for $n = 1$, $dn/dt = 1/\tau$.

Alternatively, this equation is equivalent to $dn/dt \times 1/n = 1/\tau$, and the latter relation shows that the equation is equivalent to the existence of an invariant quantity: $dn/dt \times 1/n$ which is equal to $1/\tau$ for all values of n. Doubling n thus requires doubling dn/dt. In this sense, the joint transformation $dn/dt \rightarrow 2dn/dt$ and $n \rightarrow 2n$ is a symmetry, that is to say a transformation that leaves invariant a key aspect of the system. This transformation leads from one time point to another. Discussing symmetries of equations is a method to show their meaning. Here, in a sense, the size of the population does not matter. Symmetries can also be multi-scale, for example fractal analysis is based on a symmetry between the different scales that is very fruitful in biology [9, 10].

Randomness may be defined as unpredictability in a given theoretical frame and is more general than probabilities. Probabilities may also be analyzed on the basis of symmetries. To define probabilities, two steps have to be performed. The modeler needs to define a space of possibilities and then to define the probabilities of these possibilities. The most meaningful way to do the latter is to figure out possibilities that are equivalent, that is to say symmetric. For example, in a homogeneous environment, all directions are equivalent and thus would be assigned the same probabilities. A cell, in this situation, would have the same chance to choose any of these directions assuming that the cell's organization is not already oriented in space, *see* also **Note 6**. In physics, a common assumption is to consider that states which have the same energy have the same probabilities.

Now there are several ways to write equations, independently of their deterministic or stochastic nature:

- Symmetry-based writing is exemplified by the model of exponential growth above. In this case, the equation has a genuine meaning. Of course, the model conveys approximations that are not always valid, but the terms of the equation are biologically meaningful. This also ensures that all mathematical outputs of the model may be interpreted biologically.
- Equations may also be based on a mathematical reasoning that provides a legitimacy to their form but restricts their biological interpretations. For example, many mathematical functions may be approximated around 0 by the sum $ax + bx^2 + \ldots$. As a result, a usual way to model a population which constraints itself is the following

$$\frac{dn}{dt} = \frac{n}{\tau} - \frac{n^2}{k\tau}$$

$$\frac{dn}{dt} = \frac{n}{\tau}\left(1 - \frac{n}{k}\right)$$

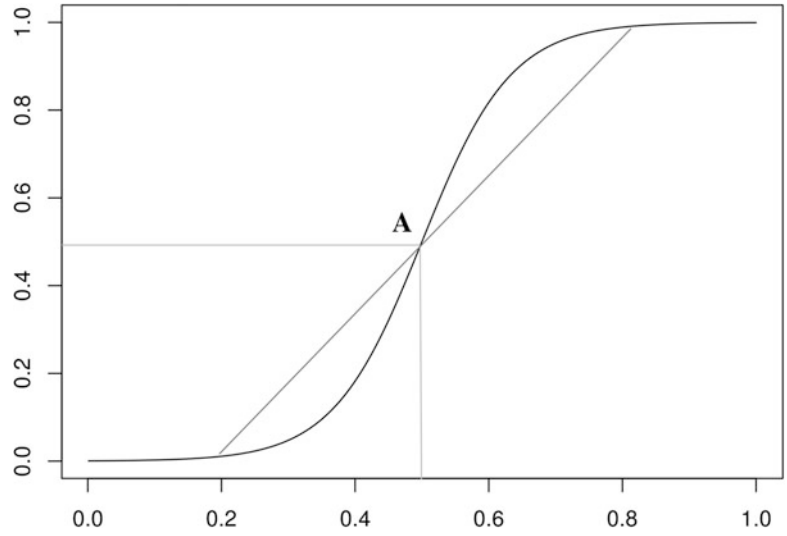


Fig. 1 The logistic function. This function is often used to model a growth with constraints leading to a saturation. However, this function possess a center of symmetry, A, which implies that the initial exponential growth is exactly equivalent to the way the growth saturates. This is biologically problematic: there is an initial lag phase and the saturation trigger causes that are not significant in the initial growth leading for example to cell death [12]

where k is the maximum of the population. Let us remark that we have written the equation in two different forms, we come back on this in **Note 7**. The solution of this equation is the classical logistic function.

Note however that this equation has symmetries that are dubious from a biological viewpoint: the way the population takes off is identical to the way it saturates because the logistic equation has a center of symmetry, *A* in Fig. 1, *see* also [11].

- The last way to write equations is called heuristic. The idea is to use functions that mimic quantitatively and to some extent qualitatively the phenomenon under study. Of course, this method is less meaningful than the others, but it is often required when the knowledge of the underlying phenomenon is not sufficient.

2.4 Theoretical Principles

Theoretical principles are powerful tools for writing equations that convey biological meaning. Let us provide a few examples.

- Cell theory implies that cells come from the proliferation of other cells and excludes spontaneous generation.
- Classical mechanics aims to understand movements in space. The acceleration of an object requires that a mechanical force is exerted on this object. Note that the principle of reaction states that if A exerts a force on B, then B exerts the same force with opposite direction on A. Therefore, there is an equivalence between "A exerts a force" and "a force is exerted on A" from

the point of view of classical mechanics. The difficulty lies in the forces exerted by cells as cells can consume free energy to exert many kinds of forces. Cells are neither an elastic nor a bag of water, they possess agency that leads us to the next point.

- As explained in the introduction, the reference to a default state helps to write equations that pertain to cellular behaviors. There are many aspects that contribute to cellular proliferation and motility. The writing of an equation such as the logistic model is not about all these factors and should not be interpreted as such. Instead, it assumes proliferation on the one side and one or several factors that constrain proliferation on the other side.

3 Methods

3.1 Model Writing

Model writing may have different levels of precision and ambition. Models can be a proof of concept, that is to say the genuine proof that some hypotheses explain a given behavior or even proofs of the theoretical possibility of a behavior. Proof of concept does not include a complete proof that the natural phenomenon genuinely behaves like the model. On the opposite end of the spectrum, models may aim at quantitative predictions. Usually, it is good practice to start from a crude model and after that to go for more detailed and quantitative analyses depending on the experimental possibilities.

We will now provide a short walkthrough for writing an initial model:

- Specify the aims of the model. Models cannot answer all questions at once, and it is crucial to be clear on the aim of a model before attempting to write it. Of course, these aims may be adjusted afterward. The scope of the model should also depend on the experimental methods that link it to reality.
- Analyze the level of description that is mandatory for the model to explain the target phenomenon. Usually, the simplest description is the better. When cells do not constrain each other, describing cells by their count n is sufficient. By contrast, if cells constrain each other, for example if they are in organized 3d structures it can be necessary to take into account the position of each individual cell which leads to a list of positions $\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots$. Note that in this case the state space is far larger than before, *see* **Note 8**. A fortiori, it is necessary to represent space to understand morphogenesis. Note that the notion of level of description is different from the notion of scale. A level of description pertains to qualitative aspects such as the individual cell, the tissue, the organ, the organism, etc. By contrast, a scale is defined by a quantity.

- List the theoretical principles that are relevant to the phenomenon. These principles can be properly biological and pertain to cell theory, the notion of default state, biological organization, or evolution. Physico-chemical principles may also be useful such as mechanics or the balance of chemical reactions.
- List the relevant states and parameters. These quantities are the ones that are expected to play a causal role that pertains to the aim of the model. This list will probably not be definitive, and will be adjusted in further steps. In all cases, we cannot emphasize enough that aiming for exhaustivity is the modeler's worst enemy. Biologists need to take many factors into account when designing an experimental protocol, it is a mistake to try to model all of these factors.
- The crucial step is to propose mathematical relations between states and their changes. We have described in Subheadings 2.2 and 2.3 what kinds of relation can be used. Usually, these relations will involve supplementary parameters whose relevance was not obvious initially. Let us emphasize here that the key to robust models is to base it on sufficiently solid grounds. A model where all relations are heuristic will probably not be robust. As such, figuring out the robust and meaningful relations that can be used is crucial.
- The last step is to analyze the consequences of the model. We describe this step with more details below. What matters here is that the models may work as intended, in which case it may be refined by adding further details. The model may also lead to unrealistic consequences and not lead to the expected results. In these latter cases, the issue may lie in the formulation of the relations above, in the choice of the variables or in oversimplifications. In all cases the model requires a revision.

Writing a model is similar to the chess game in that the anticipation of all these steps from the beginning helps. The steps that we have described are all required but a central aspect of modeling is to gain a precise intuition of what determines the system's behavior. Once this intuition is gained, it guides the specification of the model at all the steps. Reciprocally, these steps help to gain such an intuition.

3.2 Model Analysis

In this section, we will not cover all the main ways to analyze model since this subject is far too vast and depends on the mathematical structures used in the models. Instead, we will focus on the outcome of model analyses.

3.2.1 Analytic Methods

Analytic methods consist in the mathematical analysis of a model. They should always be preferred to simulations when the model is tractable, even at the cost of using simplifying hypotheses.

- Asymptotic reasoning is a fundamental method to study models. The underlying idea is that models are always a bit complicated. To make sense of them, we can look at the dynamics after enough time which simplifies the outcome. For example, the outcome of the logistic function discussed above will always be an equilibrium point, where the population is at a maximum. Mathematically, "enough" time means infinite time, hence the term asymptotic. In practice, "infinite" means "large in comparison with the characteristic times of the dynamics," which may not be long from a human point of view. For example, a typical culture of bacteria reaches a maximum after less than day. Asymptotic behaviors may be more complicated such as oscillations or strange attractors.
- Steady-state analysis. In fairly complex situations, for example when both space and time are involved, a usual approach is to analyze states that are sustained over time. For example, in the analysis of epithelial morphogenesis, it is possible to consider how the shape of a duct is sustained over time.
- Stability analysis. A very common analytic method is to find equilibria, that is to say situations where the changes stop ($dx/dt = 0$ for all state variable x). For example, $dn/dt = (n/\tau)(1 - n/k)$ has two equilibria for $n = k$ and $n = 0$. Stability analysis looks at the consequences of equation near an equilibrium point. Near the equilibrium value n_e, $n = n_e + \Delta n$ where Δn is considered to be small. Δn small means that Δn dominates Δn^2 and all other powers of Δn, *see* also **Note 9**. The reason for that is simple: if $\Delta n = 0.1$, $\Delta n^2 = 0.01$,...

  Near 0, $n = 0 + \Delta n$ and $dn/dt \simeq \Delta n/\tau$. The small variation Δn leads to a positive dn/dt therefore, this variation is amplified and this equilibrium is not stable. We should not forget the biology here. For a population of cells or animals of a given large size, a small variation is possible. However, a small variation from a population of size 0 is only possible through migration because spontaneous generation does not happen. Nevertheless, this analysis shows that a small population, close to $n = 0$, should not collapse but instead will expand.
  Near k, let us write $n = k + \Delta n$

$$\frac{dn}{dt} = \frac{(k+\Delta n)}{\tau}\left(1 - \frac{(k+\Delta n)}{k}\right) = \frac{(k+\Delta n)}{\tau}\left(\frac{-1\Delta n}{k}\right)$$

$$\frac{dn}{dt} = -\frac{\Delta n}{\tau} - \frac{\Delta n^2}{\tau k} \simeq -\frac{\Delta n}{\tau}$$

In this case, the small variation Δn leads to a negative feedback; therefore, the equilibrium is stable.

- Special cases. In some situations, qualitatively remarkable behaviors appear for specific values of the parameters. Studying these cases is interesting per se, even though the odds for parameters to have specific value are slim without an explicit reason for this parameter to be set at this value. However, in biology the value of some parameters is the result of biological evolution and a specific value can become relevant when the associated qualitative behavior is biologically meaningful [13, 14].
- Parameter rewriting. One of the major practical advantages of analytical methods is to prove the relevance of parameters that are the key to understanding the behavior of a system. These "new" parameters are usually combinations of the initial parameters. We have implicitly done this operation in Subheading 2.3. Instead of writing $an + bn^2$ we have written $n/\tau - n^2/k\tau$. The point here is to introduce τ the characteristic time for a cell division and k which is the maximum size of the population. By contrast, a and especially b are less meaningful. These key parameters and their meaning are an outcome of models and at the same time should be the target of precise experiments to explore the validity of models.

3.2.2 Numerical Methods—Simulations

Simulations have a major strength and a major weakness. Their strength lies in their ability to handle complicated situations that are not tractable analytically. Their weakness is that each simulation run provides a particular trajectory that cannot a priori be assumed to be representative of the dynamical possibilities of the model.

In this sense, the outcome of simulations may be compared to empirical results, except that simulation is transparent: it is possible to track all variables of interest over time. Of course, the outcome of simulations is artificial and only as good as the initial model.

Last, there is almost always a loss when going from a mathematical model to a computer simulation. Computer simulations are always about discrete objects and deterministic functions. Randomness and continua are always approximated in simulations and mathematical care is required to ensure that the qualitative features of simulations are features of the mathematical model and not artifacts of the transposition of the model into a computer program. A subfield of mathematics, numerical analysis, is devoted to this issue.

3.2.3 Results

We want to emphasize two points to conclude this section.

First, it is not sufficient for a model to provide the qualitative or even quantitative behavior expected for this model to be correct. The validation of a model is based on the validation of a process and of the way this process takes place. As a result, it is necessary to explore the predictions of the model to verify them experimentally. All outcomes that we have described in Subheading 3.2.1 may be

used to do so on top of a direct verification of the assumptions of the model themselves. Of course, it is never possible to verify everything experimentally; therefore, the focus should be on aspects that are unlikely except in the light of the model.

Second, modeling focuses on a specific part and a specific process. However, this part and this process take place in an organism. Their physiological meaning, or possible lack thereof, should be analyzed. We are developing a framework to perform this kind of analysis [4, 15], but it can also be performed informally by looking at the consequences of the part considered for the rest of the organism.

4 Notes

1. In biology, behavior usually has an ethological meaning and evolution refers to the theory evolution. In the mathematical context, these words have a broader meaning. They both typically refer to the properties of dynamics. For example, the behavior of a population without a constraint is exponential growth.
2. Parameters that play a role in an equation are defined in two different ways. They are defined by their role in the equation and by their biological interpretation. For example, the division rate τ corresponds to the division rate of the cells without the constraint that is represented by k. τ may also embed constant constraints on cell proliferation, for example chemical constraints from the serum or the temperature. Thus, τ is what physicists call an effective parameter it carries implicit constraints beyond the explicit constraints of the model.
3. A state may be composed of several quantities, let us say k, n, m. It is possible to write the state by the three quantities independently or to join them in one vector $X = (k,n,m)$. The two viewpoints are of course equivalent but they lead to different mathematical methods and ways to see the problem. The second viewpoint shows that it is always valid to consider that the state is a single mathematical object and not just a plurality of quantities.
4. The notion of organization in the sense of a specific interdependence between parts [4] implies that most parameters are a consequence of others parts, at other time scales. As a result, modeling a given quantity as a parameter is only valid for some time scales, and is acceptable when these time scales are the ones at which the process modeled takes place.

5. The choice between a model based on discrete or on continuous time is base on several criteria. For example, if the proliferation of cells is synchronized, there is a discrete nature of the phenomenon that strongly suggests representing the dynamics in discrete time. In this case, the discrete time corresponds to an objective aspect of the phenomenon. On the opposite, when cells divide at all times in the population, a representation in continuous time is more adequate. In order to perform simulations, time may still be discretized but the status of the discrete structure is then different than in the first case: discretization is then arbitrary and serves the purpose of approximating the continuum. To distinguish the two situations, a simple question should be asked. What is the meaning of the time difference between two time points. In the first case, this time difference has a biological meaning, in the second it is arbitrary and just small enough for the approximation to be acceptable.
6. Probabilities over continuous possibilities are somewhat subtle. Let us show why: let us say that all directions are equivalent, thus all the angles in the interval [0,360] are equivalent. They are equivalent, so their probabilities are all the same value p. However, there are an infinite number of possible angles, so the sum of all the probabilities of all possibilities would be infinite. Over the continuum, probabilities are assigned to sets and in particular to intervals, not individual possibilities.
7. There are many equivalent ways to write a mathematical term. The choice of a specific way to write a term conveys meaning and corresponds to an interpretation of this term. For example, in the text, we transformed $dn/dt = n/\tau - n^2/k\tau$ because this expression has little biological meaning. By contrast, $dn/dt = (n/\tau)(1 - n/k)$ implies that when n/k is very small by comparison with 1, cells are not constraining each other. On the opposite, when $n = k$ there is no proliferation. The consequence of cells constraining each other can be interpreted as a proportion $1\text{-}n/k$ of cells proliferating and a proportion n/k of cells not proliferating. Now, there is another way to write the same term which is: $dn/dt = n/(\tau/(1 - n/k))$. Here, the division time becomes $\tau/(1 - n/k)$ and the more cells there are, the longer the division time becomes. This division time becomes infinite when $n = k$ which means that cells are quiescent. These two interpretations are biologically different. In the first interpretation, a proportion of cells are completely constrained while the other proliferate freely. In the second, all the cells are impacted equally. Nevertheless, the initial term is compatible with both interpretations and they have the same consequences at this level of analysis.

8. The number of quantities that form the state space is called its dimension. The dimension of the phase space is a crucial matter for its mathematical analysis. Basically, low dimensions such as 3 or below are more tractable and easier to represent. High dimensions may also be tractable if many dimensions play equivalent roles (even in infinite dimension). A large number of heterogeneous quantities (10 or 20) are complicated to analyze even with computer simulations because this situation is associated with many possibilities for the initial conditions and for the parameters making it difficult to "probe" the different qualitative possibilities of the model.
9. It is very common in modeling to use the words "small" and "large." A small (resp. large) quantity is a quantity that is assumed to be small (resp. large) enough so that a given approximation can be performed. For example, a large time in the context of the logistic equation means that the population is approximately at the maximum k. Similarly, infinite and large are very close notions in most practical cases. For example, a very large capacity k leads to $dn/dt = (n/\tau)(1 - n/k) \simeq n/\tau$ which is an exponential growth as long as n is far smaller than k.

References

1. Beeman D (2013) Hodgkin-Huxley Model, Encyclopedia of computational neuroscience. Springer, New York, NY, pp 1–13. https://doi.org/10.1007/978-1-4614-7320-6_127-3
2. Descartes R (2016) Discours de la méthode. Flammarion, Paris
3. Montévil M, Mossio M, Pocheville A, Longo G (2016) Theoretical principles for biology: variation. Prog Biophy Mol Biol 122(1):36–50. https://doi.org/10.1016/j.pbiomolbio.2016.08.005
4. Mossio M, Montévil M, Longo G (2016) Theoretical principles for biology: organization. Prog Biophys Mol Biol 122(1):24–35. https://doi.org/10.1016/j.pbiomolbio.2016.07.005
5. Noble D (2010) Biophysics and systems biology. Philos Trans R Soc A Math Phys Eng Sci 368(1914):1125. https://doi.org/10.1098/rsta.2009.0245
6. Sonnenschein C, Soto A (1999) The society of cells: cancer and control of cell proliferation. Springer Verlag, New York
7. Soto AM, Longo G, Montévil M, Sonnenschein C (2016) The biological default state of cell proliferation with variation and motility, a fundamental principle for a theory of organisms. Progr Biophys Mol Biol 122 (1):16–23. https://doi.org/10.1016/j.pbiomolbio.2016.06.006
8. Montévil M, Speroni L, Sonnenschein C, Soto AM (2016b) Modeling mammary organogenesis from biological first principles: cells and their physical constraints. Prog Biophys Mol Biol 122(1):58–69. https://doi.org/10.1016/j.pbiomolbio.2016.08.004
9. D'Anselmi F, Valerio M, Cucina A, Galli L, Proietti S, Dinicola S, Pasqualato A, Manetti C, Ricci G, Giuliani A, Bizzarri M (2011) Metabolism and cell shape in cancer: a fractal analysis. Int J Biochem Cell Biol 43 (7):1052–1058. https://doi.org/10.1016/j.biocel.2010.05.002
10. Longo G, Montévil M (2014) Perspectives on organisms: biological time, symmetries and singularities. Lecture notes in morphogenesis. Springer, Dordrecht. https://doi.org/10.1007/978-3-642-35938-5
11. Tjørve E (2003) Shapes and functions of species–area curves: a review of possible models. J Biogeogr 30(6):827–835. https://doi.org/10.1046/j.1365-2699.2003.00877.x
12. Hoehler TM, Jorgensen BB (2013) Microbial life under extreme energy limitation. Nat Rev

Microbiol 11(2):83–94. https://doi.org/10.1038/nrmicro2939

13. Camalet S, Duke T, Julicher F, Prost J (2000) Auditory sensitivity provided by self-tuned critical oscillations of hair cells. Proc Natl Acad Sci U S A 97(7):3183–3188. https://doi.org/10.1073/pnas.97.7.3183
14. Lesne A, Victor J-M (2006) Chromatin fiber functional organization: some plausible models. Eur Phys J E Soft Matter 19 (3):279–290. https://doi.org/10.1140/epje/i2005-10050-6
15. Montévil M, Mossio M (2015) Biological organisation as closure of constraints. J Theor Biol 372:179–191. https://doi.org/10.1016/j.jtbi.2015.02.029

Chapter 5

The Search for System's Parameters

Alessandro Giuliani

Abstract

The analysis of biological data asks for a delicate balance of content-specific and procedural knowledge; this is why it is virtually impossible to apply standard mathematical and statistical recipes to systems biology.

The separation of the important part of information from singular (and largely irrelevant) details implies a continuous interchange between biological and statistical knowledge. The generalization ability of the models must be the principal focus of system's parameter estimation, while the multi-scale character of biological regulation orients the modeling style toward data-driven strategies based on the correlation structure of the analyzed systems.

Key words Principal component analysis, Overfitting, Soft modeling, Biological regulation

1 Sloppiness and Overfitting: The Hidden Risks of Precision

In their brilliant paper [1], James Sethna and colleagues focus on the separation between a "stiff" and a "sloppy" part present in any modeling effort in science. The authors demonstrate this statement building upon the eigenvalue distribution of the Fisher Information Matrix [1, 2] Any model (from physics to biology) presents the same distribution pattern. This pattern (Fig. 1) shows a clear gap between relatively coarse grain but effective models (top eigenvalues correspondent to a high impact of parameter modifications on the predicted values, stiff part of the model) and a plethora of largely irrelevant (sloppy) model parameter combinations. The "sloppy" parameter combinations (lower eigenvalues), whose modifications have a scarce or null effect on the actual fitting, can drastically reduce model generalization ability, so that the quest for maximal fit has the drawback of the generation of complicated models with no relevant increase in prediction power.

Figure 1 can be interpreted (without appreciable loss of generality) as referred to models of the kind: $\Upsilon = f(x_1, x_2, x_3 \ldots x_n)$ where Υ is the dependent variable we want to model in terms of n independent variables $(x_1 \ldots x_n)$, the parameters are the

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_5,

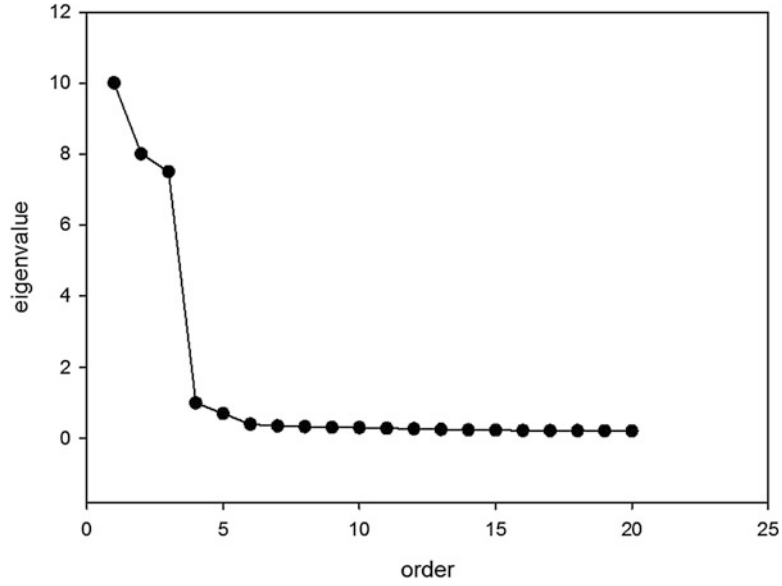


Fig. 1 The figure reports the trend of FIM eigenvalues for different system parameter combinations. Few top eigenvalues correspondent to the "stiff" part of the model go together with a plethora of largely irrelevant (sloppy) parameter combinations

coefficients of the independent variables combinations in the fitted model.

While each parameter "per se" can have a great uncertainty, this does not mean that any parameter can be varied independently of the others and the resulting model can still fit the data. Fisher Information Matrix (FIM) is a sort of covariance matrix [2] reporting the mutual between parameters dependence: this covariance structure stems from the fact the data place constraints on combinations of parameters.

The eigenvalue distribution of the FIM thus corresponds to an estimate of the relative strength of the constraints exerted by the data on the "free variation" (uncertainty) of the model parameters. Top eigenvalues in Fig. 1 correspond to the parameters combinations relevant for data fitting (and consequently for their interpretation), low-level eigenvalues correspond to irrelevant aspects. In the case reported in Fig. 1 only the top three eigenvalues are sufficiently discriminated by the measurement error to be of practical use; this suggests the presence of very general (and robust) ordering principles going together with many irrelevant details (minor eigenvalues).

Machine learning experts are aware of this phenomenon since long time: it is the so-called overfitting [3] effect. The basic issue is that each set of experimental data is affected by noise, being the main noise sources experimental errors and (still more dangerous) the selection bias, i.e., the fact that any data set is relative to a

specific choice of statistical units whose degree of mirroring of the reference population is unknown [4] . As an example of the selection bias, we can think of a model trying to correlate the folding rate to some structural properties of protein molecules: the data set we train the model upon is forcedly a limited selection of the entire protein Universe (even for the simple fact we do not know the 3D structure of all the proteins). This determines a mixing of "data set specific" and "valid on the entire protein Universe" features across all the considered variables. Statistical theory of sampling, while very powerful in the case of well-defined reference population, is largely out-of-scope in a large part of modern biological research where we use a particular model system with the goal of generalizing the results outside its realm (e.g., from cell lines to the entire organ).

Thus, we remain with the problem of "degree of generalization" (and then actual meaning) of our findings: the "overfitting" effect gives us an empirical guidance in this respect. In machine learning experiments [3] it was noted that, after a given level of accuracy (e.g., degree of correlation between the predicted and observed values) the generalization ability (performance of the model on a test set not used for model building) starts to decline.

This decline stems from the fact that, after a certain percentage of accuracy, the fitting procedure starts to model noise, the more the parameters that can be adjusted to fit the data, the faster the entrenchment of the procedure into the modeling of "data set specific" (and thus not-generalizable) details.

In [3], the authors demonstrate that the overfitting problem can be faced by reducing the dimensionality of the system, which corresponds to reducing the degrees of freedom of modeling procedure. This means that (like suggested by Fig. 1) to rely on more information is not necessarily a good thing. This statement could sound paradoxical in these times of ever-increasing computational power and of huge data sets (*see* http://omics.org/index.php/Alphabetically_ordered_list_of_omes_and_omics) but is crucial to pay attention on the above issues if we want to avoid a sort of thermal death of science.

2 Meaningful Syntheses

In his seminal 1901 paper [5], Karl Pearson synthetically defined the main goal of Principal Component Analysis (PCA): "*In many physical, statistical and biological investigations it is desirable to represent a system of points in plane, three or higher dimensioned space by the 'best fitting' straight line or plane.*" The need to collapse multidimensional information scattered over different (and sometimes heterogeneous) descriptors into a lower number of relevant dimensions is one of the main pillars of scientific knowledge and, as

we said above, the best antidote against irrelevance of our findings [6].

Pearson continues: "*In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as independent, those on the left as dependent variables.*" This implies that the minimization of the sum of squared distances only deals with the dependent (y) variable. The variance along independent (x) variable, being the consequence of the choice of the scientist (e.g., dose, time of observation...), is supposed to be strictly controlled and thus does not enter in the evaluation of the "fit" of the model.

The novelty of PCA lies in a different look at reality, much more adherent with the actual situation of systems biology where the traditional distinction between the independent and dependent variables is blurred. Karl Pearson [5] recognized this point as crucial:

> *In many cases of physics and biology, however, the 'independent' variable is subject to just as much deviation or error as the 'dependent' variable, we do not, for example, know x accurately and then proceed to find y, but* ***both x and y are found by experiment or observation****.*

This new attitude is the core of the peculiar "best fitting" procedure set forth by Pearson. Figure 2 reports on the left the original plot of Karl Pearson and on the right the classical regression scheme [6].

In PCA (left panel) the distances to minimize are perpendicular to the model of the data (the straight-line correspondent to the first principal component of x,y space), while in the classical regression model (right panel) the distances are perpendicular to the x axis, because the only uncertainty taken into account refers to y. This apparently minor geometrical detail encompasses a sort of revolution in the style of doing science [6]. The "real thing" (the structure to be approximated by least squares approach) is no more the "results as such" (the actual values of the observables that we know are an intermingled mixture of "general" and "singular" information) but their "meaningful syntheses" correspondent to the principal components.

The discriminatory principle at the basis of PCA has to do with a classical information theory axiom [7, 8]: the signal (meaningful, general) part of information carried by the data corresponds to the correlated variance (the flux of variability shared by different variables).

This choice has a physical counterpart in the dynamics of complex systems [9]: the uncorrelated part of information corresponds to the so-called noise floor, i.e., to the minor components of a data set.

Principal components are both the "best summary," in a least square sense, of the information present in the data cloud, and the

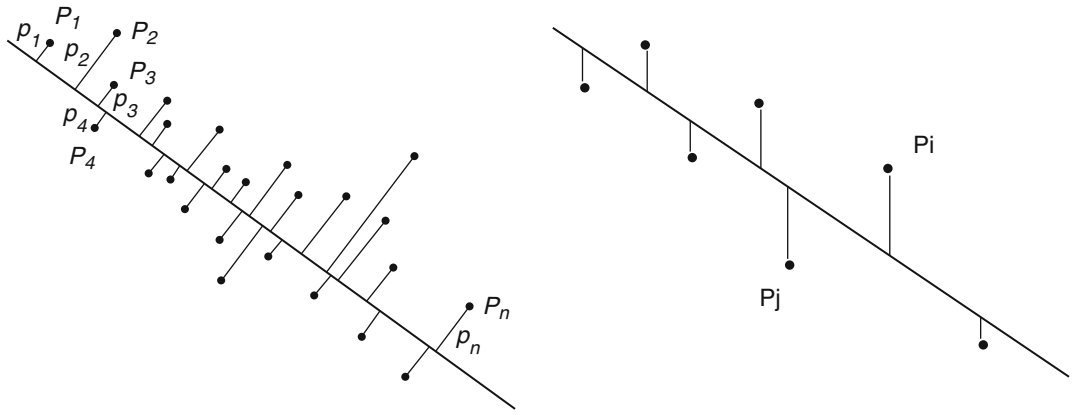


Fig. 2 The least-square estimation in PCA has as reference point the component (*left panel*) and the distances of the observed values from their estimates refer to the bi-dimensional space. In classical regression procedures (*right panel*), the neat separation between independent and dependent variables implies the distances to be minimized refer to the *Y* (dependent variable) axis

directions along which the between variables correlation is maximal. This implies the projection of the original data set into a reduced dimensionality space spanned by the major (i.e., having the higher eigenvalues) components allows us to maximize the probability of concentrating on the meaningful part of the information. In some (very general) sense, a PCA solution corresponds to an explanatory model of the system at hand [10], but this is only the beginning of the game, the search for "systems parameter" does not end with PCA.

3 "Subjective" Judgment

PCA, even being the by far more common method to generate a meaningful synthesis from complex multidimensional data, is not an obliged way for modeling. The peculiar features of biological information must be taken into account in order to define a realistic notion of "systems parameters" (and consequently of what can be considered a satisfactory explanation of a phenomenon).

Biological data are unescapably discrete: we face a numerable set of independent vectors, each corresponding to a specific observation (animal, person, cell sample, etc.) and we are asked to find some kind of regularities among these discrete events.

This fact both imposes a completely new perspective with respect to the prevalent style of thought in mathematical modeling that considers the sketching of a continuous differential equation as the "paradigmatic form" of a scientific analysis of a phenomenon and creates a totally new concept of "what is in common" among different research fields.

The material common to all the discrete approaches is the raw data matrix: each exploration of nature ends up into a matrix having as rows (statistical units) the objects of the analysis and as columns (variables) the descriptors of such objects.

This imposes the crucial choices of the "inclusion criteria" and "preferred scale" of analysis. This kind of problems are identical for an epidemiologist that must choose the inclusion criteria for the individuals entering a case-control or a double-blind trial, and for a biologist that must choose the cell populations to submit to a microarray study or the hierarchical level of an ecological study (species, genera, etc.).

In the same time, the scientists must be aware that the success of their analyses crucially depends upon the choice of the variables to consider both in terms of "what to measure" and "at what scale" [9]. In nonlinear time series analysis methods like Recurrence Quantification Analysis (RQA, 11) this translates into the choice of "embedding dimension," "windowing," "choice of metrics," "choice of recurrence thresholds, etc.," in other techniques like PCA or Multidimensional Scaling (MDS, 12) these choices correspond to the definition of data set, standardization, metrics. Similar considerations hold for cluster analysis techniques [13] regarding the number of clusters and/or the choice between a hierarchical or non-hierarchical approach.

The need for these subjective choices is normally seen as a limitation or, in any case, a lack of rigor when, on the contrary, it is a crucial advantage with respect to more "blind to the content" methods because it allows a rich and fruitful relation between the "analysis tool" and the studied system.

4 A Molecular Biology Example

In order to understand the nature of such a relation between "subjective judgment" (based on content knowledge) and data analysis (based on procedural knowledge), we will base upon an investigation on the molecular mechanism of DNA repair in gastric cancer patients [14].

The hypothesis under scrutiny is the existence of an inverse correlation between mismatch repair (MMR) mode ("marked" by the expression level of gene MLH1) and base excision repair (BER) (marked by the expression level of DNA polymerase B (PolB) gene). To gain insight into possible crosstalk of these two repair pathways in cancer, we analyzed human gastric adenocarcinoma AGS in the presence of a DNA damage agent (Methyl Methane Sulphonate MMS).

DNA repair is a process involving different enzymes whose complex relations ends up into different DNA repair efficiencies and consequently mutation load for the cells.

Thus, we contemporarily measure the gene expression level of 24 enzymes involved in the process selected in order to get a good coverage of the main DNA repair pathways [14] in primary cell cultures relative to 35 patients of gastric cancer in order to get a quantitative experimental "summary" of the above complex network.

The statistical units of the raw data matrix are the patients and the 24 DNA repair enzymes are the variables, whose intermingled network of interaction determines the experimentally observed correlation structure among the expression values of the genes coding for the selected enzymes.

The idea at the basis of PCA is that each single observable (enzyme expression in different patient cell lines in our case) derives its particular value from a combination of hidden independent factors impinging on it. The hidden factors are "the real things," the observables are the probes of such factors. This is the same case of a chemical mixture, whose observed spectrum comes from the combination of a set of elementary spectra relative to the molecules composing the mixture, the different spectral peaks are the "probes," while the molecules are the "real things." The molecules in the mixture are the components and their relative concentration corresponds to the percentage of variance explained by each component. This is more than analogy: spectroscopic apparatuses used in analytical chemistry have embedded a PCA procedure to deconvolve the obtained spectra.

The observed values of different enzymes correspond to a weighted summation over the contribution coming from their participation to the different pathways (components). As stated in the previous paragraph, this implies a total revolution of the way we look at nature: enzyme expression values are a consequence (and not a cause) of the processes (pathways) going on in the cells.

PCA applied to the data set gave rise to a "bona fide" three-component solution accounting for the "signal" part of information; in Table 1 the distribution of percentage of variation explained across the components is reported.

It is worth noting how (analogously to Fig. 1) the PCA modeling of the data encompasses three "top" eigenvalues and a long tail of minor components accounting for the "uncorrelated" (noisy) part of information. The three-component solution explains the 65% of total information; the components are extracted in order of variance explained: PC1 accounts for 41% of total variance, PC2 for 13%, PC3 for 10%. That is to say, there are three main order parameters (correlated flux of variations) organizing the mutual correlations between gene expressions, now we must "give a name" to these pathways. In the case of chemical mixtures, this process is automatic and derives from the knowledge of the typical spectrum of each molecular species, in our case we must rely on the loadings of each enzyme on the extracted

Table 1
The table reports the distribution of variance explained across the principal components. The components are in decreasing order of relevance (Eigenvalue) that in turn is normalized in terms of proportion of explained variance (Proportion). The difference between the variance explained by subsequent components is reported in the field "Difference" while "Cumulative" corresponds to the variance explained by the cumulative solutions at increasing dimensionality (number of considered components). Bolded values mark the accepted global solution (bona fide signal)

Component	Eigenvalue	%Explained variance	%Cumulative
1	9.9	41	41
2	3.3	13	54
3	2.4	10	65
4	1.4	6	71
5	1.3	6	77
6	1.0	4	81
7	0.9	3	84

components. This interpretation stems from the previous knowledge of the phenomenon at hand, and thus naturally involves "subjective judgment." This judgment builds upon the loading matrix (Table 2) reporting the Pearson correlation coefficients between original variables (probes, gene expression values) and extracted components. In Table 2, the most relevant loading for each component is bolded.

Looking at the loading pattern, we immediately discover all the enzymes have positive and, with only a few exceptions, very high correlations with PC1. This implies PC1 is a sort of "global repair activation": the cell lines with higher PC1 scores are the ones with higher repair enzyme expression, independently of the particular mechanism involved. This is an a posteriori result, we did not impose the existence of such "global repair activity," and it spontaneously emerges from experimental data. This corresponds to the fact that the mutant agent (MMS) activates the DNA repair machinery as a whole, this activation ends up into an increase in expression levels of all the genes codifying for DNA repair enzymes. Given the different patients have different levels of activation of the entire DNA repair machinery; the "global activation" corresponds to the most relevant factor in terms of variance explained.

In statistical jargon, a component like PC1 with all loadings of the same sign is a "size" component. This name indicates different PC1 scores correspond to "general changes" shared by all the considered variables, it corresponds to what physicists call "mean field." Observing a size component as the most relevant one in terms of variance explained is a signature of a very strongly connected system behaving as an integrated whole, in physics

Table 2
The table reports the Pearson correlation coefficients (component loadings) between original variables (enzymes) and components

Gene name	PC1	PC2	PC3
apex1	0.49658	0.28549	−0.08580
brca1	**0.80568**	−0.36552	−0.00639
brca2	**0.82445**	−0.21156	−0.03834
ercc1	**0.62712**	**0.44942**	−0.00154
fen1	**0.82015**	**−0.41467**	0.04431
lig1	**0.77139**	**−0.40052**	−0.08932
lig3	**0.74382**	−0.13401	−0.38229
lig4	**0.59043**	**0.53775**	−0.22779
mbd4	**0.55002**	0.37782	−0.23190
mlh1	0.35765	0.29410	**−0.76100**
mpg	0.27652	**0.43364**	**0.54428**
mre11a	**0.70047**	0.11065	**−0.53014**
msh2	**0.89619**	−0.16038	0.14345
msh3	**0.61953**	**0.41855**	−0.15099
msh6	**0.63565**	**−0.53948**	0.08440
ogg1	0.33020	0.26190	0.06547
pms2	**0.82538**	0.21081	−0.11963
polb	**0.58223**	0.33495	**0.58244**
rad51	**0.72720**	**−0.40550**	0.27327
smug1	**0.53041**	−0.19056	−0.29937
ung	**0.62931**	−0.39195	0.28066
xpc	**0.53029**	**0.42125**	0.22514
xrcc1	0.34438	−0.25749	0.10138

Bolded values point to the variables more relevant for component interpretation

terms this means that the motion of the center of mass is the main component of the observed dynamics.

PC2 instead has both positive and negative loadings (a "shape" component in statistical terms), this points to a specific balance between enzymes expression, independent of the global activity (all components are independent of each other by construction) and pointing to the fact the repair enzymes differentially participate in distinct mechanisms (pathways).

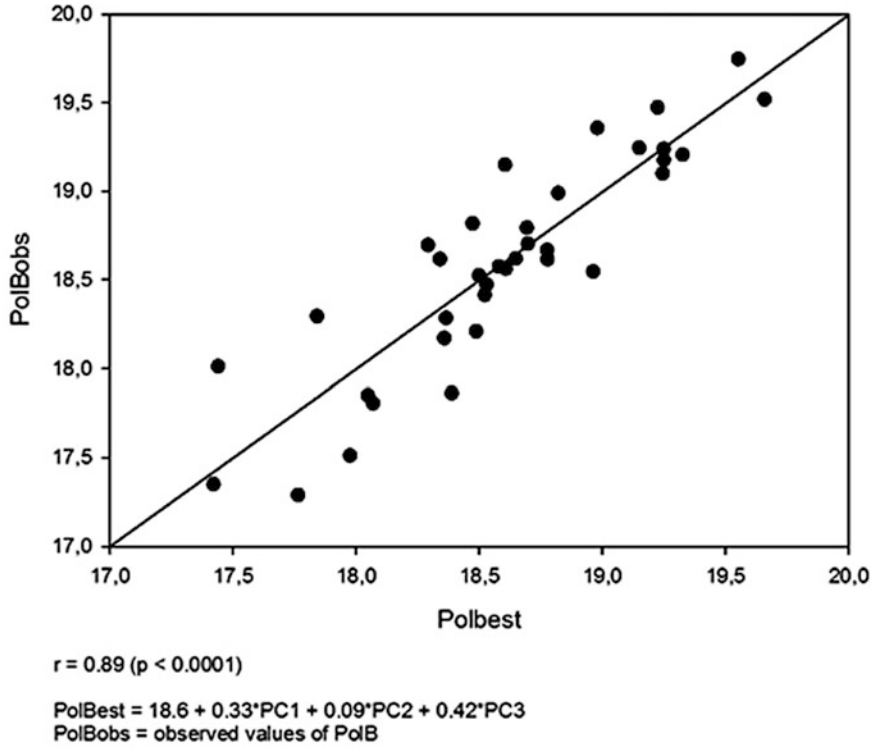


Fig. 3 The observed values of PolB expression level come from a weighted summation of the contributes of the first three principal components. The PolB expression is a mixture of 33 parts of PC1, 9 parts of PC3, and 42 parts of PC3. Each component corresponds to an independent "regulation flux" of DNA repair response, genes are not the "actual players" but probes of underlying hidden factors

The same "balance" character is present in PC3. The two genes most affected by PC3 (higher loadings in absolute value) are PolB and Mlh1. These two genes have opposite sign correlations with PC3, this implies that this "hidden" pathway encompasses a balance between these two genes expression: high values of PolB correspond to low values of Mlh1 and vice versa that corresponds to our initial hypothesis. From what we said before, this negative correlation between the two genes is not true in every context and thus does not appear in terms of direct correlation coefficient between the relative expression level as such, this balance only holds as for the specific pathway correspondent to PC3.

Genes are "probes" of processes (components) working "on behind," thus the actual values of expression can be expressed as a function of component scores. In Fig. 3 the PolB expression values are modeled as a weighted summation of component scores (vector points are patients, Y values the experimentally observed expression values of PolB, while the abscissa corresponds to the PolB values estimated as weighted summation on components).

It is worth noting the high fit of the model ($r = 0.89$) and that the main contribution (higher regression coefficient) to PolB comes from the third component, the main component (PC1, repair system as a whole) has a statistically significant influence on PolB ($p < 0.0001$) but its contribution to the PolB value (0.33) is lower than the PC3 contribution (0.42). PC2 instead has a negligible influence on PoLB determination. The fit of the equation does not reach the complete deterministic reconstruction of PolB ($r = 1.00$) because we use only the first three components, that in any case are the relevant ones, being the component from fourth onward only modeling noise (experimental error). The above equation returns a clear quantitative estimation of the relevant parameters of the system at hand: the PolB expression dynamics is modeled in terms of relative contributions of mutually independent "fluxes of variation." Correspondent to general 'repair activity' (PC1) and 'specific repair mode' (PC3).

The same procedure is applied to MLH1 obtaining a very high fit as well ($r = 0.90$, $p < 0.0001$).

This allows us to concentrate on the "pure PC3 driven" correlation between the two enzymes; this can be done by subtracting the actual expression values of the two genes by their estimation based upon PC1 and PC2, i.e.,

MLH1 (pc3specific) = MLH1 – MLH1 est (PC1, PC2)
PolB (pc3 specific) = PolB – PolB est (PC1, PC2)

Where MLH1 and PolB are the raw (observed) variables, while MLH1 est (PC1, PC2) and PolB est (PC1, PC2) are the least squares estimation of MLH1 and PolB respectively, by means of PC1 and PC2 scores namely MLH1 est = 19.31 + 0.485 (PC1) + 0.644 (PC2), Pearson $r = 0.65$ ($p < 0.0001$); Polβ est = 18.60 + 0.326(PC1) + 0.092(PC2), Pearson $r = 0.56$ ($p < 0.002$).

Given the components are independent of each other by construction, the subtraction of the PC1, PC2 contribution from MLH1 and PolB actual values only keeps alive the PC3 (signal) and noise (minor components) contributions. This allows checking for the statistical significance of the hypothesized inverse correlation of MMR (marked by MLH1) and BER (marked by PolB); this corresponds to asking for a statistically significant correlation holding between MLH1(pc3specific) and PolB(pc3specific). This was actually the case ($r = -0.61$ $p < 0.001$) demonstrating noise (minor components) still allows recognizing the PC3 pathway influence on the two gene expression causing their negative correlation (mutual balance).

5 Conclusions

System's parameter estimation in biology asks for a continuous feedback between biological and procedural information, the data analysis by no way can be considered as a "separately optimized" set of procedures to be applied to a set of experimental results. The focus must be on the underlying (and largely unknown) network linking the different players (in our example different gene expressions) of the system at hand. This network, as such, is the only relevant "causative agent" with the experimental observables acting as probes of the coordinated motion of the underlying network. This peculiar situation (Warren Weaver in a famous 1948 paper [15] named "organized complexity") asks for a completely different style of reasoning with respect to the classical approach of biologists used to a neat dependent/independent variables discrimination and considering the observables as autonomous players in the game.

Complexity can be a blessing and not a curse if we learn how to manage it resisting to the temptation of the direct consideration of "all the agents involved" in model construction.

The most fruitful way is letting the network to suggest us (e.g., by the application of unsupervised techniques like PCA) where to look avoiding the overfitting/irrelevance traps.

References

1. Transtrum MK et al (2015) Perspective: sloppiness and emergent theories in physics, biology and beyond. J Chem Phys 143:01091
2. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
3. Srivastava N et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
4. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29(6–7):476–488
5. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Lond Edinb Dubl Phil Mag J Sci 2(11):559–572
6. Giuliani A (2017) The application of principal component analysis to drug discovery and biomedical data. Drug Discov Today 22 (7):1069–1076
7. Soofi E (1994) Capturing the intangible concept of information. J Am Stat Assoc 89 (428):1243–1254
8. Pascual M, Levin SA (1999) From individuals to population densities: searching for the intermediate scale of nontrivial determinism. Ecology 80(7):2225–2236
9. Broomhead DS, King GP (1986) Extracting qualitative dynamics from experimental data. Physica D 20(2–3):217–236
10. Benigni R, Giuliani A (1994) Quantitative modeling and biology: the multivariate approach. Am J Phys Regul Integr Comp Phys 266(5):R1697–R1704
11. Marwan N et al (2007) Recurrence plots for the analysis of complex systems. Phys Rep 438 (5):237–329
12. Kruskal JB (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29(1):1–27
13. Anderberg MR (2014) Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks, vol 19. Academic, Cambridge
14. Simonelli V et al (2016) Crosstalk between mismatch repair and base excision repair in human gastric cancer. Oncotarget 5. 10.18632/oncotarget.10185
15. Weaver W (1948) Science and complexity. Am Sci 36:536–549

Chapter 6

Inverse Problems in Systems Biology: A Critical Review

Rodolfo Guzzi, Teresa Colombo, and Paola Paci

Abstract

Systems Biology may be assimilated to a symbiotic cyclic interplaying between the forward and inverse problems. Computational models need to be continuously refined through experiments and in turn they help us to make limited experimental resources more efficient. Every time one does an experiment we know that there will be some noise that can disrupt our measurements. Despite the noise certainly is a problem, the inverse problems already involve the inference of missing information, even if the data is entirely reliable. So the addition of a certain limited noise does not fundamentally change the situation but can be used to solve the so-called ill-posed problem, as defined by Hadamard. It can be seen as an extra source of information. Recent studies have shown that complex systems, among others the systems biology, are poorly constrained and ill-conditioned because it is difficult to use experimental data to fully estimate their parameters. For these reasons was born the concept of sloppy models, a sequence of models of increasing complexity that become sloppy in the limit of microscopic accuracy. Furthermore the concept of sloppy models contains also the concept of un-identifiability, because the models are characterized by many parameters that are poorly constrained by experimental data. Then a strategy needs to be designed to infer, analyze, and understand biological systems. The aim of this work is to provide a critical review to the inverse problems in systems biology defining a strategy to determine the minimal set of information needed to overcome the problems arising from dynamic biological models that generally may have many unknown, non-measurable parameters.

Key words Systems biology, Inverse problems, Sloppy models, Identifiability, Reverse engineering

1 Introduction

1.1 The Rationale of Systems Biology and the Inverse Problems

Systems biology is a relatively young discipline that considers the cells as holistic entities. In his paper *Sequences and Consequences* Brenner [1], a founding father of molecular biology, strikes at what he sees as the heart of the goal of systems biology. After reminding us that the systems approach seeks to generate viable models of living systems, Brenner goes on to say that: "Even though the proponents seem to be unconscious of it, the claim of systems biology is that it can solve the inverse problem of physiology by deriving models of how systems work from observations of their behavior. It is known that inverse problems can only be solved

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_6,

under very specific conditions. A good example of an inverse problem is the derivation of the structure of a molecule from the X-ray diffraction pattern of a crystal. The universe of potential models for any complex system like the function of a cell has very large dimensions and, in the absence of any theory of the system, there is no guide to constrain the choice of model."

Then every systems biology project essentially results in a model that tries to solve the problem of divining reality from experimental data. However, a model is not reality, it is an imperfect picture of reality constructed from bits and pieces of data. In addition, data in biological measurements are often noisy with large error and incomplete. Then systems biology may fall in the inverse problems that Brenner points out.

This means, by one side, that models derived from systems biology might be useful, and often this is a sufficient requirement for using them, despite they might likely leave out some important feature of the system. By the other side one of the major challenges in inverse problems is to find a minimal set of parameters that can describe the system under examination or to extract from the models the information included in the system. Ideally those parameters should be sensitive to variation so that one constrains the parameter space describing the given system.

However since the objective is to describe the interactions of distinct molecular entities (for example, proteins, transcripts, or regulatory sites), which give rise to particular cellular behaviors, the current models consist of sets of linear or nonlinear ordinary differential equations involving a high number of states (e.g., concentrations or amounts of the components of the network) and a large number of parameters describing the reaction kinetics.

Unfortunately, in most cases the parameters introduced into the set of equations are completely unknown and/or only rough estimates of their values are available. Therefore, their values are usually estimated from time-series experimental data. The so-called parameter estimation problem is then formulated as an optimization problem where the objective is to find the parameter set so as to minimize a given cost function that relates model predictions and experimental data, e.g. the least squares function or a similar cost index. Furthermore since parameter estimation problems in dynamic models of biochemical systems are characterized by limited observability, large number of parameters and a limited amount of noisy data, the solution of the problem is in general challenging and, even when using robust and efficient optimization methods, computationally expensive.

A particularly promising example is the use of *sloppy models* developed by Sethna and collaborators [2] in which parameter combinations rather than individual parameters are varied and those combinations which are most tightly constrained are then picked as the right ones. Then model building cycle requires the

reconciliation of the underlying hypothesis with experimental data. In the context of systems biology, this implies, in most of the cases, the necessity of identifying unknown kinetic parameters by data fitting. In this concern, Gutenkunst [3] and Gutenkunst et al. [4–6] suggest that dynamic systems biology models are universally sloppy and, thus, parameters cannot be uniquely estimated.

Karlsson et al. [7], Anguelova et al. [8], Raue et al. [9], Oana et al. [10] however, have shown how models regarded as sloppy are structurally identifiable: it means that, in principle, parameters can be given unique values. In the case of structural identifiability, it is only a matter of the experimental constraints and noise that the quality of the parameter estimates may be limited. In this sense, can be analyzed how sloppiness is affected by the experimental setup and experimental noise and can be illustrated, with a number of examples related to biochemical networks, how sloppy models are indeed practically identifiable.

Results indicate that sloppiness does not mean that parameters cannot be estimated and a complete identifiability analysis provides the tools to estimate ranges of parameters which are coherent with experimental data and can then be used to assess quality of predictions.

The notion of identifiability of systems is fundamentally a problem of uniqueness of solutions for specific attributes of certain classes of mathematical models. The identifiability problem usually has meaning in the context of unknown parameters of the model. It is clearly a critical aspect of the modelling process, especially when the parameters are analogs of physical attributes of interest and the model is needed to quantify them.

A parameterization of a subclass of dynamic systems will be called identifiable if, for any finite but sufficiently long time series of observed input–output trajectories, there exists a unique element in the subclass of systems which represents those observations.

2 Materials

2.1 Forward Model Formulation

In systems biology the forward model, in general, is represented from an ensemble of chemical reactions, see, for instance, Dilão and Muraro [11] or Shapiro et al. [12]. Then we may write:

$$\nu_{i1}A_1 + \cdots + \nu_{im}A_m \xrightarrow{r_i} \mu_{i1}A_1 + \cdots + \mu_{im}A_m \tag{1}$$

where $i = 1, \ldots, n$. The A_j, for $j = 1, \ldots, m$, represent, as for example, chemical substances. The constants ν_{ij} and μ_{ij} are the stoichiometric coefficients, in general, non-negative integers, and the constants r_i are the rate constants. If $\nu_{ij} = \mu_{ij} > 0$, the corresponding substance A_j is a catalyst, while if $\mu_{ij} > \nu_{ij} > 0$, A_j is an autocatalyst.

Under the hypothesis of homogeneity of the solution where reactions occur, the mass action law asserts that the time evolution of the concentrations of the chemical substances is described by the system of ordinary differential equations:

$$\frac{dA_j}{dt} = \sum_{i=1}^{n} r_i \left(\mu_{ij} - \nu_{ij}\right) A_1^{\nu_{i1}} \cdots A_m^{\nu_{im}} \tag{2}$$

where $j = 1, \ldots, m$, and the symbols represent both the chemical substance and its concentration; r_i is the rate constant. The rate equations 2 are derived under the following assumptions.

1. Chemical reactions. When they occur, are due to elastic collisions between the reactants.
2. Homogeneity of the reacting substances in the solution.
3. Thermal equilibrium of the solution.

At the atomic and molecular scale, chemical reactions between molecules can occur only if molecules collide or approach each other to small distances where bounding forces become meaningful. These chemical bounding forces are of electrical or quantum origin, and at distances larger than the mean free path they become less important when compared with the kinetics associated with the molecular motion. As chemical reactions only occur if the chemical substances involved collide, the vector fields associated with the right-hand side of Eq. 2 are in general quadratic, representing binary collisions. Higher order polynomial vector fields are possible but, at the microscopic level, they are associated to triple or higher order collisions, a situations that occurs with a very low probability.

The equation 2 can also be written in the matrix form,

$$\frac{dx_j}{dt} = \Gamma\omega(A) \tag{3}$$

where Γ is the $n \times m$ matrix and $A^T = (A_1, \ldots, A_m)$

$$\omega(A) = \begin{pmatrix} r_1(\mu_{1j} - \nu_{1j})A_1^{\nu_{i1}} & \ldots & A_m^{\nu_{1m}} \\ \ldots. & \ldots. & \ldots \\ r_n(\mu_{nj} - \nu_{nj})A_1^{\nu_{n1}} & \ldots & A_m^{\nu_{nm}} \end{pmatrix} \tag{4}$$

in general $n \neq m$. Associated with the differential equations one has the conservation laws:

$$\frac{d}{dt}(A\nu_k) = 0 \tag{5}$$

The input of genetic regulatory networks contains the list of transcriptional activators and repressors of the network. If, in general, the genes are catalytic substances presented in any genetically controlled biological process, the usual threshold concept in

biology is a bifurcation phenomenon of the model equations. These bifurcations are tuned by the conservation law constants of the equations, resulting from the catalytic role of genes.

2.2 Identifiability, Observability, and Sloppiness

In several applications, models' parameters of dynamic systems are not directly measurable but only indirectly accessible through inputs and measurements outputs. Furthermore signals are time varying and are subjected to some applied perturbations. Then a fundamental question to be answered before some methods are selected is if the model structure in question is identifiable. Structural identifiability is a model property that ensures that parameters can be globally or locally determined from knowledge of the input–output behavior of the system. Sedoglavic [13, 14] presents a probabilistic semi-numerical algorithm for testing the local structural identifiability of a model. A-priori non-identifiability may be caused by over-parameterization of the model that includes its observation function; while a-posteriori non-identifiability is generally due to lack of information on the available data. Sloppy models are, however, often unidentifiable, i.e., characterized by many parameters that are poorly constrained by experimental data. In principle, however, these two concepts, identifiability and sloppiness, are distinct.

In general we can consider a state variable with time invariant parameters defined by the following algebraic system

$$\sum \begin{cases} \dot{x}(t) &= f(x(t), u(t), \theta), \quad x(0) = x^0(\theta) \\ y(t) &= g(x(t), u(t), \theta) \end{cases} \tag{6}$$

where $x(t) \in \mathcal{R}^n, u(t) \in \mathcal{R}^m, \theta(t) \in \mathcal{R}^d, y(t) \in \mathcal{R}^p$ and f and g are rational functions of x, u, θ. Higher-order derivatives of the output with respect to time y^{ν} can be obtained by repeated use of the chain rule and replacing $\dot{x}$ using the system dynamics (also known as extended Lie-derivative along f)

$$\begin{aligned} y &= g \\ \dot{y} &= \frac{\partial g}{\partial x} f + \frac{\partial g}{\partial u} \dot{u} = \mathcal{L}_f g \\ \ddot{y} &= \mathcal{L}_f (\mathcal{L}_f g) = \mathcal{L}_f^2 g \\ &\vdots \\ y^{\nu} &= \mathcal{L}_f^{\nu} g \end{aligned}$$

with $\nu = n + d - 1$ and where $\mathcal{L}_f = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} + \sum_{i=0}^{\infty} u^{i+1} \frac{\partial}{\partial u^{(i)}}$ is the formal Lie derivation.

The output derivatives may be expressed in terms of the state and parameters and the inputs and its derivatives as $\mathcal{Y} = \mathcal{Y}(x\theta)$ that can be uniquely solved for x and θ if the Jacobian

$$J(x,\theta) = \frac{\partial \mathcal{Y}(x,\theta)}{\partial(x,\theta)} = \begin{pmatrix} \frac{\partial y}{\partial x_1}\cdots & \frac{\partial y}{\partial x_n}\cdots & \frac{\partial y}{\partial \theta_1}\cdots & \frac{\partial y}{\partial \theta_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y^{nu}}{\partial x_1}\cdots & \frac{\partial y^{nu}}{\partial x_n}\cdots & \frac{\partial y^{nu}}{\partial \theta_1}\cdots & \frac{\partial y^{nu}}{\partial \theta_n} \end{pmatrix} \tag{7}$$

the elements of the Jacobian matrix equals the coefficients of the formal Taylor's series expansion around $t = 0$ of the output sensitivity derivatives with regard to initial conditions and parameters.

The sensitivity equations are:

$$\overset{\star}{\sum} : \begin{cases} \sum : \dot{x} &= f(x,u,\theta), & x(0) = x^0 \\ \frac{d}{dt}\frac{\partial x}{\partial x_i^0} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial x_i^0}, & \frac{\partial x}{\partial x_i^0}(0) = 1_n \\ \frac{d}{dt}\frac{\partial x}{\partial \theta_i} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta_i} + \frac{\partial f}{\partial \theta_i}, & \frac{\partial x}{\partial \theta_i}(0) = 0_d \end{cases} \tag{8}$$

The system $\overset{\star}{\sum}$ can be solved iteratively generating truncated power series solutions of desired order. Insertion into the output sensitivity expressions gives truncated power series:

$$\begin{aligned} \frac{d}{dx_i^0}y(t) &= \frac{\partial g}{\partial x}(x(t),u(t),\theta)\frac{\partial x}{\partial x_i^0}(t) \\ \frac{d}{d\theta_i}y(t) &= \frac{\partial g}{\partial x}(x(t),u(t),\theta)\frac{\partial x}{\partial \theta_i}(t) + \frac{\partial g}{\partial \theta_i}(x(t),u(t),\theta) \end{aligned} \tag{9}$$

We may summarize the possible strategy with the following diagram:

$$\begin{array}{ccc} \sum & \longrightarrow & \left(\frac{\partial(\mathcal{L}_f^i g)_{0\le i\le \nu}}{\partial(x,\theta)}\right) \quad Symbolic \\ \Big\downarrow & & \Big\downarrow \; \begin{array}{l} x \to \tilde{x} \\ \theta \to \tilde{\theta} \\ u(t) \to \tilde{u}(t) \end{array} \\ \overset{\star}{\sum} & \longrightarrow & \left(\frac{\partial(\mathcal{L}_f^i g)_{0\le i\le \nu}}{\partial(x,\theta)}\right) \quad Numeric \\ \begin{array}{l} x \to \tilde{x} \\ \theta \to \tilde{\theta} \\ u(t) \to \tilde{u}(t) \end{array} & & \end{array}$$

2.3 Inverse Problem

Let's consider the computation of an approximation to a solution of a nonlinear operator equation

$$F(x) = y \tag{10}$$

where F: $X \rightarrow \Upsilon$ is an ill-posed operator between Hilbert spaces X, Υ. The inverse problem is to identify the model parameters observed at various time under different experimental conditions if only noisy data y^{δ} are given. Now denoting by x the parameter vector to be determined and with y^{δ} the available noisy data the inverse problem can be formulated with:

$$\left\|y^{\delta} - F(x)\right\|_{\Upsilon}^{2} \rightarrow \min_{x \in X} \tag{11}$$

In recent years, many of the well-known methods for linear ill-posed problems have been generalized to nonlinear operator equations [15].

The iterative methods by Tikhonov regularization is obtained by minimizing the Tikhonov functional

$$\begin{aligned} J_{\alpha}(x) &= \left\|y^{\delta} - F(x)\right\|^{2} + \alpha\left\|x - \overline{x}\right\|^{2} \\ x_{\alpha}^{\delta} &= \operatorname{argmin}_{x} J_{\alpha}(x) \end{aligned} \tag{12}$$

The advantage of Tikhonov regularization is that convergence of the method, i.e. $x^{\delta} \rightarrow x^{\dagger}$ for $\delta \rightarrow 0$ and an appropriate parameter choice $\alpha = \alpha(\delta)$ holds under weak assumptions to the operator. However, the difficulties for Tikhonov regularization are a proper choice of the regularization parameter and the computation of the minimizer of the Tikhonov functional [15]. Other forms of regularization are: Maximum entropy [15] and Bounded variation [16].

2.4 Sloppy Models and Fisher Information Matrix

Dynamic systems biology models involve many kinetic parameters, the quantitative determination of which could be extracted from a fit long before the experimental data constrained the parameters, even to within orders of magnitude. This pattern was attributed to a low sensitivity to model's parameter also revealed by the fact that sensitivity eigenvalues were roughly evenly spaced over many decades. Consequently, the model behavior depended effectively on only a few *stiff* parameter combinations.

Sloppiness is particularly relevant to biology, because the collective behavior of most biological systems is much easier to measure in vivo than the values of individual parameters. Gutenkunst [3], analyzing 17 system biology models drawn from the BioModels database [17], an online repository of models encoded in the Systems Biology Markup Language, have shown that exist models that are poorly constrained and/or ill-conditioned because it is difficult to use experimental data to derive their parameters. These models were called sloppy.

The change in model behavior as parameters y_i varied from their published values $\hat{y}_i$ (SBML) [18] by the average squared change in molecular species. This is accomplished by defining a cost function that quantifies how different the model output is for a

given set of parameters from the experimental data. The precise function is a sum of squared differences between the model and the data, scaled by the experimental error:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i(p) - d_i)^2}{\sigma_i^2} \tag{13}$$

The species are normalized by the term σ that is equal to the maximum value of species across the conditions considered. Thus, less noisy signals are more weighted, and all measurements are brought to the same scale. Since the signals are only available at discrete time points the errors at each measurement time point are summed. When the values of the objective functions attained, as for example, for model A and model B differ only slightly, it is not clear which one of the models is better suited to fitting the benchmark problem.

Then a local approximation method to estimate of the confidence intervals of the parameters should be applied. This may be the Fisher-Information-Matrix [19]. To analyze each model's sensitivity to parameter variation, one considers the Hessian matrix

$$H_{j,k}^{\chi^2} = \frac{\partial^2 \chi^2(r)}{\partial \log r_j \partial \log r_k} \tag{14}$$

that corresponds to approximating the surfaces of constant model to an N_p dimensional ellipsoids, where N_p is the number of the parameters of the model. The principal axes of the ellipsoids are the eigenvectors of the Hessian matrix and the width of the ellipsoids along each principal axis is proportional to one over the square root of the corresponding eigenvalue. The narrowest axes are called *stiff*, and the broadest axes *sloppy*.

Expanding the second derivative:

$$\frac{\partial^2 \chi^2(r)}{\partial \log(p_i) \partial \log(p_j)} = \sum_{i=1}^{N} \left(\frac{\partial r_k}{\partial \log(p_i)} \frac{\partial r_k}{\partial \log(p_j)} + r_k \frac{\partial^2 r_k}{\partial \log(p_i) \partial \log(p_j)} \right) \tag{15}$$

we see that the second term can be dropped in the case of a near-perfect fit when each r_k is small. Denoting the matrix of first derivatives as the Jacobian,

$$J_{kj} = \frac{\partial r_k}{\partial \log(p_j)} \tag{16}$$

we can then make the approximation $H \approx J^T J$. In the Bayesian statistics field this matrix is known as the Fisher Information Matrix

and the Jacobian is referred to as the Design Matrix for the linearized approximation of the full model [20].

Sloppy models are characterized by a logarithmic hierarchy of Fisher Information Matrix eigenvalues while unidentifiable models have in general small eigenvalues. Even though sloppiness and parameter identifiability are closely related, they should be considered as two distinct concepts [21]: sloppy models may be both identifiable and not identifiable and the same appears for not sloppy models.

3 Methods

Here we present an experiment based on a three-step approach: identifiability, sloppiness, and inverse problems. The aim of this experiment is to have the elements to analyze a biology system and to know if what we are analyzing is complete. If so, then we are able to retrieve the main elements of the system.

3.1 The Forward Model

We have selected a simpler case of mitotic oscillator analyzed by Tyson [22], but other models could be analyzed using the same approach.

Tyson states that:

- The proteins cdc2 and cyclin form a heterodimer (maturation promoting factor) that controls the major events of the cell cycle.
- A mathematical model for the interactions of cdc2 and cyclin is constructed.
- Simulation and analysis of the model show that the control system can operate in three modes: as a steady state with high maturation promoting factor (MPF) activity, as a spontaneous oscillator, or as an excitable switch.
- Solutions depend on the values assumed by the ten parameters in the model.
- Nothing is known experimentally about appropriate values for these parameters.
- The focus is on two parameters: k4, the rate constant describing the autocatalytic activation of MPF by dephosphorylation of the cdc2 subunit, and
- k6, the rate constant describing breakdown of the active cdc2-cyclin complex.

The time differential equations are:

$$\begin{aligned}
\frac{dC2}{dt} &= -k8notP \times C2(t) + k9 \times CP(t) + k6 \times M(t) \\
\frac{dCP}{dt} &= k8notP \times C2(t) - k9 \times CP(t) - k3 \times CP(t)\Upsilon(t) \\
\frac{dM}{dt} &= -k5notP \times M(t) - k6 \times M(t) + (k4prime + k4 \times M(t)^2) \times pM(t) \\
\frac{dpM}{dt} &= k5notP \times M(t) - (k4prime + k4 \times M(t)^2) \times pM(t) + k3 \times CP(t)\Upsilon(t) \\
\frac{d\Upsilon}{dt} &= k1aa - k2 \times \Upsilon(t) - k3 \times CP(t) \times \Upsilon(t) \\
\frac{d\Upsilon P}{dt} &= k6 \times M(t) - k7 \times \Upsilon P(t)
\end{aligned} \tag{17}$$

where t is the time and k_i, the rate constant for step $i (i = 1, \ldots, 9)$. aa means amino acids. There are six time-dependent variables: the concentrations of cdc2 ($C2$), cdc2-P (CP), preMPF = P-cyclin-cdc2-P (pM), active MPF = P-cyclin-cdc2 (M), cyclin (Υ), and cyclin-P (ΥP). Active MPF is described by the function F([M]). The equation where active MPF is described is given by the function M that is equal to $(k4prime + k4 \times M)^2$, where $k4prime$ is the rate constant when [active MPF] = 0 and $k4$ is the rate constant when [active MPF] = [CT], where [CT] = total cdc2. Tyson assumes $k4 \gg k4prime$. Related data are reported in Table 1.

The graphic plot is shown in Fig. 1.

Table 1
Data value used in the numerical solution

Parameters	Value
k1aa	0.015 min^{-1}
k2	0 min^{-1}
k3	200 min^{-1}
k4	10–1000 min^{-1}
k4prime	0.018 min^{-1}
k5notP	0
k6	0.1–10 min^{-1}
k7	0.6 min^{-1}
k8notP	$\gg k9$ (10^6)
k9	$\gg k6$ (10^3)

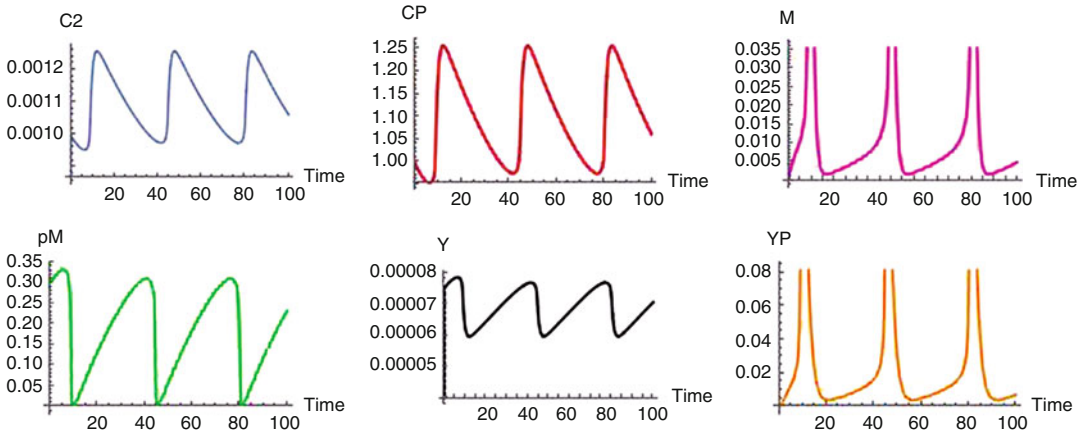


Fig. 1 Dynamical behavior of the components of *cdc2-cyclin* model

3.2 Identifiability

In general a mathematical model helps to better understand the biological phenomenon studied. It enables experiments to be specifically designed to make predictions of certain characteristics of the system that can then be experimentally verified. It summarizes the current body of knowledge in a format that can be easily communicated. Then the model will be conditioned by elements which may have an impact on the questions addressed by the users.

The mathematical assumptions are defined from the network architecture and from the modelling framework like deterministic or stochastic laws, partial differential equations, etc.

A crucial step is that to define a number of unknown non-measurable parameters that can be determined by means of experimental data fitting, the so-called identification. Raue et al. [9] report several methods to be used for identifying parameters. Among others the DAISY approach basic idea is that of manipulating algebraic differential equations as polynomials depending also on derivatives of the variable. This algorithm permits to eliminate the non-observed state variables from the system of equations and to find the input–output relation of the system [23]. The EAR approach developed by Fraunhofer Chalmers [7, 8] is based on a method for local algebraic observability [13]. PL approach checks for non-identifiability by posing a parameter estimation problem using real or simulated data. The central idea is that non-identifiability manifests as a flat manifold in the parameter space of the estimation problem, e.g. the likelihood function.

Here we use the Sedoglavic's approach, modified by Fraunhofer Chalmers as previously mentioned, applied to the Tyson's model. When the same system has the same input but one less output, i.e. $n - 1$ states are measured, for example pM, results reported in Table 2 show that the $k4$, $k4prime$, and $k5notP$ are unidentifiable. On the contrary all data are uniquely identifiable.

Table 2
Structurally identifiable parameters and parameter combinations

Parameter	Note
k6	Is uniquely identifiable
k7	Is uniquely identifiable
k9	Is uniquely identifiable
k8notP	Is uniquely identifiable
k2	Is uniquely identifiable
k3	Is uniquely identifiable
k1aa	Is uniquely identifiable
k4	Unidentifiable
k4prime	Unidentifiable
k5notP	Unidentifiable

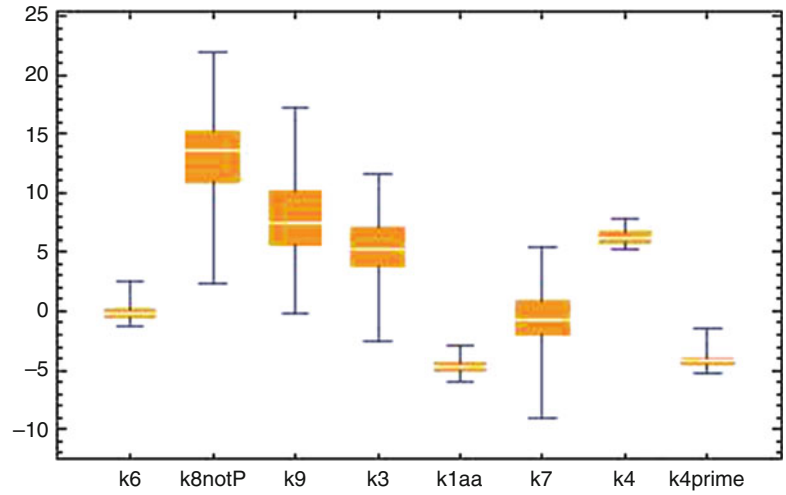


Fig. 2 Parameters confidence intervals in a semilog box plot

3.3 Sloppiness

In order to detect the level of sloppiness of the parameters involved in the Tyson' model we applied the procedure designed by Gutenkunst et al. [6] only for the nonzero parameters. Results are shown in Fig. 2.

3.4 Inverse Model

The procedure of inverse problems have been developed adding a white noise on parameters. The key is, of course, the simultaneous fitting of all ten datasets. Results are reported in Table 3 with related statistics.

Table 3
Parameters retrieval, standard error, *t* statistic and *P* value

Value	Estimate	Standard error	t-Statistic	P-value
k1aa	0.015001	0.000938491	15.9841	4.01224×10^{-48}
k2	-2.37146×10^{-7}	0.00522666	−0.0000453724	0.999964
k3	200.001	0.960852	208.149	$1.049133302926 \times 10^{-558}$
k4	180.	1.4801	121.613	$6.100160101569 \times 10^{-423}$
k4prime	0.0179989	0.000843644	21.3347	1.81877×10^{-75}
k5notP	3.1582×10^{-7}	0.00397082	0.0000795352	0.999937
k6	1.	0.00343817	290.853	$1.911259397235 \times 10^{-644}$
k7	0.6	0.00281254	213.33	$5.507756176045 \times 10^{-565}$
k8notP	1.00001×10^{6}	5864.59	170.517	$6.067726632465 \times 10^{-508}$
k9	999.999	2.45854	406.745	$8.26166736095 \times 10^{-73}$

3.5 Inverse Bifurcation

Bifurcation analysis has proven to be a powerful method for understanding the qualitative behavior of gene regulatory networks [24]. In addition to the more traditional forward problem of determining the mapping from parameter space to the space of model behavior, the inverse problem of determining model parameters to result in certain desired properties of the bifurcation diagram provides an attractive methodology for addressing important biological problems. For a certain range of a bifurcation parameter, three steady states coexist, whereas outside this interval only a single steady state exists. This implies that slowly varying parameters can induce a sudden jump. Manipulating the Tyson's model, assuming the total cdc2 $CT = C2 + CP + pM + M = \text{const}$ and $CP = (1 - w)CT - C2$ and $C2 \ll 1$ one rewrites Eq. 17 as $u = [M]/[CT]$, $\nu = ([\Upsilon] + [pM] + [M])/[CT]$, $w = ([pM] + [M])/[CT]$, and $y = [\Upsilon T]/[CT]$ and that the first three equations can be solved independently of the fourth because y does not appear in the first three equations; and because $k3[CT] \gg \max\{kiaa, k2, k6\}$, w changes very rapidly compared to changes in ν, so $w = \nu$ as long as $0 < \nu < 1$. Thus, the cdc2-cyclin model reduces to a pair of nonlinear ordinary differential equations

$$\begin{aligned} \dot{u}[t] &= k4(\nu - u)(\alpha + u^2)(\alpha + u^2) - k6u \\ \dot{\nu}[t] &= (k1aa - k6u) \end{aligned} \tag{18}$$

where $\alpha = k4prime/k4$. Applying the mathematics rules shown in Appendix under Bifurcation, results of such model are shown in Fig. 3.

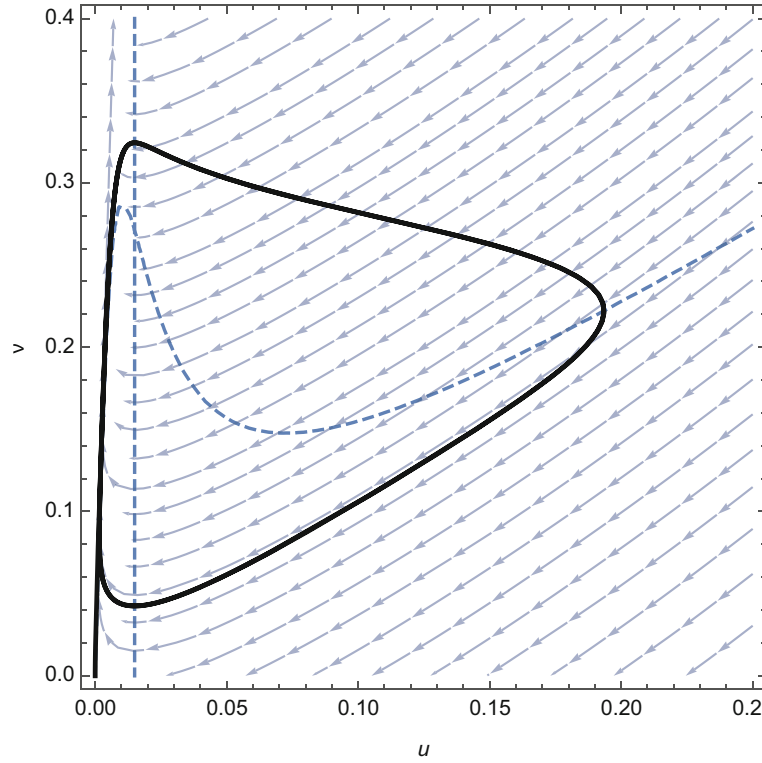


Fig. 3 A phase portrait of the Tyson's model

Lu et al. [25], applying a methodology based on the inverse map from the space of bifurcation diagrams to biological space parameters, has shown that many questions of biological interest may be formulated by the inverse bifurcation defined as an optimization problem which involves minimal distances to bifurcation manifold.

4 Conclusions

Inverse problems and sloppy models are only part of the more general problem approach of inverse problems applied to systems biology. Despite many mathematical tools are available for specific problems posed by systems biology studies, as reviewed among others by Villaverde and Banga [26], the core problem of how to derive system level organization from experimental results remains largely unsolved. Accordingly, the modelling of biological systems is far to be treated as a complete engineered system, thus the use of the reverse engineering approach appears to be a difficult road because different perspectives of systems biology coexist. Large-

scale dynamic biological models generally may have many unknown, non-measurable parameters and their tuning may appear unrealistic. Nevertheless if on one side some authors have shown, among others we cite Paci et al. [27], that it is possible to explore the biological networks by reverse engineering provided the analysis of classifying the nodes in the network is defined as a whole, on the other side, other authors, among whom we cite Villaverde et al. [28], have compared different parameter estimations methods with the aim to have a benchmark for optimal experimental design. Then in this paper we have tried to define the minimal approach to understand how a biological system can be studied without forcing, with tuned parameters, the biological nature of process involved.

Acknowledgements

The authors thank Giuseppe Macino and Lorenzo Farina for their inspiring discussions on the whole analysis presented in this study.

Appendix A: Notes

A.1 Deterministic Solutions

This appendix can be an aide to characterize the system biology, thanks to mathematical models which may consist of a system of differential equations involving state variables and parameters. If the variables of such system do not depend explicitly on time, the system is said to be autonomous. Under certain assumptions defined below such system may be considered as an autonomous dynamical system. As time does not occur explicitly in equations, solution of a system of differential equations may be projected in a space called phase-space in which the behavior of the state variables is described.

The plot of the solution in such space is called phase portrait. The bifurcation of a system of differential equations, i.e., of an autonomous dynamical system is concerned with changes in the qualitative behavior of its phase portrait as parameters vary and more precisely, when such a bifurcation parameter reaches a certain value, called critical value. Thus, bifurcation theory is of great importance in dynamical systems study because it indicates stability changes, structural changes in a system, etc. So, plotting the solution of autonomous dynamical system according to the bifurcation parameter leads to the construction of a bifurcation diagram. Such diagram provides knowledge on the behavior of the solution: constant, periodic, nonperiodic, or even chaotic. As there are many kinds of behaviors of solutions there are many kinds of bifurcations. The Hopf bifurcation corresponds to periodic solutions and period doubling bifurcation, or period doubling cascade, which is one of the routes to chaos for dynamical systems.

A.2 Bifurcation Concepts

A bifurcation occurs when a small smooth change made to the parameter values (the bifurcation parameters) of a system causes a sudden qualitative or topological change in its behavior. Generally, at a bifurcation, the local stability properties of equilibria, periodic orbits or other invariant sets changes. It has two types; *Local bifurcations*, which can be analyzed entirely through changes in the local stability properties of equilibria, periodic orbits or other invariant sets as parameters cross through critical thresholds; and *Global bifurcations*,which often occur when larger invariant sets of the system collide with each other, or with equilibria of the system. They cannot be detected purely by a stability analysis of the equilibria (fixed or equilibrium points, see the next section).

In dynamical systems, only the solutions of linear systems may be found explicitly. Unfortunately, real life problems can generally be modelled only by nonlinear systems The main idea is to approximate a nonlinear system by a linear one (around the equilibrium point).

A.3 Linear Stability Analysis

Bifurcations indicate qualitative changes in a system's behavior. For a dynamical system $\frac{dy}{dt} = f(y, \lambda)$, bifurcation points are those equilibrium points at which the Jacobian $\frac{\partial f}{\partial y}$ is singular. For definition consider a nonlinear differential equation

$$\dot{x}(t) = f(x(t), u(t)), \tag{19}$$

where f is a function mapping $R \times R^3 \rightarrow R^n$. A point $\overline{x}$ is called an equilibrium point if there is a specific $\overline{u} \in R^m$ such that

$$f(x(t), u(t)) = 0_n. \tag{20}$$

Suppose $\overline{x}$ is an equilibrium point (with the input $\overline{u}$). Consider the initial condition $x(0) = \overline{x}$, and applying the input $u(t) = \overline{u}$ for all $t \leq t_0$, then resulting solution $x(t)$ satisfies

$$x(t) = \overline{x}, \tag{21}$$

for all $t \leq t_0$. That is why it is called an equilibrium point or solution.

Linear stability of dynamical equations can be analyzed in two parts: one for scalar equations and the other for two dimensional systems

1. Linear stability analysis for scalar equations

   To analyze the Ordinary Differential Equations (ODE)

   $$\dot{x} = f(x) \tag{22}$$

   locally about the equilibrium point $x = \overline{x}$, we expand the function $f(x)$ in a Taylor series about the equilibrium point $\overline{x}$. To emphasize that we are doing a local analysis, it is customary

to make a change of variables from the dependent variable x to a local variable. Now let:

$$x(t) = \overline{x} + \epsilon(t), \tag{23}$$

where it is assumed that $\epsilon(t) \ll 1$, so that we can justify dropping all terms of order two and higher in the expansion. Substituting $x(t) = \overline{x} + \epsilon(t)$ into the Right-Hand Side (RHS) of the ODE yields

$$\begin{aligned} f(x(t)) &= f(\overline{x} + \epsilon(t)) = f(\overline{x}) + f(\overline{x})\epsilon(t) + f(\overline{x})\frac{\epsilon^2(t)}{2} + \cdots \\ &= 0 + f(\overline{x})\epsilon(t) + O(\epsilon^2), \end{aligned} \tag{24}$$

and dropping higher order terms, we obtain

$$f(x) \approx f(\overline{x})\epsilon(t). \tag{25}$$

Note that dropping these higher order terms is valid since $\epsilon(t) \ll 1$. Now substituting $x(t) = \overline{x} + \epsilon(t)$ into the Left-Hand Side (LHS) of the ODE,

$$\epsilon'(t) = f(\overline{x})\epsilon(t). \tag{26}$$

The goal is to determine if we have growing or decaying solutions. If the solutions grows, then the equilibrium point is unstable. If the solution decays, then the fixed point is stable. To determine whether or not the solution is stable or unstable we simply solve the ODE and get the solution as

$$\epsilon(t) = \epsilon_0 \exp((\overline{x})\epsilon(t)), \tag{27}$$

where ϵ_0 is a constant. Hence, the solution is growing if $\dot{f}(\overline{x}) > 0$ and decaying if $\dot{f}(\overline{x}) < 0$. As a result, the equilibrium point is stable if $\dot{f}(\overline{x}) < 0$, unstable if $\dot{f}(\overline{x}) > 0$.

A first-order autonomous ODE with a parameter r has the general form $dx/dt = f(x, r)$. The fixed points are the values of x for which $f(x, r) = 0$. A bifurcation occurs when the number or the stability of the fixed points changes as system parameters change. The classical types of bifurcations that occur in nonlinear dynamical systems are produced from the following prototypical differential equations:

- saddle: $dx/dt = r + x^2$. A saddle-node bifurcation or tangent bifurcation is a collision and disappearance of two equilibria in dynamical systems. In autonomous systems, this occurs when the critical equilibrium has one zero eigenvalue. This phenomenon is also called fold or limit point bifurcation. An equilibrium solution (where $x = 0$) is simply $x = p\backslash\mathrm{pm}\sqrt{r}$. Therefore, if $r < 0$, then we have no real solutions, if $r > 0$, then we have two real solutions.

We now consider each of the two solutions for $r > 0$, and examine their linear stability in the usual way. First, we add a small perturbation:

$$x = \overline{x} + \epsilon.$$

Substituting this into the equation yields

$$\frac{d\epsilon}{dt} = \left(r - \overline{x}^2\right) - 2\overline{x}\epsilon - \epsilon^2, \tag{28}$$

and since the term in brackets on the RHS is trivially zero, therefore

$$\frac{d\epsilon}{dt} = -2\overline{x}\epsilon,$$

which has the solution

$$\epsilon(t) = A\exp(-2\overline{x}t).$$

From this, we see that for $x = +\sqrt{r} \mid x \mid \to 0$ as $t \to \infty$ (linear stability); for $x = -\sqrt{r} \quad |x| \to 0$ as $t \to \infty$ (linear instability).

In a typical "bifurcation diagram," therefore, the saddle node bifurcation at $r = 0$ corresponds to the creation of two new solution branches. One of these is linearly stable, the other is linearly unstable.

Let's do the same for the next

- transcritical: $dx/dt = rx - x^2$
- supercritical pitchfork super: $dx/dt = rx - x^3$
- subcritical pitchfork sub: $dx/dt = rx + x^3$

and easily we find the relative stability and instability.

2. Linear stability analysis for systems
Consider the two-dimensional nonlinear system

$$\begin{aligned} \dot{x} &= f(x, y), \\ \dot{y} &= g(x, y), \end{aligned} \tag{29}$$

and suppose that $(\overline{x}, \overline{y})$ is a steady state (equilibrium point), i.e., $f(\overline{x}, \overline{y}) = 0$ and $g(\overline{x}, \overline{y}) = 0$. Now let's consider a small perturbation from the steady state $(\overline{x}, \overline{y})$

$$\begin{aligned} x &= \overline{x} + u, \\ y &= \overline{y} + v, \end{aligned} \tag{30}$$

where u and v are understood to be small as $u \ll 1$ and $v \ll 1$. It is natural to ask whether u and v are growing or decaying so that x and y will move away form the steady state or move towards the steady states. If it moves away, it is called unstable equilibrium point, if it moves towards the equilibrium point,

then it is called stable equilibrium point. As in scalar equations, by expanding the Taylor's series for $f(x, y)$ and $g(x, y)$;

$$\begin{aligned}\dot{u} = \dot{x} &= f(x, y)\\ &= f(\overline{x} + u, \overline{y} + v)\\ &= f(\overline{x}, \overline{y}) + f_x(\overline{x}, \overline{y})u + f_y(\overline{x}, \overline{y})v + \textit{higher order terms}\ldots\\ &= f_x(\overline{x}, \overline{y})u + f_y(\overline{x}, \overline{y})v + \textit{higher order terms}\ldots.\end{aligned} \tag{31}$$

Similarly,

$$\begin{aligned}\dot{v} = \dot{y} &= g(x, y)\\ &= g(\overline{x} + u, \overline{y} + v)\\ &= g(\overline{x}, \overline{y}) + g_x(\overline{x}, \overline{y})u + g_y(\overline{x}, \overline{y})v + \textit{higher order terms}\ldots\\ &= g_x(\overline{x}, \overline{y})u + g_y(\overline{x}, \overline{y})v + \textit{higher order terms}\ldots.\end{aligned} \tag{32}$$

Since u and v are assumed to be small, the higher order terms are extremely small, we can neglect the higher order terms and obtain the following linear system of equations governing the evolution of the perturbations u and v,

$$\begin{bmatrix}\dot{u}\\ \dot{v}\end{bmatrix} = \begin{bmatrix}f_x(\overline{x}, \overline{y}) & f_y(\overline{x}, \overline{y})\\ g_x(\overline{x}, \overline{y}) & g_y(\overline{x}, \overline{y})\end{bmatrix}\begin{bmatrix}u\\ v\end{bmatrix}$$

where the matrix

$$\begin{bmatrix}f_x(\overline{x}, \overline{y}) & f_y(\overline{x}, \overline{y})\\ g_x(\overline{x}, \overline{y}) & g_y(\overline{x}, \overline{y})\end{bmatrix}$$

is called Jacobian matrix J of the nonlinear system, where the raws of the Jacobian are the derivatives computed in the steady state. The above linear system for u and v has the trivial steady state $(u, v) = (0, 0)$, and the stability of this trivial steady state is determined by the eigenvalues of the Jacobian matrix at the equilibrium point $(0, 0)$ where $J(0, 0)$ give the eigenvalues by solving the characteristic equation $\det(J - \lambda I) = 0$, where I is the identity matrix and λ are the eigenvalues.

As a summary,

- Asymptotically stable. A critical point is asymptotically stable if all eigenvalues of the jacobian matrix J are negative, or have negative real parts.
- Unstable. A critical point is unstable if at least one eigenvalue of the jacobian matrix J is positive, or has positive real part.

- Stable (or neutrally stable). Each trajectory move about the critical point within a finite range of distance.
- Definition(Hyperbolic point). The equilibrium is said to be hyperbolic if all eigenvalues of the jacobian matrix have nonzero real parts.
- Hyperbolic equilibria are robust (i.e., the system is structurally stable). Small perturbations of order do not change qualitatively the phase portrait near the equilibria. Moreover, local phase portrait of a hyperbolic equilibrium of a nonlinear system is equivalent to that of its linearization. This statement has a mathematically precise form known as the Hartman-Grobman. This theorem guarantees that the stability of the steady state $(\overline{x}, \overline{y})$ of the nonlinear system is the same as the stability of the trivial steady state $(0, 0)$ of the linearized system.
- Definition(Non-Hyperbolic point). If at least one eigenvalue of the Jacobian matrix is zero or has a zero real part, then the equilibrium is said to be non-hyperbolic. Non-hyperbolic equilibria are not robust (i.e., the system is not structurally stable). Small perturbations can result in a local bifurcation of a non-hyperbolic equilibrium, i.e., it can change stability, disappear, or split into many equilibria. Some refer to such an equilibrium by the name of the bifurcation.

A.4 Applications to Two Nonlinear Equations System

In the study of nonlinear dynamics, it is useful to first introduce a simple system that exhibits periodic behavior as a consequence of a Hopf bifurcation. The two-dimensional nonlinear and autonomous system given by

$$\begin{aligned} \dot{x} &= f_1(x, y) = -x + ay + x^2 y, \\ \dot{y} &= f_2(x, y) = b - ay - x^2 y \end{aligned} \tag{33}$$

has this feature. These equations describe the autocatalytic reaction of two intermediate species x and y in an isothermal batch reactor, when the system is far from equilibrium. In this context, the steady state referred to below is a pseudo steady state, and is applicable when the precursor reactant is slowly varying with time.

The unique steady state is given by $x_S = b$ and $y_S = b/(a + b^2)$. This steady state is at the position of the green dot in the phase portrait diagram. It appears as the intersection of the dotted blue and green curves, which are the level curves given by $f_1(x, y) = 0$ and $f_2(x, y) = 0$.

The stability of steady state to small disturbances can be assessed by determining the eigenvalues of the Jacobian J

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix} \tag{34}$$

A.5 Hopf Bifurcation

Using the following procedure it is possible to compute the Hopf bifurcation parameter value of two or three-dimensional dynamical systems. Since the two-dimensional procedure may be obtained by a simple reduction, the three-dimensional procedure is only presented. A Hopf bifurcation occurs when a complex conjugate pair of eigenvalues crosses the imaginary axis.

The phase portrait with the vector field of directions around the critical point (x_S, y_S) may be simply obtained. In addition, the eigenvalues of J, the trace $trace(J)$, the determinant $|\ J\ |$, and $\Delta = trace(J)^2 - 4|J|$ may be computed as far as the time series corresponding to $x(t)$ and $y(t)$, respectively and the real and imaginary parts of the two eigenvalues λ_1 and λ_2. Then to observe a Hopf bifurcation may be defined the parameters.

A.6 Nonlinear Equations System

Newton's method can be used to solve systems of nonlinear equations. Newton-Raphson Method for two-dimensional Systems.

To solve the nonlinear system $\mathbf{F(X)} = 0$, given one initial approximation $\mathbf{P}_0$, and generating a sequence $\mathbf{P}_k$ which converges to the solution $\mathbf{P}_i$ i.e. $\mathbf{F(X)} = 0$.

Suppose that $\mathbf{P}_k$ has been obtained, use the following steps to obtain $\mathbf{P}_{k+1}$.

1. Evaluate the function

$$\mathbf{F}(\mathbf{P}_k) = \begin{pmatrix} f_1(p_k, q_k) \\ f_2(p_k, q_k) \end{pmatrix} \tag{35}$$

2. Evaluate the Jacobian

$$\mathbf{J}(\mathbf{P_k}) = \begin{pmatrix} \frac{\delta}{\delta x} f_1(p_k, q_k) & \frac{\delta}{\delta x} f_1(p_k, q_k) \\ \frac{\delta}{\delta y} f_2(p_k, q_k) & \frac{\delta}{\delta y} f_2(p_k, q_k) \end{pmatrix} \tag{36}$$

3. Solve the linear system

$$\mathbf{J}(\mathbf{P_k})\Delta\mathbf{P}_k = -\mathbf{F}(\mathbf{P_k}) \quad \text{for} \quad \Delta\mathbf{P} \tag{37}$$

4. Compute the next approximation

$$\mathbf{P}_{k+1} = \mathbf{P}_k + \Delta\mathbf{P}_k \tag{38}$$

A.7 Singular Value Decomposition

Every $\mathbf{A}$, matrix $m \times n$, $m \geq n$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \tag{39}$$

where $(.)^T$ denotes the transposed matrix and $\mathbf{U}$ is $m \times n$ matrix, $\mathbf{V}$ is $n \times n$ matrix satisfying

$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_n \tag{40}$$

and $\Sigma = E[\sigma_1, \ldots, \sigma_n]$ a diagonal matrix.

These σ_i's, $\sigma_1/g \geq \sigma_2 \geq, \ldots, \sigma_n \geq 0$ are the square root of the nonnegative eigenvalues of $\mathbf{A}^T\mathbf{A}$ and are called as the singular values of matrix A. As it is well known from linear algebra, see i.e., Press et al. [29] singular value decomposition is a technique to compute pseudoinverse for singular or ill-conditioned matrix of linear systems. In addition this method provides least square solution for overdetermined system and minimal norm solution in case of undetermined system.

The pseudoinverse of a matrix $\mathbf{A}$, $m \times n$ is a matrix $\mathbf{A}^+$, $n * \times m$ satisfying

$$\begin{aligned}\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}, \mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+, (\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A}, (\mathbf{A}\mathbf{A}^+)^* \\ = \mathbf{A}\mathbf{A}^+\end{aligned} \tag{41}$$

where $(.)^*$ denotes the conjugate transpose of the matrix.

Always exists a unique $\mathbf{A}^+$ which can be computed using SVD:

1. If $m >= n$ and $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, then

$$\mathbf{A}^+ = \mathbf{V}\Sigma^{-1}\mathbf{U}^T \tag{42}$$

   where $\Sigma^{-1} = E[1/\sigma_1, \ldots, 1/\sigma_n]$

2. If $m < n$, then compute the $(\mathbf{A}^T)^+$, pseudoinverse of $\mathbf{A}^T$ and then

$$\mathbf{A}^+ = ((\mathbf{A}^T)^+)^T \tag{43}$$

A.8 Newton-Raphson Method with Pseudoinverse

The idea of using pseudoinverse in order to generalize of Newton method is not new but has been suggested by different authors, among others we may cite Haselgrove [30]. It means that in the iteration formula, the pseudoinverse of the Jacobian matrix will be employed,

$$x_{i+1} = x_i - J^+(x_j)f(x_i) \tag{44}$$

A.9 Stochastic Resonance

Since several times the system shows a bistability one can consider the following stochastic equation:

$$dX = \frac{\partial U(X,t)}{\partial U(X)}dt + \eta dW_t \tag{45}$$

where dW_t stands for a Wiener process and η represents the noise level.

Now consider potentials of the form $U(X, t) = U_o(X) + \epsilon X\cos(2\pi t/\tau)$, composed of a stationary part U_o with two minima at X^- and X^+ and a periodic forcing with amplitude ϵ and period τ. If ϵ is small enough, X will oscillate around either X^- or X^+, without ever switching to the other.

But what happens if one increases the noise amplitude η? Then there is some probability that X will jump from one basin to the other. If the noise level is just right, X will follow the periodic forcing and oscillate between $X^-$and X^+ with period τ. This is what we mean by stochastic resonance.

In more general terms, there is *stochastic resonance* whenever adding noise to a system improves its performance or, in the language of signal processing, increases its signal-to-noise ratio. Note that the noise amplitude cannot be too large or the system can become completely random.

A.10 Stochastic Solutions

The deterministic dynamics of populations in continuous time are traditionally described using coupled, first-order ordinary differential equations. While this approach is accurate for large systems, it is often inadequate for small systems where key species may be present in small numbers or where key reactions occur at a low rate. The Gillespie stochastic simulation algorithm (SSA) [31] is a procedure for generating time-evolution trajectories of finite populations in continuous time and has become the standard algorithm for these types of stochastic models. It is well known that stochasticity in finite populations can generate dynamics profoundly different from the predictions of the corresponding deterministic model. For example, demographic stochasticity can give rise to regular and persistent population cycles in models that are deterministically stable and can give rise to molecular noise and noisy gene expression in genetic and chemical systems where key molecules are present in small numbers or where key reactions occur at a low rate. Because analytical solutions to stochastic time-evolution equations for all but the simplest systems are intractable, while numerical solutions are often prohibitively difficult, stochastic simulations have become an invaluable tool for studying the dynamics of finite biological, chemical, and physical systems.

The Gillespie stochastic simulation algorithm (SSA) is a procedure for generating statistically correct trajectories of finite well-mixed populations in continuous time. The trajectory that is

produced is a stochastic version of the trajectory that would be obtained by solving the corresponding stochastic differential equations.

A.11 The Gillespie SSAs

The SSA assumes a population consisting of a finite number of individuals distributed over a finite set of discrete states. Changes in the number of individuals in each state occur due to reactions between interacting states.

Given an initial time t_0 and initial population state $X(t_0)$, the SSA generates the time evolution of the state vector $X(t)$, $(X_1(t), \ldots, X_N(t))$ where $X_i(t)$, $i = 1, \ldots, N$, is the population size of state i at time t and N is the number of states. The states interact through M reactions R_j where $j = 1, \ldots, M$ denotes the jth reaction. A reaction is defined as any process that instantaneously changes the population size of at least one state. Each reaction R_j is characterized by two quantities. The first is its state-change vector $\nu_j = (\nu_{1j}, \ldots, \nu_{Nj})$, where ν_{ij} is the population change in state i caused by one R_j reaction. In other words, if the system is in state x, assuming $x = X(t)$, and one R_j reaction occurs, the system instantaneously jumps to state $x + \nu_j$. The second component of R_j is its propensity function $a_j(x)$ which is the probability of one R_j reaction occurring in the infinitesimal time interval $[t; t + dt]$.

Appendix B: Software Tools

Beyond personal script to manage data, we have also used and tested several public domain softwares, based on Qt and Python. Here we suggest those we consider more flexible and accurate.

- **COPASI** is a Qt software application for simulation and analysis of biochemical networks and their dynamics. COPASI is part of de.NBI, the "German Network for Bioinformatics Infrastructure". It is a stand-alone program that supports models in the SBML standard and can simulate their behavior using ODEs or Gillespie's stochastic simulation algorithm; arbitrary discrete events can be included in such simulations.http://copasi.org/
- **PyCoTools** a COPASI based tool, in Python, for parameter estimation and identifiability.https://github.com/CiaranWelsh/PyCoTools
- **SloppyCell** is a Python software environment for simulation and analysis of biomolecular networks, mainly developed for sloppy models.http://sloppycell.sourceforge.net/
- **Pycellerator** provides python libraries, a command line interface, and an ipython notebook interface for Cellerator arrow notation.https://github.com/biomathman/pycellerator

- **Inverse problem solving tools for solving inverse problems**. An open-source Python library for modelling and inversion in geophysics.http://www.fatiando.org/v0.1/api/inversion.html
- **PyDSTool** is a sophisticated and integrated simulation and analysis environment for dynamical systems models of physical systems (ODEs, DAEs, maps, and hybrid systems and bifurcation). PyDSTool is platform independent, written primarily in Python with some underlying C and Fortran legacy code for fast solving. PyDSTool supports symbolic math, optimization, phase plane analysis, continuation and bifurcation analysis, data analysis, and other tools for modelling—particularly for biological applications.http://www.ni.gsu.edu/~rclewley/PyDSTool/FrontPage.html

We have not had any problems to install and run the above mentioned software, but check the different releases be compatible. The python version was 2.7.

References

1. Brenner S (2010) Sequences and consequences. Philos Trans R Soc B Biol Sci 365:207–212. doi:10.1098/rstb.2009.0221
2. Sethna J. http://sethna.lassp.cornell.edu/research/what_is_sloppiness
3. Gutenkunst RN (2008) Sloppiness, modelling and evolution in biochemical networks. Thesis, Cornell University, Ithaca
4. Gutenkunst RN, Casey FP, Waterfall JJ, Myers CR, Sethna JP (2007) Extracting falsifiable predictions from sloppy models. Reverse engineering biological networks: opportunities and challenges in computational methods for pathway inference. Ann N Y Acad Sci 1115:203–211. doi:10.1196/annals.1407.003
5. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP (2007) Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol 3:1871–1878. doi:10.1371/journal.pcbi.0030189
6. Gutenkunst RN, Casey FP, Waterfall JJ, Atlas JC, Kuczenski RS (2007) SloppyCell. http://sloppycell.sourceforge.net
7. Karlsson J, Anguelova M, Jirstrand M (2012) An efficient method for structural identifiability analysis of large dynamic systems. In: 16th IFAC proc. vol., pp 941–946. doi:dx.doi.org/10.3182/20120711-3-BE-2027.00381
8. Anguelova M, Karlsson J, Jirstrand M (2012) Minimal output sets for identifiability. Math Biosci 239:139–153. doi:10.1016/j.mbs.2012.04.005
9. Raue A, Karlsson J, Saccomani MP, Jirstrand M, Timmer J (2014) Comparison of approaches for parameter identifiability analysis of biological systems. Bioinformatics 30:1440–1448. doi:10.1093/bioinformatics/btu006
10. Oana CT, Banga JR, Balsa-Canto E (2011) Structural identifiability of systems biology models: a critical comparison of methods. PloS One 6:e27755. doi:10.1371/journal.pone.0027755
11. Dilão R, Muraro D (2010) A software tool to model genetic regulatory networks. Applications to the modeling of threshold phenomena and of spatial patterning in Drosophila. PLoS One 5(5). doi:dx.doi.org/10.1371/journal.pone.0010743
12. Shapiro BE, Levchenko A, Wold BJ, Meyerowitz EM, Mjolsness ED (2003) Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction modeling. Bioinformatics 19(5):677–678. doi:10.1093/bioinformatics/btg042
13. Sedoglavic A (2002) A probabilistic algorithm to test local algebraic observability in polynomial time. J Symb Comput 33:735–755. http://dx.doi.org/10.1006/jsco.2002.0532
14. Sedoglavic A (2007) Reduction of algebraic parametric systems by rectification of their affine expanded lie symmetries. In: Proceedings of 2nd international conference on algebraic biology, 2–4 July 2007. doi:10.1007/978-3-540-73433-8_20

15. Guzzi R (2012) Introduction to inverse methods with applications to geophysics and remote sensing (in Italian). Earth sciences and geography series. Springer, New York. http://www.springer.com/it/book/9788847024946
16. Ambrosio L, Dal Maso G (1990) A general chain rule for distributional derivatives. Proc Am Math Soc 108:691–702. doi:10.1090/S0002-9939-1990-0969514-3
17. Li C, Donizelli M, Rodriguez N et al (2010) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol 4:92. doi:10.1186/1752-0509-4-92
18. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531. doi:10.1093/bioinformatics/btg015
19. Orr HA (2006) The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. J Theor Biol 238:279–285. doi:10.1016/j.jtbi.2005.05.001
20. Waterfall JJ (2006) Universality in multiparameter fitting: sloppy models. Ph.D. thesis, Cornell University, Ithaca, New York
21. White A, Tolman M, Thames HD, Withers HR, Mason KA, Transtrum MK (2016) The limitations of model-based experimental design and parameter estimation in sloppy systems. PLoS Comput Biol 12:e1005227. doi:10.1371/journal.pcbi.1005227
22. Tyson JJ (1991) Modeling the cell division cycle: cdc2 and cyclin interactions. Proc Natl Acad Sci USA 88:7328–7332
23. Bellu G, Saccomani MP, Audoly S, D'Angiò L (2007) DAISY: a new software tool to test global identifiability of biological and physiological systems. Comput Methods Prog Biomed 88:52–61. doi:10.1016/j.cmpb.2007.07.002
24. Song C, Phenix H, Abedi V, Scott M, Ingalls BP, Kaern M, Perkins TJ (2010) Estimating the stochastic bifurcation structure of cellular networks. PLoS Comput Biol 6. doi:10.1371/journal.pcbi.1000699
25. Lu J, Engl HW, Schuster P (2006) Inverse bifurcation analysis: application to simple gene systems. Algorithms Mol Biol 1:11. doi:10.1186/1748-7188-1-11
26. Villaverde AF, Banga JR (2014) Reverse engineering and identification in systems biology: strategies, perspectives and challenges. J R Soc Interface 11. doi:10.1098/rsif.2013.0505
27. Paci P, Colombo T, Fiscon G, Gurtner A, Pavesi G, Farina L (2017) SWIM: a computational tool to unveiling crucial nodes in complex biological networks. Sci Rep 7:44797. doi:10.1038/srep44797
28. Villaverde A, Henriques D, Smallbone K, Bongard S, Schmid J, Cicin-Sain D, Crombach A, Saez-Rodriguez J, Mauch K, Balsa-Canto E, Mendes P, Jaeger J, Banga JR (2015) BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology. BMC Syst Biol 9:8. doi:10.1186/s12918-015-0144-4
29. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, Cambridge
30. Haselgrove CB (1961) The solution of nonlinear equations and of differential equations with two-point boundary conditions. Comput J 4:255–259
31. Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55. doi:10.1146/annurev.physchem.58.032806.104637

Chapter 7

Systems Biology Approach and Mathematical Modeling for Analyzing Phase-Space Switch During Epithelial-Mesenchymal Transition

Chiara Simeoni, Simona Dinicola, Alessandra Cucina, Corrado Mascia, and Mariano Bizzarri

Abstract

In this report, we aim at presenting a viable strategy for the study of Epithelial-Mesenchymal Transition (EMT) and its opposite Mesenchymal-Epithelial Transition (MET) by means of a Systems Biology approach combined with a suitable Mathematical Modeling analysis. Precisely, it is shown how the presence of a metastable state, that is identified at a mesoscopic level of description, is crucial for making possible the appearance of a phase transition mechanism in the framework of fast-slow dynamics for Ordinary Differential Equations (ODEs).

Key words Epithelial-mesenchymal transition, Metastable states, Systems biology, Mesoscopic description, Mathematical modeling, Multiscale differential equations, Slow–fast dynamics, Stability analysis

1 Introduction

1.1 Complex Systems and Phase Transition

Cell transition from a phenotype into another constitutes a critical event during development, differentiation, and eventually the onset of degenerative diseases, like cancer. Phenotypic differentiation involves several changes at molecular, physiological, and morphological level. Yet, rather than a progressive process, such transformation behaves like a *first order phase transition*, also involving the overall system in a coherent and global (phase) change.

A phase of a thermodynamic system and the states of matter typically have uniform physical properties. During a phase transition, certain properties of the given medium change, as a result of the variation of some external conditions (for example, temperature, pressure, or others). In Physics, first order phase transitions are characterized by a discontinuity in one or more state variables, and those we are particularly interested in also imply a change in

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_7,

entropy values [1]. By analogy, in biological systems, among the most reliable potential functions which describe such transitions, the *Gibbs free energy* plays a key role since its variations in response to the control parameters are usually mirrored by changes of the entropy.

Despite a number of factors have been demonstrated to participate into cell transitions—including stochastic genetic expression, physical and chemical forces—the cell differentiating process is still poorly understood.

The dynamics of a complex living system can be described at different levels of organization. The current mainstream posits that the lower level, that is the molecular one, exerts a privileged and even unique causative role in shaping how and why the basic units of life, cells and tissues, behave and develop [2]. The prevailing approach postulates that cell fate specification occurs as a deterministic process. In response to intrinsic and/or extrinsic chemical signals, a coordinated change in gene expression patterns drives the cell population into a specific differentiating pathway. This deterministic model has been widely criticized given that gene expression patterns are physiologically stochastic, and fluctuations increase even dramatically when the system (i.e., the cell population) is facing a critical transition from one stable differentiated state into another [3].

To reconcile the wide variability occurring at the microscale (i.e., molecular level) with the deterministic achievement of stable differentiated phenotypes, the concept of *epithelial plasticity* has been introduced into the explanatory scheme [4]. This definition strives to capture two remarkable properties of living systems, namely resilience (robustness) to perturbations and extreme sensitivity to even small fluctuations of the environmental conditions.

A recurrent metaphor for the complex developmental path of cell systems across different phenotypic states is given by the *Waddington landscape*. In this model, cell phenotypes are depicted as stable attractors, also named as "valleys," while metastable or unstable states represent unstable attractors and are named as "hills" [5]. In view of the Mathematical Modeling of biological phase transitions we attempt at formalizing, a comment is in order about the semantic misunderstanding concerning the definition of *metastable states*. Actually, the geometrical characterization of such critical points is better illustrated by the denomination "saddle," and we shall employ the classical stability theory of dynamical systems [6] for the analytical study of the mathematical equations aiming at reproducing the biological experiments.

Stable states are usually identified by specific gene expression patterns and gene regulatory networks (GRNs) architecture. Indeed, the phase-space is reconstructed by computing GRNs from data provided by high-throughput experiments. However, because GRNs are typically intricate and contain highly nested

feedback and feedforward loops that give rise to complex dynamics, it is difficult to elucidate cell behavior from these regulatory circuitries. Moreover, regulation of gene expression is currently no longer considered the causal factor driving cell differentiation [7]. A compelling body of evidence has shown that higher order factors efficiently constrain, and ultimately drive, processes occurring at lower scales [8, 9]. Such results have questioned the classical causative paradigm, deeply rooted into a reductionist, bottom-up approach. In addition, the nonlinear interplay among factors belonging to different levels is highly sensitive to even smaller fluctuations in the initial conditions, or in other environmental parameters, thus providing the system with unexpected and unpredictable properties. This is why higher levels of matter aggregation display *emerging properties* that cannot be anticipated by fundamental laws or by analyzing single components, although the underlying enzymatic-genetic networks in a cell population also support the emergence of macroscopic structures.

Instead of focusing on the role of individual genes, proteins, or pathways in biological phenomena, the aim of Systems Biology is to characterize the ways in which essential molecular parts interact with each others to determine the collective dynamics of the system as a whole.

Furthermore, regulation of the cell journey across the Waddington landscape may shed light to the emergence of *complexity*, and even into biological evolution. Indeed, it seems that complex forms of "organized" behavior in living matter emerge from the competition between different forms of order, rather than between species [10]. Therefore, as longer as conceptual categories such as order and complexity are involved in these processes, parameters like entropy and dissipative structures should be properly considered in any model of cell phenotypic commitment (refer to Subheading 3.3).

Thereby, to grasp physical emergent processes—namely, those occurring during phenotypic transitions, where the biological system is involved and changes coherently as a whole—we must look at the *mesoscopic level/scale*. By analogy with Physics, this is strongly affected by fluctuations around the average and subject to a probabilistic behavior. Indeed, it is mostly from such macroscopic changes that diseases, and especially cancer, are diagnosed.

1.2 The Mesoscopic Framework

The mesoscopic scale is the realm comprised between the nanometer and the micrometer, where "wonderful things start to occur that severely challenge our understanding" [11]. That is to say, at the mesoscopic level nonlinear effects, as well as non-equilibrium processes, are more likely to be appreciated and "captured" [12]. Within that framework, both chemo-physical forces and boundary constraints can be deemed acting as causative factors, even if this property—the causal role—should be ascribed mostly to

the very specific nonlinear dynamics to which the different system components are subjected.

In Biology, the mesoscopic level usually entails both cells and tissues, and scientific investigation requires capturing pivotal features of these constituents. That approach also implies integrating different levels by focusing on parameters that display self-similarities at different scales (fractal dimension represents a paradigmatic case in point [13]). Through such a strategy, one would likely establish strict correlations between the local processes and the global structure of the living beings, by connecting every level with each other. It is worth noting that the topology (i.e., the geometrical three-dimensional distribution) of the interacting components plays a critical role in shaping biological processes. Therefore, quantitative morphological analysis of both cells and tissues architecture has recently regained much interest, given that "the organization becomes cause in the matter" [14].

Furthermore, the mesoscopic framework shall provide an acceptable solution to the *tyranny of scales problem*, still a challenge to reductive explanations in both Physics and Biology [15]. The problem refers to the scale-dependency of physical and biological behaviors, that often forces researchers to combine different models relying on different scale-specific mathematical strategies and boundary conditions. On the other hand, the mesoscopic approach outlines how coordinated (i.e., ordered) macroscale features and properties—including fractal morphology, cell population connectivity and motility, cytoskeleton rearrangement—arise from the collective behavior of microscale variables.

Those issues can be efficiently addressed by adopting a formalism (conceptual premises and framework) borrowed from the phase-space theory [16]. Indeed, the phenotypic differentiation is strongly reminiscent of phase transitions we observe in physical and chemical systems, and it is in fact formally equivalent when the nonlinear dynamics features are properly taken into account [17]. From a mathematical point of view, the nonlinearity is mandatory to support the existence of multiple stationary states with various types of stability properties [6].

By analogy with phase transitions observed in inanimate matter, specific qualities of the biological system should be viewed as *order parameters*, and then their modifications are appreciated under the variation of a number of *control parameters*. As happens in Physics, also in Biology control parameters induce coherent changes in the system by involving it as a whole, that is to say by affecting "pleiotropically" a number of hypothetical targets (molecules and pathways, as well as cellular structures).

The transition from a state of order to a new one appears at the point of instability (bifurcation point), where the increased fluctuation in some of the order parameters leads to a transformation of the cell system, that displays long-range correlations and is self-

similar at all scales of physical observation [18]. Order parameters, like the physical observables, thus enable in capturing the nonlinear dynamics of the system. Moreover, a model based on those parameters shall overcome shortcomings represented by bottom-up modeling, on which reductionist approach usually relies. We strive to identify control parameters that drive the system to instability when approaching their critical values, and the resultant changes in the order parameters that correspond to the major physical modifications in the system under study.

The relevance of control parameters, usually belonging to description levels higher than the molecular one, has recently been vindicated by studies showing that cancer can be "reversed" through physical manipulation of the microenvironment [19]. For instance, it has been demonstrated that cell fate commitment in microgravity is largely dependent on the removal of physical (i.e., gravity) constraints [20]. Overall, such data strongly indicate that the stochastic nonlinear dynamics governing processes at the molecular level can be efficiently and deterministically "constrained" and "ordered" by higher biophysical cues. The classical *principle of causality* is herewith addressed by taking into consideration those higher factors driving the system dynamics, hence recognized as control parameters, including external chemical stimuli, physical forces, environmental constraints, and so forth.

Therefore, our central hypothesis is that the phenotypic transition may be described as a dynamical phase transition by considering only few system parameters and according to a multiscale approach. That model would allow capturing the critical points of the whole process to which further focused investigations are likely to unveil pivotal targets, eventually useful for therapeutically efficient intervention. The ultimate goal is to obtain a physicochemical description of cell transition that could be translated into carcinogenesis studies, as cancer can be considered a "developmental process gone awry" [21].

1.3 Epithelial-Mesenchymal Transition as Metastable State

Cells undergoing a phenotypic switch need preliminarily to enter into a metastable state, thus "destabilizing" their previous stable differentiated state. This destabilization is consistent with a first order critical transition, since suddenly opening access to new stable states—evoking a *tipping point* in the terminology of catastrophe theory [22, 23]. In correspondence to these points, the system experiences a wide fluctuation of many inherent parameters, including gene expression patterns [24].

A paradigmatic case in point is represented by the Epithelial-Mesenchymal Transition (EMT). Epithelial cells normally interact through specialized structures—mainly relying on E-cadherin-based "bridges"—as well as with basement membrane via their basal surface, thus being distributed within the surrounding space in a characteristic (fractal) manner. EMT is the biological process

allowing such polarized cells to undergo multiple biochemical and/or structural changes that enable them to assume a mesenchymal cell phenotype, which includes enhanced migratory capacity, invasiveness, elevated resistance to apoptosis, and greatly increased production of Extra-Cellular Matrix (ECM) components [25].

This transition occurs in a sufficiently dense population of cells (refer to Subheading 2.2) and involves the replacement of one group of cells—which originally adhere to each other forming a differentiated tissue—by another group of cells characterized by a highly heterogeneous and more motile aggregate. As such, EMT is a system process given that it is usually referred to a cell population sample, and can be assessed only at this level. Therefore, from a conceptual point of view, a Systems Biology approach is required to properly investigate EMT dynamics.

The transition from epithelial- to mesenchymal-cell characteristics encompasses a wide spectrum of inter- and intra-cellular changes, also involving the relationship among cells and with their microenvironment, thus representing a true modification of the whole system. It is remarkable that such transformation is reversible under specific environmental constraints, and it should be considered like a phase transition compatible with a mathematical formalization exhibiting a *hysteresis loop* (*see* Fig. 1a).

Indeed, the reverse process, known as Mesenchymal-Epithelial Transition (MET), has also been reported [26], and promising studies on the "beneficial" effects of some external stimuli for inducing MET are in progress (*see* Fig. 2). Additionally, the recent discovery that MET is required for transforming somatic cells into pluripotent stem cells suggests that the intersection between EMT and MET is a fundamental crossroad for cell fate decisions [27].

Although such processes involve an overwhelming number of molecular factors and cellular structures [25], at the mesoscopic level a discrete number of parameters suffices for depicting the transition. Those parameters, mostly relying on (quantitative) changes entailing cell morphology and its dynamical relationships with the neighborhood, can be suitably considered as order parameters.

In this report, we aim at illustrating a methodological pathway for the phenomenon of phase-space transitions during cell fate specification, when a system passes from a stable state to another through a metastable bridge, having in mind the paradigmatic case of the EMT and MET. In that context, Mathematical Modeling provides an inherent *texture* for reality with the specific target of nonlinear dynamics of diffuse information systems [28]. Also it is required to formalize external fields and boundary conditions which are determinant for the system dynamics, and to appreciate subtler system variations to predict more sophisticated behaviors (symmetry breaking, equilibria transition, etc.). Mathematical

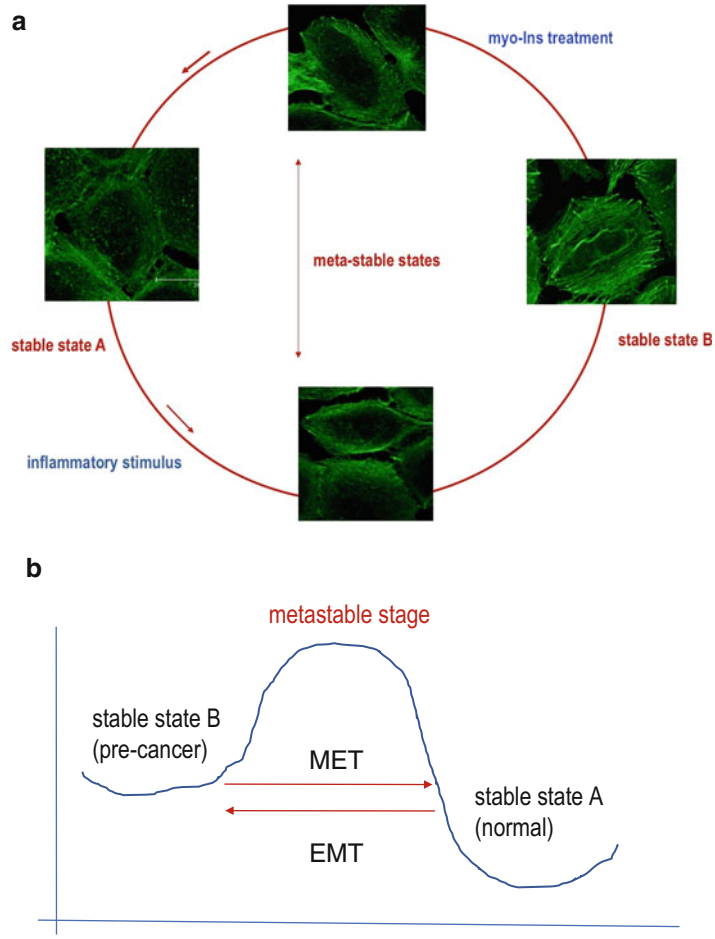


Fig. 1 The EMT-MET schema. (**a**) Inflammatory stimulus and myo-Ins treatment effects on the EMT-MET process. (**b**) An intermediate metastable state is necessary to accomplish a phase transition

Modeling may ultimately help recognizing critical factors and steps in promoting tumor reversal.

Two methodological directions are conceivable. Firstly, applied mathematics for identifying and measuring the *attractor manifolds* for different equilibria by extrapolating information from experimental data. In that respect, the mathematical formulation of the biological problem helps in facilitating measurement quantifications rather than its qualification. Secondly, real-time multi-scale modeling to give evidence of phenomena with cumulative effects, for example models with *memory terms* and search for *precursive factors* to phase-space transitions. That approach can be performed at all description levels, from cells to organs passing through tissues, in order to induce medical actions starting from the theoretical analysis of precursive factors before the system moves too far from

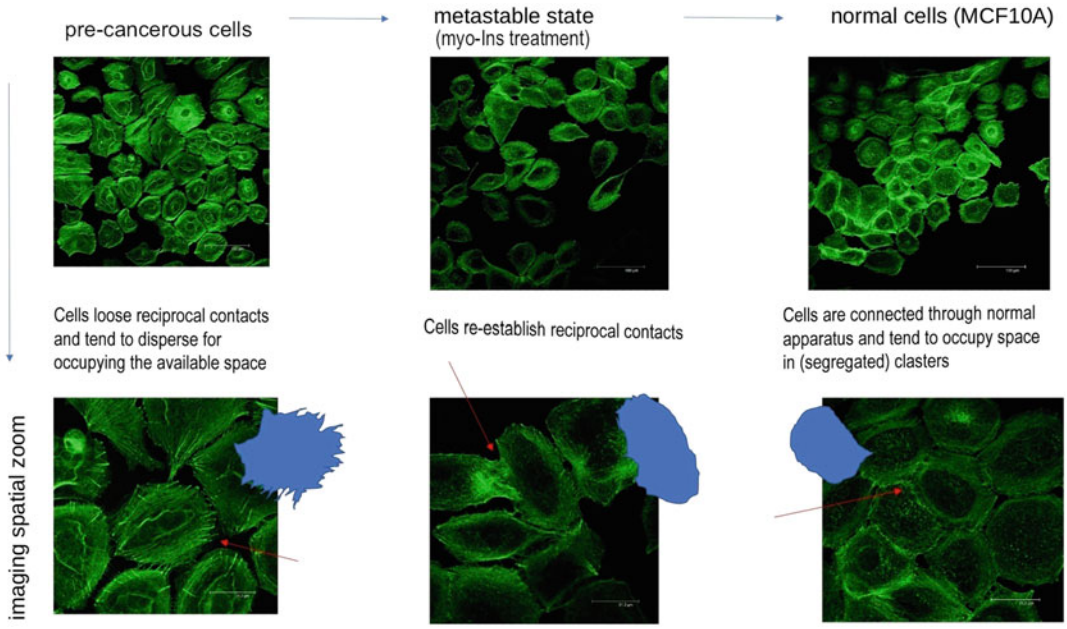


Fig. 2 Phenotypic reversal through myo-Ins-induced MET; schematic cell shape profiles are depicted as extracted from images, highlighting changes occurring during phenotypic transition

the healthy (stable) equilibrium (*see* Fig. 1b). It is worthwhile stressing that a stable dynamics should not be confused with a system in a stationary stable phase (namely, when nothing significant happens). Indeed, the former may anyway undergo a wide range of fluctuations without losing its stability. This means that a stable dynamics is characterized by resilience (robustness) with respect to external perturbations, given that it is located in the manifold of a stable attractor. On the contrary, a stationary stable system lies in a phase where no apparent dynamical changes occur.

Mathematical Modeling is asked to develop criteria to guide the interpretation of the observations in making "causes" and "effects" to raise from experiments (*see* Fig. 3). One wishes to identify lower order changes that are precursory to phase transitions inside the biological systems. In fact, identifying the metastable state during a complex biological process is a challenging task, because the state of the system may show neither apparent changes nor clear phenomena before a critical transition. Therefore, recognizing specific steps by means of additional mathematical variables which vary gradually could help, not only in identifying markers of transformation for early diagnosis, but also in determining drug targets.

The interaction between Systems Biology and Mathematical Modeling would have no hope of generating a virtuous circle, if not for the emergence of a new actor on stage: the computer. The performance development of modern computers has permitted to test models even remotely approachable in the past, through

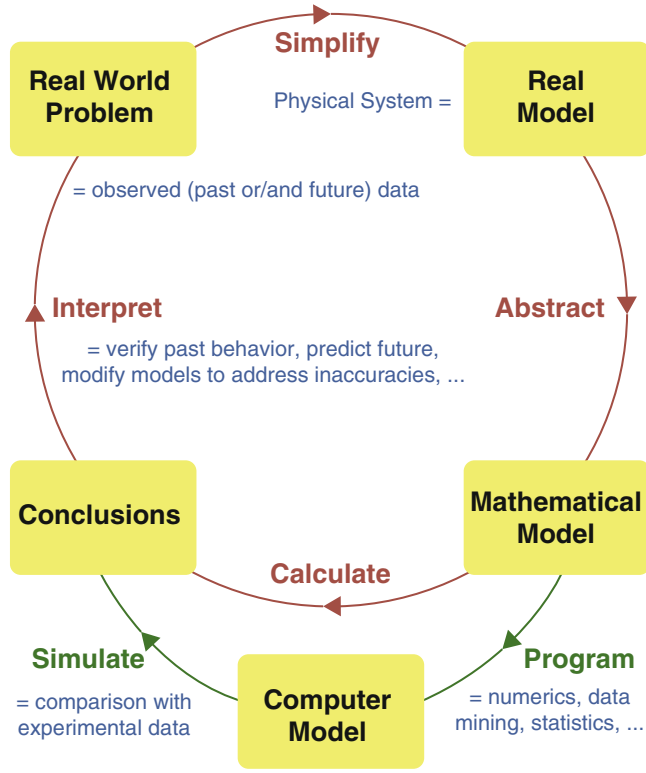


Fig. 3 Mathematical Modeling pathway for adaptive design of biological experiments

suitable numerical implementations [29]. By means of numerical algorithms, mathematical models so complex that they are not amenable of any rigorous analysis can be handled. In the biological field, we come even to coin a third experimental type class, adding to the experiments in vivo and in vitro also those in silico, with specific reference to computer simulations.

2 Material and Methods

2.1 The Experimental Setting

The experimental model is constituted by normal breast cells (MCF10A) that are exposed to micromolar concentrations of Transforming growth factor-β (Tgf-β), a well-known pro-inflammatory molecular effector [30]. As a result, MCF10A cells undergo a clear EMT within about 5 days—although preliminary effects can be appreciated already after 24–48 h—by modifying their shape, the cytoskeleton architecture, the degree of inter-cellular relationships (with a significant reduction in E-cadherin based junctions), as well as their motility striving to occupy any available space (*see* Fig. 2). The Tgf-β-induced EMT should be considered a precursive

step towards full transformation into fibrosis or an even worse (cancer) phenotype [31].

On the other hand, EMT is still a reversible process, which also exhibits an "intermediate" metastable state, that can be switched backward by appropriate changes in the control parameters. Indeed, by adding myo-Inositol (myo-Ins) treatments, the Tgf-β-induced EMT is almost reversed into a MET within 24–48 h (*see* Fig. 1a).

In what concerns the technical aspects of cell culture and reagents, the MCF10A breast cells line was purchased from the American Type Tissue Culture Collection (ATCC) and then cultured in a DMEM/F12 medium supplemented with 5% horse serum, 10 μg/mL insulin, 0. 5 μg/mL hydrocortisone, 20 ng/mL EGF and 100 ng/mL cholera toxin. The cells were accompanied by 100 IU/mL penicillin and 100 μg/mL streptomycin, and kept in 5% CO2 and humidified atmosphere at 37 °C. Recombinant human Tgf-β1 was purchased from PeproTech and myo-Inositol was obtained from Lo.Li.pharma. About 3000 cells/well were originally plated, in a complete medium, onto micro cover glasses. Once at sub-confluent concentration, the cells were treated with 1 μL/mL of Tgf-β1. After about 5 days, during which EMT occurred, the cells were stimulated with 4 mM of myo-Inositol for 24 h. As regards immunofluorescence, cellular morphology and F-actin ultrastructure have been investigated by adding phalloidin (Alexa Fluor 488) staining after cellular fixation with 4% paraformaldehyde and membrane permeabilization with ethanol and acetone in 1:1 ratio, and then visualized through confocal microscopy.

2.2 Control Parameters

According to our experimental setting, cell-phase transition is triggered by two molecular signaling factors, acting essentially in opposite ways: Transforming growth factor-β is a well-known inducer of EMT, while myo-Inositol has recently been demonstrated to be capable of inducing MET, thus counteracting the EMT opposite transformation [32].

The myo-Ins, a cyclic carbohydrate with six hydroxyl groups, is among the oldest components of living beings, undergoing complex evolutionary modifications ultimately leading to the current multiplicity of functions for Ins-containing molecules in eukaryotes [33]. While myo-Ins has no effect on normal (stable) cells, it significantly inhibits EMT in cells exposed to pro-inflammatory stimulation, as such provided by Tgf-β. This finding clearly suggests that myo-Ins effects start becoming apparent only at the bifurcation point, where the system undertakes the phase transition through a metastable state, near to *symmetry breaking points* [34].

Therefore, according to the formalism of phase-space transitions, both Tgf-β and myo-Ins can be managed as control parameters.

The cell density should also be considered an "environmental" constraint. Indeed, experiments performed at different densities typically exhibit significant differences in terms of their results [35]. Changes in the cell density may actually influence cell-to-cell adhesion (thus modifying the overall *connectivity* of the cell population), stiffness and tensegrity response of the cell cluster (by modulating the mechano-transduction of a number of biophysical cues), and ultimately the shape acquired by cells [36].

2.3 Order Parameters

As we have previously discussed, order parameters are measurable physical observables that allow representing the biological phenomenon. At the mesoscopic scale, a careful examination of the Tgf-β-induced EMT makes possible to extract a few key order parameters, which characterize crucial aspects of the experiments, including:

- *Downregulation of E-cadherin* (with reduced density values along the membrane border). Indeed, E-cadherin downregulation is a hallmark of the EMT and it constitutes a pre-requisite for cells committed towards transformation [37]. E-cadherin parameter evaluation epitomizes how different levels of observation are interconnected each other: on one side, E-cadherin can be quantified as an inter-molecular parameter (concentrations measured by western-blot assay within the cells); on the other side, E-cadherin distribution in discrete regions inside the cell can be appreciated by confocal (quantitative) microscopy, thus permitting its understanding as structural element. The combination of both these methods allows assessing the functional meaning of even subtle E-cadherin fluctuations. In addition, the correlation of raw E-cadherin concentration data with its specific localization inside the cell (in the membrane or cytosol domain) could actually provide the link between the sought molecular and structural levels of observation.

  E-cadherin also participates, altogether with a number of other factors, in the formation of cellular adhesion structures. In particular, its downregulation is responsible for reduced number of *cell-to-cell adhesion foci* experimentally observed [38]. Reduction in structural inter- and intra-molecular characteristics is among the most relevant cues that inhibit the constitution of a tissue and promote cells scattering in the available space.
- *Shape changes and fractal dimension*. Modification of the cell form usually entails the loss of apical-basolateral cell polarity, ultimately leading to substrate detachment. Cells detached from the substrate, as well as from their neighboring, are free to acquire new configurations and skills, including motility (*see* Fig. 4). Cell shape can be quantitatively assessed by means of a fractal approach and, in particular, *fractal dimension* (FD) is a well-suited marker of cell malignancy and motility [39].

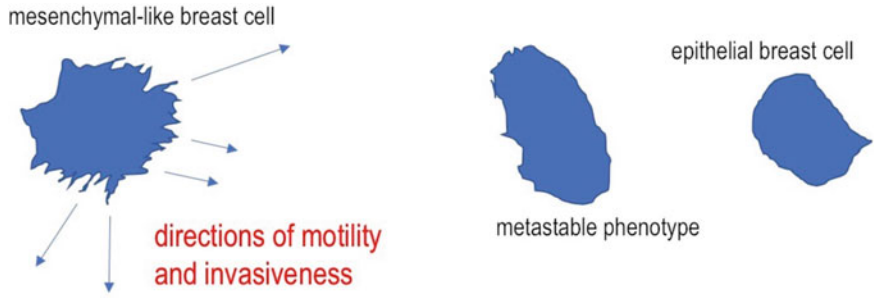


Fig. 4 Parallel changes in cell shape and phenotypic plasticity highlighting how the more migrating/invading cells display higher fractal dimension

Moreover, the fractal dimension epitomizes the *morphological complexity* of the overall cell system, by referring to the minimal required information for its description [40] and, at the same time, it allows to directly tackling the problem of multiple hierarchical levels. However, although the fractal dimension usually provides an indirect measure for the system entropy values, as extracted from quantitative morphological analysis, in the present context this value should not be confused with the internal entropy of the cell population.

In general, quantitative assessment of even subtle morphological changes has been proven to be predictive of further cell fate differentiation. In this respect, high-throughput time-lapse microscopy is a powerful tool for studying cell differentiation and bright-field imaging has been used to track and reconstruct cellular genealogies, namely through fluorescence-based recognition of molecular lineage markers [41]. However, molecular lineage markers are only available for few specific cell types, that are often already differentiated, thus hindering the early identification of differentiating cells. On the other hand, a few attempts have already been made to extract and exploit the information embedded in confocal microscopy images for prospective detection of lineage commitment.

- *Cytoskeleton rearrangement and stress fibers.* The cytoskeleton (CSK), especially through F-actin remodeling, promotes both new shape configurations and selective activation of a number of genetic and biochemical pathways. Overall, changes in CSK can be appreciated by means of an integrated parameter, named *coherency*, that extracts the relative strength of the edges of structures compared to their surroundings. Therefore, it should be considered like a measure of "the global alterations in the organization of the F-actin" [42].

  Appearance of stress fibers is quantitatively assessed, as for other CSK components, through confocal microscopy. Stress fibers have been shown to play an important role in cellular

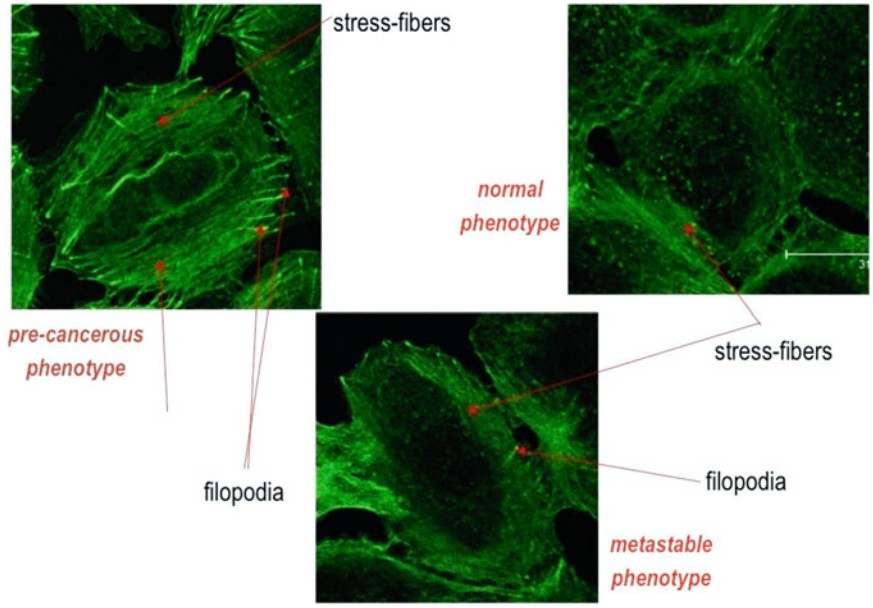


Fig. 5 Different CSK configurations supporting distinct cell phenotypes

contractility, migration and invasiveness, especially during EMT [43]. This process ultimately ends up in the emergence of filopodia and pseudopodia, indicating increased motility and invading capacity (*see* Fig. 5). These structures are mechanistically linked to CSK and to the cell membrane, allowing cells to perform many specialized functions (invasion of the ECM, motility, exploration of the surrounding space).

An important fact is that the aforementioned parameters are "independent" each other, and they cannot be replaced from one another. Yet, they are not exclusive given that various order parameters, tightly correlated with the same features we are looking at, could have been also taken into account (refer to Subheading 4, **Note 1**) and shall further be incorporated for improving the Mathematical Modeling.

It is worthwhile stressing that, in our experimental setting, normal cells in culture are usually "confined" into clusters and they do not display significant spreading. Moreover, during the first 24–48 h of culture, the mitotic and apoptotic rates do not change significantly, so that the cell density (cells per area) can reliably be considered as fixed. Together with the fact that density influences in a crucial way the experimental development, this justifies its role as control parameter.

3 Modeling

3.1 The Mathematical Framework

Next, we attempt at listing crucial features of the mathematical model to reproduce the biological problem described above.

- *ODEs and time-discrete approximation.* The experiments are essentially time-dependent, and changes of the distribution of cells in space and number (density) could be considered constant, at a first modeling stage. That assumption would be satisfied in agreement with the experimental conditions we set for our model, in particular low values of fetal bovine serum (FBS) added to the culture medium (refer to Subheading 4, **Note 2**). Indeed, low FBS concentration implies that cells are only minimally stimulated, and thus display negligible growth rate and migratory capabilities.

  Therefore, the mathematical models are constituted by systems of Ordinary Differential Equations (ODEs) with the eventual presence of stochastic terms [44]. In addition, time-discrete approximations could be developed, in order to perform numerical simulations for comparison with the experimental data (refer to Subheading 4, **Note 3**). As a matter of fact, since the evaluations of the biological process are typically conducted at discrete time instants, one could also directly formalize time-discrete models (i.e., recurrence equations) from which appropriate ODEs are deduced by taking times-continuous limits [45].
- *Space dependency.* Nevertheless, space dependency is relevant: since our target is to "revert" potentially malignant cells earlier, before they acquire a migrative and invasive phenotype, the space rather plays the role of an external parameter in the sense that important properties of the cell population manifest a space dependency (density, lacunarity, critical malignant features, etc.) although without transport terms and/or spatial gradients.

  Moreover, the experimental setting presupposes initial conditions with cells uniformly distributed and synchronized over the culture support, but however slight differences in the cell cycle cannot be avoided, and thus space inhomogeneities have to be taken into account.
- *Slow–fast dynamics.* The transition time for EMT and MET is typically very short with respect to the overall lifetime of the biological system. This translates into the fact that the corresponding mathematical model should exhibit a slow–fast decomposition [46].

  More precisely, we require that the differential equations incorporate a small parameter $\tau \geq 0$ governing the time-scale, so that, for infinitely small values of such parameter, namely as $\tau \to 0^+$ (the so-called *singular perturbation limit*), we recover the qualities of a first order phase transition. A major consequence of

this approach is that the ODEs system still hold for strictly positive values of the time-scale parameter, hence providing a reliable description also for *second order phase transitions* with $\tau \gneqq 0$.

- *Multi-scale approach.* Phase transitions are described by means of a multi-scale model. Some observable parameters are actually averages of microscopic quantities and can be further mirrored by the behavior of lower order parameters. Within its general structure, our mathematical formalization does not restrain from taking into account genetic or other microscopic factors (GRNs, for example). Systems Biology considers external forces which are integrated to the various levels for having effects on the cells, then the feedbacks inside the system are essential ingredients for adequate models. Several mathematical strategies allow to relate passages from different space-time levels and different scales can be effectively included: hydrodynamical limits from cells to tissues, integro-differential equations for memory terms and non-local issues, and asymptotic analysis, among others.
- *Entropy and fractal analysis.* In biological systems, fluctuations in the amount of entropy can be equated, at a first glance, to variations of the Gibbs free energy. In turn, changes in entropy values can be tracked by evaluating modifications in the fractal properties of the cell system [47, 48]. Various formulae for the fractal dimension of biological systems are in fact defined based on entropy functions [49]. It is worth recalling that entropy evaluation always depends on the scale of measurement, thus resulting in *uncertainty*, whilst the fractal dimension is independent of (discrete) measurement scales.

  From a mathematical point of view, we aim at identifying a global (space- and time-dependent) function, the so-called *Lyapunov functional*, accounting for the overall "stress" of the dynamical process [6], and try to determine the points where this function experiences a symmetry breaking so that the system starts transiting towards metastable states (refer to Subheading 1.3). The *variational analysis* of auxiliary quantities different from the order parameters, which have eventually varied when the system leaves an equilibrium, would provide the precursive signature of a phase-space transition.

3.2 Formal Equations

Let us consider the vector (i.e., collection) of physical variables $V = (E, F, C)$, where E, F, C stand for *system-averaged values* of E-cadherin, fractal dimension, and coherency, respectively. We assume that the dynamics of the cell system is justly characterized by the time evolution of these quantities. The choice of those order parameters for reproducing the biological experiments is not

exclusive, and the same mathematical formalism could also be adopted for other observable quantities (refer to Subheading 4, **Note 1**).

Then, the experimental setting is translated into a set of first order ODEs for the instantaneous time variation of the order parameters, which is denoted by

$$\frac{dV}{dt} = \left(\frac{dE}{dt}, \frac{dF}{dt}, \frac{dC}{dt}\right)$$

and should be interpreted as time-derivative in mathematical language.

For time t varying between 0 and about 5 days (starting and ending of the biological experiment), the differential model reads

$$\frac{dV}{dt} = \Phi(V; S) \tag{1}$$

for some (vector-valued) structural function $\Phi = (\Phi_1, \Phi_2, \Phi_3)$ describing the biological mechanism underlying the dynamical process, and with S representing the external stimuli (i.e., control parameters), that include inflammatory factors, myo-Ins, cell density, physical constraints, and other eventual terms. Equation 1 can be rewritten in scalar components as

$$\begin{cases} \dfrac{dE}{dt} = \Phi_1(E, F, C; S) \\ \dfrac{dF}{dt} = \Phi_2(E, F, C; S) \\ \dfrac{dC}{dt} = \Phi_3(E, F, C; S) \end{cases} \tag{2}$$

and it must be complemented with appropriate initial conditions $E(0) = E_0$, $F(0) = F_0$ and $C(0) = C_0$ to be deduced from the experimental measures for E_0, F_0, and C_0. On the other hand, since S embodies the control parameters, it should be considered as a known function which may be constant or rather time- and space-dependent (for example, if growth factors or treatments are administered at specific discrete temporal instants or/and in a spatial non-homogeneous way to the population of cells).

Concerning the space dependency, we choose a two-dimensional reference domain $\Omega \subset \mathbb{R}^2$ corresponding, for example, to a *Petri dish* or any technical support where the cell culture is analyzed (*see* Fig. 6). In principle, similar statements hold in the physical three-dimensional space.

Due to the high number of cells involved in the biological trials, a tissue-like behavior emerges for the whole system, and thus the hypothesis of a space-continuous description is pertinent. Hence, system-averaged values of E, F, and C can be defined in terms of the

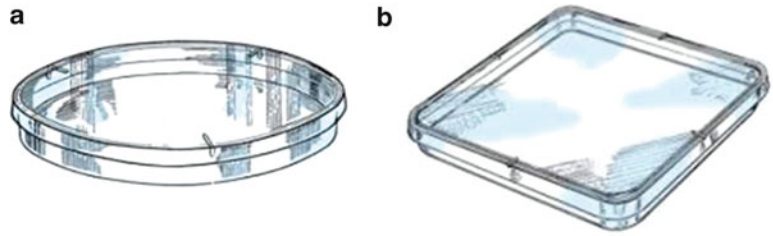


Fig. 6 Two examples of cell culture plates. (**a**) A circular Petri dish. (**b**) A squared Petri dish

corresponding cell-related "densities" as the following spatial integrals

$$\begin{aligned} E(t) &= \frac{1}{|\,\Omega\,|}\int_{\Omega} e(t,\mathbf{x})\; d\mathbf{x}, \\ F(t) &= \frac{1}{|\,\Omega\,|}\int_{\Omega} f(t,\mathbf{x})\; d\mathbf{x}, \\ C(t) &= \frac{1}{|\,\Omega\,|}\int_{\Omega} c(t,\mathbf{x})\; d\mathbf{x}, \end{aligned} \tag{3}$$

with $|\Omega|$ denoting the area of the experimental domain. Here, for time $t \geq 0$ and position $\mathbf{x} \in \Omega$, the functions e, f, and c describe the density of E-cadherin, fractal dimension, and coherency, respectively, and they are introduced to take into account the microscopic features of the cell system.

This constitutes a first instance of multi-scale approach since different levels of observation—specifically, from cells to tissues—are mathematically related. Indeed, a model similar to Eq. 2 can be formulated also at the microscopic scale, namely

$$\begin{cases} \dfrac{de}{dt} = \varphi_1(e,f,c;S) \\ \dfrac{df}{dt} = \varphi_2(e,f,c;S) \\ \dfrac{dc}{dt} = \varphi_3(e,f,c;S) \end{cases} \tag{4}$$

so that the macroscopic equations 2 are recovered through space-averaged integrals Eq. 3 provided that the structural functions φ_1, φ_2, and φ_3 in Eq. 4 are properly designated. Although intrinsically coherent with a multi-scale framework, such procedure could be extremely intricate to be performed in practical cases, especially when the control parameters S are space-dependent. Nevertheless, unlike the *global/macroscopic* order parameters E, F, and C which are naturally defined for the whole system by extracting information from the corresponding *local/microscopic* densities e, f, and c (refer to Subheading 2.3), the control parameters S are more

efficiently established directly at a higher order (i.e., mesoscopic) level, without the necessity of moving down to the microscopic scale. Obviously, that strategy does not exclude from considering the microscopic processes induced on the cells by the presence of those external stimuli—including genetic expression, physical and chemical molecular forces—by formulating explicit forms for φ_1, φ_2, and φ_3 in Eq. 4, but this is not mandatory for the success of our approach.

The minimal requirement for the vectorial model Eq. 1 or, equivalently, for its component-wise version Eq. 2 to represent an acceptable candidate for modeling EMT processes is that they display three stationary solutions, two stable states, and one unstable/metastable state (*see* Fig. 1). This forces the function Φ and its components Φ_1, Φ_2 and Φ_3 to satisfy some essential structural conditions, in order to ensure that

$$\Phi(V;S)=0 \quad \Leftrightarrow \quad V\in\{A,B,M\} \quad \text{for any } S \tag{5}$$

or, equivalently,

$$\begin{cases} \Phi_1(E,F,C;S)=0 \\ \Phi_2(E,F,C;S)=0 \\ \Phi_3(E,F,C;S)=0 \end{cases} \Leftrightarrow \quad (E,F,C)\in\{A,B,M\} \quad \text{for any } S \tag{6}$$

for some (distinct) vectors $A=(E^A, F^A, C^A)$, $B=(E^B, F^B, C^B)$ and $M=(E^M, F^M, C^M)$ corresponding to biologically relevant equilibria.

Additional conditions guaranteeing stability for A and B, and metastability for M, have also to be satisfied (refer to Subheading 3.3).

Since the equilibrium system Eqs. 5–6 is multi-dimensional, the phase-space exhibits a non-trivial geometrical landscape (*see* Fig. 7), and transitions can occur with *sudden change* of values concerning only some variables (like for *contact discontinuities* in continuum physics [50]).

3.3 A Tutorial Example

We consider a simplified model consisting of a (nonlinear) system with two coupled first order ODEs for the variables u and w, that is

$$\tau\frac{du}{dt}+u=w, \qquad \frac{dw}{dt}+w=\lambda g(u), \tag{7}$$

where the external constraints are given by constant parameters τ, $\lambda > 0$ and g is a known structural function whose properties are detailed later on. In comparison with the general model Eqs. 1–2, order and control parameters correspond to $V=(u, w)$ and $S=(\tau, \lambda)$, respectively, and

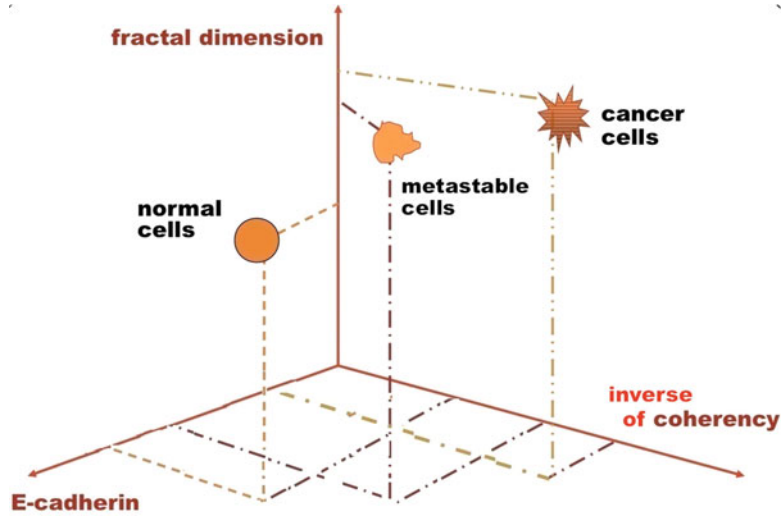


Fig. 7 Hypothetical three-dimensional space-phase diagram depicting the performance of order parameters

$$\Phi_1(u, w; \tau, \lambda) = \frac{1}{\tau}(w - u), \qquad \Phi_2(u, w; \tau, \lambda) = \lambda g(u) - w. \quad (8)$$

We attempt at formulating a hypothetical interpretation of the dynamical process Eq. 7 in terms of biological observations, assuming that u represents E-cadherin boundary values and w stands for the coherency, which is connected with relative E-cadherin density values along the membrane border with respect to its overall concentration. Then, the specific expression for Φ_1 encodes the fact that u—describing the E-cadherin boundary distribution of the cell population—tends to conform to the behavior of w—accounting for the system coherency—in a (typically fast) time-scale of order τ. Similarly, the expression for Φ_2 entails the convergence of w towards $\lambda g(u)$ in a (slower) time-scale of order 1.

According to the abstract calculations in Eqs. 5–6, that now translate into

$$\begin{cases} \Phi_1(u, w; \tau, \lambda) = 0 \\ \Phi_2(u, w; \tau, \lambda) = 0 \end{cases}$$

for the specific functions Eq. 8, the stationary solutions to Eq. 7 are given by the points (u, w) which are located at the intersection of the curves

$$w = u \quad \text{and} \quad \lambda g(u) = w \quad (9)$$

laying on the phase-plane (i.e., the two-dimensional projection of the phase-space). As a consequence, the set of equilibria for the dynamical system Eq. 7 is characterized, for any fixed $\lambda > 0$, as the zeros of the function (*see* Fig. 8b)

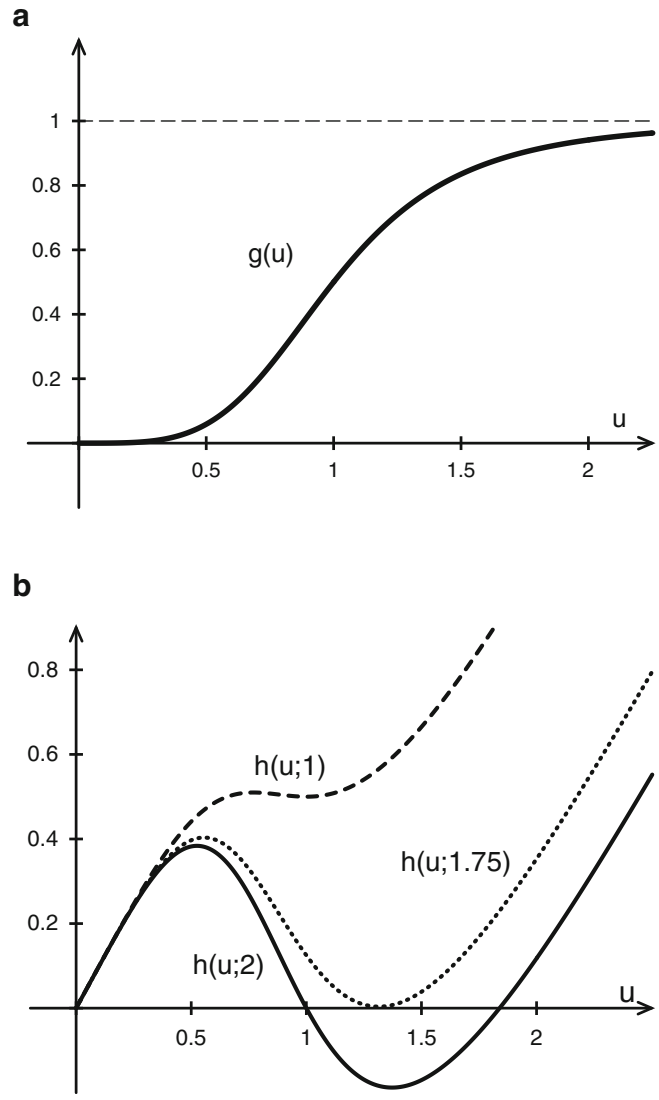


Fig. 8 The graphs of the structural functions g and h. (**a**) A typical S-shaped function g with $p = 1$ and $\nu = 1$. (**b**) The corresponding function h for distinct values of λ ($\lambda = 1$ *dashed*, $\lambda = \lambda_c = 1.75$ *dotted*, $\lambda = 2$ *continuous*)

$$h(u;\lambda) := u - \lambda g(u)$$

and the *stability* properties are deduced by analyzing its first order derivative (namely, the first order variation of h with respect to u), that is

$$\frac{dh}{du}(u;\lambda) = 1 - \lambda \frac{dg}{du}(u). \tag{10}$$

Actually, for the particular case of system Eq. 7, this approach is equivalent to the standard *spectral analysis* (refer to Subheading 4, **Note 4**). More precisely, if the derivative $\frac{dh}{du}(u;\lambda)$ is positive, the

equilibrium is stable; otherwise, if it is negative, the equilibrium is unstable/metastable. Under the assumption that g is non-decreasing (namely, $\frac{dg}{du}(u) \geq 0$ for any u) and such that $g(0) = \frac{dg}{du}(0) = 0$, the origin of the phase-plane $(u, w) = (0, 0)$ is a solution to Eq. 9 and, moreover, it is a stable equilibrium because $\frac{dh}{du}(0; \lambda) = 1$ for any $\lambda > 0$ from Eq. 10.

In terms of biological experiments, the stationary state $(0, 0)$ satisfying the above conditions could be associated with the original (unperturbed) phase of the system (normal cells). Besides, due to the nonlinearity of the function g, the mathematical description Eq. 7 also incorporates the existence of other biological equilibria—different from $(0, 0)$—corresponding to further phases of the cell system during EMT or/and MET (refer to Subheading 1.3). Indeed, according to Eq. 9, any eventual subsequent intersection between the curve $w = \lambda g(u)$ and the straight line $w = u$ gives raise to additional equilibria, alternating stable and unstable/metastable states in the case of simple zeros of h (which occur under the generic assumption that $h(u; \lambda) = 0$ implies $\frac{dh}{du}(u; \lambda) \neq 0$, that is the so-called *transversality condition*).

Then, we conjecture that g behaves like an S-shaped function, meaning that g is convex in the interval $(0, p)$ and concave in its complement $(p, +\infty)$ for some $p > 0$, and its values are bounded from above, so that $g(+\infty) = \ell$ for some threshold $\ell > 0$ (*see* Fig. 8a). Therefore, two distinct ranges of values for the control parameter λ can be considered, leading to quite different emerging scenarios (*see* Fig. 8b) classified as follows:

- *small* λ, corresponding to a unique equilibrium, given by $(u, w) = (0, 0)$;
- *large* λ, that is consistent with the presence of three intersection points (i.e., equilibria).

These two regimes are separated by a (non-generic) critical value $\lambda = \lambda_c > 0$ which produces only two distinct equilibria.

In view of the previous analysis, one infers that model Eq. 7 is, at the same time, minimal and reliable. Indeed, small values of λ (i.e., $\lambda < \lambda_c$, *see* Fig. 9a) illustrate a biological situation where the external physical constraints—for example, inflammatory factors or myo-Ins treatments—are too weak for determining any phase transition, hence the system remains in its original (healthy or pre-cancerous) configuration. On the other hand, for large λ (i.e., $\lambda > \lambda_c$, *see* Fig. 9b) the mathematical system supports phase transitions alternating stable and metastable states, and the possibility of simulating EMT or/and MET with the typical "destabilization mechanism" introducing a metastable state (refer to Subheading 1.3). The importance of identifying, and also quantifying, the critical threshold λ_c appears, in particular, when medical actions have finally to be undertaken, because the control parameters can

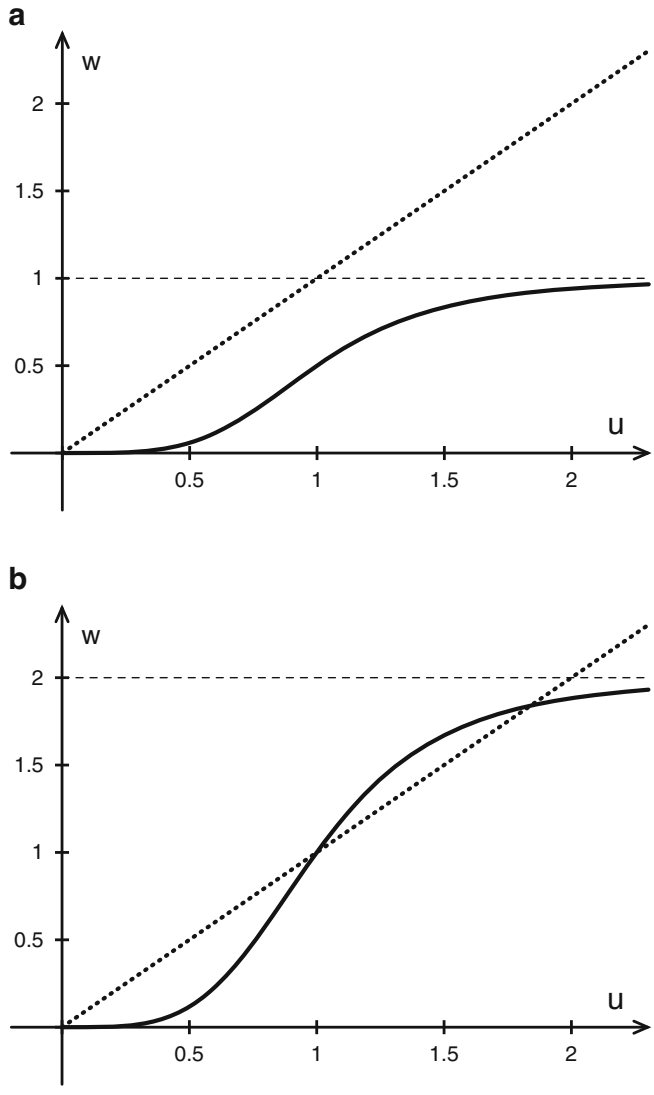


Fig. 9 Relative positions of the straight line $w = u$ (*dotted*) and the curve $w = \lambda g(u)$ (*continuous*) corresponding to different values of λ. (**a**) Case $\lambda = 1 < \lambda_c$ with $p = 1$, $\ell = 1$. (**b**) Case $\lambda = 2 > \lambda_c$ with $p = 1$, $\ell = 1$

be adjusted to predictively regulate the system response. It is worthwhile noticing that multiple equilibria could be generated by assuming other forms for the function g, thus allowing to establish effective mathematical models for reproducing a wide variety of biological dynamics.

The *fast-dynamics* of Eq. 7 is described by a new time variable s related to the old one t by the relationship $s = t/\tau$, which corresponds to the following differential equality for the instantaneous time variation:

$$\frac{d}{ds} = \tau \frac{d}{dt}.$$

This procedure consists in rescaling the time variable in order to zoom in on the system process during the first temporal period. Passing from t to s, the system of ODEs becomes

$$\frac{du}{ds} + u = w, \qquad \frac{dw}{ds} + \tau w = \tau \lambda g(u),$$

and its limit as $\tau \to 0^+$ is formally given by

$$\frac{du}{ds} + u = w, \qquad \frac{dw}{ds} \approx 0. \tag{11}$$

The second equation in Eq. 11 does trivially express the fact that w is, at first glance, independent of s (i.e., constant). Therefore, the corresponding approximated solutions to Eq. 11 are

$$\frac{du}{ds} + u \approx w_0, \qquad w \approx w_0,$$

and, by applying classical results on explicit solutions to linear ODEs [51], one ultimately obtains

$$u(s) \approx w_0 + (u_0 - w_0)e^{-s}, \qquad w(s) \approx w_0,$$

for some initial conditions $u(0) = u_0$ and $w(0) = w_0$. Finally, in a fast time-scale ($s \to +\infty$ or, equivalently, $\tau \to 0^+$) the solution gets closer and closer to the straight line $w = u$. The "fate" of the system is not yet decided, but it appears to be dictated only by the variable u, whose dynamics is determined at a slower time-scale. Coming back to Eq. 7 and putting formally $\tau = 0$, we deduce that the slow dynamics is described by the reduced system

$$u = w, \quad \frac{dw}{dt} + w = \lambda g(u), \tag{12}$$

which corresponds to the scalar equation

$$\frac{du}{dt} + u - \lambda g(u) = 0. \tag{13}$$

The characterization of the equilibria for Eqs. 12–13 and their stability analysis is precisely what has been performed above, recalling that $h(u; \lambda) = u - \lambda g(u)$.

A remarkable fact is that system Eq. 7 possesses an alternative representation consisting of a single second-order differential equation, which can be obtained by differentiating the first equation of Eq. 7 with respect to time t and taking advantage of the second equation, so that

$$\tau \frac{d^2u}{dt^2} + \frac{du}{dt} = \frac{dw}{dt} = \lambda g(u) - w,$$

and then first equation is used again to obtain

$$\tau \frac{d^2u}{dt^2} + (1+\tau)\frac{du}{dt} + u - \lambda g(u) = 0, \tag{14}$$

which is known as the *one-field equation*. Multiplying Eq. 14 by the first order derivative $\frac{du}{dt}$ and applying the *chain rule* give

$$\underbrace{\tau \frac{d^2u}{dt^2}\frac{du}{dt}}_{\frac{d}{dt}\left\{\frac{\tau}{2}\left(\frac{du}{dt}\right)^2\right\}} + (1+\tau)\left(\frac{du}{dt}\right)^2 + \underbrace{\{u - \lambda g(u)\}\frac{du}{dt}}_{\frac{d}{dt}\left\{\frac{1}{2}u^2 - \lambda G(u)\right\}} = 0, \tag{15}$$

where G denotes a *primitive* of g (i.e., a function such that $\frac{dG}{du}(u) = g(u)$ for any u). Equation 15 shows the *dissipative structure* of the dynamics: indeed, this can be rewritten as

$$\frac{d}{dt}\left\{\frac{\tau}{2}\left(\frac{du}{dt}\right)^2 + \frac{1}{2}\,u^2 - \lambda G(u)\right\} = -(1+\tau)\left(\frac{du}{dt}\right)^2, \tag{16}$$

where the quantity under the time-derivative has a negative variation, and thus decreases in time. Therefore, the term

$$\frac{\tau}{2}\left(\frac{du}{dt}\right)^2 + \frac{1}{2}\,u^2 - \lambda G(u)$$

is a *Lyapunov functional* for Eq. 14, and thus it is designated for being an *intrinsecal entropy* for the dynamical process (refer to Subheading 3.1).

4 Notes

1. *Order parameters.* In principle, we could have chosen different molecular parameters in the place of E-cadherin. However, besides the specific relevance of E-cadherin during EMT, most of these parameters cannot be considered as being independent with respect to the E-cadherin. For example, the N-cadherin—a paradigmatic marker of mesenchymal transformation—increases or decreases exactly in opposite way to E-cadherin. Similarly, the Focal Adhesion Kinases (FAK) or β-catenin membrane density, are, in some way, related to the E-cadherin. By including these parameters, no eloquent "information" would be further added to the model.

2. *Cell culture protocols.* Currently, a number of artifacts frequently biases cell culture models. For instance, cells are typically stressed by high concentrations of growth factors, which are added to the culture medium to promote sustained proliferation. As a matter of fact, this "accelerated" growth regimen could likely overcome regulatory loops by introducing into the system an additional, unwarranted and usually overlooked, control parameter (i.e., external stimulus). Therefore, we conditioned MCF10A cells growing in a medium supplemented with low FBS levels (1%) to avoid undue metabolic and proliferative consequences. Moreover, low-FBS regimen—without impairing cell viability—kept cell density in a quasi-stationary state for at least 24–48 h, with minimal change in cell population count (refer to Subheading 2.3).
3. *Numerical algorithms.* Especially to reproduce the outcome of in vitro experiments, it is pertinent to have recourse to scalar-valued equations settled on a two-dimensional domain $\Omega \subset \mathbb{R}^2$ with regular boundary (*see* Fig. 6), although the approach developed in this report straightforwardly extends to systems in the three-dimensional space. In order to perform numerical simulations for comparison with the experimental data, time-discrete approximations have to be developed, and spatial finite differences on staggered grids can be applied for dealing with the space dependency [52, 53]. The *Runge-Kutta method* is particularly suitable for the numerical simulation of time-evolution differential equations [45]. In general, time-implicit schemes are quite computationally inefficient for complex problems and, indeed, high-order Runge-Kutta time-integration solvers are important tools for improving the resolution of explicit simulations. On the other hand, the importance of designing spatially compact difference operators is motivated by the requirement of an optimal implementation in parallel computers [54, 55]. In fact, since the nearest-neighbor communication standard is extremely fast with the need of small amounts of local storage in the sub-processors (as only few values are involved to update the numerical solution at each grid point), even very large models becomes feasible, thanks to the massive number of threads especially in GPU-based computing devices [56]. For the sake of completeness, we mention that a modern C++ library for numerically solving ODEs is available at http://www.odeint.com—which is compatible with running on CUDA GPUs programming architecture through the Thrust interface available at http://thrust.github.io
4. *Linearized operator and spectral analysis.* The dynamical system Eq. 7 is nonlinear because of the presence of the nonlinear term $g(u)$ inside the second equation, which is responsible for the existence of multiple non-trivial equilibria (refer to

Subheading 3.2). The linearization at an equilibrium point (u, w) gives a system for the first order perturbation (ξ, η) that is

$$\frac{d\xi}{dt} = \frac{1}{\tau}(-\xi + \eta), \qquad \frac{d\eta}{dt} = \lambda \frac{dg}{du}(u)\xi - \eta,$$

or, in vectorial form,

$$\frac{d}{dt}\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \mathbb{A}\begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

where the matrix

$$= \begin{pmatrix} -\frac{1}{\tau} & \frac{1}{\tau} \\ \lambda \frac{dg}{du}(u) & -1 \end{pmatrix}$$

is known as the *jacobian matrix*. Spectral analysis is based on the computation of the eigenvalues (and, specifically, on their sign) of, which are the roots of the *characteristic polynomial* given by

$$\begin{aligned} p(\mu) := \det(\mathbb{A} - \mu I) &= \left(-\frac{1}{\tau} - \mu\right)(-1 - \mu) - \frac{\lambda}{\tau}\frac{dg}{du}(u) \\ &= \mu^2 + \left(1 + \frac{1}{\tau}\right)\mu + \frac{1}{\tau}\left(1 - \lambda \frac{dg}{du}(u)\right). \end{aligned} \tag{17}$$

Denoting by μ_1 and μ_2 the zeros of the above polynomial, the following representation holds

$$p(\mu) = (\mu - \mu_1)(\mu - \mu_2) = \mu^2 - (\mu_1 + \mu_2)\mu + \mu_1\mu_2,$$

and therefore, comparing with Eq. 17, we deduce that

$$\mu_1 + \mu_2 = -\left(1 + \frac{1}{\tau}\right), \qquad \mu_1\mu_2 = \frac{1}{\tau}\left(1 - \lambda \frac{dg}{du}(u)\right).$$

Recalling that $\frac{dh}{du}(u;\lambda) = 1 - \lambda \frac{dg}{du}(u)$ from Eq. 10, if $\frac{dh}{du}(u;\lambda)$ is positive, the product $\mu_1\mu_2$ of the two roots is positive—indicating that they have the same sign—and their sum $\mu_1 + \mu_2$ is negative—indicating that they are both negative—so that the equilibrium state (u, w) is stable. Complementarily, if $\frac{dh}{du}(u;\lambda)$ is negative, one root is positive and the other is negative, consistently with the appearance of a saddle point, or, in other words, a metastable equilibrium. The above fact is a special form of the more general *Routh–Hurwitz criterion* [57].

5 Conclusions and Perspectives

Reproducibility of the results presented in this report has been assessed by means of triplicate, independent experiments. Indeed, Tgf-β induced EMT is always obtained after 5 days of treatment, involving up to 90% of cells as recorded by molecular and morphological analyses. Similarly, myo-Ins induced MET occurs after 24 h by involving up to 85% of transformed cells. In addition, the analysis of the mathematical models may suggest new features for the experimental setting, also by means of numerical simulations for enlarged models obtained by adding terms, factors and mechanisms which are further developments of the biological experiments. One could also postulate auxiliary order parameters, for example mathematical derivatives of the principal functions, to earlier predict phase transitions with the ultimate target of designing external controls to prevent such transitions.

It is extremely important that mathematical models capture the emergence of dynamics at higher levels, since the behavior of the system is not merely the result of the collective evolution of its isolate components, but it proceeds from the effect of (global) constraints. This emphasizes the intrinsic limits of studying biological phenomena on the basis of purely microscopic experiments (indeterminateness of measurements, instantaneous time, etc.) and, therefore, a multi-scale model (with some parameters derived from the microscopic analysis) is better suited from a methodological point of view. In that context, the so-called *emerging properties* are interpreted as *systemic averages* of microscopic behaviors (for example, the effects of the inositol on the density of breast tissues has been measured before understanding its microscopical chemical reactions).

Acknowledgements

The results presented in this report have been obtained in the framework of the *Working Group on Phase Transitions in Biology through Mathematical Modeling*, settled at the Systems Biology Group Lab—http://www.sbglab.org—Sapienza University of Rome, Italy.

References

1. Landau L, Lifschitz EM (1980) Course of theoretical physics: statistical physics, vol 5, 3rd edn. Elsevier, Amsterdam
2. Bizzarri M, Cucina A, Conti F, D'Anselmi F (2008) Beyond the oncogene paradigm: understanding complexity in carcinogenesis. Acta Biotheor 56(3):173–196
3. Huang S, Ingber DE (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Dis 26(1):27–54

4. Tam WL, Weinberg RA (2013) The epigenetics of epithelial-mesenchymal plasticity in cancer. Nat Med 19(11):1438–1449
5. Waddington CH (1957) The strategy of the genes. George Allen & Unwin, Ltd., London
6. Chicone C (2006) Ordinary differential equations with applications. Texts in applied mathematics, vol 34, 2nd edn. Springer, New York
7. Paldi A (2012) What makes the cell differentiate. Prog Biophys Mol Biol 110(1):41–43
8. Giuliani A (2010) Collective motions and specific effectors: a statistical mechanics perspective on biological regulation. BMC Genomics 11(Suppl1):S2
9. Bravi B, Longo G (2015) The unconventionality of Nature: biology, from noise to functional randomness. In: Calude C, Dinneen M (eds) Unconventional computation and natural computation (UCNC 2015). Lecture notes in computer science, vol 9252, pp 3–34. Springer, Cham
10. Heylighen F (2001) The science of self-organization and adaptivity. In: The encyclopedia of life support systems, vol 5(3), pp 253–280. EOLSS, Oxford
11. Coleman P (2007) Frontier at your fingertips. Nature 446:379
12. Laughlin RB, Pines D, Schmalian J, Stojkovic BP, Wolynes P (2000) The middle way. Proc Natl Acad Sci USA 97(1):32–37
13. Bizzarri M, Giuliani A, Cucina A, D'Anselmi F, Soto AM, Sonnenschein C (2011) Fractal analysis in a systems biology approach to cancer. Semin Cancer Biol 21(3):175–182
14. Strohman RC (2000) Organization becomes cause in the matter. Nat Biotechnol 18:575–576
15. Green S, Batterman R (2017) Biology meets Physics: reductionism and multi-scale modeling of morphogenesis. Stud Hist Phil Biol Biomed Sci 61:20–34
16. Ma S (1976) Modern theory of critical phenomena. Advanced book program. W.A. Benjamin, Reading
17. Davies PC, Demetrius L, Tuszynki JA (2011) Cancer as dynamical phase transition. Theor Biol Med Model 8:1–30
18. Bak P (1996) How nature works. Springer, New York
19. Bissell MJ, Inman J (2008) Reprogramming stem cells is a microenvironmental task. Proc Natl Acad Sci USA 105(41):15637–15638
20. Masiello MG, Cucina A, Proietti S, Palombo A, Coluccia P, D'Anselmi F, Dinicola S, Pasqualato A, Morini V, Bizzarri M (2014) Phenotypic switch induced by simulated microgravity on MDA-MB-231 breast cancer cells. Biomed Res Int 2014. ID652434
21. Soto AM, Maffini MV, Sonnenschein C (2008) Neoplasia as development gone awry: the role of endocrine disruptors. Int J Androl 31 (2):288–293
22. Arnold VI (1986) Catastrophe theory, 2nd edn. Springer, New York
23. Gladwell M (2000) The tipping point: how little things can make a big difference. Little, Brown and Company, Boston
24. Mojtahedi M, Skupin A, Zhou J, Castaño IG, Leong-Quong RYY, Chang H, Trachana K, Giuliani A, Huang S (2016) Cell fate decision as high-dimensional critical state. PLoS Biol 14 (12):e2000640
25. Thiery JP, Sleeman JP (2006) Complex networks orchestrate epithelial-mesenchymal transitions. Nat Rev Mol Cell Biol 7(2):131–142
26. Yao D, Dai C, Peng S (2011) Mechanism of the mesenchymal-epithelial transition and its relationship with metastatic tumor formation. Mol Cancer Res 9(12):1608–1620
27. Esteban MA, Bao X, Zhuang Q, Zhou T, Qin B, Pei D (2012) The mesenchymal-to-epithelial transition in somatic cell reprogramming. Curr Opin Genet Dev 22(5):423–428
28. Quarteroni A (2009) Mathematical models in science and engineering. Not Am Math Soc 56 (1):10–19
29. Hwu WW (2011) GPU computing gems. Emerald & Jade Editions. Applications of GPU computing series. Morgan Kaufmann, Elsevier, Burlington
30. Xu J, Lamouille S, Derynck R (2009) TGF-beta-induced epithelial to mesenchymal transition. Cell Res 19(2):156–172
31. Barriere G, Fici P, Gallerani G, Fabbri F, Rigaud M (2015) Epithelial mesenchymal transition: a double-edged sword. Clin Transl Med 4(14):1–6
32. Dinicola S, Fabrizi G, Masiello MG, Proietti S, Palombo A, Minini M, Harrath AH, Alwasel SH, Ricci G, Catizone A, Cucina A, Bizzarri M (2016) Inositol induces mesenchymal-epithelial reversion in breast cancer cells through cytoskeleton rearrangement. Exp Cell Res 345(1):37–50
33. Bizzarri M, Fuso A, Dinicola S, Cucina A, Bevilacqua A (2016) Pharmacodynamics and pharmacokinetics of inositol(s) in health and disease. Expert Opin Drug Metab Toxicol 12 (10):1181–1196
34. Anderson PW (1972) More is different. Nature 177(4047):393–396

35. Sarrió D, Rodriguez-Pinilla SM, Hardisson D, Cano A, Moreno-Bueno G, Palacios J (2008) Epithelial-mesenchymal transition in breast cancer related to the basal-like phenotype. Cancer Res 68(4):989–997
36. Steinberg MS (1986) Cell surfaces in development and cancer. Springer, New York
37. Peinado H, Olmeda D, Cano A (2007) Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? Nat Rev Cancer 7(6):415–428
38. Maeda M, Johnson KR, Wheelock MJ (2005) Cadherin switching essential for behavioral NOT morphological changes during an epithelium to mesenchyme transition. J Cell Sci 118 (Pt 5):873–887
39. Pasqualato A, Palombo A, Cucina A, Mariggiò MA, Galli L, Passaro D, Dinicola S, Proietti S, D'Anselmi F, Coluccia P, Bizzarri M (2012) Quantitative shape analysis of chemoresistant colon cancer cells: correlation between morphotype and phenotype. Exp Cell Res 318 (7):835–846
40. Chaitin GI (1974) Information-theoretic computational complexity. IEEE Trans Inf Theory 20(1):10–15
41. Hoppe PS, Schwarzfischer M, Loeffler D (2016) Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. Nature 535(7611):299–302
42. Weichsel J, Herold N, Lehmann MJ, Kräusslich HG, Schwarz US (2010) A quantitative measure for alterations in the actin cytoskeleton investigated with automated high-throughput microscopy. Cytometry A 77 (1):52–63
43. Tojkander S, Gateva G, Lappalainen P (2012) Actin stress fibers – assembly, dynamics and biological roles. J Cell Sci 125(8):1855–1864
44. Oksendal B (2000) Stochastic differential equations. Springer, Berlin
45. Quarteroni A, Sacco R, Saleri F (2007) Numerical mathematics. Texts in applied mathematics, vol 37, 2nd edn. Springer, Berlin
46. O'Malley RE (1991) Singular perturbation methods for ordinary differential equations. Applied mathematical sciences, vol 89. Springer, New York
47. Zmeskal O, Dzik P, Vesely M (2013) Entropy of fractal systems. Comput Math Appl 66 (2):135–146
48. Spillman WB, Robertson JL, Huckle WR, Govindan BS, Meissner KE (2004) Complexity, fractals, disease time, and cancer. Phys Rev E 70:061911
49. Chen Y (2016) Equivalent relation between normalized spatial entropy and fractal dimension. Available via arXiv.org > physics > arXiv:1608.02054. http://arxiv.org/abs/1608.02054. Accessed 4 May 2017
50. Dafermos C (2005) Hyperbolic conservation laws in continuum physics, 2nd edn. Springer, Berlin
51. Vrabie II (2004) Differential equations. An introduction to basic concepts, results and applications. World Scientific, River Edge
52. Richtmyer RD, Morton KW (1994) Difference methods for initial-value problems, 2nd edn. Robert E. Krieger Publishing Co. Inc., Malabar
53. LeVeque RJ (2007) Finite difference methods for ordinary and partial differential equations. Steady-state and time-dependent problems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia
54. Sanders J, Kandrot E (2010) CUDA by example: an introduction to general-purpose GPU programming, NVIDIA Corporation. Addison-Wesley, Upper Saddle River
55. Karniadakis GE, Kirby RM II (2003) Parallel Scientific Computing in C++ and MPI: a seamless approach to parallel algorithms and their implementation. Cambridge University Press, Cambridge
56. Murray L (2012) GPU acceleration of Runge-Kutta integrators. IEEE Trans Parallel Distrib Syst 23:94–101
57. Gantmacher F (1959) Applications of the theory of matrices. Interscience, New York

Chapter 8

Parameters Estimation in Phase-Space Landscape Reconstruction of Cell Fate: A Systems Biology Approach

Sheyla Montero, Reynaldo Martin, Ricardo Mansilla, Germinal Cocho, and José Manuel Nieto-Villar

Abstract

The thermodynamical formalism of irreversible processes offers a theoretical framework appropriate to explain the complexity observed at the macroscopic level of dynamic systems. In this context, together with the theory of complex systems and systems biology, the thermodynamical formalism establishes an appropriate conceptual framework to address the study of biological systems, in particular cancer.

The Chapter is organized as follows: In Subheading 1, an integrative view of these disciplines is offered, for the characterization of the emergence and evolution of cancer, seen as a self-organized dynamic system far from the thermodynamic equilibrium. Development of a thermodynamic framework, based on the entropy production rate, is presented in Subheading 2. Subheading 3 is dedicated to all tumor growth, as seen through a "phase transitions" far from equilibrium. Subheading 4 is devoted to complexity of cancer glycolysis. Finally, some concluding remarks are presented in Subheading 5.

Key words Biological phase transition, Tumor growth, Entropy production rate, Dissipation function, Metabolic rate, Fractal dimension, Glycolitic oscillations

Abbreviations

ATP	Adenosine triphosphate
ATPase	ATP-consuming processes
DPG	2-Phosphoglycerate
ENO	Enolase
F6P	Fructose-6-phosphate
G3P	Glyceraldehyde-3-phosphate
GAPDH	Gliceraldehyde-3-phosphate dehydrogenase
GLUT 1 and 3	Glucose transporters
HIF1	Hypoxia-inducible transcription factor
HK	Hexokinase
HPI	Hexose phosphate isomerase
Lac	Lactate
MCT	Monocarboxylate transporter
$NADH/NAD^+$	Nicotinamide adenine dinucleotide

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_8,

NHE	Exchanger NA^+/H^+
PFK-1	Phosphofructokinase type 1
PGK	3-Phosphoglycerate kinase
PYK	Pyruvate kinase
Pi	Inorganic phosphate
Pyr	Pyruvate
ROS	Reactive oxygen species
SNAIL	Transcription factor
SP1	Transcription factor
TPI	Triose phosphate isomerase

1 Why Is Systems Biology So Relevant to Cancer Studies?

General systems theory is about of the scientific exploration of "wholes" and "wholeness" that, not so long ago, were considered to be metaphysical notions transcending the boundaries of science. Novel concepts, methods, and mathematical fields have recently developed to deal with them. At the same time, the interdisciplinary nature of concepts, models, and principles applying to "systems" provides a possible approach toward the unification of science [1].

One of the theoretical foundations of Systems Biology is tightly associated with the "Modern Systems Theory" which was significantly influenced by two Austrian scientists, the biologist *Paul A. Weiss* and the philosopher and theoretical biologist Ludwig von Bertalanffy [2]. Toward the end of the 1960s, Bertalanffy published the concept of a general systems theory, where one of the key statements can be summarized as the following: "*Biology is autonomous and life itself cannot be reduced to disciplines in physics or chemistry or to physico-mechanistic relationships. As consequence, biology needs to be described from a differentiated and methodological point of view, where integrative and holistic aspects play the key role*". As a consequence of this view, Bertalanffy postulated the *organismic biology*, describing the entity of biological units as a main feature that, in its whole, is more than just the sum of all (single) units [2].

Efforts to define Systems Biology through a rational path toward the integration of multidisciplinary, multi-hierarchical levels of analysis have been disappointing. As a result, the concept of "Systems Biology" remains as a somewhat nebulous idea [3].

In 2005, two principal streams can be recognized within Systems Biology [3]: The Pragmatic Systems Biology and Theoretical Systems Biology. The first emphasizes the use of large-scale molecular interactions ("omic" approach), aimed at building complex signaling networks by applying mathematical modeling and thus showing how cells make decisions based on the "information" that flows through their networks.

Otherwise, Theoretical Systems Biology, according to which both theoretical and methodological approaches in biological research must be radically changed. That statement has recently been underscored [3]. Theoretical Systems Biology recognizes that complex physiological and adaptive phenomena take place at biological levels of organization higher than the subcellular one [3].

We define systems biology as the study of complex interactions in biological systems and the emergent properties that arise from such interactions. In the field of cancer, systems biology aims at developing an increasingly holistic view of cancer development and progression [4].

A system-level understanding of a biological system can be derived from insight into four key properties [5]:

1. System structures. These include the network of gene interactions and biochemical pathways, as well as the mechanisms by which such interactions modulate the physical properties of intracellular and multicellular structures.
2. System dynamics. How a system behaves over time under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as phase portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors. Bifurcation analysis traces time-varying change(s) in the state of the system in a multidimensional space where each dimension represents a particular concentration of the biochemical factor involved.
3. The control method. Mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease.
4. The design method. Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error. Progress in any of the above areas is reviewed at [5].

We are therefore facing a significant intellectual challenge: how to include chaotic and nonlinear, unpredictable processes into our comprehension of Biology. This task will likely improve our understanding of complexity of the real world, no longer confined to simplified and idealized phenomena. Systems Biology entails investigating phenomena in terms of how the objects are related, rather than what their compositions are [3].

It has been emphasized by von Bertalanffy that, "according to definition, the second law of thermodynamics applies only to closed systems, it does not define the steady state." The extension and generalization of thermodynamical theory has been carried through

by Prigogine. As Prigogine states, “classical thermodynamics is an admirable but fragmentary doctrine” [6].

This fragmentary character results from the fact that it is applicable only to states of equilibrium in closed systems. It is necessary, therefore, to establish a broader theory, comprising states of non-equilibrium as well as those of equilibrium. Thermodynamics of irreversible processes and open systems leads to the solution of many problems where, as in electrochemistry, osmotic pressure, thermodiffusion, Thomson and Peltier effects, etc., classical theory proved to be insufficient [6].

A living complex system is thermodynamically open and is characterized by a nonlinear dynamics, allowing it to have a history: this means that the present behavior of the system is in part determined by its past behavior. Such a system displays both sensitivity and resilience (robustness) with respect to the perturbations exerted by internal and/or external stimuli. In addition, living systems are characterized by both local and long-range interactions (non-locality), as well as by complex interactions between molecules and structures that make their determination “non-separable” (i.e., “entangled”), according to an analogy remnant of quantum mechanics [3].

A systems biology approach using new modeling techniques and nonlinear mathematics is needed to investigate gene-environment interactions and improve treatment efficacy. A broader array of study designs will also be required, including prospective molecular epidemiology, immune competent animal models, and in vitro/in vivo translational research that more accurately reflects the complex process of tumor initiation and progression [7].

Systems biology approaches are helping to understand the mechanisms of tumor progression and design more effective cancer therapies [4].

Cancer is a disease based on malfunctioning of the system properties of parts of biology. We hence identify it as a Systems Biology disease. Indeed, progress in cancer research toward cancer therapy may develop faster if cancer is not researched only in terms of Molecular Biology but rather in terms of Systems Biology [8].

Cancer is a heterogeneous and highly robust disease that represents the worst-case scenario of entire system failure: a fail-on fault where malfunction components are protected by mechanisms that support robustness in normal physiology [1, 2]. It involves hijacking the robustness mechanisms of the host. The survival and proliferation capability of tumor cells are robustly maintained against a range of therapies, due to intra-tumor genetic diversity, feedback loops for multidrug resistance, tumor–host interactions [9].

What is needed is to provide a conceptual framework able to integrate some entrenched aspects, such as complexity, hierarchical structured levels of observation, geometrical relationships,

nonlinear dynamics, network modeling, influence of biophysical constraints, operating on different scales, rather than solely focusing on building numerical mathematical or computer models [10].

Cancer is systemic by nature and reductionist approaches have failed to improve treatment and understanding substantially. Despite the variability in the nature of diseases related to cancer, it is expected that systems biology can make essential contributions to [11]:

- The identification of early biomarkers for a noninvasive prognosis of tumor development.
- Personalized medicine by building computer models predicting different stages of the disease.
- Improving treatment of later stages by comparing biochemical networks and gene expression levels in primary tumors and metastases [11].

Computational and mathematical approaches used in systems biology are highly versatile; a few categories of general methodologies have emerged for specific purposes in cancer research. One class is integrative statistical analysis of large-scale cancer multiomics and clinical data. These unbiased data-driven analyses have identified key biological processes underlying cancer pathogenesis, prognostic biomarkers, and predictive signatures for drug response [4].

Another class is mathematical modeling of interaction networks such as intracellular signaling pathways or extracellular crosstalk's between tumor and the microenvironment. These models have proved useful at unraveling mechanisms of drug resistance and in optimizing combinatorial targeted therapy. Furthermore, evolutionary models that simulate tumor growth and progression have provided important insights into the evolution dynamics of tumor and have led to the discoveries of more effective dosing schedules. Overall, the application of systems biology approaches has led to substantial improvements in our understanding of cancer initiation and progression and to the discovery and implementation of more effective anticancer therapeutic strategies [4].

The cancer complexity has been probed at the genomic, protein, post-translational, and tissue levels [4]. The architecture of signal transduction pathways is not where the complexity of cancer ends. Being parts of the cell, signaling networks are affected by additional levels of organization, for instance, as many proteins are not uniformly distributed over the cell. Areas with high protein concentrations might lead to macromolecular crowding and cause steep spatial gradients of activated signaling proteins. Numerous interactions at the supra-cellular level make the cancer system even more complex [8].

2 Thermodynamics Framework

From the formalism of the classical thermodynamics [12] entropy production can be evaluated through the variation of Gibbs's free energy dG_{Tp} when the system evolves subjected to the constraints the temperature T and the pressure p constants as

$$\delta S_i = -\frac{1}{T} dG_{Tp} \tag{1}$$

The temporal variation of the expression of Eq. (1) represents the entropy production rate as

$$\frac{\delta S_i}{dt} = -\frac{1}{T}\frac{dG_{Tp}}{dt} \tag{2}$$

where $\frac{\delta S_i}{dt} \equiv \dot{S}_i$ represents the entropy production rate. The term $\frac{dG_{Tp}}{dt}$ can be developed by means of the chain rule as a function of the degree of advance of the reaction ξ as

$$\frac{dG_{Tp}}{dt} = \left(\frac{\partial G}{\partial \xi}\right)_{Tp}\frac{d\xi}{dt} \tag{3}$$

where $\left(\frac{\partial G}{\partial \xi}\right)_{Tp}$, according to De Donder and Van Rysselberghe [13], represents the affinity $A = -\left(\frac{\partial G}{\partial \xi}\right)_{Tp}$, and the term $\frac{d\xi}{dt}$ is the reaction rate $\dot{\xi}$.

The rate of entropy production (Eq. 3) can be written as

$$\frac{\delta S_i}{\partial t} = \dot{S}_i = \frac{1}{T} A\dot{\xi} = -\frac{1}{T}\Delta G\dot{\xi} \tag{4}$$

where $A = -\Delta G$. The affinity A can be evaluated from the isotherm of the reaction [14] by the equation

$$A = RT \ln K_C - RT\sum_{i=1}^{k} \nu_k \ln C_k = RT \ln\left(\frac{K_C}{\prod C_k^{\nu_k}}\right) \tag{5}$$

where $K_C = \frac{k_f}{k_b}$ is the Guldberg-Waage constant; k_f, k_b are the specific rate constants of the direct and inverse reaction steps f, b, respectively; C_k is the concentration of the kth specie; and the ν_k are the stoichiometric coefficients that are taken, by agreement, as positive for the products and negative for the reactants. Therefore, Eq. (5) can be written as

$$A = RT \ln\left(\frac{k_f \prod C_{k(f)}^{\nu_{k(f)}}}{k_b \prod C_{k(b)}^{\nu_{k(b)}}}\right) \tag{6}$$

The rate of reaction $\dot{\xi}$ can be written as

$$\dot{\xi} = \left(\dot{\xi}_f - \dot{\xi}_b\right) = k_f \prod C_{k(f)}^{\nu_{k(f)}} - k_b \prod C_{k(b)}^{\nu_{k(b)}} \tag{7}$$

where $\dot{\xi}_f, \dot{\xi}_b$ are the velocities of the direct and reverse reaction passes, respectively. Substituting expressions (7) and (6) in to (5) is obtained:

$$\dot{S}_i = \left(\dot{\xi}_f - \dot{\xi}_b\right) \ln \frac{\dot{\xi}_f}{\dot{\xi}_b} \geq 0 \tag{8}$$

Formula (8) [14] is fulfilled regardless of whether the network of chemical reactions is close or far from equilibrium and also ends the controversy related to the divorce between classic thermodynamics and chemical kinetics.

Using the expression for the Gibbs equation, $G \equiv H - TS$, the affinity A can be evaluated as

$$A = -\left(\frac{\partial H}{\partial \xi}\right)_{Tp} + T\left(\frac{\partial S}{\partial \xi}\right)_{Tp} \tag{9}$$

The term $\left(\frac{\partial H}{\partial \xi}\right)_{Tp}$ represents the heat of process, q_{Tp}. Sometimes, it is possible to neglect the term $\left(\frac{\partial S}{\partial \xi}\right)_{Tp}$, due to $\left|q_{Tp}\right| \gg \left(\frac{\partial S}{\partial \xi}\right)_{Tp}$ [12]. Taking into account (Eq. 4), we get that the entropy production rate can be rewritten as

$$\frac{\delta S_i}{\partial t} = \dot{S}_i = \frac{1}{T} A\dot{\xi} \approx \frac{q_{Tp}\dot{\xi}}{T} = \frac{1}{T}\left(\frac{\delta q}{dt}\right)_{Tp} \tag{10}$$

Formula (10) is an approximation to the entropy production rate of a living organism [12]. Equation (10) can be rewritten according to Zotin [15] using the metabolic rate $\left(\frac{\delta q}{dt}\right)_{Tp} \equiv \dot{q}$ as follows:

$$\frac{\dot{S}_i = \dot{q}_{O_2} + \dot{q}_{Gl}}{T} \tag{11}$$

where $\dot{q}_{O_2}, \dot{q}_{Gl}$ are the metabolic rates of oxygen consumption $\dot{q}_{O_2}$, due to oxidative phosphorylation (OxPhos) and due also to glycolysis $\dot{q}_{Gl}$, respectively. Under aerobic conditions $\dot{q}_{Gl}$ is negligible, except in cancerous cells where the glycolysis is the main process [16].

Sometimes, it is convenient [12] to use the so-called dissipation function, $\Psi \equiv T\dot{S}_i$, introduced by Lord Rayleigh. According to Eq. (11) the dissipation function can be rewritten as

$$\Psi \equiv T\dot{S}_i = \dot{q}_{O_2} + \dot{q}_{Gl} \tag{12}$$

In the tumor cells the glycolysis is the main process [16], thus Eq. (12) can be written as

$$\dot{S}_i \approx \frac{\dot{q}_{Gl}}{T} \tag{13}$$

The appearance of new structures in nature far from thermodynamic equilibrium appears similar to a "phase transition," generically called bifurcation [17]. Due to the nonlinear nature of the dynamical system and the feedback processes, the fluctuations grow and amplify at the macroscopic level which leads to the appearance of the system's self-organization and consequently the complexity that it exhibits [18].

The seminal work of Landau [19] established the theoretical foundation of the phase transition. According to Landau formalism, the potential Φ or "Landau potential" is defined in terms of the variables that describe the system γ, and the order parameter μ [20].

The order parameter is defined empirically. As shown in a previous work, in the case of cancer evolution [21], the order parameter μ is taken as the difference between the fractal dimension of normal d_f^N and tumor cells d_f^C, thus $\mu = d_f^N - d_f^C$, where $\mu = 0$ in the symmetric phase (normal cells) and $\mu \neq 0$ for tumor cells. The order parameter μ is called the "morphological degree of complexity" [22].

The Landau potential Φ can be written as

$$\Phi(\gamma, \mu) = \Phi_0(\mu) + \alpha(\gamma)\mu^2 + \beta(\gamma)\mu^4 + \cdots \tag{14}$$

In the neighborhood of the transition point $\mu = 0$:

$$\left(\frac{\partial \Phi(\gamma, \mu)}{\partial \mu}\right)_\gamma = 4C\mu^3 + 2A\mu = 0 \tag{15}$$

the stability condition is fulfilled as

$$\left(\frac{\partial^2 \Phi(\gamma, \mu)}{\partial \mu^2}\right)_\gamma = 2A + 12C\mu^2 > 0 \tag{16}$$

Considering Eqs. (15) and (16), there are three possibilities:

1. $\gamma > \gamma_C \Rightarrow A > 0$,
2. $\gamma < \gamma_C \Rightarrow A < 0$,
3. By continuity we have: $\gamma = \gamma_C \Rightarrow A = 0$.

Although a satisfactory description is achieved through the dissipation function Ψ of tumor growth [22], Landau formalism is limited since it is based on a mean field theory [23]; in other words, it does not consider correlations or fluctuations.

For this reason, it is convenient to use instead, in order to describe the process of tumor growth the so-called Lyapunov function [24]. At the end of the nineteenth century, Lyapunov developed a method for studying the stability of equilibrium positions that bears his name. This method allows knowing the global stability of the dynamics of a system [25]. Let p be a fixed point, steady state, of the flow $\frac{dx}{dt} \equiv \dot{x} = f(x)$. A function $V(x)$ is called a

Lyapunov function for p if for certain neighborhood N of p the following conditions hold:

1. $V(x) > 0$ for every $x \neq p$ in N and $V(p) = 0$
2. The Eulerian derivative, $\frac{dV(x)}{dt} < 0$ for every x in N.

In this way, it can be stated for all $t \geq t_0$ that p is globally asymptotically stable, i.e., the system evolves to a minimum of the function $V(x)$.

Thus, we have that the entropy production per unit time meets the necessary and sufficient conditions for Lyapunov function [26]. Indeed:

$$\dot{S}_i = f(\Omega) > 0 \tag{17}$$

where Ω is the vector of control parameters. The Eulerian derivative (Eq. 17) must meet

$$\frac{d\dot{S}_i}{dt} = \frac{\partial \dot{S}_i}{\partial \Omega} \frac{d\Omega}{dt} < 0 \tag{18}$$

3 What Can Be Learned from a Phase Transition in Tumor Growth?

The thermodynamics formalism of irreversible processes [17], systems biology [3], and complex systems theory [27] offers a theoretical framework appropriate for the characterization of the emergence and evolution of cancer. In such a sense, in a previous work [28], we presented a conceptual definition of the cancer, which is expressed as follows: "*Cancer is a generic name given to a complex network of interactions of malignant cells, which have lost their specialization and control over normal growth. This network of malignant cells could be considered a nonlinear dynamical system, self-organized in time and space, far from thermodynamic equilibrium, exhibiting high complexity* [29], *robustness* [30], *and adaptability* [31]".

Tumor cells show two aspects of robustness: functional redundancy, due to cellular heterogeneity, and those arising because of feedback control systems [28]. The robustness of tumor cells allows a system to maintain its functionality against various external and internal perturbations [28]. The control of the robustness of the tumor cells constitutes a potential strategy for the development of drugs and therapies.

The evolution of a tumor runs through three basic stages: avascular, vascular, and metastasis [32]. In the avascular stage, the tumor grows to a state known as "dormant" state [33], with a microscopic nature ($\sim$1 mm diameter). This dormant stage can remain silent for a long time and is not macroscopically perceptible.

For reasons that are still unknown [34], the tumors that are in the dormant state leave this state and begin a process of

angiogenesis, vascular growth. We conjectured about it, in a previous work [35], where apparent fluctuations related to the joint action of the host and immune system on tumor cells may cause an adverse outcome causing a type of stochastic resonance effect. This leads to a change in the fractal dimension of the tumor interface and consequently a certain number of active cells in the tumor interface could escape which could lead to vascular growth. It is an acceptable explanation of why a tumor in a latent phase, stationary state [36], can go to a critical state, reach macroscopic dimensions, vascular phase, and subsequently invade distant organs, metastasis, despite actions of the immune system, and the host [37, 38].

It is important to note that whereas it is possible to predict when tumors reach latency, as in a first approximation occurs, it is virtually impossible to predict when a tumor metastasizes [39], given the highly random character of the action of the immune system and the host.

In the vascular phase, the tumor acquires macroscopic dimensions, invading much of the host and adjacent organs, that is to say, the nonequilibrium self-organized tumor system acquires a higher level of hierarchy and apparently robust as it is known that in most cases after surgical removal there of, micro-metastasis is found [40].

The process of metastasis [41] appears abruptly as a reminiscent of hard, first-order transitions, in these cases, the chances of survival are lower compared to the previous stages, because they exhibit a higher robustness and a higher level of hierarchy. The tumor now competes with the different levels of hierarchical and functional organization of the body (those which play vital roles), so it is considered like a cancer tumor, given its ability to metastasize [42].

The pioneering work of Prigogine and Lefever [43], considering the stability of tumor growth in the presence of cytotoxic cells, revealed that cancer growth can be described by a phase transition, like a second-order phase transition. Moreover, Delsanto et al. [44] developed a dynamical system model for the analysis of phase transition from tumor growth to latency, while Solé showed that tumors have a behavior close to a limit of instability, as well as the tumor cell populations [45]. Davies et al. [46] argued that the transition from health cells to malignant cells can be described by a phase transition of the first order.

In order to describe the dynamics of avascular process based empirically on the evidence discussed above, we have proposed the following heuristic mechanism [21] sustained by a chemical network model

$$
\begin{aligned}
&(1)\ \ N + x \rightarrow 2x \\
&(2)\ 2x \rightarrow \text{ncp} \\
&(3)\ H + x \rightarrow \text{ncp}
\end{aligned}
$$

Steps 1, 2 are related to the process of mitosis and apoptosis, where x represents the population of proliferating tumor cells in an avascular phase, and the reactions are associated with the mitosis k_{m} and apoptosis k_{ap} constants. The N represents the population of normal cells, H the population of the host cells [38]. The numbers of members of N , H are considered constants. Finally, ncp represent non-cancerous products.

The dynamic behavior of the chemical network model for avascular growth is given by the following ordinary differential equation [21]:

$$\frac{dx}{dt} = \gamma x - x^2, \tag{19}$$

where $\gamma = Nk_{\mathrm{m}} - H$. The exact solution of Eq. (19) is $x(t) = \frac{\gamma}{1+\mathrm{c}^{-\gamma t}}$. The stability analysis of Eq. (19) shows that at $\gamma = 0$ a transcritical bifurcation [47] takes place and the system evolves (avascular growth) toward a stable steady state, known as the dormant state [48]. Hence, as we postulated in previous works [21, 49], this process resembles a "second-order" phase transition, whose biological implication is clear: the difficulty of early detection of cancer.

Such formalism discussed above [21] can be extended to the study of the dynamics of prostate tumor cell lines, LNCaP and PC3 [49]. The system in question is a 2D region with characteristic length L in which initially there are very few tumor cells. The cell density increases with time due to the proliferation of these cells (*see* Fig. 1).

The morphology observed in this region has fractal nature as a result of the stochastic nature of the mitosis and apoptosis processes that occur at the level of single cells [28]. Thus we have

$$x = L^{d_{\mathrm{f}}}, \tag{20}$$

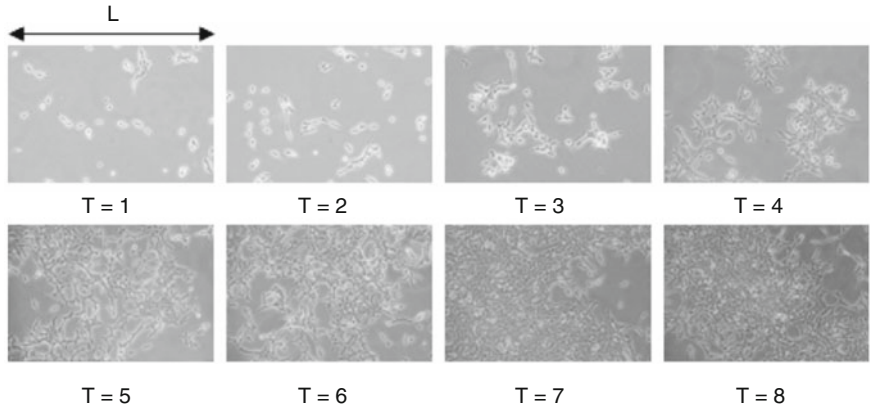


Fig. 1 In vitro growth of prostate tumor cell line LNCaP. *T* is time in days. Reprinted from [49]

where x as we have said, represents the population of proliferating tumor cells in the region of characteristic size L and d_f is the fractal dimension of the contour. The temporal variation of Eq. (20) is given by

$$\frac{d \ln x}{dt} = \frac{\partial d_f}{\partial t} \ln L + d_f \frac{\partial \ln L}{\partial t}. \tag{21}$$

We show that the temporal variation of the total number of cells N can be described equivalently through a logistic or Gompertz dynamical equation [21], such that

$$\frac{d \ln d_f}{dt} = -\ln d_f + \alpha \ln K, \tag{22}$$

where $K = d_{f(0)} e^{\frac{\widehat{A}}{\alpha}}$ is the carrying capacity and $\widehat{A}$ and α are empirical constants in the Gompertz equation [50]. Equation (22) has the analytical solution

$$d_f(t) = k\left(\exp\left(-\frac{\widehat{A}}{\alpha}\right)\right)^{e^{-t\alpha}} \tag{23}$$

In Fig. 2, the time dependence of the fractal dimension d_f is shown for the dynamical behavior of the prostate tumor cell lines, LNCaP and PC3 respectively.

Unlike other cells lines showing fractal behavior [38], the prostate tumor cell lines, LNCaP and PC3 exhibit a multifractal behavior.

The evolution of a tumor, through the three basic stages avascular, vascular, and metastasis [32], could be generalized throughout the following heuristic mechanism [21] based on a chemical network model (see Fig. 3).

The present model (Fig. 3) is a qualitative representation of the dynamics of the development of cancer, based on the above experimental evidence, without claim to suggest any type of therapy. The x , y , z represents the population of proliferating tumor cells in an avascular, vascular, and metastases phases respectively; N represents the population of normal cells, H the population of the host cells [38] is considered constant and I the population of immune cells (T lymphocytes, CTL and natural killer, NK cells [39]) is considered the control parameter. Finally, ncp represents a non-cancerous product.

In the Fig. 3, the reactions 1, 3 and 2, 4, 6 are related to the process of mitosis and apoptosis of the proliferating tumor cells respectively; 5 and 7 correspond to the action of the host H [38]. Finally, reactions 8 and 9 show the action of the population immune cells. It is assumed that the cells of type z are generated during cell apoptosis y (reaction 4) representing the occurrence of micrometastases [40], that is to say, the spread of cancer cells from

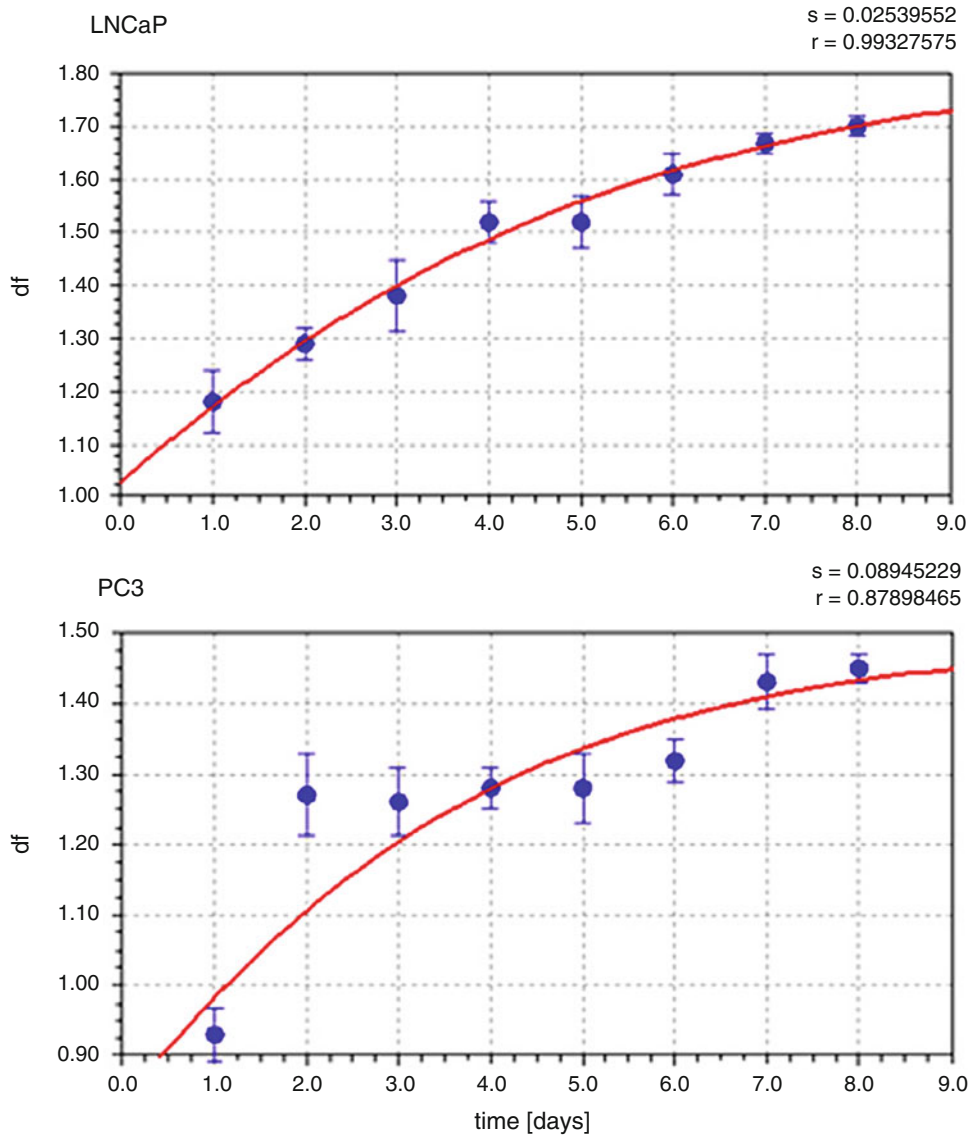


Fig. 2 Dependence of fractal dimension d_f with the growth time of the tumor cell lines: LNCaP ($\widehat{A} = 0.2524, \alpha = 0.1466, K = 1.8384$) and PC3 ($\widehat{A} = 0.3335, \alpha = 0.1944, K = 1.4924$). Reprinted from [49]

the primary site with the secondary tumors being too small to be clinically detected.

The model proposed (*see* Fig. 3) contains as a particular case the avascular growth discussed previously [21]. Starting from the proposed model (Fig. 3) and under the Law of Mass Action, the system of ODEs is obtained:

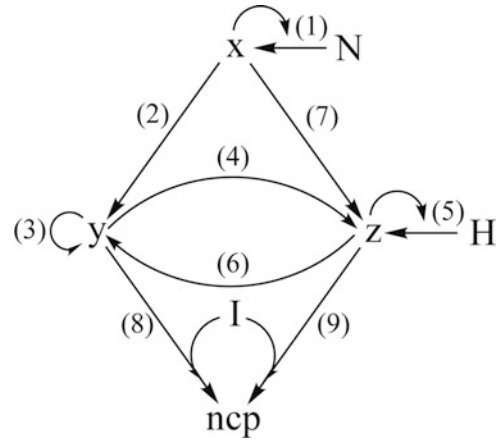


Fig. 3 Chemical network for the model for cancer growth

$$\begin{aligned}\frac{dx}{dt} &= x(2N - x) - Hxz,\\ \frac{dy}{dt} &= y(4 - 0.14y) + 0.5x^2 - Iy - 0.5Hyz + 0.001z^2,\\ \frac{dz}{dt} &= Iz + 0.07y^2 + 0.5Hyz - 0.002z^2,\end{aligned} \tag{24}$$

The numerical values of the constants were taken from the model of chemical network studied in a previous work [83]. Obtaining fixed points, stability analysis and bifurcation was performed using the standard procedure [24, 51, 52] using as a control parameter the population of immune cells I (T lymphocytes, CTL and natural killer, NK, cells [39]).

Lyapunov exponents were calculated using the classical Wolf algorithm, FORTRAN language [53]. In relation to the calculation of the Lyapunov exponents, note that:

1. It is important to choose an appropriate algorithm, usually the Wolf algorithm [53] that yields good results.
2. The length of the time series used must contain at least 2000 values to achieve good results [14].
3. In the calculation it must be verified that there is no appreciable difference between the sum of the Lyapunov exponents and the divergence of the flow.

The LZ complexity [54, 55] was calculated using the proposed algorithm by Lempel and Ziv, Lyapunov dimension D_{L}, also known as dimension Kaplan-York [56], it was evaluated across the spectrum of Lyapunov exponents λ_j as

$$D_{\mathrm{L}} = j + \frac{\sum_{i=1}^{j} \lambda_i}{\left|\lambda_{j+1}\right|}, \tag{25}$$

where j is the largest integer number for which $\lambda_1 + \lambda_2 + \cdots + \lambda_j \geq 0$. The results are summarized in Table 1.

Table 1
Stability, and complexity for the system of ODEs (Eq. 24) for different values of the control parameter *I* ($N = 5$, $H = 3$). Reprinted from [83]

I	Eigenvalues of the Jacobian matrix	Lyapunov exponents λ_j	LZ complexity	D_L
4 Ss_s stable focus	$-7.2 \times 10^{-2} - 5.3i$ $-7.2 \times 10^{-2} + 5.3i$ -7.8	-0.0720024 -0.0740862 -7.82744	–	0
3 Limit cycle (saddle-foci)	$+8.5$ $-3.3\text{e-}001 - 1.6i$ $-3.3\text{e-}001 + 1.6i$	~0.00 -0.227573 -5.89284	0.03589	1
2 Saddle-foci	$+6.3$ $-1.3 \times 10^{-1} - 1.9i$ $-1.3 \times 10^{-1} + 1.9i$	~0.00 -0.30561 -4.05829	0.03888	1
1 Saddle-node	$+10$ $+3.0$ -1.0	~0.00 -0.779727 -1.87158	0.04187	1
0.7 Limit cycle	$+3.5$ $-4.3 \times 10^{-2} - 1.5i$ $-4.3 \times 10^{-2} + 1.5i$	~0.00 ~0.00 -2.10445	0.06580	2
0.4 Shilnikov's chaos	$+2.8$ $-3.3 \times 10^{-2} - 1.2i$ $-3.3 \times 10^{-2} + 1.2i$	$+0.0519588$ ~0.00 -1.71075	0.08972	2.03

For the modeling of the chemical network model which appears in Fig. 3, COPASI v. 4.6.32 software was used. On the other side, numerical integration was performed of the system of ODEs Eq. (24) through implementation of the Gear algorithm for hard equations ("stiff"), in Fortran with double precision and tolerance of 10^{-8} [57]. It is important to note that most systems of ordinary differential equations, derived from biological systems, have rigidity, i.e., stiffness, which is why it is advisable to use numerical methods such as predictor corrector, for example the algorithm of Gear [57]. For bifurcation diagram, the package TISEAN 3.01 [58] was used for Poincare maps, correlation dimension, and power spectrum.

In Fig. 4 dynamical behavior of the proposed chemical network model for different values of the control parameter I is shown. When the value of control parameter I is smaller them the EDO system Eq. (24) exhibit the maximum complexity.

For "high" values of immune surveillance ($I = 4$) it exists a population of proliferating tumor cells, where x cells are the predominant, which corresponds to the phase state where a growth avascular stable steady state is reached, the tumor grows to a state known as dormant state [33, 35].

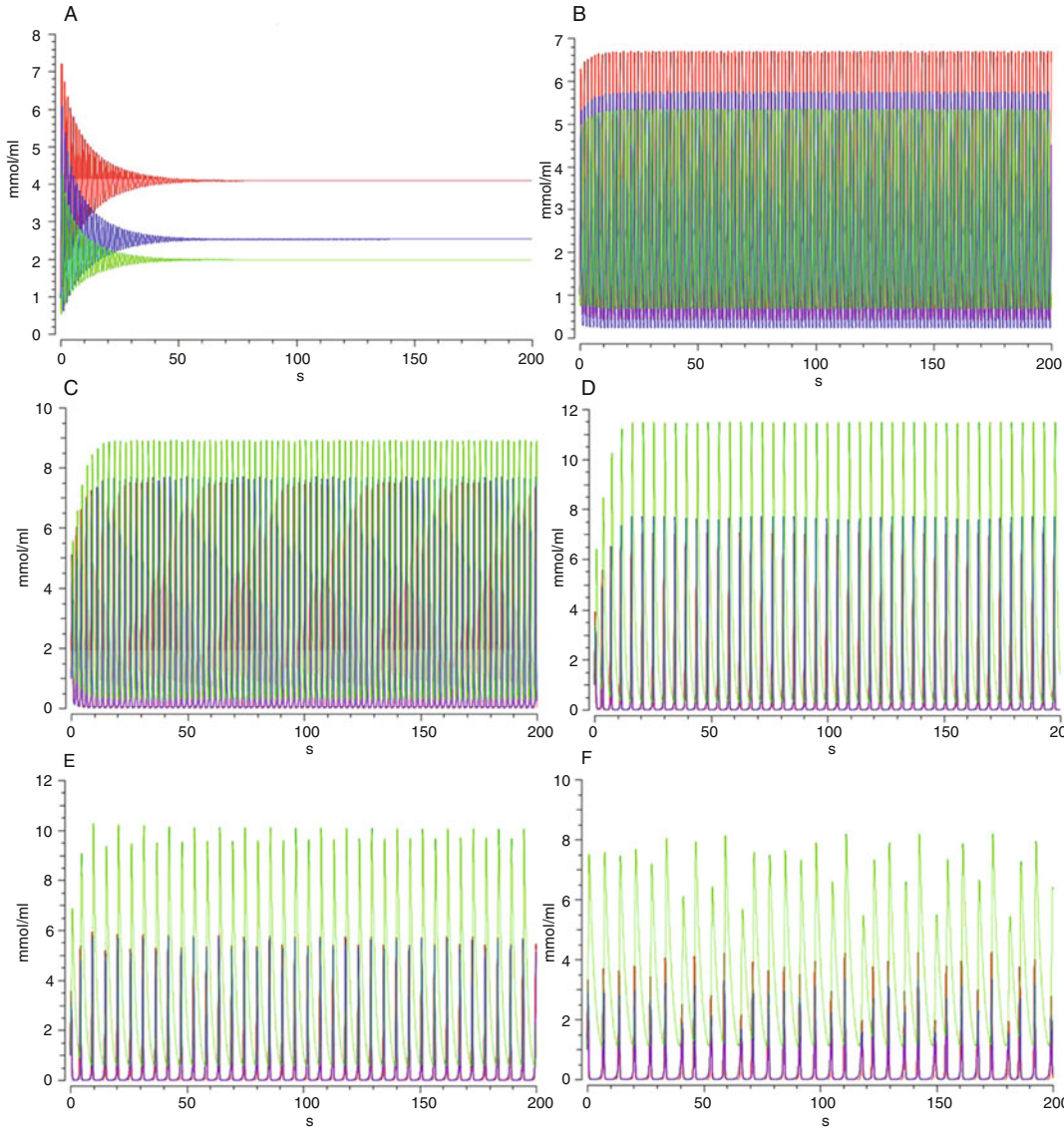


Fig. 4 Dynamical behavior of the proposed model (Eq. 24) for different values of the control parameter I ($N = 5$, $H = 3$): (**a**) stationary state ($I = 4$), (**b**) limit cycle ($I = 3$), (**c**) limit cycle ($I = 2$), (**d**) limit cycle ($I = 1$), (**e**) limit cycle ($I = 0.7$), (**f**) chaos ($I = 0.4$); *red* (x), *blue* (y) and *green* (z). Reprinted from [83]

For the critical value $I \approx 3.71$ it appears as a phase transition of the type "first-order" through a supercritical Andronov-Hopf bifurcation as described in [24, 52], giving rise to a limit cycle, periodic oscillations, and self-organizing system outside the thermodynamic equilibrium. It is further noted as the maximum concentration values of the population of proliferating tumor cells increase significantly.

It is noteworthy, as can be seen in Fig. 4, as far as it decreases the value of the control parameter I, although the system exhibits

periodic oscillations, the maximum concentration of z cells increases, and in fact they become predominant.

From the control parameters $I \approx 0.65$ a cascade of bifurcations is triggered by the route of saddle-focus Shilnikov's bifurcation [59] and even when the maximum values of the population of proliferating tumor cells decrease, the prevalence of metastatic cells z continues.

It shows how, for lower outcomes of the critical value of the control parameter (*see* Table 1, $I = 0.4$ and Fig. 4f), related to immune surveillance, tumor cells exhibit random behavior, as Shilnikov's chaos [59] (*see* attractor and return map in Fig. 5).

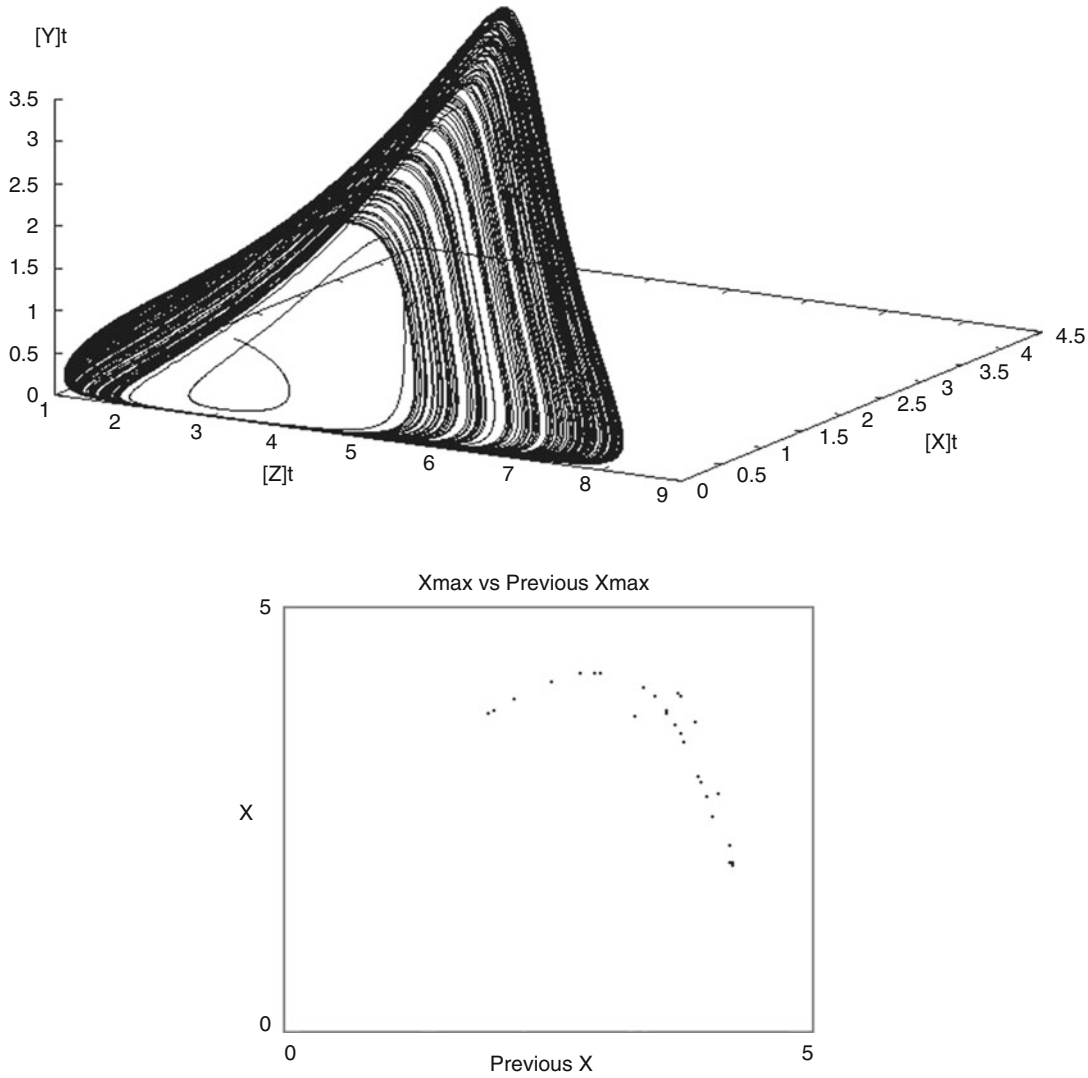


Fig. 5 Chaotic attractor and return map obtained from Eq. (19), with $I = 0.4$. Reprinted from [83]

It is important to emphasize how the homoclinicity in the Shilnikov scenario for Shilnikov condition: $-\Re e_3 \gg \Re e_\pm > 0$ appears. This suggests a rapid injection of flow in the plane of the stable manifold with a relatively slow winding. The striking fact is that other authors [60, 61] have developed models, based on different assumptions, which exhibit similar behavior. In addition, there is sufficient experimental evidence [62–66] in relation to the complexity of the dynamics of cancer.

Because the process of metastases presents chaos like Shilnikov's has important biological implications. On one hand, the high sensitivity of the system to initial conditions determines the inability to make long-term predictions in relation to the evolution of the disease, i.e., the end forecasts are improbable (poor prognosis). In addition, it gives a high degree of robustness [67, 68], hence the level of success in the treatment is low, especially at the stage of metastases [69]. Additionally, the information created during the evolutionary process of cancer cannot be destroyed [70], which is manifested in clinical recurrence (relapse) of cancer after a while that has apparently been "removed."

It is well known that the tumor progression in prostate cancer following radical prostatectomy occurs in 20–40% of patients, even though in pathological stage, approximately 40% of patients who undergo surgical removal of non-small cell lung cancer without overt metastases relapse within 24 months after surgery, in breast cancer, 20% of clinically disease-free patients relapse 7–25 years after mastectomy [33, 40].

Therefore, the tactic in the treatment of cancer lies not only in their physical elimination also must change the information created [71, 72]. Recently, a study [73] has shown that pancreatic cancer cells can be coaxed to revert back toward normal cells by introducing a protein called E47. The protein E47 binds to specific DNA sequences and control genes involved in growth and differentiation.

In other words, the growth of a primary tumor from a microscopic level, avascular phase, to the macroscopic level, vascular phase and subsequent onset of metastases, is not simply an accumulation of malignant cells but is a process exhibiting: (1) a non-linear dynamic, (2) self-organization out of thermodynamic equilibrium, (3) showing that a high degree of robustness, complexity, and level of hierarchy; which leads to the creation of new information and learning capacity.

The metastases process is associated with a biological phase transition type "first-order" through supercritical Andronov-Hopf bifurcation, coming out of limit cycle and subsequently through a cascade of bifurcations of the type saddle-foci Shilnikov bifurcations. This could explain the epithelial-to-mesenchymal transition during metastases [74].

The diagnosis of tumor proliferation capacity and invasion capacity is a very complex issue, because these terms include many factors such as the tumor aggressiveness, which is related to the tumor growth rate $\dot{\xi}$, and the tumor invasion capacity, which is associated with the fractal dimension d_f [75] among other factors. We have that $\dot{S}_i$ can be approximately determined (*see* Eq. (8)) from the rate of mitosis and apoptosis, ψ , η as

$$\dot{S}_i \approx R(\psi - \eta) \ln \frac{\psi}{\eta} \tag{26}$$

where $(\psi - \eta) \equiv \dot{\xi}$ is a tumor growth rate.

The fractal dimension d_f quantifies the tumor malignancy, in other words, the tumor capacity to invade and infiltrate healthy tissue [75]. In fact, the fractal dimension, as has been pointed out by other authors [76], can be considered "... *a quantitative shape descriptors which possess thermodynamics meaning and they could provide insights into the complexity score of the observed system....*"

As shown in previous works [35], the fractal dimension d_f can be given as a function of the quotient between mitosis ψ and apoptosis η rates, which quantify the tumor aggressiveness as

$$d_f = \left(\frac{5 - \frac{\psi}{\eta}}{1 + \frac{\psi}{\eta}} \right) \tag{27}$$

Substituting (Eq. 27) into (Eq. 26) we obtain that the entropy production rate can be expressed as a function of fractal dimension d_f as

$$\dot{S}_i = R\dot{\xi} \ln \left(\frac{5 - d_f}{1 + d_f} \right) \tag{28}$$

In equation (28) two properties fundamentals of the tumor growth are exhibited: the rate of growth $\dot{\xi}$, which is associated with their invasiveness capacity and a morphological characteristic as the fractal dimension d_f of the tumor interface that as we have said, quantifies the capacity of the tumor to invade and infiltrate healthy tissue, that is, its complexity. This represents the "degree of malignancy," which quantifies the capacity of the tumor invasiveness and infiltration into the healthy tissue.

An increment of d_f is associated with a reduction of the values of the quotient mitosis/apoptosis and to higher values of entropy production rate and this is consistent with the findings of Luo [77].

Starting from Eqs. (28) and (13) and substituting in (12) the following results hold:

$$\Psi = TS_{\cdot i} = \dot{q}_{\mathrm{Gl}} = TR\dot{\xi} \ln \left(-\frac{1}{d_f + 1}(d_f - 5) \right) \tag{29}$$

Equation (29) shows a relationship between the entropy production per unit time, the dissipation function of metabolic rate,

Table 2
Dissipation function for different human tumor cell lines, $T = 310$ K , $R = 8.31$ J/mol K. Reprinted from [22]

Human tumor cell lines	Growth rate[a] $\dot{\xi}\ \left[\frac{\mu m}{h}\right]$	d_f[a]	Metabolic rate $\dot{q}_{GI}\ \left[\frac{kJ\,\mu m}{mol\,h}\right]$
AT5 Primary human foreskin fibroblasts Human	8.72	1.23	11.97
C-33a Cervix carcinoma Human	6.40	1.25	8.46
HT-29 Colon adenocarcinoma Human	1.93	1.13	2.98
HT-29 M6 Mucus secreting HT-29 cells Human	1.85	1.12	2.88
HeLa Cervix carcinoma Human	1.34	1.30	1.64
Saos-2 Osteosarcoma Human	0.94	1.34	1.09

[a][38]

the fractal dimension, and the tumor growth rate. A common property of invasive cancers is their altered glucose metabolism. The glycolytic rate in cultured cell lines seems to correlate with tumor aggressiveness [78]. That an increase in proliferation and a reduction in apoptosis trigger a maximum of ATP consumption by the tumor cells has been already confirmed [79]. The tumor metabolome—mainly characterized by the glycolytic phenotype—confers an advantage to the evolving cancer cell population and contributes to their tissue invasion and the metastasis spreading [80].

The dissipation function could be evaluated through Eq. (29) from the data reported by Brú [38] for different tumors cell lines (*see* Table 2).

During the avascular development the growth rate remains constant [35], it can be proved from Eq. (29) that the metabolic rate exhibits a maximum with respect to the fractal dimensions (*see* Fig. 6).

As other authors have suggested, the thermodynamic dissipative function is correlated with both the glucose metabolism and cell shape [76]. We suggest that the latter could interfere with the metabolic pathways.

In previous works [26, 81, 82] we have shown that the rate of entropy production is a Lyapunov function (*see* Eq. (18)), in fact we extended this formalism to the development of cancer [21, 22, 49, 83]. The dissipation function is related to the entropy production rate through Eq. (29), showing that the dissipation function $\Psi = f\left(\dot{\xi}, d_f\right) \geq 0$ is a function of the fractal dimension d_f and of tumor growth rate $\dot{\xi}$. Taking the Eulerian derivative of Eq. (12), we have

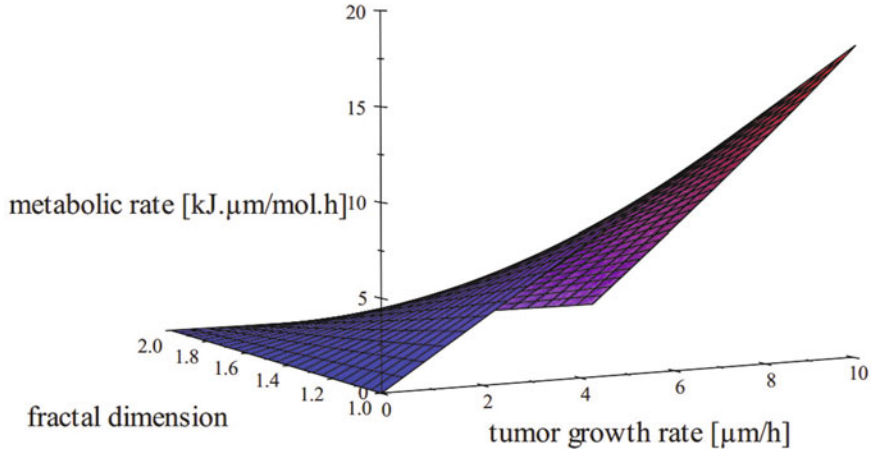


Fig. 6 Relationship between metabolic rate $\dot{q}_{\mathrm{GI}}$, fractal dimension d_{f}, and tumor growth rate $\dot{\xi}$. Reprinted from [22]

$$\frac{d\Psi}{dt} = \left(\frac{\partial \Psi}{\partial \xi}\right)_{d_{\mathrm{f}}} \frac{d\dot{\xi}}{dt} + \left(\frac{\partial \Psi}{\partial d_{\mathrm{f}}}\right)_{\dot{\xi}} \frac{d(d_{\mathrm{f}})}{dt} \tag{30}$$

Taking tumor growth rate $\dot{\xi}$ constant as in [38] and taking fractal dimension d_{f} as a control parameter, Eq. (30) can be rewritten as

$$\frac{d\Psi}{dt} = \left(\frac{\partial \Psi}{\partial d_{\mathrm{f}}}\right)_{\dot{\xi}} \frac{d(d_{\mathrm{f}})}{dt}, \tag{31}$$

In previous work [49] we obtained the analytical solution of the Gompertz equation for LNCaP tumor cell line (*see* Eq. (23)).

Substituting Eq. (23) into Eq. (29) it can be verified that $\left(\frac{\partial \Psi}{\partial d_{\mathrm{f}}}\right)_{\ddot{\xi}_{\mathrm{n}}} < \ddot{\xi}_{\mathrm{n}} 0$, and as $\frac{d(d_{\mathrm{f}})}{dt} \geq 0$ [49] we have

$$\frac{d\Psi}{dt} \leq 0 \tag{32}$$

This shows that the dissipation function Ψ is a Lyapunov function of the control parameter d_{f}.

4 Complexity of Cancer Glycolysis

In the last years, cancer glycolysis has been a target in oncology research [78]. Most of tumor cells show a high glycolytic rate compared with normal cells. This phenomenon is known in the literature as Warburg's effect and is observed without an increase in the tricarboxylic acid cycle (TAC) or the electron transport chain (ETC) rate [84]. The significant increase in glycolysis rate observed in tumors has been recently verified, yet few oncologists or cancer

researchers understand the full scope of Warburg's work [78, 84] despite its great importance. Altered energy metabolism is proving to be as widespread in cancer cells as many of the other cancer-associated traits that have been accepted as hallmarks of cancer [16]. The regulation of metabolism, relevant to senescence process, would be a key to improve and identify new anti-cancer therapies in the future.

In cancerous cells the glycolysis is the main process (*see* Eq. (13)). Then according to Luo [77] the affinity of glycolytic processes yields 52 kcal/mol and 1 mol of glucose produces 2 mol of ATP. Accordingly, the affinity decrease is

$$A_{\mathrm{Gl}} = 52 - 7.3 \times f_{\mathrm{C}}\ [\mathrm{kcal/mol(glucose)}] \tag{33}$$

In the cancer cell (the modifying factor f_{C} comes from the dependence of standard free energy on the pH value of cancerous cells), namely, $f_{\mathrm{C}} = 0.83$ [85], hence we have

$$A_{\mathrm{Gl}} = 39.9\ [\mathrm{kcal/mol(glucose)}] = 6.65[\mathrm{kcal/mol(O_2)}] \tag{34}$$

Taking into account Eq. (10), Eq. (13) can be rewritten as

$$\dot{S}_i \approx \frac{\dot{q}_{\mathrm{Gl}}}{T} = \frac{1}{T} A_{\mathrm{Gl}} \dot{\xi}_{\mathrm{RR}} \tag{35}$$

where $\dot{\xi}_{\mathrm{RR}}$ is the respiration rate (RR) of cancer cells [86], and T is the temperature in the physiological conditions used experimentally [87]: $T = 310$ K. The entropy production rate, Eq. (35), was determined for 16 human tumor cells from the data reported by Moreno-Sánchez et al. [88] for the respiration rate for different cancer cells (*see* Table 3).

Tumor growth rate is usually characterized by the tumor volume doubling time (DT). The term DT was introduced 50 years ago and a graphical method was proposed for its estimation [89]. The doubling time is the period of time required for a quantity to double in size or value.

As can be seen in Table 3, the entropy production rate exhibits no global correlation with doubling time. On the other hand, if similar cells were compared (as HepG2 and Hep 3B or breast carcinoma cells or leukemia cells) it can observed that the entropy production rate increases with decreasing tumor volume doubling time, and so increases tumor aggressiveness.

It is well known [93] that the aneuploid HTB-126 cell line originated from a human mammary epithelial carcinoma grows aggressively in soft agar, exhibits invasive properties at in vitro matrigel outgrowth assays, and causes tumors that extensively metastasize in immunodeficient nude mice. The human MCF-7 breast epithelial cell line is estrogen receptor positive and retains many of the biochemical and phenotypic characteristics of normal mammary epithelial cells. It has only a very limited ability to grow in soft agar, is noninvasive in matrigel assays, and does not form

Table 3
Entropy production rate for different human tumor cells

Human tumor cells	Doubling time [90] (h)	Respiration rate [88] $\left(\frac{\text{nmol}O_2}{\text{min mg cell}}\right)$	$\left(\frac{\dot{S}_i \times 10^{-9}}{\frac{\text{kcal}}{\text{K min mg cell}}}\right)$
HeLa (human cervix carcinoma)	48	3–6.5	0.102
MCF-7 (human breast carcinoma)	60 [91]	7	0.150
HTB-126 (human breast carcinoma)	40 [91]	28.5	0.611
SH-SY5Y (human neuroblastoma)	>55	10	0.215
SF188 (human glioblastoma)	–	5.5	0.118
H460 (human lung cancer cells)	42–60	6	0.129
HCC4017 (non-small-cell lung cancer)	–	12.4	0.266
HL-60 (human promyelocytic leukemia cells)	About 40	2–3.4	0.058
U937 (human histocytic leukemia cells)	30–40	1.5–3.3	0.052
Jurkat (human acute lymphoblastic leukemia cells)	25–35	3.6	0.077
KMS20 (human myeloma)	–	30	0.644
HepG2 (human hepatoma cells)	50–60	6.7	0.144
Hep3B (human hepatoma cells)	40–50	9.6	0.206
LNCAP (human prostate cancer)	About 60	18.7	0.401
PC3 (human prostate cancer)	50	45	0.965
143B (human osteosarcoma)	–	4.9	0.105
AGS (human gastric cancer cell line)	20 [92]	8	0.172

For doubling time the reference is [90], except for those whose reference is either individually given in doubling time column or no value at all is given. Reprinted from [22]

tumors in nude mice. Indeed, MCF-7 are widely recognized to present only some malignant features, showing a low aggressive behavior [94, 95]. The cell line HTB-126 shows a much greater value for the entropy production rate compared to a MCF7 (*see* Table 3).

The human hepatoma cells HepG2 and Hep3B have very different morphology and even differ in their growth rates. HepG2 cells are slowly growing with average doubling time around 48 h whereas Hep3B grows faster with doubling time of 24 h. These cell lines differ in *p*53 functionality/mutation. Hep3B is *p*53-null, while HepG2 has wild-type. Also, Hep3B contain

hepatitis B, while HepG2 cells do not. Generally, HepG2 cells are highly susceptible to chemical agents and drugs, while Hep3B cells are more resistant because of *p*53 deficiency [96]. As can be seen (Table 3), the cell line Hep3B shows a much greater value for the entropy production rate compared to a HepG2.

The cell line PC3, which is known to have an increased invasive ability and is more aggressive as compared to the LNCaP [97, 98], shows a much greater value for the entropy production rate compared to a LNCaP. These results corroborate, from the new point of view of thermodynamics, what other studies have shown for the cell line PC3, that is, it is more malignant, aggressive and has a higher metastatic potential and is more resistant to treatment compared to LNCaP [99]. It was found that the cellular line PC3 exhibits a greater value of the entropy production rate compared to LNCaP. As a fact, a similar result was found by us in previous work [49].

These results show how the entropy production rate can be a useful tool to quantify the robustness and it may be used as a quantitative index of the metastatic potential of tumors. (As can be seen, this analysis suggests that the tumor cell lines that exhibit a greater value of the entropy production rate have an increased invasive ability and so they have more aggressive capacity. As a matter of fact, the entropy production rate can be the useful tool to quantify the robustness and it may be used as a quantitative index of the metastatic potential of tumors.)

4.1 The Influence of Hypoxia, Normoxia, and pH in Tumor Cells

The metabolism of cancer as a therapeutic target is under constant investigation, along with the search for small molecules that are capable of specifically inhibiting the key metabolic pathways associated with cell growth [100]. Attenuation or inhibition of glycolysis has been very helpful in preventing cancer development, demonstrating that glycolysis is essential for proliferation, invasion, and metastasis [100, 101]. Identification of the most important reactions involved in the regulation of the glycolytic pathway is a useful strategy to define therapeutic targets in oncology, and thus can be a crucial step in cancer drug development. In addition, research on the influence of intracellular pH in the robustness of the glycolytic mechanism in neoplasic cells is equally important to understand the cancer biology.

In previous studies we showed the way entropy production per time can be used to select the main steps in a complex chemical reactions network, as Belousov-Zhabotinsky reaction [102]. Moreover, we also showed that entropy production is a specific fingerprint of the behavior of tumor, related to cancer robustness and prognostic of the disease [28]. The entropy production per time unit $\dot{S}_i$ with T, p fixed, disregarding diffusive and viscous effects of each reaction of the glycolytic pathway, was assessed in [17] as Eq. (4).

The Gibbs free energy of the kth reaction is written [103] as

$$\Delta G_k = \Delta G_k^{\oplus}(T, \mathrm{pH}, I) + RT \sum_i \nu_i \ln c_i, \tag{36}$$

where ν_i, c_i represent the stoichiometric coefficients and concentrations respectively of the involved biomolecules in each reaction and $\Delta G_k^{\oplus}(T, \mathrm{pH}, I)$ is the standard Gibbs free energy adjusted taking into account its dependence of temperature, pH and ionic force I [104, 105], in the physiological conditions used experimentally [87]: $T = 310.15$ K, $I = 0.18$ M and pH = 7.

To calculate the rectified standard Gibbs free energy ($\Delta G_k^{\oplus}$) eq. (37) was used,

$$\begin{aligned}
\Delta G_{k\,(T,\mathrm{pH},I)}^{\oplus} &= \sum_n \Delta G_{n\,(T,\mathrm{pH},I)}^{\oplus} \\
\Delta G_{n\,(T,\mathrm{pH},I)}^{\oplus} &= \frac{T}{298.15}\Delta G_n^{\theta} + \left(1 - \frac{T}{298.15}\right)\Delta H^{\theta} \\
&\quad + (N_{\mathrm{H}} RT \ln 10)\mathrm{pH} - \frac{RT\alpha\left(z^2 - N_{\mathrm{H}}\right)\sqrt{I}}{\left(1 + 1.6\sqrt{I}\right)_{1/2}}
\end{aligned} \tag{37}$$

where α is the Debye-Hückel constant $\alpha = 1.20078\left(\frac{\mathrm{kg}}{\mathrm{mol}}\right)^{}$, z is the specie charge, R is the universal gases constant 8.31 J/(mol K), and N_{H} is the average number of hydrogen atoms bond to the specie.

The metabolic models of cancer glycolysis used for the studies were previously reported [87] for HeLa tumor cell under conditions of normoxia (95% O_2), hypoxia (0.1–0.5% O_2), and AS-30D cells (rat hepatoma). Figure 7 shows the results of total entropy

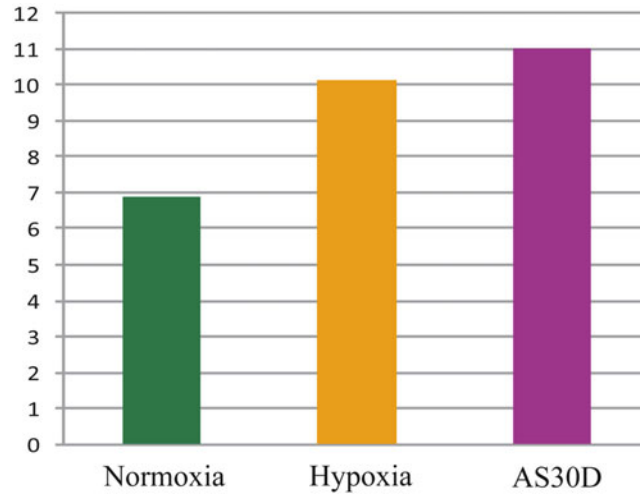


Fig. 7 Total entropy production rate [J/K min]10^{-3} for AS-30D and HeLa cell in normoxic and hypoxic conditions

production rate and glycolytic fluxes for HeLa cellular cell lines under conditions of normoxia and hypoxia and for AS-30D cells.

As can be noticed in Fig. 7, the entropy production per unit of time of the glycolysis pathway in HeLa cell line in hypoxic conditions is higher than in normoxia. This indicates, not only how is the glycolysis process favored under low oxygen conditions, but also that under these conditions it becomes more robust [106].

It is known that cancerous tumors under conditions of hypoxia become more resistant and more aggressive [107]. On the other hand, for the hepatoma AS-30D, the glycolytic flow (29 mmol/min) is identical to HeLa under conditions of hypoxia, while the $\dot{S}_i$ is slightly higher in hepatoma AS-30 (11.028 [J/K min]10^{-3}) compared to HeLa in hypoxic conditions (10.1 [J/K min]10^{-3}). The total value of the entropy production rate for HeLa cells is doubled under Hypoxia conditions (HeLa normoxia = 6866 [J/K min]10^{-3}) as shown in Fig. 7.

The increase in $\dot{S}_i$ under conditions of hypoxia may be due to overexpression of transcription regulatory mechanisms favoring the glycolytic pathway in low oxygen concentration such as HIF-1 (hypoxia-inducible transcription factor). Metabolism of hypoxia elicits multiple adaptive changes in cellular metabolism, which are coordinated by HIF-1 and serve to maintain redox homeostasis by increasing glycolysis and attenuating respiration. Decreased oxidative metabolism is necessary to limit production of reactive oxygen species (ROS) by the mitochondrial electron transport chain under hypoxic conditions [108].

HIF-1α stimulates glucose uptake necessary to compensate for the relative inefficiency of glycolysis, by upregulating glucose membrane transporters, GLUT1 and GLUT3. In addition, HIF-1α upregulate glycolytic enzymes such as hexokinases and phosphoglycerate kinase 1 and inhibit mitochondrial respiration by activating the transcription of pyruvate dehydrogenase kinase (PDK), which in turn phosphorylates and inactivates pyruvate dehydrogenase (PDH).

Also, HIF-1α upregulates lactate dehydrogenase-A to promote lactate production and regenerate NAD^+ [109]. HIF-1 maintains intracellular pH by increasing lactate and proton efflux through expression of monocarboxylate transporter (MCT)-4, carbonic anhydrase 9, and Na^+–H^+ exchanger-1 [108].

From the results for the HeLa cell line under conditions of normoxia and hypoxia and AS-30D cells, the fundamental reactions in the glycolysis process were identified (Fig. 8).

The fundamental postulate follows: those reactions that exhibit a higher value of $\dot{S}_i$ are considered fundamentals in the process [81]. This statement could be considered an extension of the "Principle of Maximum Entropy" [110]. The entropy production rate was normalized in percent using as a baseline the highest value.

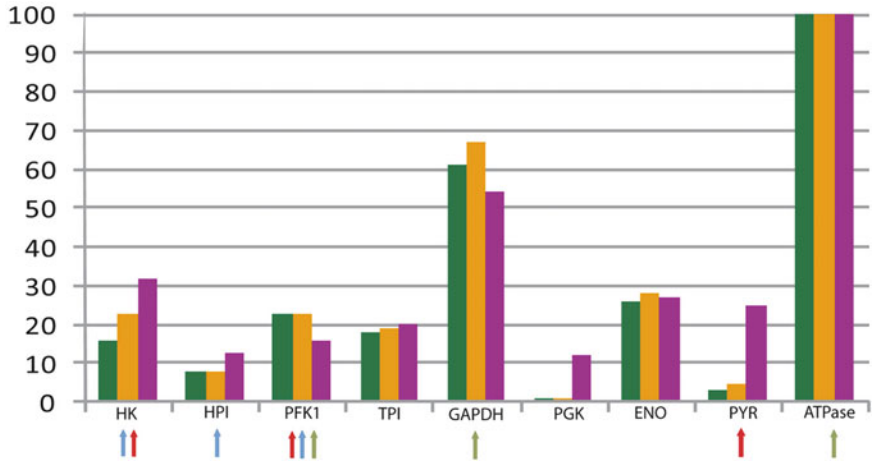


Fig. 8 Entropy production rate normalized values for fundamental reactions, in HeLa tumor cells in Normoxia (*green*) and Hypoxia (*orange*) conditions and AS-30D (*purple*) tumor cell. Reaction reported as fundamental by sensitivity analysis (*green arrow*), metabolic control analysis (*blue arrow*) and biochemistry control points (*red arrow*)

We identified 9 reactions of the 20 that are part of the glycolisis model (*see* Fig. 8), for both HeLa and AS-30D cells. Three with those found by sensitivity analysis (PFK1, GAPDH, ATPase), and also collected the so-called control points (HK, PFK1, PYR) [111]. The greater value of $\dot{S}_i$ was found for both cell lines in the reaction catalyzed by ATPase. This protein, ATPase, consists of a complex multi-subunit composed by a catalytic domain V1, responsible for the hydrolysis of ATP and a transmembrane domain V0 which is a proton channel [112].

The V-ATPase is overexpressed in many types of cancers and is related to the ability of these invasion and metastases [113]. The V-ATPase, a cell-specific proton pump, is expressed in the plasma membranes of human tumor cells and may have a specialized role in cell growth, differentiation, angiogenesis, and metastasis plays an important role in the control of intracellular and extracellular pH [114]. It is involved in the maintenance of a $pH_{intracelular}$ (pH_i) relatively neutral (slightly basic) and $pH_{extracelular}$ (pH_e) acid, which is known as the reverse gradient through pumping H^+ from the intracellular medium to vacuoles and the extracellular medium [115]. In cancerous tissues, the extracellular microenvironment can produce an increase in the secretion and activation of proteases and promote the degradation and remodeling of the extracellular matrix through the activation of proteolytic enzymes, which contributes to invasion and metastasis [114].

The V-ATPase is involved in the acquisition of a drug-resistant cellular phenotype and has therefore been identified as a potential target for cancer treatments [113].

As a second fundamental reaction, the reaction catalyzed by GAPDH was identified in both cell lines. This is vital for the glycolytic route because this is where the first high energy compound is formed. It is responsible for the conversion of glyceraldehyde-3-phosphate to 1,3-bisphosphoglycerate coupled with reduction of NAD^+ to NADH. It is catalyzed by the enzyme GAPDH which is considered unique and special because it has the ability to bind to NAD^+ or NADH and sometimes to DNA and RNA [116].

In addition to the role it plays in energy metabolism, GAPDH has other functions independent of glycolysis; these include cytoskeletal regulation, endocytosis, membrane fusion, tRNA transport, and DNA replication and repair [117].

Its overexpression in several types of tumors, such as breast cancer, as well as the gene that produces it, is associated with the activity of the HIF-1α transcription factor, as recently reported by [118].

Data suggest that GAPDH may play an important role in promoting metastasis, at least in colon cancer. At the molecular level, GAPDH seems to physically interact with Sp1, a key transcriptional factor known to bind to the promoter of SNAIL and enhance its expression. It seems possible that GAPDH forms a protein complex with Sp1 and enhances to expression of SNAIL, which is a transcriptional inducer of epithelial-mesenchymal transition [119].

The enolase catalyzed reaction (ENO), also identified as fundamental, shows similar proportions in the $\dot{S}_i$ for both conditions and in the two cell lines. However, their absolute values differ considerably (HeLa: normoxia: 0.777 [J/K min]10^{-3}, hypoxia: 1.155 [J/K min]10^{-3}. This is an enzyme that is highly conserved and its three isoforms (α, β, and γ) show very few kinetic differences [120].

According to one analysis [121] the isoform α, present primarily in fetal cells, is overexpressed in human liver carcinomas and the gene encoding it is one of the targets of HIF-1. The importance of this enzyme for the development of cancer is that it is capable of acting as a plasminogen receptor 6, favoring the growth and dissemination of tumor cells [87].

The reaction catalyzed by Hexokinase (HK, reaction # 2) is identified as one of the fundamental in glycolytic pathways by the calculation of the entropy production rate and is one of the biochemical checkpoints. In this reaction, glucose is activated for the following reactions by phosphorylation producing glucose 6 phosphate, with ATP as a phosphate donor [111].

HK is a limiting enzyme (control point) of the flow along the route. There are four isoforms present in mammals, HK2 expression is markedly induced in cancer cells by multiple mechanisms

and oncogenic drivers and is transcriptionally upregulated by MYC [122].

The HKII isoform was reported by Mathupala et al. [123] like "cancer's double-edged sword." On the one hand, it facilitates and safeguards the malignancy of the tumors when it is attached to the mitochondria [123, 122]. On the other hand, the activity of the mitochondrial HKII shows that it is required by growth-inducing factors of cell survival, and by protein kinase B (AKT) [123]. In addition the binding of the HKII to the mitochondrial membrane directs the inhibition of apoptosis, the mechanism by which this occurs is still unknown. HK is not only important for maintaining high glycolysis rates but is crucial for tumor survival [120].

In normal differentiated adult cells, intracellular pH is generally ~7.2 and lower than the extracellular pH of ~7.4. Deregulated pH is emerging as a hallmark of cancer [16] because cancers show a "reversed" pH gradient with a constitutively increased intracellular pH_i that is higher than the extracellular pH_e. The increase in pH_i in cancer cells seems paradoxical considering the high proliferation of these and the high glycolytic flows that generate acid metabolites. However, changes in the expression or activity of pumps and transporters in the plasma membrane facilitate efflux of H^+ and maintain a high pH_i and a low pH_e [124].

The rate of entropy production for AS30D and HeLa tumor cells under conditions of normoxia and hypoxia was evaluated for a range of pH values from 6.2 to 7.4. Figure 9 shows a marked correlation between the rate of entropy production of reaction 14 (ATPase), previously reported as the one exhibiting the highest $\dot{S}_i$ and pH_e.

As can be seen (Fig. 9), when intracellular pH_i gets lower, also does the entropy production rate, which measures the loss of the process robustness with lower intracellular pH_i.

Cytoplasmic pH has been shown to play a crucial role in multiple cell functions including the control of cell growth and proliferation and programmed cell death. Intracellular acidification has been reported to be a sign of apoptosis in a variety of cancer cells [125]. Some recent reports have shown that the intracellular pH plays a key role in determining the way cancer cells obtain energy. Thus, an alkaline pH_i drives aerobic glycolysis and an acidic pH_i drives oxidative phosphorylation [125].

The pH_i could rapidly decrease and this would be lethal for the cancer cell if it were not compensated for by the increase in the export of protons, resulting in an acidification of the extracellular medium causing the inverse gradient of pH.

A characteristic feature of cancer cells and especially highly aggressive cells is overexpression and increased activity of multiple transporters and regulatory enzymes of pH such as carbonic anhydrase which acidify the extracellular space by importing

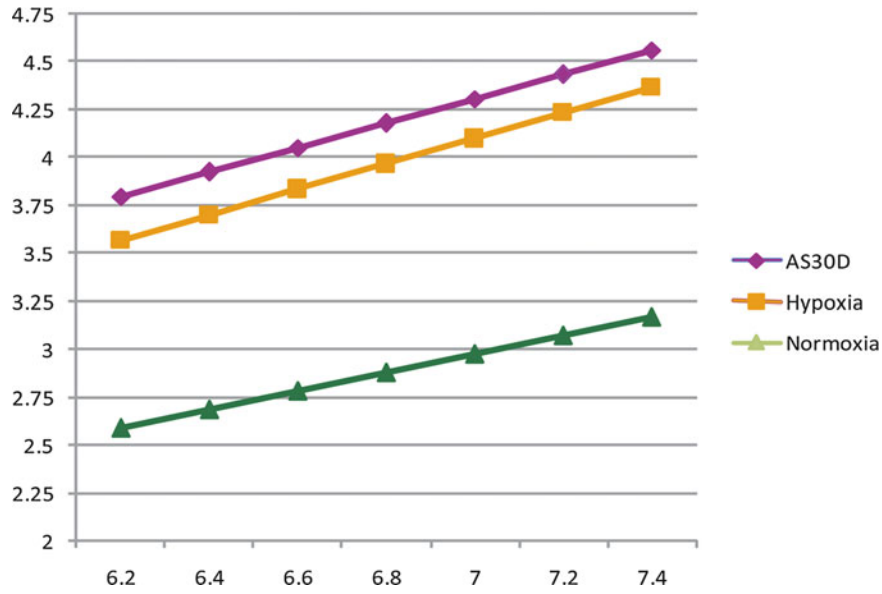


Fig. 9 Entropy production rate [J/K min]10^{-3} for reaction ATPase as a function of intracellular pH_i, for glycolysis network model AS-30D and HeLa cells

bicarbonate, the exchanger NA^+/H^+ (NHE), Which is a prominent mediator of the export of protons, as is the H^+-ATPasa vacuolar (V-ATPasa), Which hydrolyzes ATP and pumps protons [126].

A cancer-specific intracellular alkalosis represents a common final pathway in cell transformation and an antiapoptotic defensive mechanism in cancer cells. In terms of mechanisms, it is known that decreasing intracellular pH induces a cell death program, either via apoptosis or necrosis, while the elevation of cellular pH by different methods protects cancer cells by preventing them from entering the apoptotic cascade [127].

Environmental acidification, mediated by overexpression of either NHE and/or other PTs, is known to play a role in hindering DNA repair, increasing mutagenesis, and driving genomic instability in cancer [127].

Cytosolic acidification in cancer cells during apoptosis was accompanied by the activation of OXPHOS. Some recent reports have shown that the intracellular pH plays a key role in determining the way cancer cells obtain energy. Thus, an alkaline pH_i drives aerobic glycolysis and an acidic pH_i drives oxidative phosphorylation [125].

As an alternative or common strategy with the correction of pH_e acid in tumors, a large number of investigations have explored the inhibition of membrane ion pumps that maintain the pH alkaline in the intracellular medium, by exporting protons or importing bicarbonate [128].

Inhibition of proton pumps decreases the value of pH_i and this tends to increase the value of pH_e. The effect in pH_e is sustained

presumably because the acidification of pH_i suppresses the efficiency of glycolysis. Intracellular acidification associated with inhibition of proton pumps can cause a high impact on cancer behavior regardless of whether it is associated with decrease of the pH_e [129].

In fact, inhibition of proton pumps exerts an antiproliferative and pro-apoptotic effect on some cancer cell lines. On the other hand, intracellular acidification has been shown to be related to the efficacy of death by hyperthermia (42 °C). Recently, new variants of tumor therapies that use hyperthermia have been studied and it is observed that the low value of the pH_i is strongly related to the thermosensitivity [130].

It is an important possibility as future therapy to combine the use of proton transport inhibitors along with hyperthermia. Also, the pH_i is linked to the apoptotic response to tumor necrosis factor-related ligand-inducing apoptosis [131].

A key factor in drug resistance is the inverse gradient of pH. The pH_i alkaline confers cancer cells resistance to drug cytotoxicity and to the external acidic environment. A large number of studies [132] have shown that resistance to cisplatin and doxorubicin (anticancer drugs) is associated with elevation of pH_i in multiple tumor cell lines (epidermal cancer, prostate cancer, ovarian cancer, melanomas, lung cancer, and breast cancer [113]).

In this way, we see how the reaction catalyzed by ATPase represents a potential target of the glycolysis process in the treatment of cancer, not only because it is the most important in the regulation of the glycolytic route, but also because it exhibits a greater dependence with the values the pH_i takes.

4.2 How Much Damage Can the Glucose Make in Cancer?

Due to a combination of high glucose consumption rates by tumor cells and reduced tumor vascularization, the glucose concentration in the tumors can be 3–10 times lower than in normal tissues, according to the stage of its development. Therefore, tumor cells must develop strategies for their growth and survival in metabolically unfavorable environments [133].

Since the last years, cancer glycolysis has been a target in oncology research [134]. The significant increase in glycolysis rate observed in tumors has been recently verified, yet only a few oncologists or cancer researchers understand the full scope of Warburg's work [84, 134] despite, as we have said above, its great importance. Altered energy metabolism is proving to be as widespread in cancer cells as many of the other cancer-associated traits that have been accepted as hallmarks of cancer [16]. The regulation of metabolism, relevant to senescence process, would be a key to improve and identify new anti-cancer therapies in the future.

The complex systems theory and the thermodynamics formalism in the last years have shown to be a theoretical framework as

well as a useful tool to understand and forecast the evolution of the tumor growth [71, 98, 135–139].

The model used was proposed by Marin et al. [140] for the glycolytic network of HeLa tumor cell-lines growth under three metabolic states: Hypoglycemia (2.5 mM), Normoglycemia (5 mM), and Hyperglycemya (25 mM) during enough time to induce phenotypic change in cellular metabolisms. However, the growth saturation was not attained in this phase. In the other stage the cells were exposed to different glucose concentrations: 2.5, 5 y 25 mM until they reach the stationary state. The rate of entropy production was calculated using the glycolysis network model of HeLa cell lines at steady state.

The highest values of entropy production rate were observed in the hypoglycemic phenotype, which means this phenotype exhibits higher robustness [82, 141]. This can be correlated to the metabolic change induced in the HeLa cells lines grown in hypoglycemic conditions and its independence of the extracellular glucose conditions in the second face (2.5, 5 y 25 mM) until they reach the stationary state (*see* Fig. 10a).

The sustained decrease in the glucose availability can stimulate changes in the cellular phenotype. For example, KRAS mutations can increase the GLUT1 expression and that of many genes that codify the enzymes of the fundamental steps of glycolysis, like HK1, HK2, PFK-1, and LDH-A, [142]. These changes imply an increase in glycolytic flow and consequently an increase in entropy production rate (*see* Eq. (4)). Even if the extracellular glucose

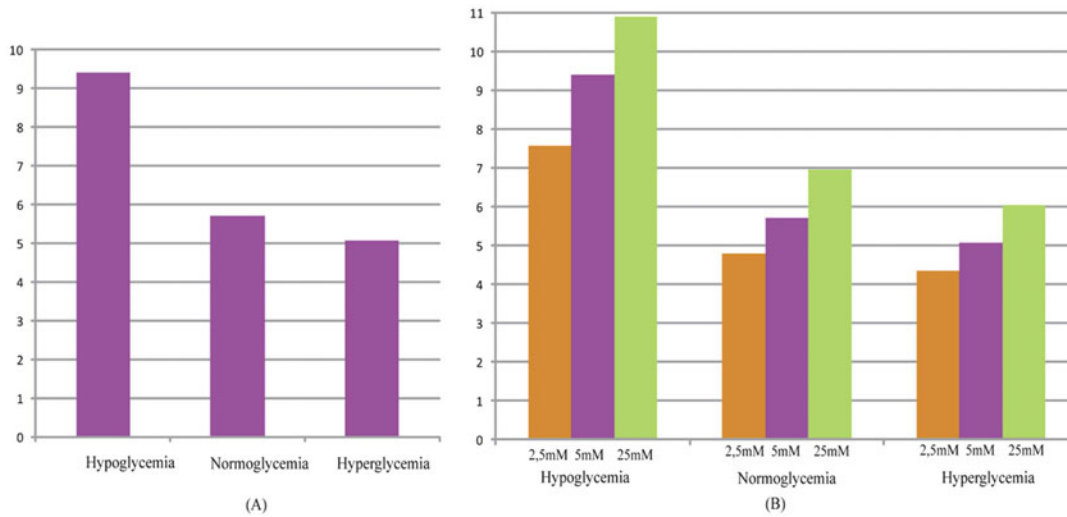


Fig. 10 Total entropy production rate [J/mM K min]10^{-3}. (**a**) For HeLa cells in different metabolic phenotypes; (**b**) For HeLa cells exposed to different glucose concentrations until they reach the stationary state in each phenotype

concentration returns to normal values, the changes can be maintained [143].

Moreover, the entropy production rate increases when exposed to higher extracellular glucose concentration in the three phenotypes (*see* Fig. 10b).

Taking the glucose concentration like a control parameter and replacing in Eq. (18), we have

$$\frac{d\dot{S}_i}{dt} = \frac{\partial \dot{S}_i}{\partial \mathrm{Glc}} \frac{d\mathrm{Glc}}{dt} < 0 \tag{38}$$

The glucose concentration decreases because it is a reactant, thus we have $d\mathrm{Glc}/dt < 0$; therefore, we must have $\partial \dot{S}_i / \partial \mathrm{Glc} > 0$ (*see* Fig. 10b). It is noted that the production of entropy per unit of time $\dot{S}_i$, evaluated through Eq. (4), is indicative of the directional character and stability of the dynamical behavior of cancer glycolysis [141].

One of the strategies used to fight the cancer has been the abrupt decrease of glucose concentration in the tumor microenvironment [134]. Cancer cells that develop accelerated glycolysis due to activation of oncogenes (including Ras, Her-2, and Akt) or due to loss of tumor suppressor function (including TCS1/2, p53, LKB1) undergo rapid apoptosis when placed in culture conditions with low glucose concentrations [134]. That is observed in Fig. 10b for the three phenotypes.

Therefore, glucose deprivation must be carried out in a shorter time than required by the tumor cells to acquire a characteristic phenotype. In this case, the hypoglycemic phenotype which, as shown in Fig. 10a, exhibits a higher entropy production rate, and consequently will have a greater robustness.

It is known that the glucose deprivation markedly enhances oxidative stress by increasing the intracellular level of ROS [122]. ROS acts as a signal transduction messenger and can promote the proliferation or cellular death of cancer cells, depending on the intra and extracellular condition of the antioxidant defense mechanisms.

Cancer cells subjected to persistent endogenous and exogenous oxidative stress were shown to develop adaptive responses, mainly related to the upregulation and activation of the antioxidant machine, that can contribute to cancer progression through an array of interconnected signals, among them, activation of RAS oncogen [144].

The bigger robustness of the hypoglycemic phenotype may be related to the increase of levels of ROS as consequence of low extracellular glucose concentration, and therefore is related to the contribution of ROS to the development and cell proliferation. The cells grown in hypoglycemic conditions could be adapted to a ROS-resistant phenotype, and this could be maintained even if the cells were later submitted to a high glucose concentration.

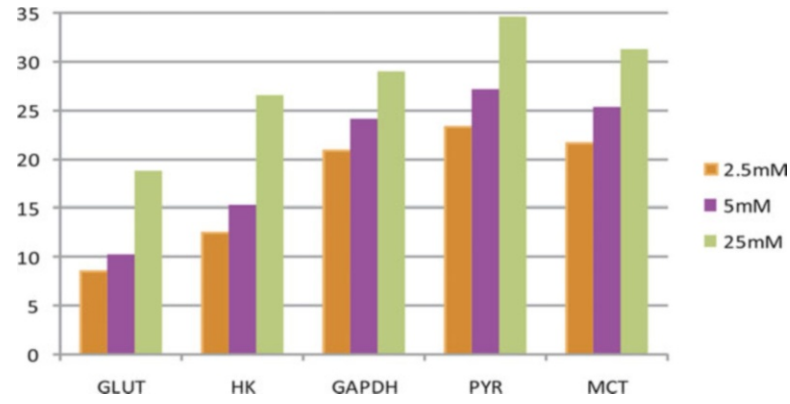


Fig. 11 Entropy production rate normalized values for reactions that show a high variation in the range of 2.5–25 mM of glucose

In Fig. 11 it is observed that the reactions that show a high $\dot{S}_i$ variation in the range of 2.5–25 mM of glucose were GLUT, MCT1, HK, GADPH y PYK. This behavior is related to the capacity of these proteins to change their flow in this range, in other words, an increase in their reaction rate and in consequence an increase in the entropy production rate (*see* Eq. (8)).

The reactions GLUT and HK exhibit high values of $\dot{S}_i$ for the glucose concentrations of 25 mM in the three phenotypes (*see* Fig. 11). This behavior was corroborated by the results of the studio of the sensibility analysis. Sensitivity analysis [140] quantitatively investigates the behavior of a system as a response to changes in parameter values.

The transporter GLUT 1 performs a fundamental roll in many steps of cancer progression. It has been demonstrated that GLUT1 may regulate proteins that play a role in early tumor growth as well as in cancer invasiveness and metastasis [145].

Several studies demonstrate that hexokinase, particularly the Type II isoform (HK II), plays a critical role in initiating and maintaining the high glucose catabolic rates of rapidly growing tumors. Thus, it appears that hexokinase and its association with mitochondrial protein complex may play important roles in the essential homeostatic processes such as glucose metabolism and apoptosis. The inhibition of HK has significant effects in the metabolism and cell survival [116].

4.3 The Return of Oscillations: Cancer Glycolysis Model

At macroscopic scales, the self-organization and the complexity exhibited by dynamic systems are manifested through oscillations sustained in time and/or space. In biological systems these oscillations are usual [146], and they not only guarantee robustness, but also allow the system to perform several functions, including control and regulation.

Oscillations in yeast metabolism are well documented in the literature [146]. However, to the best of our knowledge, there are

just two reports related to cancer glycolytic oscillations [147, 148]. Despite the low existing information about this aspect, we can assume as a hypothesis that cancer glycolysis is a self-organized process far from thermodynamics equilibrium. In other words, sustained oscillations in cancer glycolysis grant high robustness and complexity. Consequently, a strategy aimed at exploiting abnormalities in cancer glycolytic therapies would be focused on recognizing those regions in which control parameters lead to a loss in self-organization.

The first developed glycolysis models were made taking yeast as experimental reference. Glycolysis in yeast exhibits periodic and aperiodic oscillations [149]. In 1964, Higgins [150] proposed a general oscillatory mechanism for yeast glycolytic intermediaries. Higgins' model has six reactions, with PFK in the center showing this enzyme great oscillatory potential in yeast.

Sel'kov in 1968 [151] refers again to PFK oscillations, by using a simple model of a monosubstrate enzymatic reaction with substrate inhibition and product activation. In 1981, Termonia and Ross proposed a glycolysis model for yeast, slightly more extensive, which beholds the coupling between PFK and PYK and takes into consideration the inhibition of the latest [152].

In 1982, Decroly and Goldbeter [153] established a model describing the coupling between two allosteric enzymes E_1 and E_2 activated by their own products. This model embraces a series of different cases of stationary state stability, including complex behaviors, periodic oscillations, and chaotic behavior. However, that model is a generic one, intended for no specific biological system. Although it refers to PFK and PYK as an example of enzymes, that model is unable in portraying cancer as it does not consider PYK's inhibition by ATP.

In the 1990s, several models aroused, all of them referring to oscillations in yeast cells [154–156], where the glycolytic mechanism is in deep investigated alongside with its oscillatory behavior.

As noted before, in cancer, although there is proof of the existence of oscillations [63, 148], to this moment, there is no such model that, at least qualitatively, reproduces oscillatory behavior in glycolysis, which would help in understanding the observed robustness and complexity.

Based on the evidence discussed above, we propose a model based on a simple biochemical network to describe the dynamics of glycolytic oscillations in HeLa cancer cell lines (*see* Fig. 12).

The model is based on six reactions that are named after their enzymes: HK, PFK, GAPDH, PYK, ATPase, and LDH. These reactions have been selected by a sensitivity analysis and entropy production rate method [106, 157], according to the glycolytic mechanism proposed by Marín et al. [140] for HeLa tumor cells. The reactions identified as HK, PFK, and PYK are known as control points due to their condition of allosteric enzymes, which regulate

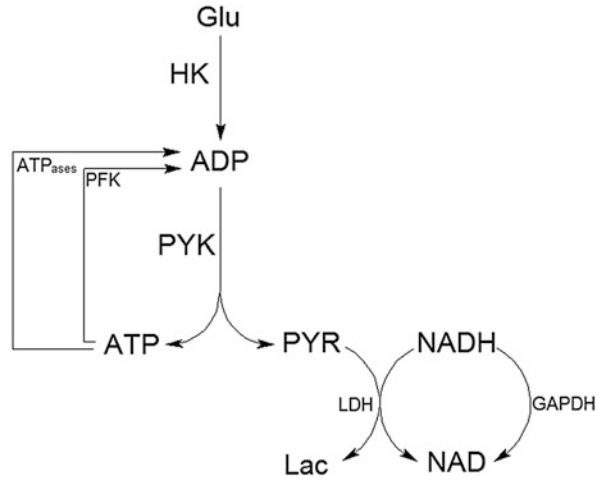


Fig. 12 Biochemical network for a mechanism for glycolysis for HeLa cell lines

the mechanism rate [158]. These processes have also been identified as main targets for cancer treatment [134, 159].

We developed our model (*see* Fig. 12) as a representation of the dynamics of the glycolysis for HeLa cell lines on the basis of the above-discussed experimental evidence. From the biochemical network model (Fig. 12), using the classic chemical kinetics method that applies the law of mass action, the following system of ordinary differential equations (ODEs) was obtained:

$$\begin{aligned}
\frac{d[\mathrm{ATP}]}{dt} &= -V_{\mathrm{PFK}} + 4V_{\mathrm{PYK}} - V_{\mathrm{ATPasess}} \\
\frac{d[\mathrm{ADP}]}{dt} &= -V_{\mathrm{HK}} + V_{\mathrm{PFK}} - V_{\mathrm{PYK}} + V_{\mathrm{ATPases}} \\
\frac{d[\mathrm{NAD}]}{dt} &= -V_{\mathrm{GAPDH}} + V_{\mathrm{LDH}} \\
\frac{d[\mathrm{NADH}]}{dt} &= V_{\mathrm{GAPDH}} - V_{\mathrm{LDH}} \\
\frac{d[\mathrm{Pyr}]}{dt} &= V_{\mathrm{PYK}} - V_{\mathrm{LDH}}
\end{aligned} \tag{39}$$

Where V_i are the reaction rate and the numerical values of the constants are obtained [160]. Fixed points, stability analysis, and bifurcation were performed using the standard procedure [161] and using as control parameters the concentration of Glucose (Glu) and Inorganic phosphate (Pi).The LZ complexity [54] was calculated using the proposed algorithm by Lempel and Ziv. Figures 13 and 14 show how the LZ complexity varies for different Glucose (Glu) and Inorganic phosphate (Pi) concentrations.

For modeling chemical network model (Fig. 12), COPASI v. 4.6.32 software was used.

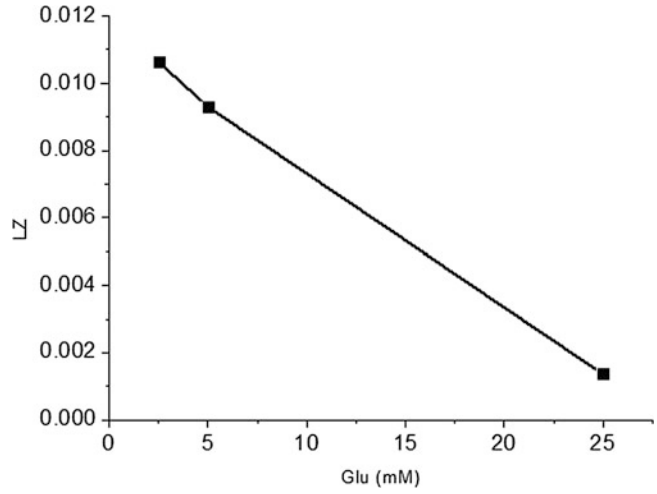


Fig. 13 Variation of LZ complexity with glucose concentration (2.5; 5 and 25 mM)

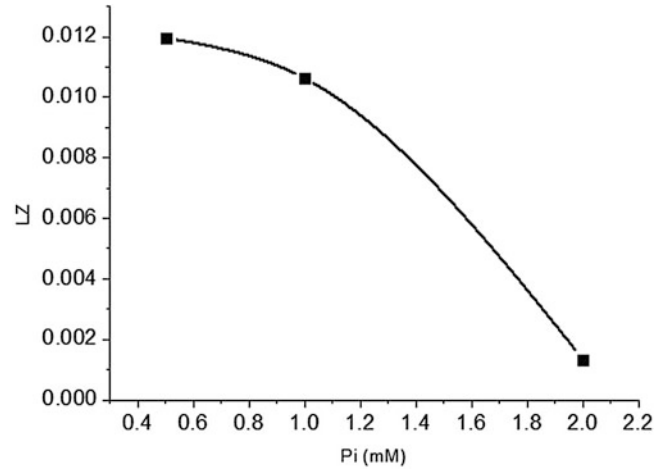


Fig. 14 Variation of LZ complexity with inorganic phosphate concentration (0.5; 1 and 2 mM)

When we perform the steady-state stability analysis, we find that the system evolves from an asymptotically stable state to a limit cycle in the form of periodic oscillations when we decrease the glucose concentration from 25 mM down. A similar behavior is observed at decreasing Pi. When we reach low values of both Glu and Pi, the system exhibits a series of unstable states that lacks physical sense.

We can see in Fig. 15 the relation between ATP and NADH at 5 mM of Glu and 0.5 mM of Pi.

An interesting fact is that the oscillations period observed (2–4 days) is close to the values experimentally found by Potter [148] in rat hepatoma (six diploids and another three rat

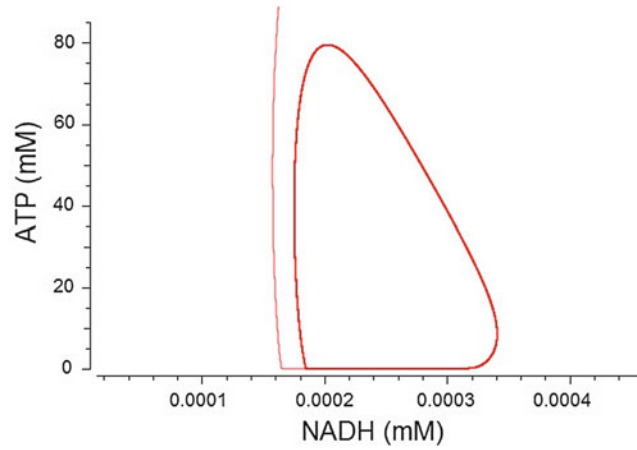


Fig. 15 Limit Cycle observed for Glucose 5 mM and Inorganic Phosphate 0.5 mM

hepatomas). The oscillation period is around 12 h and 2 days, depending on feeding conditions. The oscillations' period shown in the model is of the same order, despite that a different kind of cancer cell line has been used.

It can be easily recognized that higher complexity states (and consequently higher robustness states) are found at low Pi and Glu. When Pi concentration is increased, the system reaches less complex states. The zone in which periodic oscillations can be found grows thinner with increasing Glu, while the opposite happens for stable steady states. This finding reaffirms the hypothesis that the complexity and robustness of the system decreases when glucose concentration increases. In our model LZ complexity is thought to be a measure of the different patterns arising in a process, and therefore its increase correlates with the number of configurations the system acquires. This capability is usually deemed to characterize self-organizing systems and it is associated with increased robustness, i.e., resilience with respect to stressing events.

Contrary to what has been observed for LZ complexity, we obtain an increase in entropy production rate associated with increasing glucose concentration, which coincides with results obtained from experimental data of HeLa for different growth conditions [141]. Similar behavior is observed for Pi. Calculations of entropy rate production for different values of the control parameters show that $\frac{\partial \dot{S}_i}{\partial Pi>0}$ and $\frac{\partial \dot{S}_i}{\partial \mathrm{Glu}>0}$.

We believe that such conundrum lies its roots in the divergent grasping of the entropy production principle [162]. As we have shown in previous works [14, 81, 82] that if the entropy production rate is not used as an extremal principle [163], we should postulate that those reactions that exhibit a higher value of $\dot{S}_i$ necessarily are fundamentals in the process [81]. This statement

can be considered an extension of the "Principle of Maximum Entropy" [110].

On the other hand, we find that [26] the Prigogine Principle of "minimal entropy production" [12], understood as an extremal or variational principle, can be generalized for nonlinear systems far from thermodynamics equilibrium through the Lyapunov function [82, 81]. In this sense, the apparent contradiction abovementioned is explained, as in this context the entropy production rate is used as extremal principle.

In our case, we use as control parameter, $\Omega = f(\mathrm{Pi}, \mathrm{Glu})$, so $\dot{S}_i = f(\mathrm{Pi}, \mathrm{Glu})$. And we can rewrite Eq. (18) as

$$\frac{d\dot{S}_i}{dt} = \frac{\partial \dot{S}_i}{\partial \mathrm{Pi}} \frac{d\mathrm{Pi}}{dt} + \frac{\partial \dot{S}_i}{\partial \mathrm{Glu}} \frac{d\mathrm{Glu}}{dt} < 0 \tag{40}$$

and considering that in the glycolysis process:

$$\mathrm{Glu} + 2\mathrm{NAD} + 2\mathrm{ADP} + 2\mathrm{Pi} = 2\mathrm{Pyr} + 2\mathrm{NADH} + 2\mathrm{H}^+ + 2\mathrm{ATP} + 2\mathrm{H_2O}$$

It can be observed that both Glu and Pi act as reactants $\frac{\partial \mathrm{Pi}}{\partial t} < 0$ and $\frac{\partial \mathrm{Glu}}{\partial t} < 0$, so $\frac{\partial \dot{S}_i}{\partial \Omega \leq 0}$. Then it fulfills that (*see* Eq. (40)): $\frac{d\dot{S}_i}{dt<0}$.

Those results allow us to posit that the rate of entropy production is a Lyapunov function, i.e., shows its directional character, behaving as "first-order" phase transition, through a supercritical bifurcation of Andronov-Hopf. Conclusively, the process of cancer glycolysis exhibits sustained oscillations leading to self-organization far from thermodynamics equilibrium.

Therefore, a viable therapeutic approach exploiting abnormalities in cancer glycolytic metabolism would be finding those regions of control parameters in which self-organization is lost, leading thus to less complex steady states. That strategy would likely represent a key to improve and identify new anti-cancer therapies in the future.

5 Concluding Remarks

The integration of thermodynamics formalism of irreversible process, complex systems theory, and systems biology [164] offers a holistic view of cancer as a self-organized nonlinear dynamical system far from thermodynamic equilibrium. In this sense, we show how its evolution of tumor occurs as an emergent phenomenon, as a phase transition, could be called "biologic phase transition." In summary, it was found that:

1. A mechanism for avascular, vascular, and metastasis tumor growth based on a chemical network model. Vascular growth and metastasis appear as a hard phase transition type, as "first-order," through a supercritical Andronov–Hopf bifurcation,

emergence of limit cycle, and then through a cascade of bifurcations type saddle-foci Shilnikov's bifurcation.

2. The entropy production rate may be used as a quantitative index of the metastatic potential of tumors.
3. A relationship between the production of entropy per unit time, the fractal dimension, and the tumor growth rate for human tumors cells has been derived.
4. In cancer glycolysis under hypoxia conditions, the entropy production rate is higher than the entropy production rate of normoxia that means more complexity and robustness. This conduces to the thesis that the employ of any type of therapy should be in normoxia conditions.
5. The total entropy production rate that is shown by cancer glycolysis in the hypoglycemic phenotype is greater than those of the other states. In fact, this metabolic condition exhibits more robustness.
6. A kinetic model is proposed using experimental data for HeLa tumor cells grown in hypoxia conditions which confirms the existence of glycolytic oscillations in cancer. This allows it to self-organize in time and space far from thermodynamic equilibrium, and provides it with high robustness, complexity, and adaptability. Glycolytic oscillations in cancer emerge through an Andronov-Hopf bifurcation as a "first order" phase transition, and could be called "biologic phase transition."

We hope to provide a better understanding of the dynamics of the growth of cancerous tumors, resulting in better and more effective therapies.

Acknowledgments

Prof. Dr. A. Alzola *in memoriam*. We would like to thank Prof. M. Bizzarri for inviting us to write the Chapter. We would like to thank E. Silva for aiding us with illustrations; we would also like to thank the rest of our colleagues M. D. Mesa, D. J. Rodriguez, I. Durán, J. C. Jaime, and J. P. Pomuceno.

References

1. Bertalanffy L (1972) The history and status of general systems theory. Acad Manag J 15 (4):407–426
2. Schmitz U, Wolkenhauer O (2016) Systems medicine. Springer, New York
3. Bizzarri M, Palombo A, Cucina A (2013) Theoretical aspects of systems biology. Prog Biophys Mol Biol 112:33–43
4. Du W, Elemento O (2014) Cancer systems biology embracing complexity to develop better anticancer therapeutic strategies. Oncogene 34:3215–3225
5. Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664

6. Bertalanffy L (1950) The theory of open systems in physics and biology. Science 111:23–29
7. Knox SS (2010) From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. Cancer Cell Int 10(11):1–13
8. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J (2006) Cancer: a systems biology disease. Syst Biol 83:81–90
9. Kitano H (2013) Cancer systems biology: a robustness-based approach. In: Walhout M, Vidal M, Dekker J (eds) Handbook of systems biology. Academic, New York
10. Auffray C, Nottale L (2008) Scale relativity theory and integrative systems biology: 1. Founding principles and scale laws. Prog Biophys Mol Biol 97:79–114
11. European Science Foundation (2008) Advancing systems biology for medical applications. Science Policy Briefing 35
12. Prigogine I (1961) Introduction to thermodynamics of irreversible processes. Wiley, New York
13. De Donder T, Van Rysselberghe P (1936) Thermodynamics theory of affinity. Oxford University Press, London
14. Nieto-Villar JM, Izquierdo-Kulich E, Betancourt-Mar JA, Tejera E (2013) Complejidad y auto-organización de patrones naturales. Editorial UH, La Habana, Cuba
15. Zotin AI (1988) Thermodynamic principles and reaction of organisms. Nauka, Moscow
16. Hanahan D, Weinberg R (2011) Hallmarks of cancer: the next generation. Cell 144 (5):646–674
17. Nicolis G, Prigogine I (1977) Self organization in nonequilibrium systems. Wiley, New York
18. Nicolis G, Daems D (1998) Probabilistic and thermodynamic aspects of dynamical systems. Chaos 8(2):311–320
19. Landau LD, Lifshitz EM (1964) Curso de Física Teórica, Física Estadística, vol 5. Reverté, México
20. Landau LD (2008) On the theory of phase transitions. Ukr J Phys 53:25–35
21. Izquierdo-Kulich E, Rebelo I, Tejera E, Nieto-Villar J (2013) Phase transition in tumor growth: I Avascular development. Physica A 392(24):6616–6623
22. Betancourt-Mar JA, Llanos-Pérez JA, Cocho G, Mansilla R, Martin RR, Montero S, Nieto-Villar JM (2017) Phase transitions in tumor growth: IV Relationship between metabolic rate and fractal dimension of human tumor cells. Physica A 473:344–351
23. Ivancevic VG, Ivancevic TT (2007) High-dimensional chaotic and attractor systems, a comprehensive introduction, vol 32. Springer, New York
24. Andronov AA, Khaikin SE (1949) Theory of oscillations. Princeton University Press, Princeton
25. Andronov AA, Vit A, Chaitin C (1966) Theory of oscillators. Pergamon Press, Oxford
26. Nieto-Villar JM, Quintana R, Rieumont J (2003) Entropy production rate as a Lyapunov function in chemical systems: proof. Phys Scr 68(3):163–165
27. Nicolis G, Nicolis C (2012) Foundations of complex systems: emergence, information and prediction. World Scientific, River Edge, NJ
28. Izquierdo-Kulich E, Nieto-Villar JM (2013) Morphogenesis and complexity of the tumor patterns. In: Rubio RG (ed) Without bounds: a scientific canvas of nonlinearity and complex dynamics. Understanding complex systems. Springer, Berlin
29. Dinicola S (2011) A systems biology approach to cancer: fractals, attractors, and nonlinear dynamics. OMICS 15(3):93–104
30. Kitano H (2007) Towards a theory of biological robustness. Mol Syst 3:137
31. Rockmore D (2005) Cancer complex nature. Santa Fe Inst Bull 20:18–21
32. Roose T, Chapman SJ, Maini PK (2007) Mathematical models of avascular tumor growth. SIAM Rev 49(2):179–208
33. Enderling H, Almog N, Hlatky L (2012) Systems biology of tumor dormancy, vol 734. Springer Science & Business Media, New York
34. D'Onofrio A (2013) Multifaceted kinetics of immuno-evasion from tumor dormancy. In: Systems biology of tumor dormancy. Springer, New York, pp 111–143
35. Izquierdo-Kulich E, Nieto-Villar JM (2008) Morphogenesis of the tumor patterns. Math Biosci Eng 5(2):299–313
36. Kuznetsov VA, Knott GD (2001) Modeling tumor regrowth and immunotherapy. Math Comput Modell 33(12):1275–1287
37. Page K, Uhr J (2005) Mathematical models of cancer dormancy. Leuk Lymphoma 46 (3):313–327
38. Brú A (2003) The universal dynamics of tumor growth. Biophys J 85(5):2948–2961
39. Nasir NA, Kaiser HE (2008) Selected aspects of cancer progression: metastasis, apoptosis

and immune response, vol 11. Springer Science & Business Media, New York
40. Pantel K, Alix-Panabiéres C, Riethdorf S (2009) Cancer micrometastasis. Nat Rev Clin Oncol 6(6):339–351
41. Hielscher A, Wirtz D (2013) A physical sciences network characterization of non-tumorigenic and metastatic cells. Sci Rep 3:1449
42. Bubendorf L (2000) Metastatic patterns of prostate cancer: an autopsy study of 1,589 patients. Hum Pathol 31(5):578–583
43. Prigogine I, Lefever R (1980) Stability problems in cancer growth and nucleation. Comp Biochem Physiol 67B:389–393
44. Delsanto PP, Romano A, Scalerandi M, Pescarmona GP (2000) Analysis of a "phase transition" from tumor growth to latency. Phys Rev E 62:2547–2549
45. Solé RV (2003) Phase transitions in unstable cancer cell populations. Eur Phys J 35:117–124
46. Davies PC, Demetrius L, Tuszynski JA (2011) Cancer as a dynamical phase transition. Theor Biol Med Model 8:30
47. Strogatz SH (2000) Nonlinear dynamics and chaos. Westview, Cambridge
48. Kuznetsov VA, Makalkin IA, Taylor MA, Perelson AS (1994) Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis. Bull Math Biol 56(2):295–321
49. Llanos-Pérez J, Betancourt-Mar A, De Miguel M, Izquierdo-Kulich E, Royuela-García M, Tejera E, Nieto-Villar J Phase transitions in tumor growth: II Prostate cancer cell lines. Physica A 426:88–92
50. Izquierdo-Kulich E, Nieto-Villar JM (2007) Mesoscopic model for tumor growth. Math Biosci Eng 4(4):687–698
51. Anishchenko VS, Vadivasova TE, Okrokvertskhov GA, Strelkova GI (2003) Correlation analysis of dynamical chaos. Physica A 325 (1):199–212
52. Kuznetsov YA (2013) Elements of applied bifurcation theory, vol 112. Springer Science & Business Media, Dusseldorf, Germany
53. Wolf A, Swift JB, Swinney HL, Vastano JA (1985) Determining Lyapunov exponents from a time series. Physica D 16(3):285–317
54. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. IEEE Trans Inform Theory 23(3):337–343
55. Ziv J, Lempel A (1978) Compression of individual sequences via variable-rate coding. IEEE Trans Inform Theory 24(5):530–536
56. Frederickson P, Kaplan JL, Yorke ED, Yorke JA (1983) The Lyapunov dimension of strange attractors. J Differ Equ 49 (2):185–207
57. Gear CW (1968) The automatic integration of stiff ordinary differential equations. In: Proceedings IFIP68. North-Holland, Amsterdam, pp 187–193
58. Hegger R, Kantz H, Schreiber T (1999) Practical implementation of nonlinear time series methods: the TISEAN package. Chaos 9 (2):413–435
59. Shilnikov AL, Turaev DV, Chua LO (2001) Methods of qualitative theory in nonlinear dynamics, vol 5. World Scientific, Singapore
60. Itik M, Banks SP (2010) Chaos in a three-dimensional cancer model. Int J Bifurcat Chaos 20(01):71–79
61. El-Gohary A (2008) Chaos and optimal control of cancer self-remission and tumor system steady states. Chaos Solitons Fractals 37 (5):1305–1316
62. Mackey MC, Glass L (1977) Oscillation and chaos in physiological control systems. Science 197(4300):287–289
63. Posadas E, Criley S, Coffey D (1996) Chaotic oscillations in cultured cells: rat prostate cancer. Cancer Res 56(16):3682–3688
64. Wolfrom C, Chau NP, Maigné J, Lambert JC, Ducot B, Guerroui S, Deschatrette J (2000) Evidence for deterministic chaos in aperiodic oscillations of proliferative activity in long-term cultured Fao hepatoma cells. Cell Sci 113(6):1069–1074
65. Sedivy R, Windischberger C, Svozil K, Moser E, Breitenecker G (1999) Fractal analysis: an objective method for identifying atypical nuclei in dysplastic lesions of the cervix uteri. Gynecol Oncol 75(1):78–83
66. Landini G, Rippin JW (1993) Fractal dimensions of the epithelial-connective tissue interfaces in premalignant and malignant epithelial lesions of the floor of the mouth. Anal Quant Cytol Histol 15(2):144–149
67. Kitano H (2003) Cancer robustness: tumour tactics. Nature 426:125
68. Betancourt-Mar JA, Nieto-Villar JM (2007) Theoretical models for chronotherapy: periodic perturbations in funnel chaos type. Math Biosci Eng 4(2):177–186
69. Wells A (2006) Cell motility in cancer invasion and metastasis, vol 8. Springer Science & Business Media, New York
70. Volkenstein MV (2009) Entropy and information, vol 57. Springer Science, New York

71. Luo L (2009) Entropy production in a cell and reversal of entropy flow as an anticancer therapy. Front Phys China 4:122–136
72. Lucia U, Ponzetto A (2017) Some thermodynamic considerations on low frequency electromagnetic waves effects on cancer invasion and metastasis. Physica A 467:289–295
73. Kim S, Lahmy R, Riha C, Yang C, Jakubison BL, van Niekerk J, Itkin-Ansari P (2015) The basic helix-loop-helix transcription factor E47 reprograms human pancreatic cancer cells to a quiescent acinar state with reduced tumorigenic potential. Pancreas 44(5):718–727
74. Aceto N, Toner M, Maheswaran S, Haber DA (2015) En route to metastasis: circulating tumor cell clusters and epithelial-to-mesenchymal transition. Trends Cancer 1(1):44–52
75. Norton L (2005) Conceptual and practical implications of breast tissue geometry: toward a more effective, less toxic therapy. Oncologist 10:370–381
76. D'Anselmi F, Valerio M, Cucina A, Galli L, Proietti S, Dinicola S, Pasqualato A, Manetti C, Ricci G, Giuliani A, Bizzarri M (2011) Metabolism and cell shape in cancer: a fractal analysis. Int J Biochem Cell Biol 43 (7):1052–1058
77. Luo L et al (2006) Physicochemical attack against solid tumors based on the reversal of direction of entropy flow: an attempt to introduce thermodynamics in anticancer therapy. Diagn Pathol 1(1):43
78. Gatenby RA, Gillies RJ (2004) Why do cancers have high aerobic glycolysis? Nat Rev Cancer 4(11):891–899
79. Lucia U, Ponzetto A, Deisboeck TS (2016) Constructal approach to cell membranes transport: amending the 'Norton-Simon' hypothesis for cancer treatment. Sci Rep 6:19451
80. De Berardinis R, Lum JJ, Hatzivassiliou G, Thompson CB (2008) The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell Metab 7:11–20
81. Nieto-Villar JM, Izquierdo-Kulich E, Quintana González RL, Rieumont J (2013) Una aproximación del criterio evolutivo de Prigogine a sistemas químicos. Revista Mexicana de Física 59:527–529
82. Betancourt-Mar JA, Rodrıguez-Ricard M, Mansilla R, Cocho G, Nieto-Villar JM (2016) Entropy production: evolution criteria, robustness and fractal dimension. Rev Mex Fis 62:164–167
83. Llanos-Pérez J, Betancourt-Mar J, Cocho G, Mansilla R, Nieto-Villar JM (2016) Phase transitions in tumor growth: III Vascular and metastasis behavior. Physica A 462:560–568
84. Seyfried TN, Flores RE, Poff AM, D'Agostino DP (2014) Cancer as a metabolic disease: implications for novel therapeutics. Carcinogenesis 35(3):515–527
85. Molnar J (2005) Thermodynamic aspects of cancer: possible role of negative entropy in tumor growth, its relation to kinetic and genetic resistance. Lett Drug Des Discov 26:429–438
86. Wagner BA, Venkataraman S, Buettner GR (2011) The rate of oxygen utilization by cells. Free Radic Biol Med 51:700–712
87. Marín-Hernández A, Gallardo JC, Rodríguez S, Encalada R, Moreno R, Saavedra E (2011) Modeling cancer glycolysis. Biochim Biophys Acta Bioenerg 1807 (6):755–767
88. Moreno-Sánchez R (2014) Who controls the ATP supply in cancer cells? Biochemistry lessons to understand cancer energy metabolism. Int J Biochem Cell Biol 50:10–23
89. Collins VP, Loeffler RK, Tivey H (1956) Observations on growth rates of human tumors. J Roentgenol Radium Ther Nucl Med 76:988–1000
90. Leibniz-Institut DSMZ Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (2014) Catalogue of human and animal cell lines
91. Netherlands Translational Research Center B.V. (NTRC) Oncolines
92. American Type Culture Collection (ATCC) ATCC Human cell lines
93. Reeves R, Edberg DD, Li Y (2001) Architectural transcription factor HMGI(Y) promotes tumor progression and mesenchymal transition of human epithelial cells. Mol Cell Biol 21(2):575–594
94. Nunez AM, Jakolev S, Briand JP, Gaire M, Krust A, Rio MC, Chamon P (1987) Characterization of the estrogen-induced pS2 protein secreted by the human breast cancer cell line MCF-7. Endocrinology 121:1759–1765
95. Totpal K, Aggarwal BB (1991) Interleukin 4 potentiates the antiproliferative effects of tumor necrosis factor on various tumor cell lines1. Cancer Res 51:4266–4270
96. Chi TY, Chen CG, Lai PB (2004) Eicosapentaenoic acid induces Fas-mediated apoptosis through a p53-dependent pathway in hepatoma cells. Cancer J 10:190–200
97. Hoosein NM (1991) Involvement of urokinase and its receptor in the invasiveness of human prostatic carcinoma cell lines. Cancer Commun 3(8):255–264

98. Dozmorov MG (2009) Unique patterns of molecular profiling between human prostate cancer LNCaP and PC-3 cells. Prostate 69 (10):1077–1090
99. MuraliKrishna PS (2005) RNA interference-directed knockdown of urokinase plasminogen activator and urokinase lasminogen activator receptor inhibits prostate cancer cell invasion, survival, and tumorigenicity in vivo. J Biol Chem 280(43):36529–36540
100. López-Lázaro M (2008) The Warburg effect: why and how do cancer cells activate glycolysis in the presence of oxygen? Science 8:305–312
101. Gatenby ET, Gawlinski AF, Gmitro KB, Gillies RJ (2006) Acid-mediated tumor invasion: a multidisciplinary study. Cancer Res 66 (10):5216–5224
102. Rieumont J, Nieto JM, García JM (1997) The rate of Entropy Production as a mean to determinate the most important reactions step in Belousov Zhabotinsky reaction. Anales de Química 93:147–152
103. Izquierdo-Kulich E, Alonso-Becerra E, Nieto-Villar JM (2011) Entropy production rate for avascular tumor growth. J Mod Phys 2(06):615
104. Li X, Dash RK, Pradhan RK, Qi F, Thompson M et al (2010) A database of thermodynamic quantities for the reactions of glycolysis and the tricarboxylic acid cycle. J Phys Chem B 144:16068–16082
105. Alberty RA (2006) Biochemical thermodynamics: applications of mathematica. Wiley-Interscience, Hoboken, NJ
106. Montero S, Martin RR, Guerra A, Casanella O, Cocho G, Nieto-Villar JM (2016) Cancer glycolysis I: entropy production and sensitivity analysis in stationary state. J Adenocarcinoma 1:1–8
107. Parks SK, Chiche J, Pouysségur J (2013) Disrupting proton dynamics and energy metabolism for cancer therapy. Nat Rev Cancer 13 (9):611–623
108. Schito L, Semenza GL (2016) Hypoxia-inducible factors: master regulator of cancer progression. Trend Cancer 2(12):758–770
109. Jun JC, Rathore A, Younas H, Gilkes D, Polotsky VY (2017) Hypoxia-inducible factors and cancer. Curr Sleep Med Rep 3 (1):1–10
110. Martyushev LM, Seleznev VD (2006) Maximum entropy production principle in physics, chemistry and biology. Phys Rep 264(1):1–45
111. Nelson DL, Cox N (2013) Lehninger. Principles of biochemistry, 6th edn. W. H. Freeman and Company, New York
112. Lozupone F, Borghi M, Marzoli F, Azzarito T, Matarrese P, Lessi E (2015) TM9SF4 is a novel V-ATPase-interacting protein that modulates tumor pH alterations associated with drug resistance and invasiveness of colon cancer cells. Oncogene 34 (40):5163–5174
113. Fais S, Venturi G, Gatenby B (2014) Microenvironmental acidosis in carcinogenesis and metastases: new strategies in prevention and therapy. Cancer 33:195–108
114. Granja S, Tavares-Valente D, Queirós O, Baltazar F (2016) Value of pH regulators in the diagnosis, prognosis and treatment of cancer. Semin Cancer Biol. https://doi.org/10.1016/j.semcancer.2016.12.003
115. Sennoune SR, Bakunts K, Martinez GM, Chua-tuan JL, Kebir Y, Attaya MN, Martinez-Zaguilan R (2004) Vacuolar H+ ATPase in human breast cancer cells with distinct metastatic potential: distribution and functional activity. Am J Physiol 286:1443–1452
116. Pelicano H, Martin DS, Xu RH, Huang P (2006) Glycolysis inhibition for anticancer treatment. Oncogene 25(34):33–46
117. Icard P, Lincet H (2012) A global view of the biochemical pathways involved in the regulation of the metabolism of cancer cells. Biochim Biophys Acta 1826(2):423–433
118. Higashimura Y, Nakajima Y, Yamaji R (2011) Up-regulation of glyceraldehyde-3-phosphate dehydrogenase gene expression by HIF-1 activity depending on Sp1 in hypoxic breast cancer cells. Arch Biochem Biophys 509(1):1–8
119. Liu K, Tang Z, Huang A, Chen P, Liu P, Yang J (2017) Glyceraldehyde-3-phosphate dehydrogenase promotes cancer growth and metastasis through upregulation of SNAIL expression. Int J Oncol 50:252–262
120. Li X, Gu J, Zhou Q (2015) Review of aerobic glycolysis and its key enzymes - new targets for lung cancer therapy. Lung Cancer 6:17–24
121. Draoui N, Feron O (2011) Lactate shuttles at a glance: from physiological paradigms to anti-cancer treatments. Dis Model Mech 4:727–732
122. Hay N (2016) Reprogramming glucose metabolism in cancer: can it be exploited for cancer therapy? Nat Rev Cancer 16:635–649
123. Mathupala SP, Ko YH, Pedersen PL (2006) Hexokinase II: cancer's double-edged sword acting as both facilitator and gatekeeper of malignancy when bound to mitochondria. Oncogene 25:4777–4786

124. Webb BA, Chimenti M, Jacobson MP, Barber DL (2011) Dysregulated pH: a perfect storm for cancer progression. Nat Rev Cancer 11 (9):671–677
125. Sergeeva TS, Shirmanova MV, Zlobovskaya OA, Gavrina AI, Dudenkova VV, Lukina MM, Lukyanov KA, Zagaynova EV (2017) Relationship between intracellular pH, metabolic co-factors and caspase-3 activation in cancer cells during apoptosis. Biochim Biophys Acta. https://doi.org/10.1016/j.bbamcr.2016.12.022
126. Schwartza L, Seyfried T, Alfarouk KO, Moreira JV, Fais S (2017) Out of Warburg effect: an effective cancer treatment targeting the tumor specific metabolism and dysregulated pH. Semin Cancer Biol. https://doi.org/10.1016/j.semcancer.2017.01.005
127. Harguindey S, Stanciu D, Devesa J, Alfarouk K, Cardone RA, Polo Orozco JD, Devesa P, Rauch C, Orive G, Anitua E, Roger S, Reshkin SJ (2017) Cellular acidification as a new approach to cancer treatment and to the understanding and therapeutics of neurodegenerative diseases. Semin Cancer Biol. https://doi.org/10.1016/j.semcancer.2017.02.003
128. Swietach P, Vaughan-Jones RD, Harris AL, Hulikova A (2014) The chemistry, physiology and pathology of pH in cancer. Philos Trans R Soc B 329:1–9
129. Harguindey S, Oriveb G, Pedráz JL, Paradiso JA, Reshkind SJ (2005) The role of pH dynamics and the Na+/H+ antiporter in the etiopathogenesis and treatment of cancer. Two faces of the same coin—one single nature. Biochim Biophys Acta 1756:1–24
130. Reshkin SJ, Bellizzi A, Caldeira S, Albarani V, Malanchi I, Poignee M (2000) Na+/H+ exchanger-dependent intracellular alkalinization is an early event in malignant transformation and plays an essential role in the development of subsequent transformation-associated phenotypes. FASEB J 14:2185–2197
131. Abaza MY, Luqmani A (2013) The influence of pH and hypoxia on tumor metastasis. Autophagy 13(10):1229–1242
132. Link C (2011) Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. Cancer 27:441–464
133. Vicent EE, Sergushichev A, Griss T et al (2016) Mitochondrial phosphoenolpyruvate carboxykinase regulates metabolic adaptation and enables glucose-independent tumor growth. Mol Cell 60:195–207
134. Gatenby RA, Gillies RJ (2007) Glycolysis in cancer: a potential target for therapy. Int J Biochem Cell Biol 39(7):1358–1366
135. Lucia U (2013) Thermodynamics and cancer stationary states. Physica A 392:3648–3653
136. Lucia U (2014) Transport processes in biological systems: tumoral cells and human brain. Physica A 393:327–336
137. Lucía U (2014) Transport processes and irreversible thermodynamics analysis in tumoral systems. Physica A 410:380–390
138. Lucia U, Ponzetto A, Deisboeck TS (2016) Investigating the impact of electromagnetic fields on human cells: a thermodynamic perspective. Physica A 443:42–48
139. Lucia U, Ponzetto A, Deisboeck TS (2015) A thermodynamic approach to the 'mitosis/apoptosis' ratio in cancer. Physica A 436:246–255
140. Marín-Hernández A, López-Ramírez SY, Del Mazo-Monsalvo I, Gallardo-Pérez JC, Rodríguez-Enríquez S, Moreno-Sánchez R, Saavedra E (2014) Modeling cancer glycolysis under hypoglicemia, and the role played by the differential expression of glycolytic isoforms. FEBS J 281:3325–3345
141. Montero S, Durán I, Pomuceno-Orduñez JP, Martin RR, Mesa MD, Mansilla R, Cocho G, Nieto-Villar JM (2017) How much damage can make the glucose in cancer? J Tumor Res 3(1):116
142. Bryant KL, Mancias JD, Kimmelman AC, Der CJ (2014) KRAS: feeding pancreatic cancer proliferation. Trends Biochem Sci 39 (2):91–100
143. Yun J, Rago C, Cheong I, Pagliarini R, Angenendt P, Rajagopalan H, Schmidt K (2009) Glucose deprivation contributes to the development of KRAS pathway mutations in tumor cells. Science 325:1555–1559
144. Manda G, Nechifor MT, Neagu TM (2009) Reactive oxygen species, cancer and anticancer therapies. Curr Chem Biol 3:342–366
145. Iglesias PA, Ingalls BP (2010) Control theory and systems biology. The MIT Press, Cambridge, MA
146. Goldbeter A (1997) Biochemical oscillations and cellular rhythms: the molecular bases of periodic and chaotic behaviour. Cambridge University Press, Cambridge
147. Potter VR, Gebert RA, Pitot HC, Peraino C, Lamar C Jr, Lesher S, Morris HP (1966) Systematic oscillations in metabolic activity in rat liver and in hepatomas I. Morris Hepatoma No. 7793. Cancer Res 26 (1):1547–1560

148. Potter VR, Watanabe M, Pitot HC, Morris HP (1969) Systematic oscillations in metabolic activity in rat liver and hepatomas. Survey of normal diploid and other hepatoma lines. Cancer Res 29:55–78
149. Smolen P (1995) A model for glycolytic oscillations based on skeletal muscle phosphofructokinase kinetics. J Theor Biol 174:137–148
150. Higgins J (1964) A chemical mechanism for oscillation of glycolytic intermediates in yeast cells. Proc Natl Acad Sci U S A 51:989–994
151. Selkov EE (1968) Self-oscilations in glycolysis. Eur J Biochem 4:79–86
152. Termonia Y, Ross J (1981) Oscillations and control features in glycolysis: numerical analysis of a comprehensive model. Proc Natl Acad Sci U S A 78(5):2952–2956
153. Decroly O, Goldbeter A (1982) Birhythmicity, chaos, and other patterns of temporal self-organization in a multiply regulated biochemical system. Proc Natl Acad Sci U S A 79:6917–6921
154. Hynne F, Danno S, Sorensen PG (2001) Full-scale model of glycolysis in Saccharomyces cerevisiae. Biophys Chem 94:121–163
155. Nielsen K, Sorensen PG, Hynne F, Busse HG (1998) Sustained oscillations in glycolysis: an experimental and theoretical study of chaotic and complex periodic behavior and of quenching of simple oscillations. Biophys Chem 72:49–62
156. Teusink B, Bakker BM, Westerhoff HV (1996) Control of frequency and amplitudes is shared by all enzymes in three models for yeast glycolytic oscillations. Biochim Biophys Acta 1275:204–212
157. Guerra A, Triana L, Montero S, Martin R, Rieumont J, Nieto-Villar JM (2014) La producción de entropía en la glicólisis del cáncer. Rev Cub Fis 31:103
158. Nelson DL, Lehninger MC (2008) Principles of biochemistry, 5th edn. W. H. Freeman and Company, New York
159. Hamanaka RB, Chandel NS (2012) Targeting glucose metabolism for cancer therapy. J Express Med 209:211
160. Martín RR, Montero S, Silva E, Bizzarri M, Cocho G, Mansilla R, Nieto-Villar JM (2017) Phase transtition in tumor growth: V What can be expected from cancer glycolytic oscillation? Physica A 486:762–771
161. Anischenko VS, Vadivasova TE, Okrokvertskhov GA, Strelkova GI (2003) Correlation analysis of dynamical chaos. Physica A 325 (1):199–212
162. Bruers S (2006) Classification and discussion of macroscopic entropy production principles. arXiv preprint cond-mat/0604482
163. Rieumont J, Nieto-Villar JM, García JM (1997) The rate of Entropy Production as a mean to determinate the most important reactions step in Belousov Zhabotinsky reaction. Anales de Química Int Ed 93:141–152
164. Betancourt-Mar J, Mansilla R, Cocho G, Nieto-Villar JM (2017) What can be learned from a phase transitions in tumor growth? Insights Biomed 2(1):2

Chapter 9

Complexity of Biochemical and Genetic Responses Reduced Using Simple Theoretical Models

Kumar Selvarajoo

Abstract

Living systems are known to behave in a complex and sometimes unpredictable manner. Humans, for a very long time, have been intrigued by nature, and have attempted to understand biological processes and mechanisms using numerous experimental and mathematical techniques. In this chapter, we will look at simple theoretical models, using both linear and nonlinear differential equations, that realistically capture complex biochemical and genetic responses of living cells. Even for cases where cellular behaviors are stochastic, as for single-cell responses, randomness added to well-defined deterministic models has elegantly been shown to be useful. The data collectively present evidence for further exploration of the self-organizing rules and laws of living matter.

Key words Systems biology, Nonlinear dynamics, Biological networks, Oscillation, Modeling

1 Introduction

Recent estimates suggest that the human body consists of about 4×10^{13} cells across 200 cell types, covering different organs and tissues. Within each cell, there are an estimated 20–25,000 genes, and over 100,000 proteins and metabolites. These molecular species, the fundamental building blocks of life, are connected through a complex series of biochemical reaction networks that process external information for survival and reproduction (Fig. 1) [1]. For example, the food that an organism ingests is broken down into various biochemicals, such as carbohydrates, proteins, lipids, and nuclei acids through diverse reaction networks, either spontaneously with binding partners or directed through catalytic enzymes.

Aberrations to the highly robust biochemical networks, either through genetic mutations or non-genetically through chronic poor lifestyle habits, can lead to diseases and pre-mature death. Thus, understanding the dynamic behaviors of various biochemical networks is indispensable for biological research and good health,

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_9,

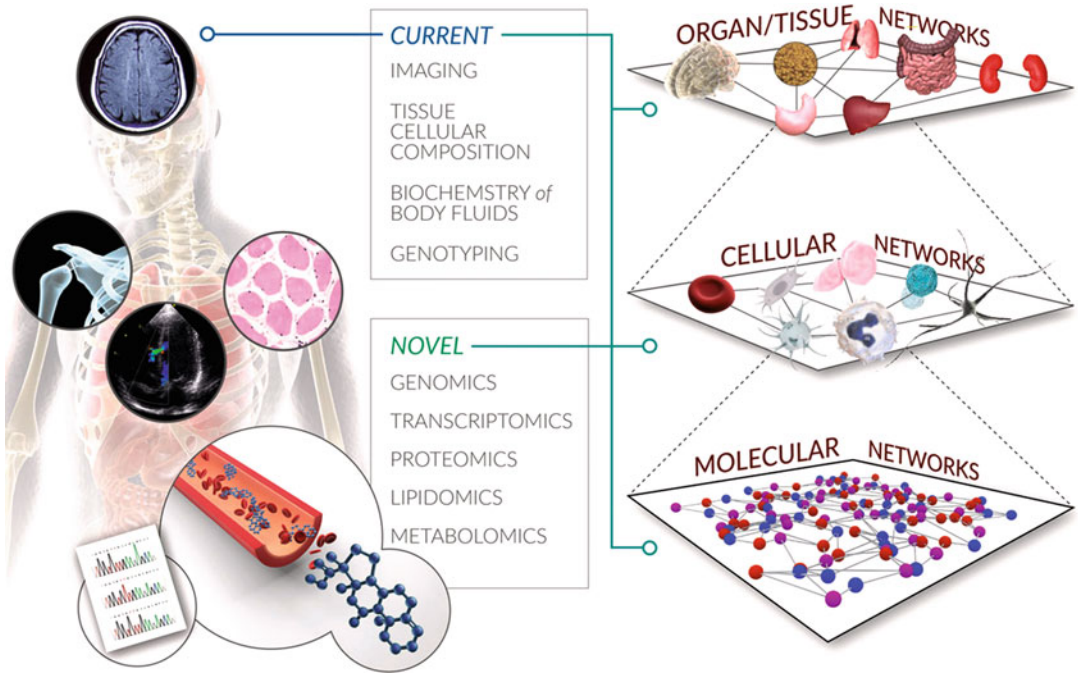


Fig. 1 Molecular species from cells, tissues, and organs are biochemically inter-connected through large-scale networks. Figure adapted from [1]

especially for overcoming diseases, where therapeutic targets can be developed to control specific biochemical aberrations. Biochemical network studies are also useful for understanding the origins of life, as there have been several studies that have reproduced complex biological properties using simple models, for example, the generation of self-organized animal skin patterns using Alan Turing's reaction-diffusion equations (*see* Subheading 5) [2]. However, since there are trillions of possible combinatorial molecular interactions with a large number of biochemicals within a cell, it is an overwhelming task to interpret cellular responses in every aspect or in entirety. Therefore, it is conceivable and necessary that a reductionist approach to investigating functional aspects of living systems is taken.

The reductionist view, although is meant to simplify complexities, can still be very useful. In aerospace science, the Navier-Stokes equations highly simplify the macroscopic streamline flow of air and do not take into account fluctuations or turbulence at the microscopic scales. However, it breaks down under highly turbulent conditions where aircraft are difficult to control, and at supersonic speeds where thermal heating becomes a challenge. Nevertheless, modeling the airflow using the Navier-Stokes equations has been fundamental for the design of modern airplanes that travel at less than the speed of sound, and has been widely adopted by the

aerospace industries. Similarly, in thermodynamics related to gas flows, the ideal gas law (the popular $PV = nRT$ equation), although poses several limitations, is still widely used in the chemical industries as a good approximation to study the behavior of many gases under various conditions. The lessons from other disciplines over the last two centuries have indicated that simple theoretical foundations under a "window" of operating conditions can be highly beneficial to model realistic behaviors [3, 4].

In biology, the goal to simplify complex biochemical networks using a reductionist view is also not new. In the early part of the last century, Victor Henri, Leonor Michaelis, and Maud Menten thoroughly investigated enzymatic biochemical reactions, in vitro, and developed the hyperbolic rate equation that we now popularly call the Michaelis-Menten enzyme kinetics. Using this method, functional networks of the energy metabolism, such as the glycolysis and the Krebs cycle, have been widely studied. However, the earlier models were plagued with the dilemma where increased accuracy in model prediction required detailed knowledge of in vivo reaction parameters that were too difficult to measure precisely. Most, if not all, studies adopted in vitro experiments to determine the parameter values of reaction species from an artificial environment where the species were deliberately purified from its physiologic neighbors. There have been various reports that claim the kinetic parameters determined through in vitro and in vivo experiments can differ by several orders of magnitudes [5]. As a result, when combining these errors into the model, the final predictions could differ by several orders of magnitude. For example, the steady-state concentration of the glycolytic metabolite 3-phosphoglycerate in *Trypanosoma brucei* was under-predicted by an order of 7 [6].

Despite the difficulty posed by adaptive living systems, numerous mathematical techniques have been developed and used to decipher major system-level properties in development, differentiation, growth, aging, and the immune response. In this chapter, fundamentals of simple linear models for understanding cellular response that follow formation and depletion waves are introduced (Subheading 2). This is followed by a review on nonlinear models developed to recapture oscillatory responses observed in cells (Subheading 3). Linear stability analysis that could be used to study the robustness of multi-stable states of living systems is also introduced (Subheading 4). A brief look at the reaction-diffusion equations for spatio-temporal pattern analysis is presented in the penultimate section (Subheading 5), followed by comments on stochasticity and heterogeneity more recently observed in single-cell systems (Subheading 6).

2 Modeling Biochemical Reactions

2.1 Mass Action Kinetics

The earliest research in biochemical reactions, around the mid-1800s, dealt with chemical equilibrium. That is, chemical systems where perturbation from a stable steady-state subsequently returns to its original state are said to be in equilibrium. For example, heating oil at room temperature and, subsequently, removing the heat results in the temperature of the oil decay to its original reading. This happens in a "closed" system where external conditions, such as pressure, energy, mass, do not enter or leave the system that is in cooling.

To illustrate mathematically, consider a simple case involving only the decay process, that is, the point onward when a heat source is removed. The temperature (T) that changes over time follows the deterministic Newton's Law of Cooling:

$$\frac{dT}{dt} = -k[T - T_0] \tag{1}$$

where T_0 is the room temperature and k is the rate of cooling. The minus sign indicates temperature drop in time. Equation 1 is called an ordinary differential equation, where the rate of change of T depends only on its concentration and a rate constant.

Now consider a closed chemical reaction of molecular species A into B:

$$A \rightarrow B$$

The arrow indicates the equilibrium state lies far to the right, that is, the reverse reaction ($B \rightarrow A$) proceeds only at an infinitesimal extent and can be ignored. For every species B formed (concentration units in mol), an A species disappears:

$$\frac{d[B]}{dt} = \frac{-d[A]}{dt} = k_1[A] \tag{2}$$

The constant k_1 is called the rate constant and it has the unit of per second. Here, the masses of chemicals are conserved, and the rate constant provides a direct measure of how fast the reaction is. The higher the k_1 value, the faster the reaction. This type of reaction is called a first-order reaction, as its rate only depends on the first power of the reactant concentration (Fig. 2A).

A second-order reaction occurs typically when two same species react to form a chemical product. An example:

$$2A \xrightarrow{k_2} A_2$$

The rate of such a reaction is proportional to the second power of the concentration of reactant, for the reaction can occur only when two molecules of the same species collide:

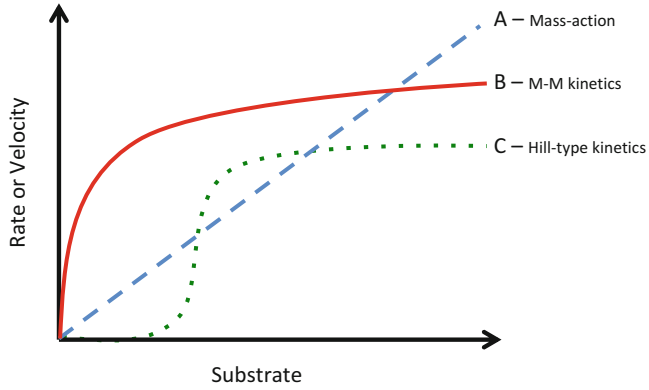


Fig. 2 Schematic of velocity (rate) versus substrate (species) concentration for (*A*) mass-action, (*B*) Michaelis-Menten kinetics, and (*C*) Hill-type allosteric reactions

$$\frac{-d[A]}{dt} = k_2[A]^2 \tag{3}$$

where k_2 is the second-order rate constant. It has dimensions of $(\text{mol/L})^{-1}\ \text{s}^{-1}$.

A chemical reaction can also be reversible, where species A can become species B and vice versa with different rate constants:

$$A \underset{k_{-1}}{\overset{k_1}{\leftrightarrow}} B$$

$$\frac{d[B]}{dt} = \frac{-d[A]}{dt} = k_1[A] - k_{-1}[B] \tag{4}$$

It follows that for any system in chemical equilibrium, the rate of an elementary reaction is proportional to the product of the concentrations of the reacting species. These types of reactions are called the mass-action kinetics, and Cato Maximilian Guldberg and Peter Waage first devised them in 1864 [7].

2.2 Enzyme Kinetics

In certain types of biochemical reactions, such as in metabolic reactions, other species can involve and aid a reaction without themselves being affected through the process. These species are usually catalytic proteins. To consider such situations, the hyperbolic rate equation, which we now popularly call the Michaelis-Menten enzyme kinetics, was introduced. This is a more sophisticated form of mass-action type reaction, which considers the role of enzymes (proteins that act as catalyst). Unlike mass-action, the kinetics of reactions saturates at higher substrate concentrations instead of ever increasing profile for the former.

For reactions that require catalytic enzymes, the mechanism to account for such reactions assumes that the species A combines with species E (catalyst or enzyme), in a reversible manner to give complex EA, which then dissociate reversibly or react irreversibly to produce species B while leaving E unchanged.

$$E + A \underset{k_{-1}}{\overset{k_1}{\leftrightarrow}} EA \overset{k_{cat}}{\rightarrow} E + B$$

In this case, the rate of B formation can be shown to be (*see* ref. 8 for detailed derivation):

$$\frac{dB}{dt} = \frac{V_{max}[A]}{[A] + K_M} \tag{5}$$

where $V_{max} = k_{cat}[E]_t$ and $K_M = \frac{k_{cat}+k_{-1}}{k_1}$.

The reaction rate increases with increasing $[A]$, approaching an asymptotic at V_{max}, when all enzymes (limiting factor) are bound to A (Fig. 2B). $[E]_t$ is the total enzyme concentration and k_{cat} is the maximum number of enzymatic reactions catalyzed per second. Subsequent work on this basic principle led to the extension of the kinetics to represent more complex scenarios, such as multi-substrate ping-pong and ternary-complex mechanisms [8].

There are certain classes of enzymes with multiple active sites that alter the reaction kinetics in complex ways. For example, the reaction activity for lower substrate concentration is inefficient while at higher concentration the activity is highly efficient, resulting in S-shape reaction rates (Fig. 2C). This usually occurs through substrate concentration-dependent conformation changes that vary the enzyme affinity.

The commonly used allosteric reactions adopt the Hill equation, which is a modified form of the Michaelis-Menten kinetics:

$$\frac{dB}{dt} = \frac{V_{max}[A]^n}{[A]^n + K_I} \tag{6}$$

where n is the Hill coefficient that describes the cooperativity, and K_I is a constant that is different to K_M. Note that negative or positive cooperativity is represented when $n < 1$ or $n > 1$, respectively. When $n = 1$, the Hill equation becomes Michaelis-Menten kinetics.

Overall, there are various forms of enzyme kinetics, depending on the types of intermediates or co-factor affecting the overall reactions. There are entire books just dealing with different types of enzyme kinetics, and the details are beyond the scope of this chapter.

2.3 From Reactions to Networks

As the development of computing power progressed significantly in the 1960s, there have been numerous efforts to model complete biological network modules, such as energy metabolic pathways and immune signaling cascades. Here, we consider a series of chemical reactions that form pathways and networks.

Consider a closed system with n species, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, that are connected through chemical reactions. Given a perturbation to one of the species in the system, the resultant changes in the

concentration of the connected species are governed by the generalized kinetic evolution equation:

$$\frac{\partial X_i}{\partial t} = F_i(X_1, X_2, \ldots, X_n), i = 1, \ldots, n \tag{7}$$

where the corresponding vector form of Eq. 7 is $\frac{\partial \boldsymbol{X}}{\partial t} = \boldsymbol{F}(\boldsymbol{X})$. $\boldsymbol{F}$ is a vector of any nonlinear function including reaction and diffusion of the species. Here, the partial, instead of ordinary, differential notation is used to consider diffusion processes, where the species' concentrations could also vary with space. For simplicity, we will now ignore the diffusion process and revert to the ordinary form, assuming a well-mixed homogenous condition. Equation 7 becomes

$$\frac{dX_i}{dt} = F_i(X_1, X_2, \ldots, X_n), i = 1, \ldots, n \tag{8}$$

For illustration with an example, let $n = 3$ and each reaction be governed by first-order mass action kinetics

$$X_1 \xrightarrow{k_1} X_2 \xrightarrow{k_2} X_3$$

Equation 8 for the above can be written as

$$\frac{dX_1}{dt} = -k_1[X_1] \tag{9}$$

$$\frac{dX_2}{dt} = k_1[X_1] - k_2[X_2] \tag{10}$$

$$\frac{dX_3}{dt} = k_2[X_2] \tag{11}$$

In a closed system, as there is no exchange of species to external environments, the total masses of all species at any time will be constant. Also, the rates of reactions, k values, are assumed constant. Therefore, the summation of the three differential equations 9 to 11 will be zero. This simplest linear system will remain stable for all positive real values of rate constants or species concentration. Figure 3a shows the concentrations of species with time for a selected initial condition and parameter values.

Let us now consider the second reaction, X_2 to X_3, utilizes an enzymatic catalyst (X_1 to X_2 remains unchanged), Eqs. 10 and 11 become:

$$\frac{dX_2}{dt} = k_1[X_1] - \frac{V_{\max}[X_2]}{K_M + [X_2]} \tag{12}$$

$$\frac{dX_3}{dt} = \frac{V_{\max}[X_2]}{K_M + [X_2]} \tag{13}$$

Although the reaction kinetics has changed and follows a hyperbolic relation, the system still remains linear and stable for

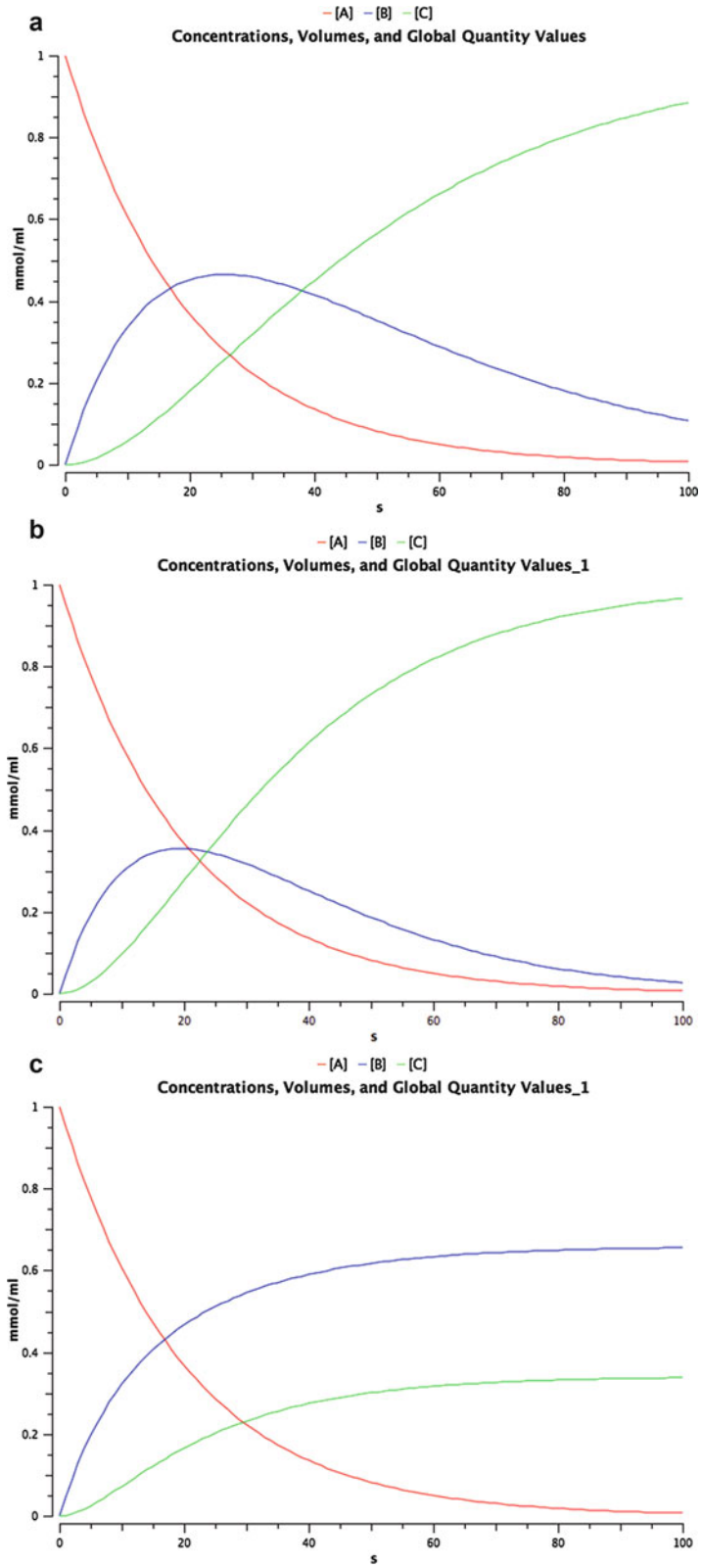


Fig. 3 Simulations for (*A*) mass-action, (*B*) Michaelis-Menten (M-M) kinetics, and (*C*) Michaelis-Menten kinetics with reversible reaction. The initial condition for $A = 1.0$, B and $C = 0$ mmol/ml for all simulations. For (**a**) $k_1 = 0.05$/s and $k_2 = 0.03$/s, (**b**) $k_1 = 0.05$/s, $K_m = 9$, $V_{max} = 0.5$ mmol/ml s and (**c**) $k_1 = 0.05$/s, $K_m = 9$, $V_{max} = 0.5$ mmol/ml s and $k_3 = 0.1$/s

any set of positive real parameter values. Note that we do not use negative or complex numbers for parameter values for biochemical systems. Figure 3b shows the concentrations of species with time.

To consider reversible reactions, from X_3 to X_2 with rate constant k_3:

$$X_1 \xrightarrow{k_1} X_2 \underset{k_3}{\leftrightarrow} X_3$$

$$\frac{dX_2}{dt} = k_1[X_1] - \frac{V_{\max}[X_2]}{K_M + [X_2]} + k_3[X_3] \tag{14}$$

$$\frac{dX_3}{dt} = \frac{V_{\max}[X_2]}{K_M + [X_2]} - k_3[X_3] \tag{15}$$

Here, the reversible reaction is assumed to follow mass-action kinetics to the previous situation. Figure 3c shows the concentrations of species with time.

In this manner, biochemical reactions can be connected to form complicated biological pathways and networks, where each reaction is represented based on the detailed information available. In the case where sufficient data are unavailable, simple mass-action kinetics can be used as a first approximation to compare simulation with experimental dynamics.

The mass-action and enzyme kinetic equations have widely been used to model reaction pathways that are known to display deterministic responses consisting of formation and depletion waves. That is, when the stimulation or perturbation of cells in vitro (in a dish) with respective biochemicals resulted in the dynamic activation profiles of intracellular molecular species that followed gradual increase from their initial state to reach peak activation levels and, subsequently, decay to their original state or a new stable state. Figure 4 gives examples observed in glycolysis, epidermal growth factor (EGF) receptor, and toll-like receptors (TLRs) signaling response. Although the kinetics can vary slightly from sample to sample or dish to dish (due to technical or cellular variability, *see* Subheading 6), the general average response profiles are very well reproducible. In other words, even in complex reaction networks, certain systems' averaged properties display simple deterministic waves.

A vast majority of cellular models today are based on mass-action kinetics, Michaelis–Menten kinetics, or even Boolean logics. In most circumstances, if not all, the investigations considered "closed" system modular approach, where the models did not include continuous exchange of materials between the internal and external environments and, hence, chemical and thermal equilibrium have been assumed. That is, the approaches often adopted well-mixed, homogenous, and isothermal environment where each reaction in the cellular network is connected through first-order,

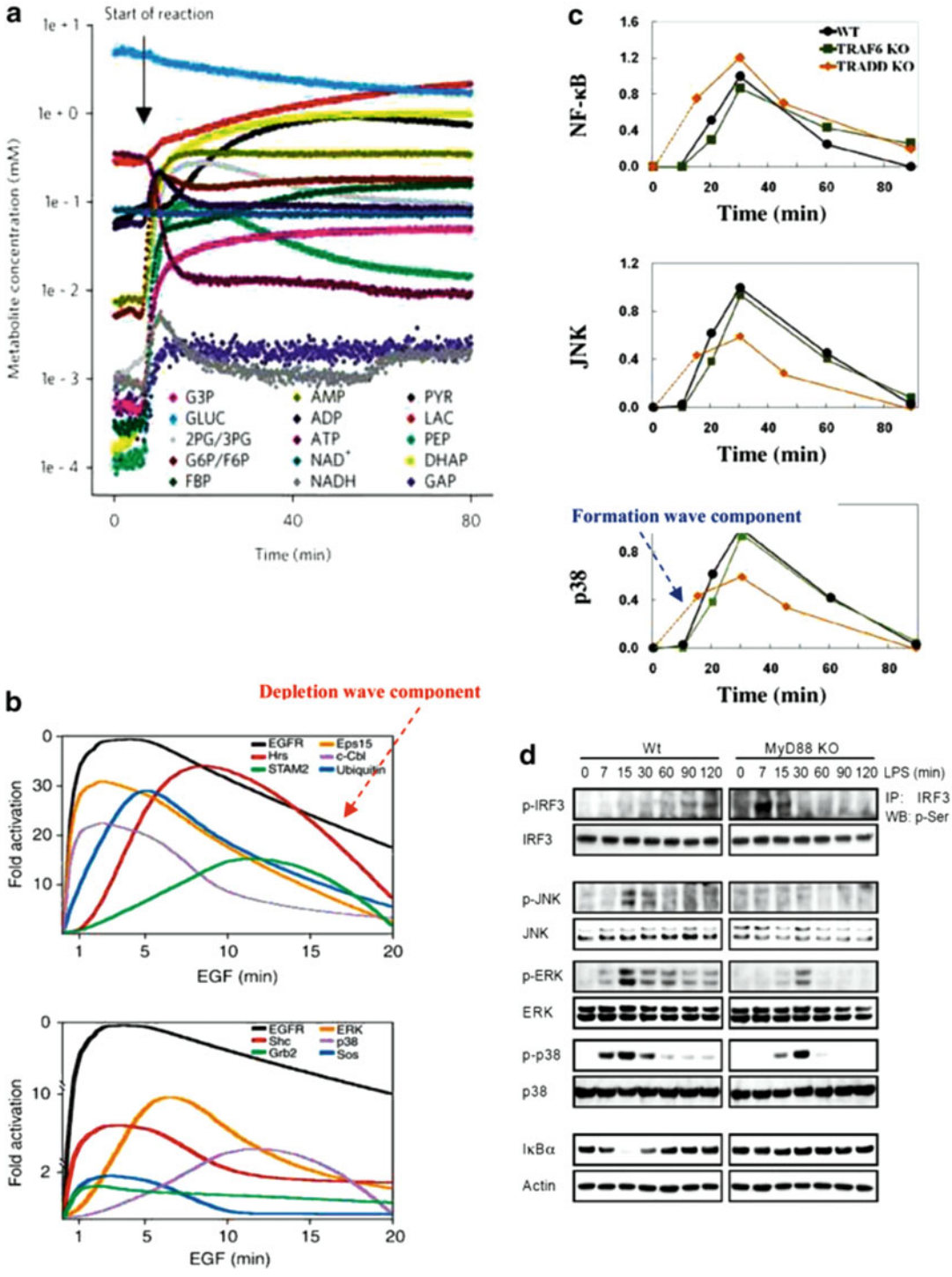


Fig. 4 Linear response observed in cellular dynamics. (**a**) Glycolysis, adapted from [9], (**b**) Epidermal growth factor (EGF) signaling, adapted from [10], (**c**) Toll-like receptor (TLR) 3 signaling, adapted from [11] and (**d**) TLR 4 signaling, adapted from [12]. All species' dynamics follow gradual increase and decrease that can be successfully modeled using mass-action or M-M kinetics

higher-order mass-action, enzyme kinetic equations, or simply Boolean logics, depending on the knowledge gained for an individual reaction.

In our research, we have used the linear mass-action approach and the law of conservation to model the dynamical responses of innate immune [11–14] and cancer response [15–17] to distinct perturbations (closed system approximation). In these signaling studies, the respective models have revealed missing molecular species, novel bypass pathways, and crosstalk mechanisms that were experimentally verified. In addition, for TNF and TRAIL signaling, the models have aided in identifying crucial molecules to regulate proinflammatory and apoptotic responses, respectively. From the successes of these studies and that of many others, it is evident that highly complex biological networks, upon external perturbation, can stably process their downstream information through linear response waves [4, 18, 19].

It is interesting to note that numerous closed system linear models have generated insightful results, especially when dealing with population-averaged metabolic or signaling networks, despite the fact that living systems are constantly exchanging matter and energy to the surroundings. As such, organisms or cells should be considered to exist far from equilibrium to achieve biological order [20]. One important example is the ability of bacteria to exchange pheromones during environmental threats, such as antibiotic treatments, to form biofilms that are highly organized structures resistant to the therapeutic intervention [21]. The biofilm example demonstrates that the cooperative behavior of organisms can be very different to their individual response. Thus, using the ergodic principle or predictive deterministic approaches to understand the majority of cellular behaviors can be questionable, and this issue has been debated from time to time.

The following sections are devoted to more complex response that do not follow the closed system approximation and require more sophisticated, yet simple, nonlinear differential equations to understand their dynamical response.

3 Nonlinear Dynamics

3.1 Periodic Oscillations

The mass-action and enzyme kinetics equations have been largely used to study stable equilibrium conditions, where steady-state levels (i.e., $\frac{dX_i}{dt} = 0$) or clear formation and depletion (linear) response waves of molecular species are observed. Under several other conditions, self-organizing oscillatory behaviors and multistable levels have been realized. Under these conditions, nonlinear differential equation approaches have been investigated.

Periodic behaviors or biological rhythms, such as sleep and menses cycles, have been evident since the evolution of life.

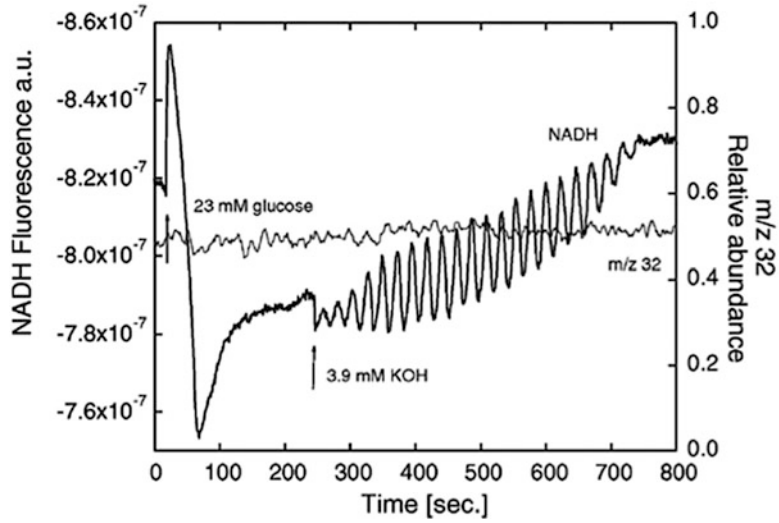


Fig. 5 Oscillatory Cellular Dynamics in Yeast. Glycolytic (NADH) oscillations are induced by adding KOH. Figure adapted from [22]

However, the observation of periodic and synchronized oscillations in laboratory yeast glycolysis in the late 1950s, subsequently, brought immense focus to biochemists interested in understanding the self-organizing response (Fig. 5) [22]. Each yeast, in a suspended culture, was able to synchronize its oscillation with other yeasts collectively to gain phase with each other. The yeast was one of the first evidence under laboratory conditions that showed order arising among cell cultures under carefully controlled biological parameters (culture volume, aeration rate, fermenter agitation rate, etc.). However, the continuous oscillations are sensitive to laboratory conditions and can be destroyed if any of the control parameters was not precisely maintained. Thus, in the words of Nobel laureate Ilya Prigogine, living organisms can be considered dissipative systems, where energy and matter are exchanged to generate order [23].

The observation of self-organizing behaviors in simple laboratory experiments led many theoretical biologists to investigate nonlinear approaches to model biological networks.

3.2 Belousov–Zhabotinsky Reaction and the Brusselator

While studying the Kreb's cycle to identify an inorganic analog, Boris Belousov, in the 1950s, accidently observed periodic spatial patterns when he mixed citric acid, bromate, and cerium with sulfuric acid solution (Fig. 6). Inspired by this, Anatol M. Zhabotinskii further worked on similar self-organizing behaviors using malonic acids, resulting in non-equilibrium thermodynamics and the establishment of a nonlinear chemical system (Belousov–Zhabotinsky or B-Z reaction), where oscillatory behavior or multi-stable states are produced autocatalytically through feedback regulation of one of its species.

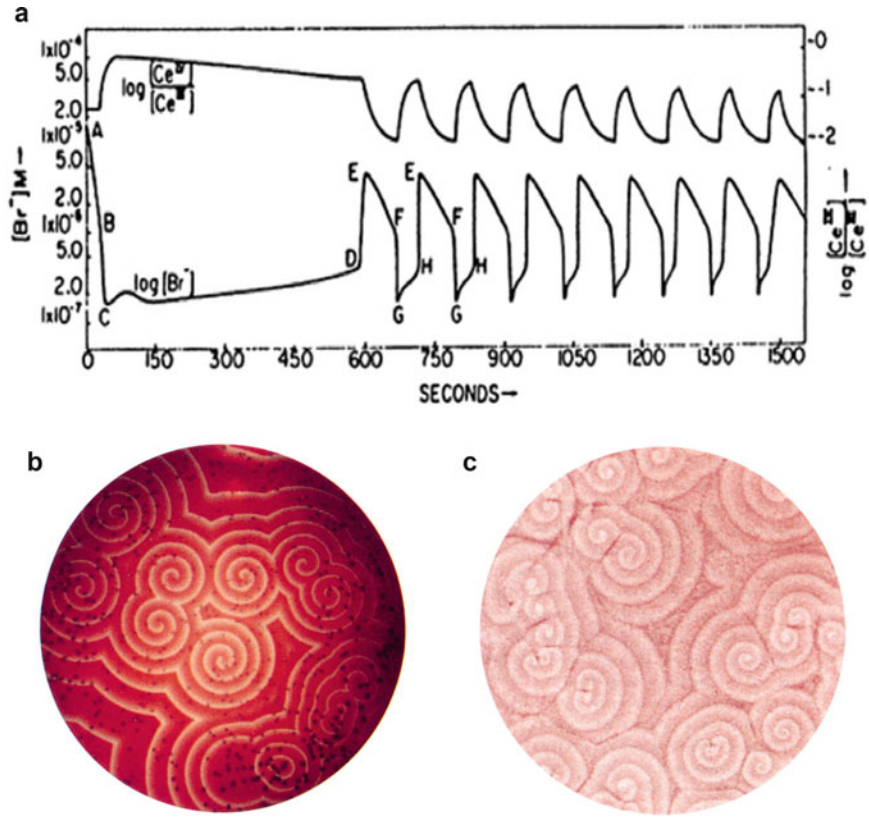


Fig. 6 Belousov–Zhabotinsky (BZ) reaction dynamics. (**a**) The oscillations of BZ species after the induction period, adapted from [24]. (**b**) BZ reaction in 2-D petri dish (*left*) and similar pattern found on *Dictyostelium discoideum*, adapted from [25]. Copyright (2006) National Academy of Sciences, U.S.A.

The B-Z reaction mechanism, a nonlinear chemical oscillator, serves as an example for non-equilibrium thermodynamics.

The Brusselator is a simplified theory to model non-equilibrium, nonlinear chemical system, developed by Ilya Prigogine and colleagues in Brussels that has been shown to reproduce the dynamics of B-Z reactions. Although the Brusselator (named after combining Brussels and Oscillator) considers six species in its four chemical reactions, the solution is given only for two autocatalytic species that can produce limit-cycle oscillations. The remaining four species act as regulators for the self-organizing behavior.

To illustrate the six-species Brusselator:

$$A \xrightarrow{k_1} X_1$$

$$2X_1 + X_2 \xrightarrow{k_2} 3X_1$$

$$B + X_1 \xrightarrow{k_3} X_2 + C$$

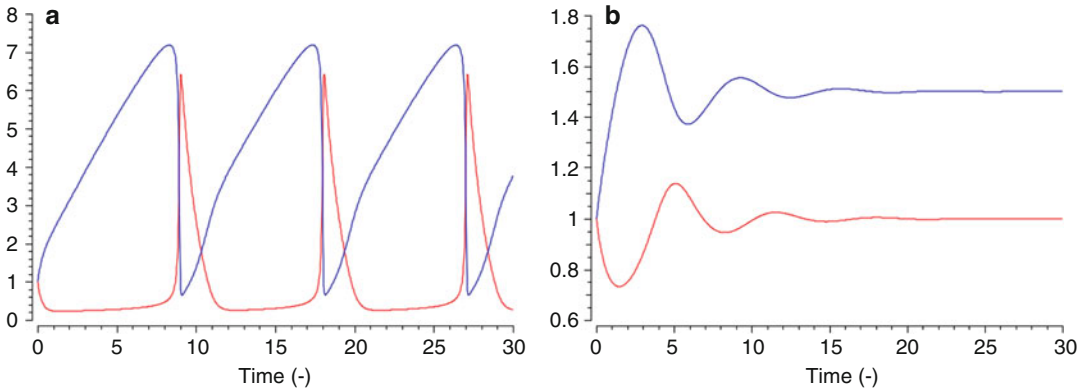


Fig. 7 Brusselator dynamics. (**a**) Limit cycle oscillations with $X_{10} = X_{20} = 1$, $A = 1$, $B = 4$, $k_1 = k_2 = k_3 = 1$, (**b**) Damped oscillations with $X_{10} = X_{20} = 1$, $A = 1$, $B = 1.5$, $k_1 = k_2 = k_3 = 1$. *x*-axis represents time and *y*-axis represents concentration in arbitrary units. X_1 is *red* while X_2 is *blue*

$$X_1 \xrightarrow{k_4} D$$

These reactions lead to the following ordinary differential equations:

$$\frac{dX_1}{dt} = k_1 A - k_2 B[X_1] + k_3[X_1]^2[X_2] - k_4[X_1] \tag{16}$$

$$\frac{dX_2}{dt} = k_2 B[X_1] - k_3[X_1]^2[X_2] \tag{17}$$

where the rate constants k_1 to k_4 and species A, B are real and positive. Note that X_1 and X_2 are reactant species in dimensionless form and, generally, $\sum_{i=1}^{2} \frac{dX_i}{dt} \neq 0$ meaning that the law of mass conservation is not observed for the Brusselator as it considers a non-equilibrium and nonlinear dissipative system. It can be shown that the Brusselator can reach equilibrium state under certain conditions, when $X_1 = A$, $X_2 = (B/A)$ and $B < 1 + A^2$ with all rate constants set to 1. Figure 7 shows stable dynamics of species X_1 and X_2 to different values of parameters, demonstrating limit-cycle and damped oscillations.

There have also been other works, subsequently, extending the Brusselator model, for example the Oregonator developed by Field and Noyes [26] which considers a third autocatalytic species, to reflect more realistic chemical dynamics of the B-Z reactions. Fundamentally, the other works are also based on non-equilibrium and nonlinear conditions, to model self-organizing chemical systems displaying emergent responses that do not show "sum of the parts" or linear dynamics.

3.3 Goodwin Model

The Brusselator was developed for chemical systems. For studying biological rhythms considering the regulation of genes and

proteins, Brian C. Goodwin developed, in 1965, a highly simplified regulatory network with a negative feedback mechanism to display oscillatory behaviors [27]. His coupled oscillator model simulated synchronous locking and sub-harmonic resonance arising from the interaction of the oscillators. By varying the coupling constants, the Goodwin model was able to show a wide range of oscillatory frequencies.

The single-oscillator Goodwin model is illustrated by

$$\frac{dX_1}{dt} = \frac{k_1}{k_2 + k_3[X_2]} - k_4 \tag{18}$$

$$\frac{dX_2}{dt} = k_5[X_1] - k_6 \tag{19}$$

where X_1 represents a mRNA while X_2 is the protein coded by X_1. k_1 and k_5 represent the formation rate constants while k_4 and k_6 represent the decay or depletion rate constants of X_1 and X_2, respectively. k_2 and k_3 control the negative regulation of X_1 independently and dependently by X_2, respectively. Figure 8a shows the suppression of X_1 on X_2 leads to periodic oscillatory dynamics.

To consider multiple crosstalk regulation between mRNAs and proteins, Goodwin proposed a coupled oscillator

$$\frac{dX_{11}}{dt} = \frac{k_{11}}{k_{21} + k_{31}[X_{21}] + k_{41}[X_{22}]} - k_{51} \tag{20}$$

$$\frac{dX_{12}}{dt} = \frac{k_{12}}{k_{22} + k_{32}[X_{21}] + k_{42}[X_{22}]} - k_{52} \tag{21}$$

$$\frac{dX_{21}}{dt} = k_1[X_{11}] - k_2 \tag{22}$$

$$\frac{dX_{22}}{dt} = k_3[X_{12}] - k_4 \tag{23}$$

As seen from Fig. 8b–d, the Goodwin's coupled oscillator can be used to produce several complex oscillatory patterns according to the parameter values chosen, compared to the one in Fig. 8a. Therefore, the model has been investigated on numerous occasions to understand emergent biological oscillatory and self-organizing responses [28–30]. However, unlike the Brusselator, the Goodwin models could not produce limit-cycle oscillation (a closed trajectory observed in phase-space plots, where at least one other trajectory spirals into it as time approaches infinity). Nevertheless, there have been subsequent modifications to this model, for example by Griffith in 1968 that overcame the limitation by making the repressor term a sigmoidal Hill coefficient larger than 8 [31].

3.4 Synthetic Biological Oscillator

To understand oscillatory behaviors in actual living systems, Elowitz and Leibler developed an artificial biological clock, called the repressilator, in the bacteria *Escherichia coli* [32]. Starting from a

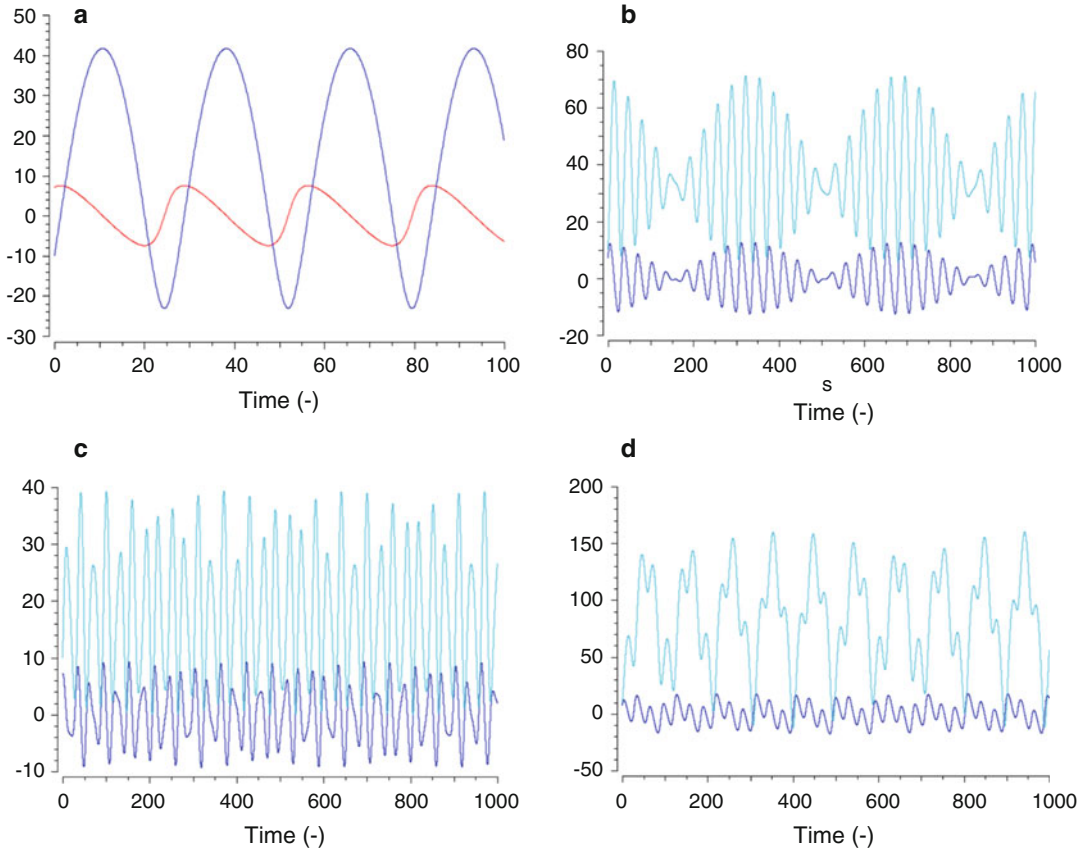


Fig. 8 Simple and Complex Goodwin dynamics. Simulations with single oscillator with parameters (**a**) $X_0 = 7$, $Y_0 = -10$, $k_1 = 72$, $k_2 = 36$, $k_3 = 1$, $k_4 = 2$, $k_5 = 1$, $k_6 = 0$, and coupled oscillators (**b**) $X_{10} = 7$, $Y_{10} = 10$, $X_{20} = 7$, $Y_{20} = 10$, $k_{11} = k_{12} = 360$, $k_{21} = 36$, $k_{22} = 43$, $k_{31} = 1.0$, $k_{32} = 0$, $k_{41} = 0.1$, $k_{42} = 1$, $k_{51} = 5$, $k_{52} = 5$, $k_1 = 0.5$, $k_2 = 0$, $k_3 = 0.6$, $k_4 = 0$, (**c**) $k_{31} = 0.5$, $k_{41} = 0.2$, $k_{51} = 8$, $k_{52} = 8$, (**d**) $k_{31} = 0.1$, $k_{32} = 1$. Only parameters that are different from (**b**) are listed for (**c**) and (**d**). *x*-axis represents time and *y*-axis represents concentration in arbitrary units. X_1 is *red* while X_2 is *blue* for (**a**), X_1 is *dark blue* while X_2 is *light blue* for **b–d**

simple six coupled linear differential equations model approximating the gene regulatory network of repressing genes, they investigated the key regulatory parameters that are required to produce unstable steady-state, resulting in regular oscillations:

$$\frac{dX_i}{dt} = X_i + \frac{k_1}{1 + \Upsilon_j^n} - k_2 \tag{24}$$

$$\frac{d\Upsilon_i}{dt} = -k_3\left(\Upsilon_j - X_i\right) \tag{25}$$

where X_i (i = *lacI*, *tetR*, *cI*) represent mRNAs (genes) and Υ_j (j = λcI, LacI, TetR) represent proteins concentrations. Figure 9a shows the oscillations in protein concentrations with time.

Subsequently, a plasmid encoding the repressilator and a reporter protein were constructed and inserted into the bacteria.

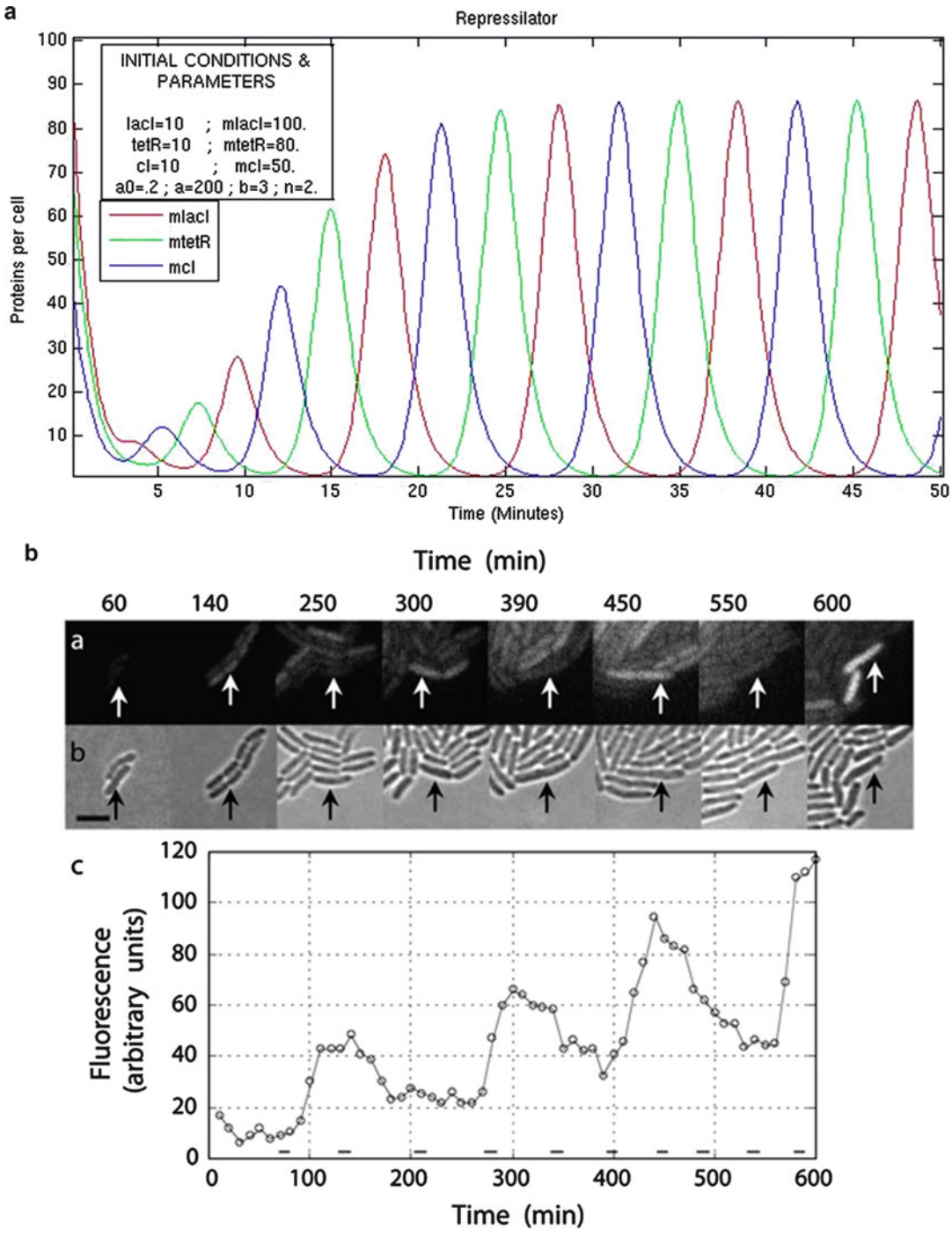


Fig. 9 A synthetic biological oscillator. (**a**) Simulations of oscillator with initial conditions and parameters, see *insert*. (**b**) Time-course experimental observation of growth and GFP fluorescence intensity in living *E. coli* inserted with the repressilator plasmid. *See* ref. 32 for details. Figure adapted from [32]

The resultant "mutated" or "synthetic" bacteria culture reproduced the predicted oscillatory dynamics of the model (Fig. 9b).

This work provided the first known documented evidence that a simple ordinary differential equations based model could be used to design artificially fused cells, whose function can be controlled in silico prior to actual experimentations. Thus, complex behavior of livings cells may be guided by simple rules. However, it is important to note that livings cells are not homogenous in their behavior, or identical in their molecular constituents. The effect of heterogeneity gives rise to variability in cell to cell behavior (*see* Subheading 6).

3.5 Bistable Response

Biological switches play an essential part in a living system's functionality. There are times when certain genes are turned on or off periodically to develop a functional response for survivability or adaptability. For example, the survival pathways of immune cells are initially activated to fight invading pathogens, and are, subsequently, cleared through the switching to the apoptotic pathway to prevent chronic inflammation [4].

Collins and colleagues [33] constructed a simple genetic toggle switch in *E. coli* to function as a controllable memory unit (Fig. 10a), with the aid of a simple differential equation model, similar to the Goodwin model with Hill coefficient:

$$\frac{du}{dt} = \frac{\alpha_1}{1 + v^{\beta}} - u \tag{26}$$

$$\frac{dv}{dt} = \frac{\alpha_2}{1 + u^{\gamma}} - v \tag{27}$$

where u and v represent the concentration, and α_1 and α_2 are effective synthesis rates, of repressors 1 and 2, respectively. β and γ are the cooperativity of repression of promotors 1 and 2, respectively (Fig. 10a).

In this system, two stable states are possible when inducers 1 and 2 are absent, (1) a "low" state where repressor 2 is transcribed when repressor 1 is repressed and, (2) a "high" state where repressor 1 is transcribed when repressor 2 is repressed (Fig. 10b).

In a stable position, controlling inducer concentration for the repressed repressor causes total transcription till the active repressor is repressed. Thus, control of inducers in the theoretical model allows switching between two stable states.

Using the knowledge obtained from their simple model, Collins and colleagues subsequently constructed two classes of toggle switch plasmids, pTAK class (thermal inducer) and pIKE class (chemical inducer), in *E. coli* and demonstrated bistability in the expression of the tagged green fluorescent protein or GFP (*see* Fig. 4 of ref. 33).

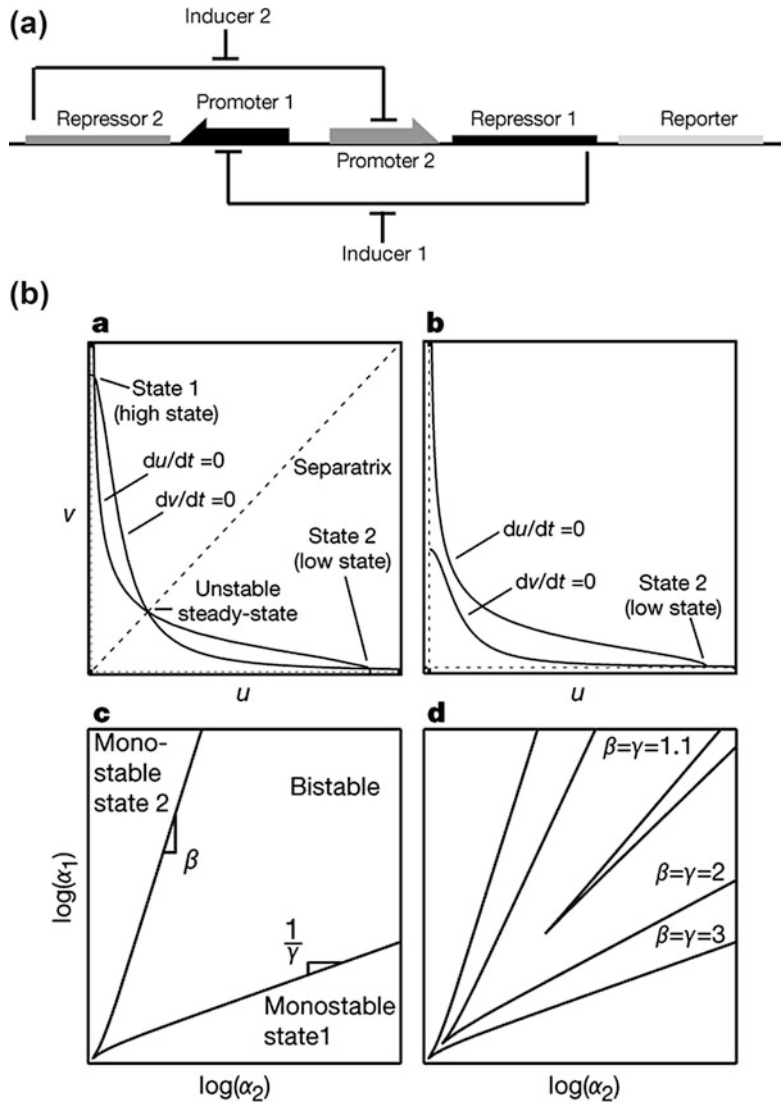


Fig. 10 A simple genetic toggle switch in *E. coli*. (**a**) Schematic of genetic toggle switch comprising inducers, repressors, and promotors, (**b**) Simulations for different ranges of parameter values in equations 26 and 27. *See* ref. 33 for details. Figure adapted from [33]

4 Stability Consideration

Most often, biological network models that attempt to understand any particular cellular process require knowing whether an equilibrium state, such as a steady-state condition or a periodic oscillation, is stable. This is because stable and unstable equilibria can play different roles in biology.

In the case of immune response, it is necessary that the host invoke a response that will neutralize the invading pathogens

deterministically and stably. Otherwise, the host will lose its battle against attacks and their species will not survive for many generations. In these situations, stable models (like the ones mentioned in Subheading 2.3) can be used to understand the response patterns. For cell differentiation to occur, on the other hand, a cell needs to move away from its initial equilibrium state to a new equilibrium state pertaining to the differentiated cell. If the initial equilibrium state is "too" stable, it will be difficult to change the cell fate. Thus, a cell has to move from a stable to an unstable equilibrium state for the induction of cell differentiation. For example, reprogramming the key transcription or Yamanaka factors allows a differentiated cell to dedifferentiate into an inducible pluripotent cell [34].

Similarly, to treat major diseases like cancer or diabetes, the therapeutic intervention aims to change the equilibrium from an "unhealthy" to a "healthy" state. However, the new equilibrium state should be stable; otherwise, the treatment will not be successful and the disease symptoms will persist. Therefore, depending on the situation, it is necessary to investigate and classify equilibrium based on stability. Thus, stability analysis can be an important aspect of biological network modeling.

Linear models, such as those that are made up of first-order mass-action or M-M kinetics, are always stable as long as their parameter values are real and positive. For oscillatory or nonlinear response, as we have seen in the Brusselator and coupled Goodwin examples, the equilibrium states can vary and can become unstable depending on the parameter values. Stability analysis, involving linearization and calculating the eigenvalues of Jacobian matrices, can be performed to check when a nonlinear model will be stable at any particular time or for a range of parameter values.

Let us consider again the linear (first-order) mass-action chain reactions depicted in Eqs. 9 to 11. In the Jacobian form, they can be written as

$$\frac{d\boldsymbol{X}}{dt} = J\delta\boldsymbol{X} \tag{28}$$

or, in Matrix form

$$\frac{d}{dt}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & -k_3 \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \tag{29}$$

To determine the stability of a system, the eigenvalues (λs) of the Jacobian matrix are evaluated. If all λs are real and negative, the reactions will reach a stable node (steady-state level) for each species; otherwise, the system can follow a stable focus or become unstable. Table 1 summarizes the conditions for different eigenvalue solutions.

Table 1
Stability criteria

Eigenvalues	Evaluated points	Type
Real and negative	Stable nodes	Convergence at all conditions
Real and positive	Unstable nodes	Divergence at all conditions
Mixed positive and negative real parts	Saddle points	Mixed or asymptotic conditions
Complex numbers with negative real parts	Stable focus	Convergence at specific conditions
Complex numbers with positive real parts	Unstable focus	Divergence at specific conditions

Determinant, i.e., det $(J\text{-}\lambda I)$, of Eq. 29, and setting it to zero to solve for λs:

$$\det\begin{bmatrix} -k_1-\lambda & 0 & 0 \\ k_1 & -k_2-\lambda & 0 \\ 0 & k_2 & -k_3-\lambda \end{bmatrix} = 0$$
$$\Rightarrow (-k_1-\lambda)(-k_2-\lambda)(-k_3-\lambda) = 0$$
$$\therefore \lambda = -k_1, -k_2, -k_3$$

We note that in biochemical reactions, the rates of reactions (k values) are never negative numbers. Thus, λs for the above condition are all negative indicating stable nodes. It can be shown that networks of any complex configurations, connected by first-order mass-action reactions, are highly stable if k values are non-negative as is for biological systems.

For systems where nonlinear differential equations are used to represent the dynamics, such as the Brusselator, the equations are first linearized using techniques such as Taylor series. Next, stability is analyzed at specific fixed points around a known equilibrium point. To illustrate, let us refer back to the generalized kinetic evolution equation in Sect. 2.3. Applying Taylor series to Eq. 7 at $\boldsymbol{X} = \boldsymbol{a}$, where $\boldsymbol{a}$ is an equilibrium point:

$$\frac{\partial \boldsymbol{X}}{\partial t} = \frac{\partial \boldsymbol{F}(\boldsymbol{X})}{\partial \boldsymbol{X}}\bigg|_{X=a} \delta\boldsymbol{X} + \frac{\partial \boldsymbol{F}^2(\boldsymbol{X})}{\partial \boldsymbol{X}^2}\bigg|_{X=a} \delta\boldsymbol{X}^2 + \frac{\partial \boldsymbol{F}^3(\boldsymbol{X})}{\partial \boldsymbol{X}^3}\bigg|_{X=a} \delta\boldsymbol{X}^3 + \ldots \quad (30)$$

where $\delta\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{a}$ is a small displacement away from the known equilibrium point at which stability is to be evaluated. Note that $\boldsymbol{F}(\boldsymbol{a}) = 0$, by definition. Since $\delta\boldsymbol{X}$ is usually small, higher order terms $\delta\boldsymbol{X}^2$, $\delta\boldsymbol{X}^3$, etc., become negligible and often ignored, leaving only the first-order term.

Considering the species vector $\boldsymbol{X}(= X_1, X_2, \ldots, X_n)$, and rate of reaction vector $\boldsymbol{F}(= F_1, F_2, \ldots, F_n)$, the Jacobian is

$$J = \begin{bmatrix} \frac{\partial F_1}{\partial X_1} & \cdots & \frac{\partial F_1}{\partial X_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial X_1} & \cdots & \frac{\partial F_n}{\partial X_n} \end{bmatrix}$$

Solving the determinant of this Jacobian will provide the linear stability of a nonlinear system near an equilibrium point.

Consider the Brusselator again. For simplicity, all rate constants are set to 1. Eqs. 16 and 17 become

$$\frac{dX_1}{dt} = A + [X_1]^2[X_2] - (1 + B)[X_1] \tag{31}$$

$$\frac{dX_2}{dt} = B[X_1] - [X_1]^2[X_2] \tag{32}$$

The Brusselator has an equilibrium point at $X_1 = A$ and $X_2 = B/A$ when Eqs. 31 and 32 are set to zero and solved. Its Jacobian at this point is therefore

$$J = \begin{bmatrix} B - 1 & A^2 \\ -B & -A^2 \end{bmatrix} \tag{33}$$

Solving the determinant of Eq. 33 reveals

$$\lambda = \frac{(B - A - 1) \pm \sqrt{(B - A - 1)^2 - 4A}}{2} \tag{34}$$

Because A is always positive, it can be shown that the Brusselator is stable when $B < A^2 + 1$ and outside this regime, instability or Hopf bifurcation can be achieved.

The phase-space plots are a simple and powerful way to observe stability of a nonlinear system. They show all possible states of a system and by observing the focus, the stability of the system can be observed. To illustrate, by tracking the time trajectories of X_1 and X_2, we can deduce the type of stability for a range of parameter values. For example, by choosing different values of A and B, we can achieve steady state or lose stability leading to oscillatory patterns (Fig. 11).

In a similar fashion, we can also determine the stability focus of the Goodwin model at the equilibrium point. Figure 12 shows the phase-space plots of the Goodwin model for a range of parameter values. From the Brusselator and Goodwin examples, understanding nonlinear dynamics requires precisely tuned parameter values, as any small variations can lead to drastic changes in dynamics or stability.

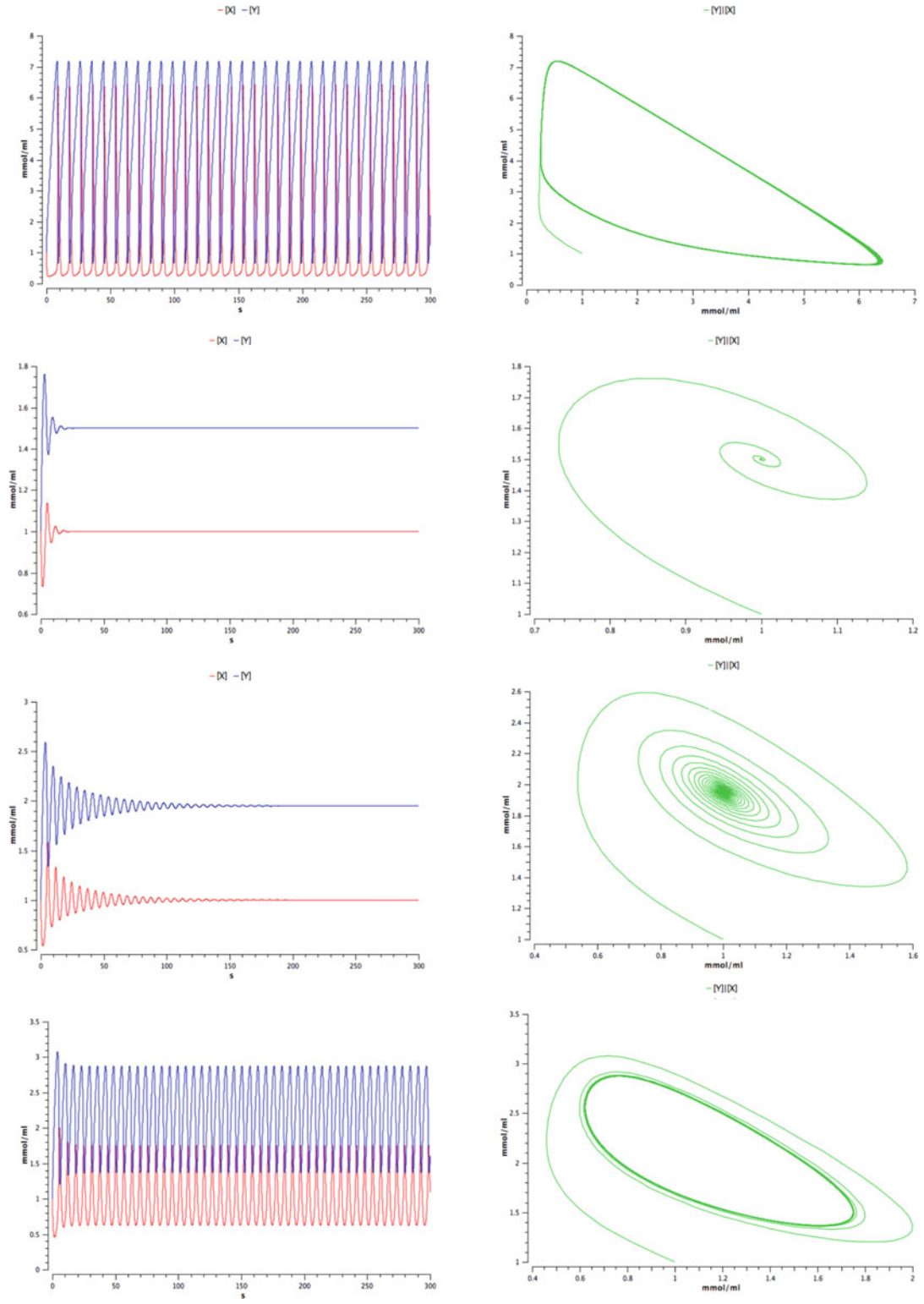


Fig. 11 Brusselator simulations and phase-space plots of model, (**a**) in Fig. 7a, (**b**) in Fig.7b, (**c**) with $B = 1.95$, (**d**) with $B = 2.2$. All other rate constants remain equal to 1.0. *x*-axis represents time and *y*-axis represents concentration in arbitrary units for *left panels*, and *x*- and *y*-axes represent concentration (phase-space plots) in arbitrary units for the *right panels*

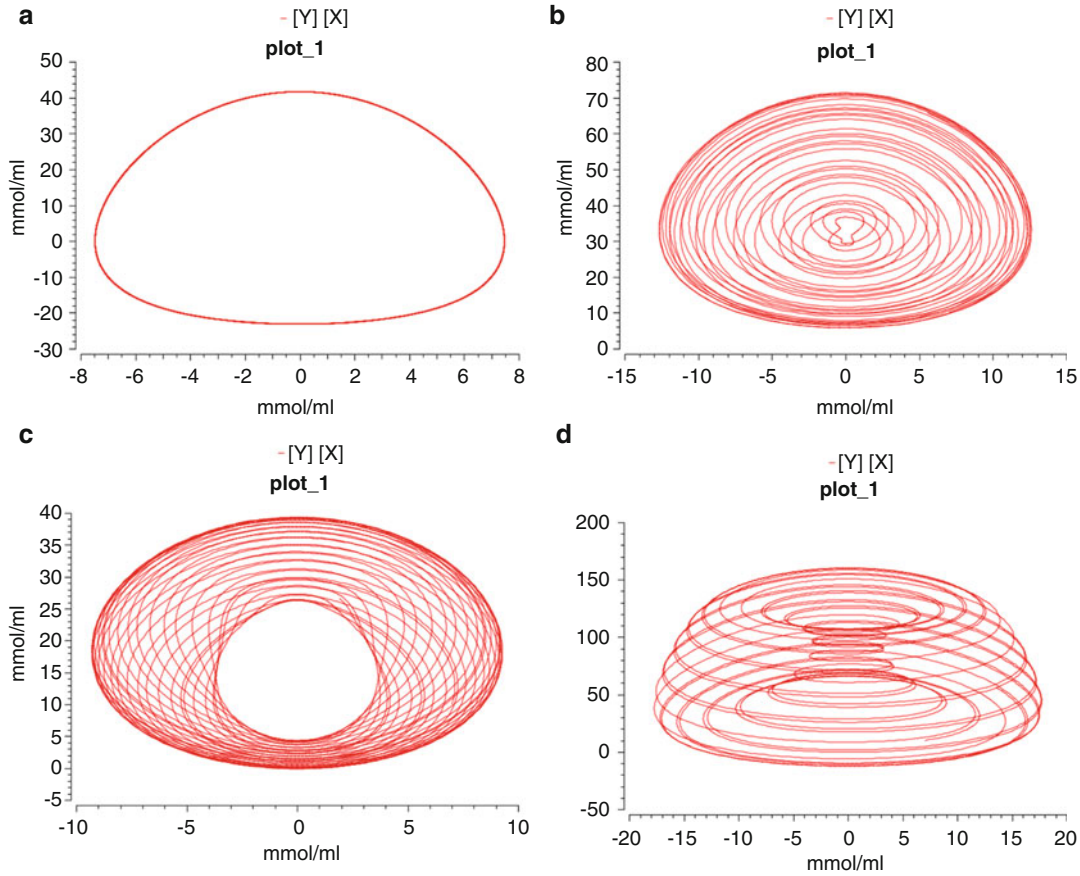


Fig. 12 Goodwin phase-space plots from model with parameters of (**a**) Fig. 8a, (**b**) Fig. 8b, (**c**) Fig. 8c, (**d**) Fig. 8d. *x*- and *y*-axes represent concentration in arbitrary units

5 Reaction-Diffusion Models

So far, we have considered only temporal biochemical reaction systems that can generate complex dynamics. In certain situations, spatial effects are also important for investigation, such as for the understanding of self-organized pattern formation during the developmental process. For example, how does a butterfly development produce beautiful symmetric shapes on its wings, the spots derived in leopard skins? Naturally, the analysis of spatial properties should carry the spatial coordinates.

The best-known theory used to study self-organized pattern formation in biology is the Alan Turing's reaction-diffusion equations [2]. It consists of two coupled reacting species:

$$\frac{\partial X_1}{\partial t} = r_1(X_1, X_2) + D_1 \nabla^2 X_1 \tag{35}$$

$$\frac{\partial X_2}{\partial t} = r_2(X_1, X_2) + D_2 \nabla^2 X_2 \tag{36}$$

where r_1 and r_2 are reaction terms, and D_1 and D_2 are diffusion coefficients of activator-repressor species X_1 and X_2, respectively.

The diffusion terms are key for a biochemical system, in far from equilibrium conditions, to undergo symmetry breaking and form macroscopic stationary patterns. Such pattern formations are often referred to as the Turing patterns, named after the scientist who was first known to have developed it in 1952 [2]. Subsequently, there have been several types of activator-repressor reaction terms that have been developed to model various spatial patterns, according to the type of patterns (Fig. 13).

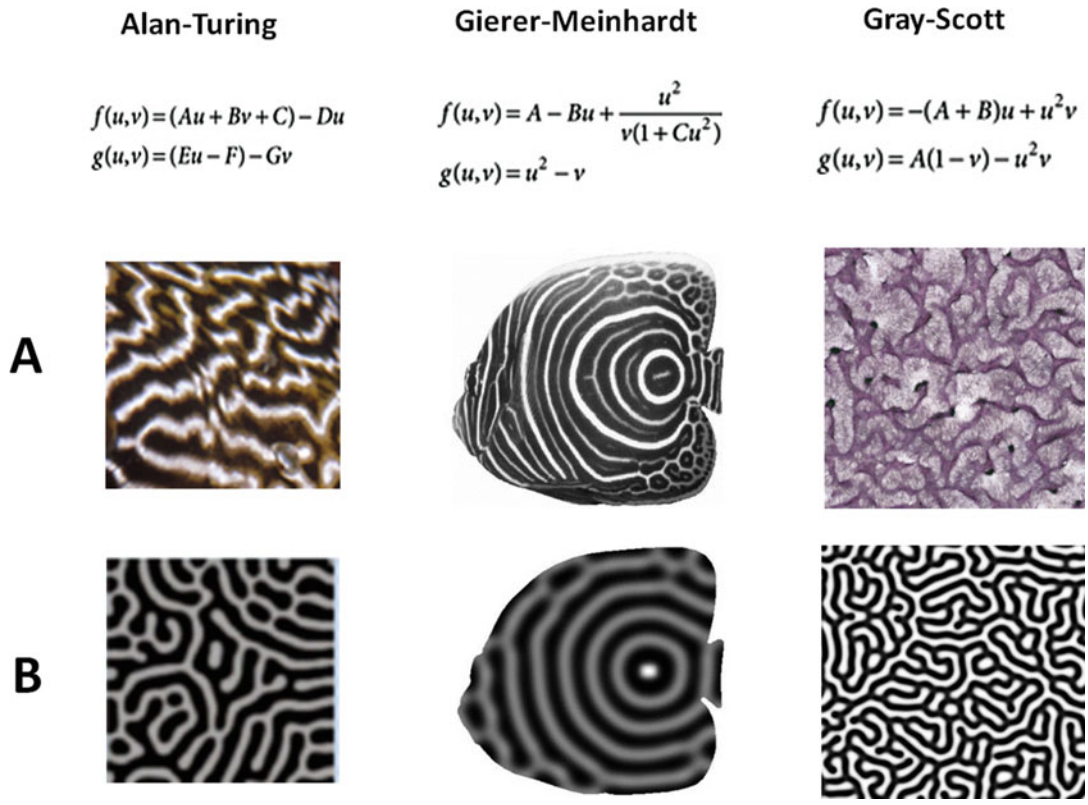


Fig. 13 Spatio-temporal patterns using reaction-diffusion equations that mimic life patterns. (**a**) Simulations from Turing [2], Gierer-Meinhardt [35], and Gray-Scott models, (**b**) actual experimental patterns found on fish skin and vascular mesenchymal cells. Figures adapted from [36, 37]

Turing patterns have also been generated using the Brusselator equations for the reaction terms [38]. Thus, depending on the form of the activator-repressor reaction-diffusion system used, diverse 2-D pattern formations can be achieved. However, the spontaneous pattern formations are due to the instability brought in, mainly, by the diffusion terms rather than the reaction terms. In other words, stable heterogeneous patterns or states arise from a homogenous field due to the instability caused when the diffusive terms are very different between the two reacting species.

6 Stochasticity and Heterogeneity

The population-wide averaging approaches, discussed so far, have been instrumental in our basic understanding of myriad deterministic biological processes such as growth, metabolism, cell signaling, and diseases. For developmental biology, on the other hand, the major challenge has been to understand how multimodal decisions are undertaken. For instance, how a single stem or progenitor cell can produce distinct lineages, which can be tilted even by small external perturbations? Also, it is intriguing how genetically identical cells can produce diverse phenotypes during cell cycle, aging, and epigenetic regulation [39]. The cooperative behavior of microorganisms, such as *Escherichia coli* and yeast, to form biofilms that enhance their survival capacity to environmental threat, is distinct from their individual activity. These observations on phenotypic diversity or individual to cooperative response cannot lend itself to population-based readout as multiple measurements of single cells across time are required in order to unravel the multifaceted decision capability of the living system.

It is now known that the single-cell heterogeneity within cell populations, measured by transcription, phosphorylation, morphology, and motility, arises from a combination of intrinsic and extrinsic elements. Stochasticity in gene or protein expression is a result of two sources of biological noise: (1) intrinsic or "uncorrelated" noise; the random nature of biochemical reactions, e.g., due to low copy numbers of intracellular molecules in a Poisson process, and (2) extrinsic or "correlated" noise; fluctuations in other cellular components or states that indirectly affect the expression of a specific gene or protein [40, 41]. However, stochasticity in mRNA and variability in protein expression are not simply due to the effect of low copy numbers on a Poisson gene regulatory process, but can also be due to the quantal or bursting nature of promoter activity (Fig. 14) [42, 43]. Moreover, by varying the rates

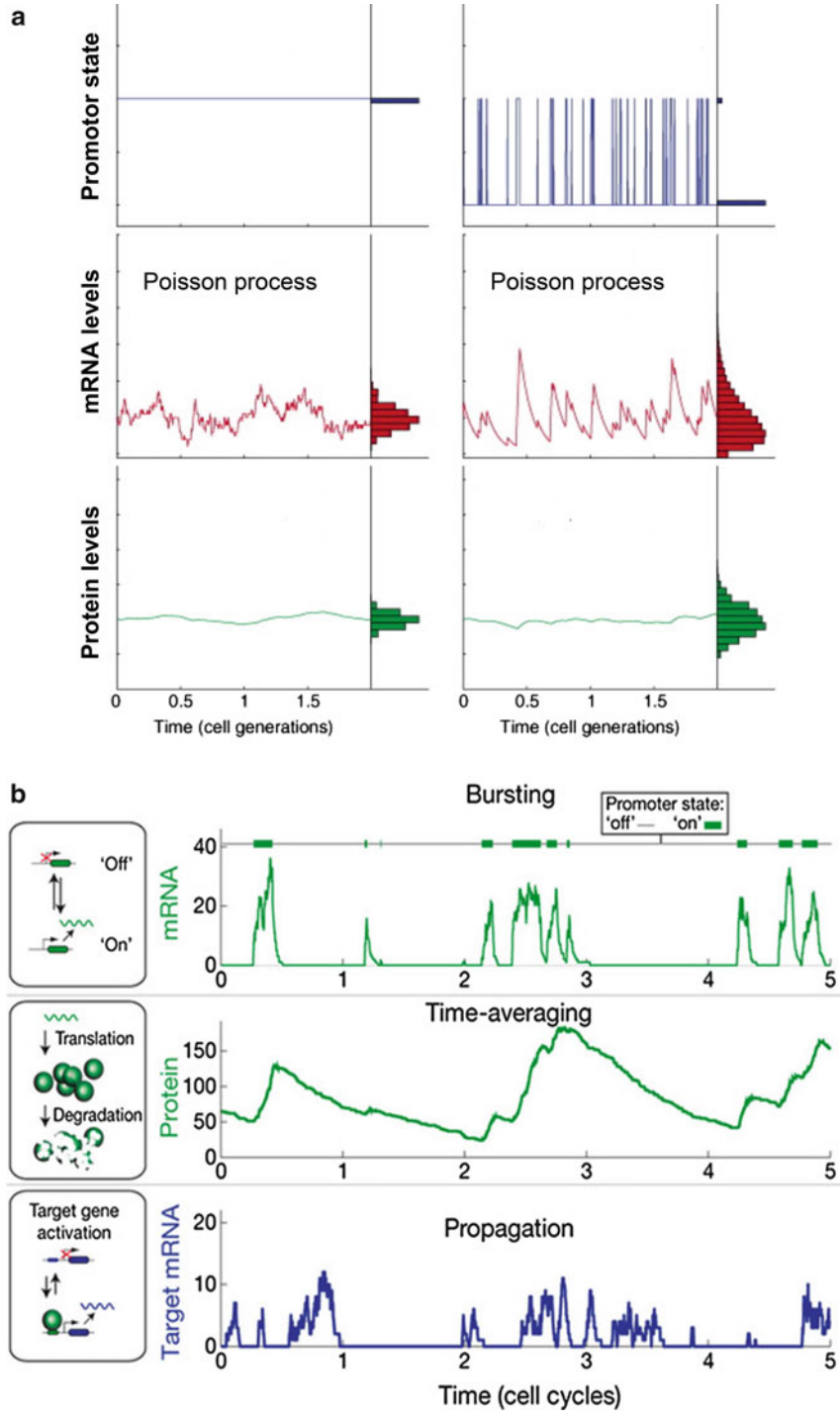


Fig. 14 Stochastic and bursting gene (mRNA) expression produces variable protein expressions. Figure adapted from [43]

of transcription and translation of a bacterial protein, it is now known that increasing transcriptional, and not translational, noise is responsible for the variability in reporter protein expressions [42, 44].

In *Saccharomyces cerevisiae*, the control of the transcription rate of GAL contributed to the heterogeneity of reporter yEGFP protein expression within a clonal population [45]. Moreover, it was shown that increasing the transcriptional noise propagation in the corresponding gene regulatory network resulted in the generation of bistable expression states of yEGFP. This pioneering work on actual cells is a good example of how the control of noise can non-intuitively regulate the diversity of molecular constituents in living systems.

In a more recent study, to understand global gene expression noise patterns of single cells during mammalian developmental stages, transcriptome-wide RNA-Seq expressions of oocytes to blastocysts were investigated using high-dimensional statistical techniques (correlation metrics, Shannon entropy, and square of coefficient of variation) [46]. Notably, gene expression variability increased sharply from 2-cell to 4-cell stage onward in both human and mouse (Fig. 15a). In addition, a phase transition in noise (square of coefficient of variation) patterns occurred between 2-cell and 8-cell stages (Fig. 15b).

Subsequently, a stochastic transcriptional model (based on deterministic ordinary differential equations with random processes) was developed and fitted the model to experimental noise patterns (Fig. 15c). From the simulation results, it was concurred that the early developmental stages were mainly dominated by low transcriptional activity dominated by Poisson noise. The increase in transcriptome-wide noise for the middle stage developmental cells was due to stochastic transcriptional amplification, which generated heterogeneity in gene expressions between individual cells. Such heterogeneity has been shown to be necessary for cell fate diversifications (*see* review in ref. 39). For the later stages, on top of a high transcriptional process, the cells possess quantal activation of most transcription factors, or are subject to more extrinsic variability such as phenotypic diversity among individual cells. These factors increase the general expression scatter and noise levels. Overall, the investigations into the transcriptome-wide expressions of the early mammalian developmental stages revealed increasing variability and noise patterns across the mammalian development process. This result may support the notion in chaos theory [39], where increasing noise generated along the embryonic development process (between 2 to 4-cell stage) may aid in generating a noisy landscape for multi-lineage cell differentiation to proceed through an underlying (still unknown) chaotic mechanism.

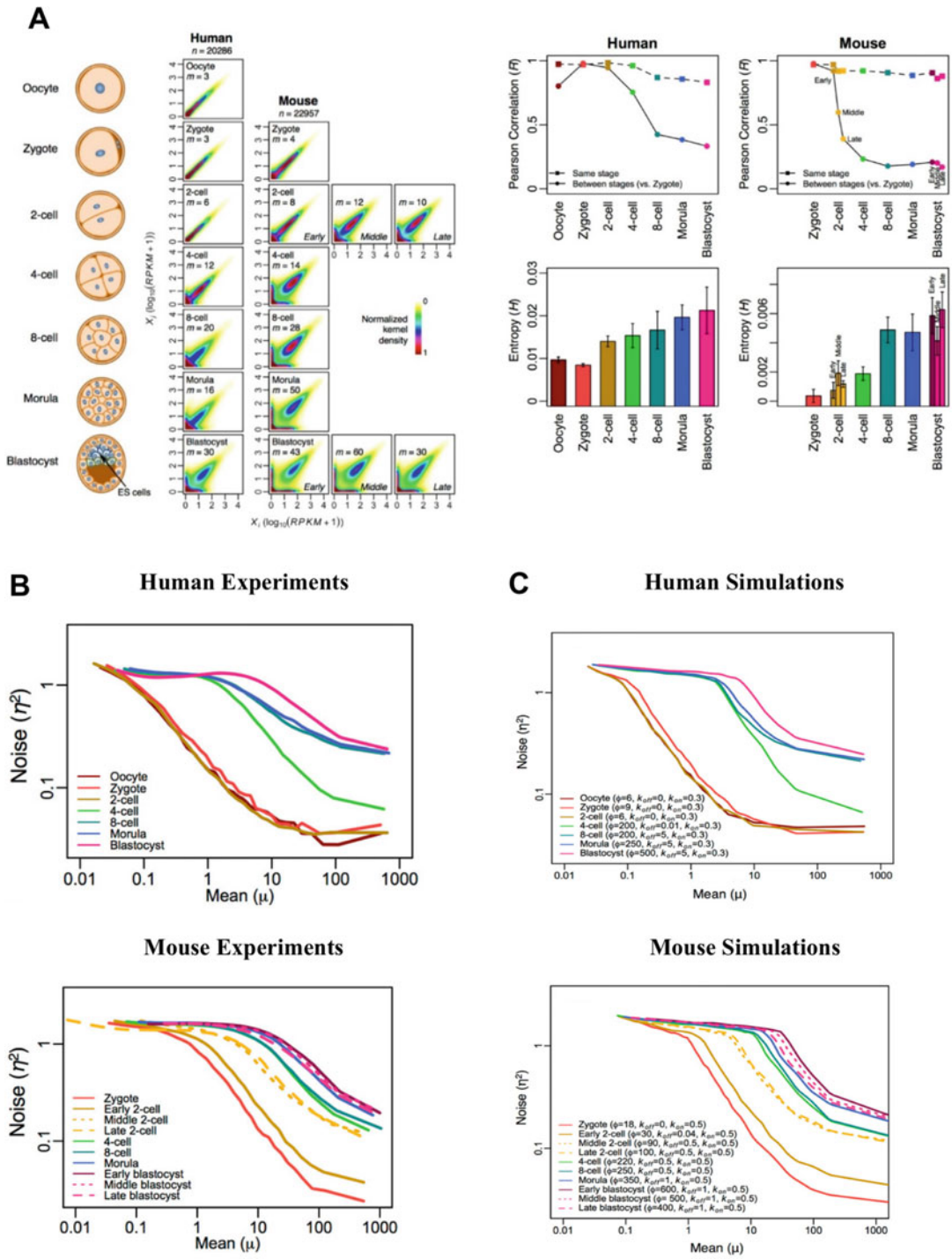


Fig. 15 (**a**) Gene expression distribution between two single cells (*left*), Pearson correlations (*top right*), and Shannon entropy (*bottom right*) of development cells from oocytes to blastocysts in human and mouse. (**b**) Experimental and (**c**) Simulated noise (η^2) versus mean (μ) expression patterns for each development stage in human (*top*) and mouse (*bottom*). Figure reproduced from [46]

7 Summary

It has been discussed for a long time whether living systems can be mathematically conceptualized using simple theories as they possess very complex dynamic and emergent behaviors, and many times display unpredictable outcomes [47]. In this chapter, we have looked at several complex response dynamics of living cells, and have shown simple biochemical models, based on linear and nonlinear differential equations, which can be used to successfully understand or interpret the data. For linear models, the reaction topology rather than kinetics plays crucial and sensitive roles [4, 19]. For nonlinear dynamics, the parameters need to be precise or the response cannot be accurately determined due to the stability issue. For single-cell response, stochastic modeling can be useful in understanding the diversifying cell fates or heterogeneous response. We, therefore, believe mathematical models will continue to play significant roles in unlocking further secrets on the complex behaviors of living systems.

References

1. Goldman AW, Burmeister Y, Cesnulevicius K, Herbert M, Kane M, Lescheid D, McCaffrey T, Schultz M, Seilheimer B, Smit A, St Laurent G III, Berman B (2015) Bioregulatory systems medicine: an innovative approach to integrating the science of molecular networks, inflammation, and systems biology with the patient's autoregulatory capacity? Front Physiol 6:225
2. Turing AM (1952) The chemical basis of morphogenesis. Philos Trans R Soc B 237:37–72
3. Selvarajoo K (2011) Macroscopic law of conservation revealed in the population dynamics of Toll-like receptor signaling. Cell Commun Signal 9:9
4. Selvarajoo K (2013) Immuno systems biology: a macroscopic approach for immune cell signaling. Springer, New York
5. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP (2007) Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol 3:1871–1878
6. Bakker BM, Michels PA, Opperdoes FR, Westerhoff HV (1997) Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes. J Biol Chem 272:3207–3215
7. Guldberg CM, Waage P (1864) Studies concerning affinity, C. M. Forhandlinger: Videnskabs-Selskabet i Christiana, 35
8. Leskovac V (2003) Comprehensive enzyme kinetics. Kluwer Academic/Plenum Pub, New York
9. Bujara M, Schümperli M, Pellaux R, Heinemann M, Panke S (2011) Optimization of a blueprint for in vitro glycolysis by metabolic real-time analysis. Nat Chem Biol 7:271–277
10. Blagoev B, Ong SE, Kratchmarova I, Mann M (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nat Biotechnol 22:1139–1145
11. Helmy M, Gohda J, Inoue J, Tomita M, Tsuchiya M, Selvarajoo K (2009) Predicting novel features of toll-like receptor 3 signaling in macrophages. PLoS One 4:e4661
12. Selvarajoo K, Takada Y, Gohda J, Helmy M, Akira S, Tomita M, Tsuchiya M, Inoue J, Matsuo K (2008) Signaling flux redistribution at toll-like receptor pathway junctions. PLoS One 3:e3430
13. Selvarajoo K (2006) Discovering differential activation machinery of the Toll-like receptor (TLR) 4 signaling pathways in Myd88 knockouts. FEBS Lett 580:1457–1464
14. Hayashi K, Piras V, Tabata S, Tomita M, Selvarajoo K (2013) A systems biology approach to suppress TNF-induced proinflammatory gene expressions. Cell Commun Signal 11:84

15. Piras V, Hayashi K, Tomita M, Selvarajoo K (2011) Enhancing apoptosis in TRAIL-resistant cancer cells using fundamental response rules. Sci Rep 1:144
16. Hayashi K, Tabata S, Piras V, Tomita M, Selvarajoo K (2015) Systems biology strategy reveals PKCδ is key for sensitizing TRAIL-resistant human fibrosarcoma. Front Immunol 5:659
17. Selvarajoo K (2017) A systems biology approach to overcome TRAIL resistance in cancer treatment. Prog Biophys Mol Biol 128:142–154
18. Selvarajoo K, Tomita M, Tsuchiya M (2009) Can complex cellular processes be governed by simple linear rules? J Bioinformatics Comp Biol 7:243–268
19. Selvarajoo K (2014) Parameter-less approaches for interpreting dynamic cellular response. J Biol Eng 8:23
20. Kaufmann S (1995) At home in the universe: the search for laws of self-organization and complexity. Oxford University Press, New York
21. Chatterjee A, Cook LC, Shu CC, Chen Y, Manias DA, Ramkrishna D et al (2013) Antagonistic self-sensing and mate-sensing signaling controls antibiotic-resistance transfer. Proc Natl Acad Sci U S A 110:7086–7090
22. Poulsen AK, Lauritsen FR, Olsen LF (2004) Sustained glycolytic oscillations—no need for cyanide. FEMS Microbiol Lett 236:261–266
23. Prigogine I (1997) The end of certainty. The Free Press, New York
24. Field RJ, Korös E, Noyes RM (1972) Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the bromate-cerium-malonic acid system. J Am Chem Soc 94:8649–8664
25. Epstein IR (2006) Predicting complex biology with simple chemistry. Proc Natl Acad Sci U S A 103:15727–15728
26. Richard JF, Noyes RM (1974) Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. J Chem Phys 60:1877–1884
27. Goodwin BC (1965) Oscillatory behaviour in enzymatic control processes. Adv Enz Reg 3:425–428
28. Ruoff P, Vinsjevik M, Monnerjahn C, Rensing L (1999) The Goodwin oscillator: on the importance of degradation reactions in the circadian clock. J Biol Rhythm 14:469–479
29. François P, Despierre N, Siggia E (2012) Adaptive temperature compensation in circadian oscillations. PLoS Comput Biol 8:e1002585
30. Zeiser S, Muller J, Liebscher V (2007) Modeling the Hes1 oscillator. J Comput Biol 14:984–1000
31. Griffith JS (1968) Mathematics of cellular control processes. I. Negative feedback to one gene. J Theor Biol 20:202–208
32. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–338
33. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403:339–342
34. Takahashi K, Yamanaka S (2016) A decade of transcription factor-mediated reprogramming to pluripotency. Nat Rev Mol Cell Biol 17:183–193
35. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. Kybernetik 12:30–39
36. http://www.uoguelph.ca/~mgarvie/turing.html
37. Chen TH, Guo C, Zhao X, Yao Y, Boström KI, Wong MN, Tintut Y, Demer LL, Ho CM, Garfinkel A (2012) Patterns of periodic holes created by increased cell motility. Interface Focus 2:457–464
38. Peña B, Pérez-García C (2001) Stability of Turing patterns in the Brusselator model. Phys Rev E Stat Nonlinear Soft Matter Phys 64:056213
39. Selvarajoo K (2012) Understanding multimodal biological decisions from single cell and population dynamics. Wiley Interdiscip Rev Syst Biol Med 4:385–399
40. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297:1183–1186
41. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci U S A 99:12795–12800
42. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6:451–464
43. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467:167–173
44. Dunlop MJ, Cox RS III, Levine JH, Murray RM, Elowitz MB (2008) Regulatory activity revealed by dynamic correlations in gene expression noise. Nat Genet 40:1493–1498
45. Blake WJ, KAErn M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. Nature 422:633–637
46. Piras V, Tomita M, Selvarajoo K (2014) Transcriptome-wide variability in single embryonic development cells. Sci Rep 4:7137
47. Bizzarri M, Palombo A, Cucina A (2013) Theoretical aspects of systems biology. Prog Biophys Mol Biol 112:33–43

Chapter 10

Systems Biology Modeling of Nonlinear Cancer Dynamics

Christian Cherubini, Simonetta Filippi, and Alessandro Loppini

Abstract

Systems Biology represents nowadays a promising standard framework for natural and human sciences to attack complicated problems involving Life. Here a particular application of such a program is discussed in the case of Cancer, by using a basic toy model for solid tumor spread for framing together two apparently different conceptual leading paradigms of Oncogenesis.

Key words Cancer, Systems biology, Computational biology, Mathematical modeling

1 Introduction

A quantitative description of Morphogenesis, i.e., the dynamical process realizing development in living beings, is nowadays becoming a central target not only for areas as Biology, Medicine, and Chemistry but also for Biophysics, Biomathematics, and Bioengineering.

Dating back from Darcy Thompson's 1917 classic monograph "On Growth and Form" [1] in fact a description of living beings in geometrical (or in wider sense mathematical) terms has appeared to be not only physically sound but also unavoidable. More recently, systemic studies on plant morphogenesis (phyllotaxis) (*see* ref. 2 for a review) have shown elegant cross bridges of group theory, geometry, and number theory used to explain the elegant structures of plants, in particular about seeds disposition and flower patterns. Also, some populations of simple living systems as fungi and amoebae form colonies whose aggregation shows complicated structures with very high levels of complexity both in time and in space. Similar patterns, often associated with spiral geometries, also appear in many biological contexts as cardiac and neural dynamics for instance [3, 4]. Strikingly, many of the phenomenologies above depicted do not belong to the realm of Life only but are also manifested by unanimated natural systems as oscillating chemical reactions, atmospheric eddies, or growing crystals for instance

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_10,

[5, 6]. The lowest common denominator of all of these systems is that they all belong to the realm of non-equilibrium thermodynamical processes [7] subject to complex bifurcations characterizing their dissipative dynamics. During several decades, many scientists have looked for the best physical and mathematical tools to adopt to describe these dynamics appropriately. All of these systems have, depending on the spatiotemporal scale involved, a discrete structure starting from atoms up to molecules, cells, cells aggregates, and even complex living beings populations. In this zooming-out tale, a quantum description rapidly seems to fade out for a classical and sometimes coherent [8, 9] one, possibly complicated by a stochastic flavor which however can still be seen as an echo of the underlying smaller scales quantum probabilistic dynamics. Such a large-scale description can be quite successfully obtained by using ordinary or partial differential equations (the latter seen as a continuum limit for discrete systems). Variations on the theme as stochastic or delay differential equations, as well as integrodifferential mathematical descriptions or maps and cellular automata, are possible (although less popular) options that can be successfully used too (for a complete review on some of these mathematical tools, we refer to refs. 10, 11). Partial differential equations of Reaction-Diffusion (RD) class [12] have a long-lasting and fruitful history in treating non-equilibrium phenomena in chemistry and biology.

The pioneer work for using RD equations in Biology is described in the 1952 classic article "The Chemical Basis of Morphogenesis" by Alan Turing [13]. Here, by using as actors some chemicals, known as morphogens (equivalently described as activators vs. inhibitors), a possible "toy model" explanation for animal coat patterns appearance as well as developmental processes as Hydra tentacles, for instance, is discussed. Turing article has been widely recognized as the source for the interpretative paradigm which has driven plenty of quantitative biology research in the past 65 years.

However, another 1952 article would have shared with Turing's the title of one of the most influential articles in quantitative biology, i.e., the classical Hodgkin and Huxley study (Nobel Prize awarded) on action potential propagation in nerves [14]. This work combined both experiments and equations (of reaction-diffusion class) to give quantitative predictions on the bioelectrical behavior of a squid giant axon.

The wise reader would be at this stage tempted to ask at which level and to which extent such a quantitative description of living beings meets all the decades (in some cases centuries) of studies in Biology and Physiology, Chemistry, and Biochemistry as well as in Medicine. Stated—say—40 years ago, such a question would have received an "almost none" answer. The investigation tools and the methodologies of all of these branches of Science were at that time

too different, and different was the technical language that scientists were using to describe the same phenomenon. In short, some decades ago Biophysics (and Biomathematics even more) was to be seen as a niche in the broad realm of Biology. However, the situations started to change when mathematical modeling appeared to be an extremely useful tool in the successfully understanding of complicated underlying dynamics as in viruses for instance, especially in the case of patients' HIV blood testing eventually leading to an AIDS diagnosis [15]. Biophysics, Biomathematics and all the other Biological Sciences needed to meet and possibly to merge somewhere. Time was ready for the Systems Biology revolution whose manifesto could be recognized to be Denis Noble's "The Music of Life" book [16].

The aim of Systems Biology is, by taking full advantage of the quite recent impressive boost in performance of computers and scientific instrumentation, to integrate different datasets from experiments with several mathematical biophysical descriptions of the same biological system. In particular, in a zoom-in and zoom-out perspective, it aims to predict, control, and possibly explain Life through an "in silico" living being. The paradigm is entirely inclusive for all the Sciences involved, i.e., Biology, Chemistry, Physics, Mathematics, Engineering, Computer Sciences, Medicine and Veterinary up to Economics, Philosophy and even Sociology. A natural question would arise at this stage: which type of biological problems would truly require the involvement of some many different competencies to be addressed and successfully explained and controlled? A natural answer comes immediately in mind: cancer for instance.

Cancer has always constituted a central problem for humankind, although in the past century only two major large-scale campaigns, in Germany in the 1930s [17] and in the US with 1971 National Cancer Act, have been initiated trying to find a cure. This ambitious target was not reached. However, the knowledge acquired became a solid background for all the subsequent studies, in particular for a Cancer Systems Biology formulation [18].

Cancer is believed to begin at cellular level developing somatic mutations which result transferred from a cell to its progeny. In this process, a failure of controls by the immune system is mainly responsible for the tumor development, and for these reasons cancer analyses require proper knowledge of Immunology. Incidentally, similar failures occur also in the case of some Virus infections as HIV for instance, so that in standard Immunology volumes [19] both these pathologies are examined, sometimes highlighting possible similarities in their dynamics.

Cancers are very diversified, although in a simplified way they can be divided into two broad families, i.e., solid and liquid ones and in the rest of this work we shall focus on the former case.

Cancer initiation appears to be a phenomenon which involves some regulatory genes, which can either enhance or inhibit tumor levels of malignancy and whose micro-scale dynamics is regulated by mesoscale and macroscale properties and processes. Indeed, biological tissues are somewhat ordered but complex structures generating forces exchanged by cells in union with extracellular matrix surrounding them. These mechanical interactions are superimposed to biochemical and electrical ones, in some cases even back-reacting dynamically as it happens in cardiac tissue for instance. All of these contributions at the end determine the geometrical features of the tissue and can in many situations define cells' final state.

Such communication pathways at the cellular and tissue levels have an important role in the global tissue organization so that their impairment results in several cases to be associated with cancer generation and progression [20, 21]. In this context, we must point out that specific chemicals known as morphostats are experimentally known [22] to drive "cell to cell" and "tissue to tissue" communications in a way somewhat very similar to Turing's activators and inhibitors activities. These morphostats present dynamically varying concentrations in space and have a great influence on the tissue expressed phenotype, although a clear understanding of their underlying mechanistic dynamics is still not present.

It is natural to look for an appropriate conceptual framework which could analyze and possibly interpret both these phenomenologies. The mainstream nowadays follows as primary candidates two theories, i.e., the Somatic Mutation Theory or SMT [23] and the Tissue Organization Field Theory or TOFT [24, 25]. The former deserves cancer generation to progressive DNA changes within a single cell. This is an approach in which the main source of cancer is the cell itself. On the other hand, in TOFT possible carcinogen factors tend to mismatch cellular communications, in a fashion similar to Turing morphogens' theory.

As standard in hard Sciences, it is mandatory often to accommodate different perspectives based on empirical facts taking advantage of an appropriate mathematical framework and computational resources.

In silico studies do not represent necessarily the starting point for mathematical modeling, however. For instance, books devoted to the study of Computational Biology [26] and Evolutionary dynamics of cancer [27] and Mathematical Biology in general [6, 28] offer different possible analytical approaches to describe cancer dynamics. Together with mathematical tools previously discussed as ordinary and partial differential equations with stochastic, delay and integrodifferential extensions, also more exotic tools as maps and cellular automata or even the recently growing field of complex networks applied to biology can be used to this aim [29–31].

Each of these has its specific advantage and limitations and must be carefully chosen depending on the particular problem that one wants to investigate. Another key ingredient to take into account is the geometry on which the model lives, whether an idealized one or a realistic domain imported in the computer from specific biomedical datasets coming from histological, NMR, CT scans, or others.

Traditional physicists and mathematicians often feel themselves more comfortable with the former case, thinking about cancer for instance as a sphere of a cube so that they can take advantage of the geometrical symmetries to facilitate the study both analytically and numerically. On the other hand engineers and computer scientists are happy with complex situations. This approach takes advantage of the incredible boost in performance of nowadays computers which use multiprocessing technology and powerful graphic processing units (also present in commercial smartphones, video games consoles, smart tv, etc.), which 15 years ago were found in particular academic or industrial contexts only. While the latter situation is more in the spirit of Systems Biology, we have to say that the formerly described use of simplified scenarios (often called toy models) can be extremely useful to orientate and interpret complex numerical simulations.

2 Experimental Research

As an example, in Fig. 1 we show the simulation at a given time of a brain cancer cells growth in a realistic NMR imported geometry.

In this work, we shall focus on a particular subset of these possibilities, i.e. the case of solid cancer dynamics described by a single nonlinear reaction-diffusion equation, mostly on the lines of previous studies by some of the authors [32]. This approach, although minimalist in its toy model nature, will accommodate both points of view of SMT and TOFT conceptually. It moreover represents a simple example of a Systems Biology activity which merges somewhat together the knowledge of mathematics, biophysics and philosophy of science.

Specifically, the proposed model for the tumor cell concentration in space and time $c(t,x,y,z) = c\left(t,\vec{x}\right)$ is governed by the following reaction-diffusion type partial differential equation ($\vec{\nabla}$ is the Laplacian operator, and $R = F(c)$ is the reaction term):

$$\frac{\partial c}{\partial t} + \vec{\nabla} \cdot \vec{J} = R$$

which, by using Fick's law for the matter flux vector $\vec{J} = -D\vec{\nabla} c$, results, in Cartesian coordinates, in:

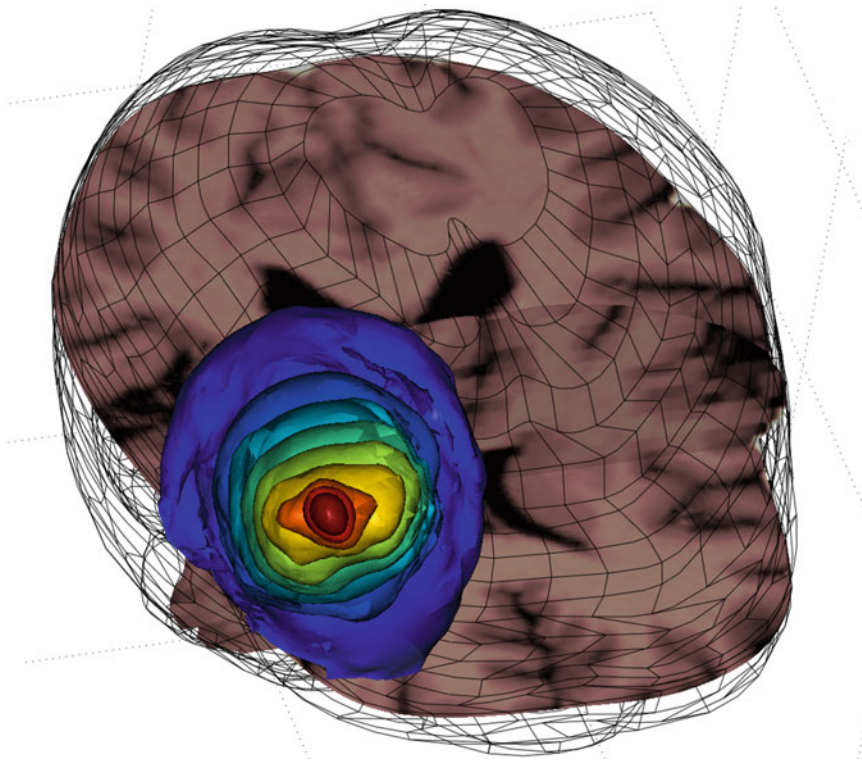


Fig. 1 Finite elements simulation at a given time of a brain cancer cells concentration concentric isosurfaces (in *different colors*) in a realistic NMR imported geometry

$$\frac{\partial c}{\partial t} = F(c) + D\left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2}\right)$$

where D is the (effective) homogeneous and isotropic diffusion coefficient resulting from a suitable smoothing of the heterogeneous character of the tissue. On the other hand $F(c)$ represents the reaction term which accounts for the generation and disappearance of cancer cells. The diffusive part of the equation by including partial derivatives couples each point of tissue to the closer ones (a communication term which is welcome in TOFT). The reaction term instead does not contain derivatives so that it deals with what happens to the point itself only (and internal cancer cell dynamics which is fine for SMT). Both terms in the equation interact to have a space-time dynamics for cancer growth. The proposed form for the function $F(c)$ is the simplest one, i.e. a cubic one:

$$F(c) = k(c - c_1)(c - c_2)(c_3 - c)$$

where c_1 , c_2 , c_3 and k are constants with $0 < c_1 < c_2 < c_3$. In dynamical systems theory, this is a basic way to obtain bistability. In fact the value of species c will settle down asymptotically in time whether to c_1 (in our case this is zero corresponding to no cell

cancer density) or to c_3 (in our case this is non-zero corresponding to appreciable cell cancer density). The quantity c_2 plays a role of a switch in this dynamics. If cancer cells concentration c gets higher than c_2, then the fate of the point of tissue is in principle determined, and a stably dense cancer scenario there arises if diffusion does not carry away cancer cells fast enough. On the other hand, if the value of c is lower, the organism can avoid this pathological situation. It is clear that in this model then c_2 (in union with diffusion which—by construction—lowers cellular population) characterizes the role of the immune system against cancer progression.

3 Results

The model previously described can be cast in a non-dimensional form, a standard procedure in any numerical situation and studied in different dimensions.

In one dimension (1D), the resulting equation is

$$\frac{\partial C}{\partial T} = \frac{\partial^2 C}{\partial X^2} + aC(1 - C)(C - \alpha)$$

where C is non-dimensional concentration, and X and T are non-dimensional space and time respectively. Parameters $a > 0$ and $0 < \alpha < 1$ describe the entire dynamics then. The 1D equation above can be both studied analytically and numerically. In the former case, one can find a traveling wave exact solution for an infinite spatial domain given by

$$C = \frac{1}{2}\left\{1 + \tanh\left[\frac{\sqrt{a}(X + VT)}{2\sqrt{2}}\right]\right\}$$

with constant non-dimensional velocity $V = \sqrt{\frac{a}{2}}(1 - 2\alpha)$. This is a traveling wave that transfers high concentrations of cancer cells in regions that had before almost zero values for this species.

These equations have been numerically investigated in 1D and 2D numerical simulations, showing how a nontrivial initial data for cancer cells leads to a tumor cells cancellation in some regions but finally to a growth in the space of tumor mass.

Similar studies can also be performed in 3D, even in the presence of more complicated situations in which the diffusion coefficient is not a constant but a function of space and time (inhomogeneous diffusion) instead, i.e., $D = D\left(t, \vec{x}\right)$.

This is what can be done for instance in studying tumor diffusion processes in the brain, a standard application of mathematical modeling extremely useful in Neurosurgery [6, 33]. We point out that a more realistic (but complicated) scenario would be the one in which the diffusion coefficient is not a scalar but a matrix (diffusion

tensor) taking into account not only heterogeneity but also anisotropy [34] associated with the biological fibered structures of tissues.

Specifically in this case the symmetric diffusion tensor results into

$$\widehat{D} = \begin{pmatrix} D_{xx}\left(t,\vec{x}\right) & D_{xy}\left(t,\vec{x}\right) & D_{xz}\left(t,\vec{x}\right) \\ D_{yx}\left(t,\vec{x}\right) & D_{yy}\left(t,\vec{x}\right) & D_{yz}\left(t,\vec{x}\right) \\ D_{zx}\left(t,\vec{x}\right) & D_{zy}\left(t,\vec{x}\right) & D_{zz}\left(t,\vec{x}\right) \end{pmatrix}$$

so that the matter flux vector $\vec{J} = -\widehat{D}\mathrm{n}\vec{\nabla}\, c$ is not directed as the concentration gradient anymore. Stated more technically, such a vector is not orthogonal to the surfaces of constant cancer cells density.

In ref. 32 however, a further relevant improvement of the formulation just discussed was presented. Taking advantage of modeling works on animal dispersal, in fact, the authors hypothesized that cancer cells would act as a predator against prey. In ecological models, this effect can be mathematically described in several ways although a significant contribution can be given by changing the diffusive dynamics as follows. If there are many animals in a given region, they tend to eat the food supplies rapidly. In this terminology, we could think about several animal species which eat the same food as grass for instance. Alternatively, we could talk about lions or others eating smaller mammals. In any case, these "eaters," after realizing that food is becoming scarce tend to reach zones where there are fewer competitors and almost settle down there. This process should be a sort of random walk type. When the random walkers are many, one can substitute the dynamics with a diffusive one. Here, however, one must take into account that this is not a conventional diffusive process but an anomalous one, because when eaters are locally many, they want to diffuse away rapidly, while when their population locally reduces, their dynamics is of "walkabout" type, i.e., a little diffusion. In other words, we are saying that the diffusion coefficient is a function of the concentration of the diffusive species itself, i.e., $D\left(c\left(t,\vec{x}\right)\right)$.

This mathematical modification has dramatic consequences because it leads to a porous media type nonlinear partial differential equation [6], which for the sake of simplicity in dimensionless form results schematically in

$$\frac{\partial C}{\partial T} = \sigma^{-m}\vec{\nabla}\,\mathrm{n}\left(C^{m}\vec{\nabla}\, C\right) + F(C).$$

In the limit of vanishing m (here σ is a constant), this process gives the previously discussed reaction-diffusion equation. On the

other hand, a nonzero m creates a nonlinear equation for which the cancer propagation front is non-smooth.

This is an interesting modification which could be tested experimentally both in vitro and in vivo. The best candidate for this type of description, always in the context of Neuro-Oncology, would be glioblastoma multiform cancers for which many data exist both for cultures and for NMR and TC biomedical images leading to a well-developed and dynamic sector of mathematical oncology (*see* ref. 35 and references therein).

4 Notes

In the spirit of Systems Biology, we must take into account now the fact that the presence of tumor masses, for instance in the brain, requires substantial refinements of the mathematical modeling presented above. A growing tumor, in fact, leads to significant mechanical effects on the underlying anatomy. Cancer moves into a limited region, i.e., the skull, and this leads to the displacement of anatomical structures, for instance, the ventricular cavities, which experimentally are well known to collapse resulting in an anomalous hydrodynamics for the brain. Moreover, cancer tends to infiltrate pre-existing fibred tissues and vascularized structures, eventually becoming necrotic and leading to an entirely new (pathological) anatomy which is revealed by subsequent patient's biomedical imaging. The role of the model is crucial here in giving a predictive snapshot on what will occur in patients.

Physiology too of course results affected in this quite complicated developmental dynamics. A Systems Biology perspective must be adopted again by merging these new aspects into a model possibly (but not necessarily) described by reaction-diffusion partial differential equations acting into space and time changing domains ruled both by coupled nonlinear solid mechanics and microfluidics (*see* for instance refs. 36, 37). The modeling could be then still enriched by taking into account the effects of chemotherapy and radiation therapy, ablation, immunological treatments, temperature changes, and many others.

A Systems Biology perspective of such a problem, in conclusion, seems to resemble a sort of constructions toy set. Each model can be expanded and enriched taking advantage of new ingredients coming from advances not only in experiments and mathematics but also in computational resources. This refinement process eventually should lead to updated "in silico" versions of the system giving new insights into such a complex problem. Biological and human sciences in this procedure surely will take advantage of a mathematical language, mostly free of ambiguity, to find new interpretative frameworks for cancer.

References

1. Thompson DW (1992) On growth and form, revised edition. Dover Publications, New York
2. Jean RV (1994) Phyllotaxis. A systemic study in plant morphogenesis. Cambridge University Press, Cambridge
3. Keener J, Sneyd J (1998) Mathematical physiology. Springer, New York
4. Bini D, Cherubini C, Filippi S, Gizzi A, Ricci PE (2010) On spiral waves arising in natural systems. Commun Comput Phys 8 (3):610–622
5. Winfree AT (2010) The geometry of biological time. Springer, New York
6. Murray JD (2002) Mathematical biology, vol vols 1 and 2. Springer, New York
7. Kondepudi D, Prigogine I (2014) Modern thermodynamics: from heat engines to dissipative structures, 2nd edn. Wiley, Hoboken, NJ
8. Loppini A, Capolupo A, Cherubini C, Gizzi A, Bertolaso M, Filippi S, Vitiello G (2014) On the coherent behavior of pancreatic beta cell clusters. Phys Lett A 378(44):3210–3217
9. Bertolaso M, Capolupo A, Cherubini C, Filippi S, Gizzi A, Loppini A, Vitiello G (2015) The role of coherence in emergent behavior of biological systems. Electromag Biol Med 34 (2):138–140
10. Saaty TL (2011) Modern nonlinear equations. Dover Publications, New York
11. Saaty TL (2010) Nonlinear mathematics. Dover Publications, New York
12. Cherubini C, Filippi S (2009) Lagrangian field theory of reaction-diffusion. Phys Rev E 80 (4):046117
13. Turing A (1952) The chemical basis of morphogenesis. Philos Trans R Soc Lond Ser B Biol Sci 237(641):37–72
14. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117(4):500–544
15. Nowak MA, May RM (2000) Virus dynamics. Mathematical principles of immunology and virology. Oxford University Press, Oxford
16. Noble D (2008) The music of life: biology beyond genes, 1st edn. Oxford University Press, Oxford
17. Proctor RN (2000) The Nazi war on cancer. Princeton University Press, Princeton, NJ
18. Werner HMJ, Mills GB, Ram PT (2014) Cancer systems biology: a peek into the future of patient care? Nat Rev Clin Oncol 11:167–176
19. Abbas AK, Lichtman AHH, Pillai S (2017) Cellular and molecular immunology, 9th edn. Elsevier, Philadelphia, PA
20. Soto AM, Sonnenschein C (2004) The somatic mutation theory of cancer: growing problems with the paradigm? BioEssays 26:1097–1107
21. Sonnenschein C, Soto AM (2008) Theories of carcinogenesis: an emerging perspective. Semin Cancer Biol 18:372–377
22. Potter JD (2007) Morphogens, morphostats, microarchitecture and malignancy. Nat Rev Cancer 7:464–474
23. Weinberg RA (1998) One renegade cell: how cancer begins. Basic Books, New York
24. Sonnenschein C, Soto AM (2000) Somatic mutation theory of carcinogenesis: why it should be dropped and replaced. Mol Carcinog 29:205–211
25. Baker SG, Soto AM, Sonnenschein C, Cappuccio A, Potter JD, Kramer BS (2009) Plausibility of stromal initiation of epithelial cancers without a mutation in the epithelium: a computer simulation of morphostats. BMC Cancer 9:89
26. Wodarz D, Komarova NL (2005) Computational biology of cancer, Lecture notes and mathematical modeling. World Scientific, Singapore
27. Nowak MA (2006) Evolutionary dynamics. Exploring the equations of life. The Belknap Press of Harvard University Press, Cambridge, MA
28. Britton NF (2003) Essential mathematical biology. Springer, New York
29. Barrat A, Barthèlemy M, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press, Cambridge
30. Giuliani A, Filippi S, Bertolaso M (2014) Why network approach can promote a new way of thinking in biology. Front Genet 5:83
31. Cherubini C, Filippi S, Gizzi A, Loppini A (2015) Role of topology in complex functional networks of beta cells. Phys Rev E 92 (4):042702
32. Cherubini C, Gizzi A, Bertolaso M, Tambone V, Filippi S (2012) A bistable field model of cancer dynamics. Commun Comput Phys 11 (1):1–18
33. Cherubini C, Filippi S and Gizzi A (2006) Diffusion processes in human brain using COMSOL multiphysics. In: Proceedings of COMSOL Users Conference of Milan, Italy. ISBN: 0-9766792-4-8

34. Crank J (1980) The mathematics of diffusion. Oxford University Press, Oxford
35. Jackson PR, Juliano J, Hawkins-Daarud A,·Rockne RC and Swanson KR (2015) Patient-specific mathematical neuro-oncology: using a simple proliferation and invasion tumor model to inform clinical practice. Bull Math Biol 77:846–856
36. Stylianopoulos T (2017) The solid mechanics of cancer and strategies for improved therapy. J Biomech Eng 139(2):1–23
37. Jain RK, Martin JD, Stylianopoulos T (2014) The role of mechanical forces in tumor growth and therapy. Annu Rev Biomed Eng 16:321–346

Chapter 11

Endogenous Molecular-Cellular Network Cancer Theory: A Systems Biology Approach

Gaowei Wang, Ruoshi Yuan, Xiaomei Zhu, and Ping Ao

Abstract

In light of ever apparent limitation of the current dominant cancer mutation theory, a quantitative hypothesis for cancer genesis and progression, endogenous molecular-cellular network hypothesis has been proposed from the systems biology perspective, now for more than 10 years. It was intended to include both the genetic and epigenetic causes to understand cancer. Its development enters the stage of meaningful interaction with experimental and clinical data and the limitation of the traditional cancer mutation theory becomes more evident. Under this endogenous network hypothesis, we established a core working network of hepatocellular carcinoma (HCC) according to the hypothesis and quantified the working network by a nonlinear dynamical system. We showed that the two stable states of the working network reproduce the main known features of normal liver and HCC at both the modular and molecular levels. Using endogenous network hypothesis and validated working network, we explored genetic mutation pattern in cancer and potential strategies to cure or relieve HCC from a totally new perspective. Patterns of genetic mutations have been traditionally analyzed by *posteriori* statistical association approaches in light of traditional cancer mutation theory. One may wonder the possibility of a priori determination of any mutation regularity. Here, we found that based on the endogenous network theory the features of genetic mutations in cancers may be predicted without any prior knowledge of mutation propensities. Normal hepatocyte and cancerous hepatocyte stable states, specified by distinct patterns of expressions or activities of proteins in the network, provide means to directly identify a set of most probable genetic mutations and their effects in HCC. As the key proteins and main interactions in the network are conserved through cell types in an organism, similar mutational features may also be found in other cancers. This analysis yielded straightforward and testable predictions on an accumulated and preferred mutation spectrum in normal tissue. The validation of predicted cancer state mutation patterns demonstrates the usefulness and potential of a causal dynamical framework to understand and predict genetic mutations in cancer. We also obtained the following implication related to HCC therapy, (1) specific positive feedback loops are responsible for the maintenance of normal liver and HCC; (2) inhibiting proliferation and inflammation-related positive feedback loops, and simultaneously inducing liver-specific positive feedback loop is predicated as the potential strategy to cure or relieve HCC; (3) the genesis and regression of HCC is asymmetric. In light of the characteristic property of the nonlinear dynamical system, we demonstrate that positive feedback loops must be existed as a simple and general molecular basis for the maintenance of phenotypes such as normal liver and HCC, and regulating the positive feedback loops directly or indirectly provides potential strategies to cure or relieve HCC.

Key words Systems biology, Endogenous molecular-cellular network hypothesis, Nonlinear stochastic dynamical system, Hepatocellular carcinoma (HCC), Stable state, Genetic mutation pattern, Positive feedback loop, Cancer therapy, Adaptive landscape

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_11,

1 Introduction

Cancer, a popular generic term for malignant neoplasms, has been defined as an abnormal mass of tissue the growth of which exceeds and is uncoordinated with that of the normal tissues and persists in the same excessive manner even after the cessation of the stimuli which evoked the change, and the new growth has virulent or adverse properties in the body [1]. Its exact nature is still not well understood. Numerous theories and hypotheses on cancer genesis and progression have been proposed during the long discourse on this disease and the attempts to cure it. From the Renaissance to the nineteenth century, scientific hypotheses about cancer had started to be formed, such as cancer as chronic irritation disease, trauma disease, infectious disease, and so on. These theories and hypothesizes are obviously subjected to their limited understanding of cancer at that time, and evidently not suitable hypotheses for today's perspective. Nevertheless, those efforts show changings of paradigms in cancer research and are very useful in the development of research programs and treatments on cancer.

During the twentieth century there was a tremendous progress in genetics, we have gained enormous knowledge on DNA's and genes increased. Cancer has then been hypothesized as a genetic disease, still the dominant dogma in the field. This genetically centric hypothesis suggested that the genesis and progression of cancer is caused by genetic alterations, carcinogenic factors caused cancer by their damages to normal genome [2, 3]. Since then, cancer research has focused on genetic and genomic aspects, such as gene sequencing [4], oncogenes [5], suppressor genes [6].

On the other hand, away from gene focus mountains of experimental evidence and theoretical analyses have suggested that genome is not the whole story on cancer genesis and progression. From the experimental side, evidences show that some other factors such as microenvironment and inflammation cannot be ignored. One pronounced instance is the seed and soil hypothesis: In 1889, the English surgeon, Stephen Pagt, concluded his analyses of cancer histories borrowing a plant analogy. It states that when a plant goes to seeding, its seeds are carried in all directions; but they can only live and grow if they fall on congenial soil [7]. The seed in the modern usage has been reinterpreted as progenitor cell, cancer stem cell, or metastatic cell, and the soil as host factors, stoma, or the organ microenvironment [7, 8]. The experimental evidence suggested that organ microenvironment cannot be ignored in cancer genesis and progression. Another more direct evidence is the study of the precursor of esophageal adenocarcinoma. By tracking its source of precursor, recent findings suggest that certain precancerous lesions, such as Barrett's, initiate not from genetic alterations but from competitive interactions between cell lineages driven

by opportunity [9]. Based on this kind of phenomena as well as others one may readily conclude that besides genome other factors such as congenial soil (microenvironment) and inflammation also play key roles in cancer genesis and progression and these factors cannot be ignored.

From both the clinic and theoretical sides, evidences also suggest that the genetic and genomic information are important but not enough. Biological systems are characterized by the stochastic dynamical phenomena and concepts such as adaptation [10, 11], robustness [12, 13],phenotype switch [14, 15]. In tumors such concepts correspond to well-documented drug resistance [16], tremendously difficult for cancer regression [17] and genesis and progression to tumor from normal tissue respectively. The phenomena and concepts arise from the complex regulatory machinery, the building blocks of the regulatory machinery including genetic switch, feedback loops [18], double-edge effect [19, 20], etc. A desirable method to make the biological system's phenomenon and concepts clear is to understand and manipulate the regulatory machinery quantitatively, it is also one of the biggest challenges for contemporary biology [21–25]. It is clear that we cannot achieve this goal just by using genetic information. Consensus starts to form that complete information of the DNA sequence of an organism will not enable us to reconstruct the regulatory machinery quantitatively because of the many gaps between the genotype and the phenotype. We need to reveal the regulatory machinery behind a biological phenomenon quantitatively, which would be of great importance for our understanding and manipulating biological phenomenon, such as cancer genesis and progression.

To meet this challenge, based on the current understanding of biological systems, the endogenous molecular-cellular network hypothesis for cancer genesis and progression has been proposed [26–28]. In the following section, the key aspects of the hypothesized theory and its implications will be reviewed and elaborated. In Subheading 2.1, we will discuss the basic elements in the hypothesis and the essential requirements. In Subheading 2.2, we applied the endogenous molecular-cellular network hypothesis in hepatocellular carcinoma (HCC). In Subheading 2.3 and 2.4, We use the endogenous molecular-cellular network hypothesis to quantify and understand the genesis and progression of HCC, which suggest that the stable states of the endogenous network can be used to represent normal liver and HCC at both the modular and molecular levels. In Subheading 2.5, we explored a genetic mutation pattern in cancer using the endogenous network hypothesis. In Subheading 2.6, we explored potential strategies to cure or relieve HCC. The similarities and differences among similar proposals, along with a few dominant cancer theories, are discussed in Subheading 3. We

conclude in Subheading 4 that the endogenous network theory may provide the best candidate theory to understand cancer genesis and progression.

2 Experimental Research

2.1 Forming Endogenous Molecular-Cellular Network Hypothesis

In this subsection, we discuss what would be needed for the endogenous molecular-cellular network hypothesis on cancer genesis and progression. Three important aspects of this hypothesis will be examined in detail.

The hypothesis can be stated as follows. In order to maintain the normal physiological function and developmental process for tissue-specific function shaped by evolution, a minimal set of fundamental functional modules or pathways, for example cell cycle, cell death, inflammation, metabolism, cell adhesion, angiogenesis, are needed; Each module can accomplish a relatively autonomous function, and cross-talks between modules allow one function to influence another. At the molecular-cellular level, it is hypothesized that the functional modules are deeply hierarchical and may be specified by important molecular and cellular agents, such as oncogenes, suppressor genes, and related growth factors, hormones, cytokines, etc. The interactions between these agents form an autonomous, nonlinear, stochastic, and collective dynamical network. We have tentatively named it as the endogenous molecular–cellular network. The endogenous network may generate many locally stable states with obvious or non-obvious biological function. Normal state and cancer state are assumed to be the stable states of the endogenous molecular-cellular network. The endogenous network may stay in each stable state for a considerably long time, and in certain condition, the stable states can switch between each other. In this hypothesis, the genesis and progression of cancer can be regarded as transition of the endogenous molecular-cellular network from the normal stable state to the cancer state [26, 27].

With this hypothesis, biological systems will be greatly simplified, and make it possible for us to grasp and manipulate the general regulatory machinery of cancer genesis and progression quantitatively. We now discuss the three basic tenets of this hypothesis, modularization, deeply hierarchy, and autonomous regulation, one by one.

1. ***Modularization:*** *Biological system is built by modules and cross-talk between modules, each module can accomplish a relatively autonomous function, and cross-talks between modules allow one function to influence another.*

A functional module is, by definition, a discrete entity whose function is separable from those of other modules [18]. The first evidence that supports this assumption is that functional modules

are transplantable and conserved in different organisms, for example, functional module cell cycle are conserved in different organisms, much of what we know about cell cycle can be traced back to basic studies in yeast cells [29, 30], cell-specific functional modules can be transplanted into different types of cells [31, 32], four stem cell transcription factor induced cell type switch to iPS cell, including morphology, proliferation, surface antigens, gene expression, epigenetic status of pluripotent cell-specific genes, and telomerase activity [31]. These evidences directly indicated that many biological functions are relatively autonomous, the separated functional modules can accomplish its own function and of course the modules cross-talk with each other.

Another evidence is the theoretical modes based on functional modules verifying their predictions match reality, such as cell cycle [13, 33], apoptosis [34]. The quantitative model, by theoretical predications and experimental validation, is the best way to test our understanding of biological systems. The match of predications and reality suggest that we can adopt the concepts of autonomous functional modules. Therefore, we argue here that the biological system is built by modules and cross-talk between modules, each module can accomplish a relatively autonomous function, and cross-talks between modules allow one function to influence another and due to this we can model the functional modules quantitatively. It also suggests that the recognition of functional "modules" as a critical level of biological organization is important for our understanding of biological systems.

2. ***Deeply hierarchy****: there are regulators that serve as decision-marking in modules to determine module's fate, then the regulators trigger the expression of a whole battery of downstream genes related to the decided fate. Modules and cross-talk between modules can be simplified and specified by interactions of important proteins, the general principles will allow us to grasp and manipulate the modules.*

Multiple experimental evidences suggest the existence of key regulators, for example two regulatory proteins CI and Cro regulate phage lambda genetic switch [35], key regulators regulate developmental process of sea urchin [36], four transcriptional factors induce human fibroblasts to pluripotent stem (iPS) cells [31]. The module components include Genes (DNA), mRNA, MicroRNA, Proteins, small molecules, etc. Among all the components proteins play special roles in the biological system, because they are the main biological function executor, they can regulate gene expression, and they are key regulators in signaling transduction. It is clear that functions cannot be understood by studying the single proteins [23, 37]. Important functions arise from the regulatory machinery. General modules and cross-talk between modules

can be specified by proteins and their interactions in simplifying modeling [38].

3. ***Autonomous network and intrinsic stable states****: the regulatory proteins and their interactions form a closed and decision-making network which is responsible for the biological stable states, we named the closed network as endogenous molecular-cellular network. The endogenous molecular-cellular network is shaped by evolution, normal tissue and tumor can be regarded as intrinsic stable states.*

There is a paradox in biological systems [38]: if mRNAs are required to synthesize proteins and proteins are required, in turn, to regulate the expression of mRNAs, then what's the cause-and-effect relationship of the interdependent components? The paradox can be interpreted as the interdependent components of biological systems forming a closed network. Mathematically, the closed network forms a nonlinear autonomous dynamical system implying many locally stable states [39, 40].

We further assume that the backbone and essential structure of the endogenous molecular-cellular network, in other words the regulatory machinery, would remain the same and is conserved, because it has been shaped by millions, or even billions, years of evolution. During the lifetime of an organism, there is a little chance of any major modification of the essential structure of the endogenous network for a viable cell [41]. And multiple evidences suggest that cancer is similar to many normal physiological processes such as wound healing, developmental process, cancer often is called un-healing wound [42], inflammation, and aberrant development [43]. Based on the evidence we reason that cancer may be an intrinsic stable state in organisms shaped by evolution, cancer state is one of the stable states of the endogenous molecular-cellular network.

2.2 Quantitative Implementation of Endogenous Molecular-Cellular Network Hypothesis in Hepatocellular Carcinoma (HCC)

Endogenous molecular-cellular network hypothesis of cancer has been proposed as an alternative picture to understand cancer [26, 27]. Hepatocellular carcinoma (HCC) is the main primary liver tumor, accounting for 85–90% of primary liver cancers diagnosed [44]. We take hepatocellular carcinoma as an example using the endogenous molecular-cellular network hypothesis to quantify and understand the genesis and progression of cancer.

First, we assumed that the biological system is built by a set of functional modules and cross-talk between the modules. According to the current understanding of cancer biology [45], a minimal set of core functional modules (Fig. 1) to describe HCC at the systemic level may include the cell cycle module, apoptosis module, metabolism module, liver-specific function module, cell adhesion module, immune response module, and angiogenesis module [18, 45–47]. Further, we assumed that the status of each functional

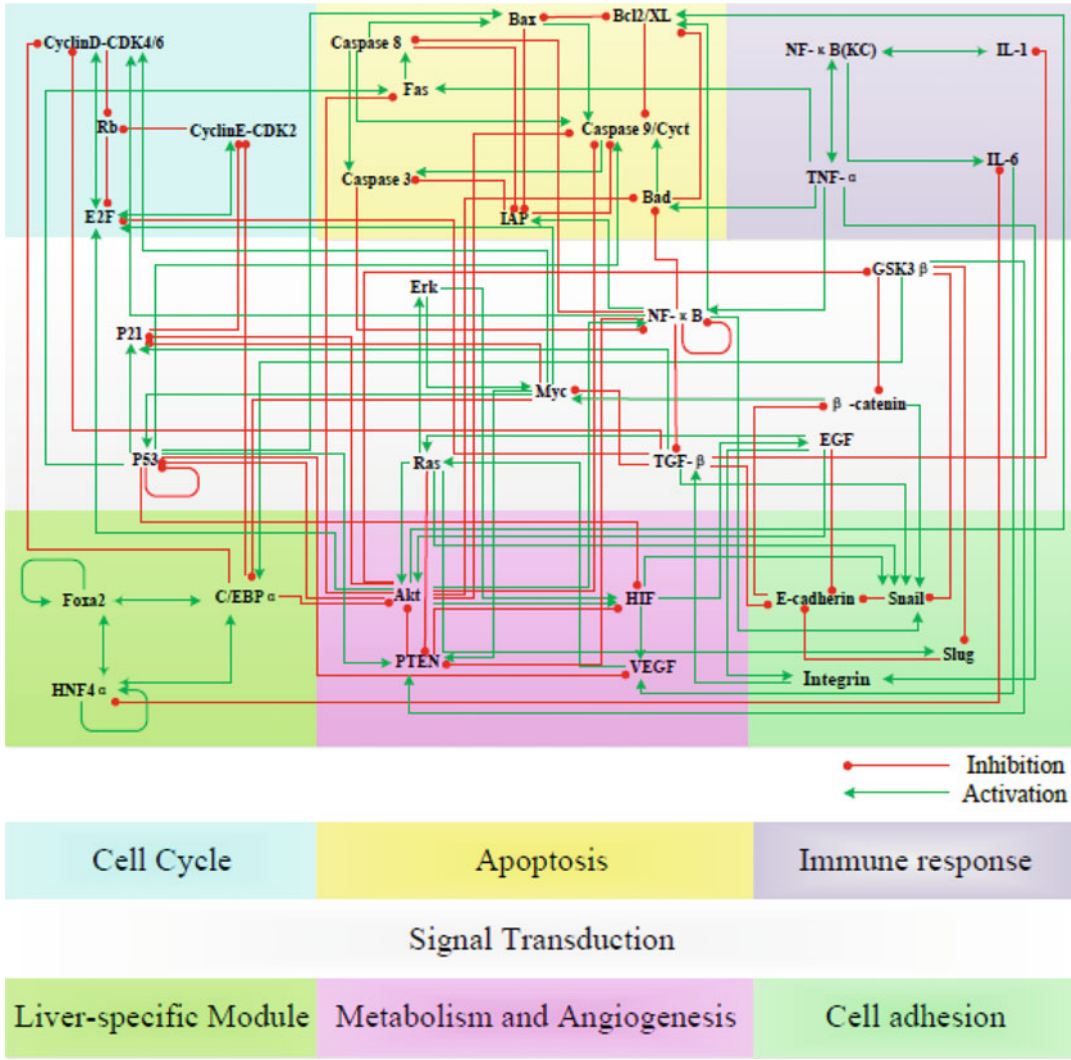


Fig. 1 Core endogenous molecular-cellular network of hepatocyte. Cell cycle, apoptosis, metabolism, hepatocyte-specific function, cell adhesion, immune response, and angiogenesis module were selected to capture the essential features of hepatocyte. Each module was simplified and specified by a set of key proteins. Interactions between the proteins, activation/up-regulation (*green line*), or inhibition/down-regulation (*red line*) were summarized from well-documented and conserved gene regulatory network and signaling transduction pathways. The core endogenous network of hepatocyte included 37 key proteins and 114 interactions

module may be regulated by a minimal set of molecular-cellular agents according to the accumulated molecular biology knowledge of liver [46, 47]. The molecular-cellular agents are assumed to be proteins, as proteins are the main cell fate decision maker via regulating signaling transduction and gene expression.

Second, the molecular-cellular agents will interact with each other. The activation/upregulation or inhibition/downregulation

among the agents was summarized from the well-documented gene regulatory network and signaling transduction pathway that suggest that the interactions have a solid biochemical basis [46, 47]. We assume that the agents and interactions among the agents form autonomous and decision-making network, which suggests that the transmission of information is not one way and there is no privileged causality in the network. To emphasize the network is formed by the interactions of the endogenous cellular-molecular agents shaped by evolution, we name the network as an endogenous molecular-cellular network. The aim to establish a liver endogenous network is to reveal the core regulatory mechanisms of liver at the systemic level. Working endogenous network of HCC has been established according to the hypothesis (Fig. 1). As the key molecular-cellular agents to regulate functional modules status and their interactions have been documented and proved to be conserved, it seems that the working endogenous network is reproducible according to the hypothesis.

It is necessary to discuss the gaps between the working endogenous network of the liver and the real liver. There is no doubt that the real liver has been greatly simplified with these assumptions above. For example, the functional modules are far more than the selected modules here; other molecular-cellular agents such as microRNA, metabolites, and other proteins are not considered explicitly in the network. Thus, the simplified and incomplete working endogenous network is by no means the only realization, and is open to further expansion and revision. However, we will show later that it is one of the simplest variants which may reproduce the main features of normal liver tissue and HCC tissue at both the modular and molecular levels. Moreover, we will show that solid predications can be reached by careful analysis even in this case of partial knowledge.

2.3 Quantification of Endogenous Network

The quantitative description of the core endogenous molecular-cellular network consists of a set of coupled differential equations [48, 49]. In Fig. 1, we use CyclinD-CDK4/6 as an example to show how to obtain the differentiation equations. Cyclin D-CDK4/6 was upregulated by transcription factor E2F and Myc, while it was downregulated by C/EBPα, p21, and GSK-3β. First, we assume the dynamical equation for the concentration or activity of CyclinD-CDK4/6 under the influence of the protein E2F, Myc, C/EBPα, p21, and GSK-3β takes the form in Eq. 1 [48–51]:

$$\frac{\mathrm{dCyclinD\text{-}CDK4/6}(t)}{\mathrm{d}t} = f(\mathrm{E2F}, \mathrm{Myc}, \mathrm{C/EBP\alpha}, \mathrm{p21}, \mathrm{GSK\text{-}3\beta}) - \frac{\mathrm{CyclinD\text{-}CDK4/6}}{\tau_{\mathrm{cyclinD\text{-}CDK4/6}}} \quad (1)$$

Here, Cyclin D-CDK4/6 means the concentration or activity of this protein. f(E2F, Myc, C/EBPα, p21, GSK - 3β) means the integrated production rate of CyclinD-CDK4, $\frac{\text{CyclinD-CDK4/6}}{\tau_{\text{cyclinD-CDK4/6}}}$ means the degradation term of CyclinD-CDK4/6, $\tau_{\text{cyclin D - CDK4/6}}$ was the degradation constant of protein Cyclin D-CDK4/6. We further assume that the activation of CyclinD-CDK4/6 needs activated E2F or Myc, and at the same time inactivated C/EBPα, p21, and GSK-3β. Then, the integrated production rate of CyclinD-CDK4/6, f(E2F, Myc, C/EBPα, p21, GSK - 3β), is quantified in Eq. 2:

$$\begin{aligned} &f(\text{E2F}, \text{Myc}, \text{C/EBP}\alpha, \text{p21}, \text{GSK-3}\beta) \\ &\quad = V_{\text{cyclinD-CDK4/6}} \times \frac{\left(\frac{[\text{E2F}]}{K_{11}}\right)^{n_{11}} + \left(\frac{[\text{Myc}]}{K_{12}}\right)^{n_{12}}}{1 + \left(\frac{[\text{E2F}]}{K_{11}}\right)^{n_{11}} + \left(\frac{[\text{Myc}]}{K_{12}}\right)^{n_{12}}} \\ &\quad \times \frac{1}{1 + \left(\frac{[\text{C/EBP}\alpha]}{K_{13}}\right)^{n_{13}} + \left(\frac{[\text{p21}]}{K_{14}}\right)^{n_{14}} + \left(\frac{[\text{GSk-3}\beta]}{K_{15}}\right)^{n_{15}}} \end{aligned} \tag{2}$$

$V_{\text{cyclinD - CDK4/6}}$ means the maximal production rates of protein CyclinD-CDk4, n_{1i} and K_{1i} are parameters of biochemical reaction, which is used to describe the detailed weight of each protein in regulating the production of Cyclin D-CDK4.

It is impossible to obtain all the parameters, fortunately as we focus on the relative concentration or activity of these agents, we can grasp the key features of the interactions and architecture by assuming appropriate parameter values. In this framework, the content or activity of each protein was normalized to range from 0 to 1, 0 means minimal content or activity, 1 means maximal. The maximal production rate, $V_{\text{cyclinD - CDK4/6}}$, and the degradation rate,$\tau_{\text{cyclinD - CDK4/6}}$, may be normalized as 1. We assume that $n_{1i} = 3$ and $K_{1i} = 8$ to grasp the key feature of activation or inhibition, we have validated that the values of n_{1i} and K_{1i} can be relaxed without affecting the conclusions we made in this paper [49]. With these assumptions, the equation may be simplified in Eq. 3:

$$\begin{aligned} \frac{\text{d[CyclinD-CDk4/6]}}{\text{d}t} &= \frac{8 \times [\text{E2F}]^3 + 8 \times [\text{Myc}]^3}{1 + 8 \times [\text{E2F}]^3 + 8 \times [\text{Myc}]^3} \\ &\times \frac{1}{1 + 8 \times [\text{C/EBP}\alpha]^3 + 8 \times [\text{p21}]^3 + 8 \times [\text{GSK-3}\beta]^3} \\ &- [\text{CyclinD-CDK4/6}] \end{aligned} \tag{3}$$

The quantitative assumptions provide a general framework to quantify endogenous network, the function types and parameters can be relaxed without affecting the major conclusions. It seems

that the relative content or activity assumptions will not affect validation, because many experimental data such as gene expression level is also relative. Other molecular-cellular agents in the endogenous molecular-cellular network of liver are quantified in a similar way, the endogenous molecular-cellular network of liver was transformed into a set of ordinary differential equations which implies some attractors underlying the molecular-cellular network of liver [49].

2.4 Normal Liver and HCC Are Stable States of Endogenous Network

According to the ordinary differential equations, we obtain five attractors that may have obvious or non-obvious biological functions in Table 1. Each attractor is specified by the relative content or activity of these molecular-cellular agents. According to the expression or the activation level of these agents we can judge the status, ON or OFF, of the functional modules in each attractor. ON means the expressional or activation level of the molecular-cellular agents supports the execution of the functional module, while OFF means the expression of activation of these agents did not support the execution of the functional module. In Table 1, we summarize the status of these functional modules, including cell cycle, metabolism, liver-specific function, cell death, cell adhesion, immune response, and angiogenesis, in each attractor. On the other hand, we also summarize the status of these functional modules in normal liver and HCC form clinical and experimental data in Table 1 [45]. Two attractors obtained from the model have a perfect match with normal liver and HCC respectively when comparing model results with clinical and experimental results. So, we preliminarily conclude that the two attractors of the model reproduce the clinical normal liver and HCC.

Table 1
Model result and clinical observations at the modular level

	Model results				Clinical observation [45]	
	A	B	C	D	Normal hepatocyte	Cancerous hepatocyte
Cell cycle	OFF	ON	OFF	ON	OFF	ON
Liver differentiation	ON	OFF	ON	OFF	ON	OFF
Apoptosis	OFF	OFF	ON	OFF	OFF	OFF
Cell adhesion	ON	OFF	OFF	OFF	ON	OFF
Proliferative metabolism	OFF	ON	OFF	ON	OFF	ON
Immune response	OFF	ON	ON	OFF	OFF	ON
Angiogenesis	OFF	ON	OFF	ON	OFF	ON

Stable states A and B reproduced the key features of normal hepatocyte and cancerous hepatocyte at the modular level

Further, we test this preliminary conclusion at the molecular-cellular level. First, according to the modeling result, we summarize the relative change of each agent from the normal liver to HCC in Table 2. Then, we summarize the relative change of each agent from the normal liver to HCC in Table 2 according to the experimental data. The experimental results are collected from two independent ways. One is the specialist knowledge which suggests the current biological understanding of each agent and its relative change from normal liver to HCC, relative changes of the agents from clinical data has also been summarized as double check [46, 47].The other way is to analyze the high-throughput microarray data, checking the transcriptional change of each agent from the normal liver to HCC [52]. The comparison shows that the modeling results and specialist knowledge have an agreement of 88%, and have an agreement of 62%, 76%, 71% with three independent HCC tissues respectively (NCBI ID: GSE33006) [52].

The working endogenous network is established from the interactions determined at different contexts and by different groups. It seems quite striking that the model and experimental results have a perfect match at the modular level, and are coherence at the molecular-cellular level. Consistency of validation suggests that the present working network of liver may reproduce the key features of normal liver and HCC. Given the heterogeneous nature of cancer and the incomplete working network, experimental results of certain molecular-cellular agents that do not appear to fit this model must be expected.

We also want to point out that as it is impossible to enumerate all the possible attractors of this high-dimensional dynamic system, the attractors obtained here are by sampling. By sampling enough times, we conclude that there are at least five attractors; however, we cannot preclude the possibility that there are other attractors. Mathematically, the structure of the endogenous molecular-cellular network model is similar to the Morse-Smale dynamical system [53]: it is structurally stable with a finite number of attractors and the transitions between attractors are possible in the presence of perturbations.

2.5 Model Prediction Reproduces Key Features of Cancer Genetic Mutation Data

It has been known that dynamical equations can be used to predict the effect of mutations. For example, a detailed analysis of mutations against experiments has been performed for the core regulatory network of phage lambda genetic switch [54, 55]. Here, we examined the molecular-level details of these two stable states not at such dynamical level, but from the relative expression level side which is less sensitive to kinetic parameters [56]. We first characterized the activity of each protein in a given stable state as activated or inactivated by setting a threshold: if the activity of a protein is greater than this threshold, we identified the protein as active, and if lower, the protein was identified as inactive. Activated

Table 2
Model result and experimental data at the molecular level

	Model results	Experimental results		
	–	HCC 1	HCC 2	HCC 3
Molecular agents (Gene symbol)	–	Agreement ratios 24/37 = 64.9%	Agreement ratios 29/37 = 78.4%	Agreement ratios 25/37 = 67.6%
Cyclin D-CDK4/6 (CCND1)	Up	Up	Up	Up
Cyclin E-CDK2 (CCNE1)	Up	Up	Up	Up
Rb (Rb1)	Down	Up	Down	Down
E2F (E2F4)	Up	Up	Up	Up
C/EBPα (CEBPA)	Down	Up	Down	Down
Foxa2 (FOXA2)	Down	Up	Down	Down
HNF4α (HNF4A)	Down	Down	Down	Down
Fas (Fas)	Up	Up	Up	Down
Bcl-2 (BCL2)	Up	Down	Un-change	Up
XIAP (XIAP)	Up	Up	Up	Up
Bax (BAX)	Un-change	Up	Down	Down
Bad (BAD)	Up	Up	Down	Down
Caspase 9 (CASP 9)/ Cytchrome c (CYCS)	Un-change	Up	Un-change	Un-change
Caspase 8 (CASP8)	Un-change	Up	Up	Up
Caspase 3 (CASP3)	Un-change	Down	Up	Down
E-cadherin (CDH1)	Down	Down	Down	Down
Snail (SNAIL1)	Up	Up	Up	Up
Slug (SNAIL2)	Up	Up	Up	Up
Integrin (ITGB2)	Up	Up	Up	Up
Akt (AKT3)	Up	Up	Up	Up
PTEN (PTEN)	Down	Down	Down	Un-change
HIF (HIF1A)	Up	Up	Up	Un-change
TNFα (TNF)	Up	Up	Down	Up
NF-κB (RELA)	Up	Down	Up	Up

(continued)

Table 2
(continued)

	Model results	Experimental results		
NF-κB (Kupffer) (RELA)	Up	Down	Up	Up
IL-1 (IL1B)	Up	Up	Up	Up
IL-6 (IL6)	Up	Up	Up	Up
EGF (EGFR)	Up	Up	Up	Up
VEGF (KDR)	Up	Un-change	Up	Down
Ras (KRAS)	Up	Up	Up	Down
ERK (MAPK1)	Up	Up	Up	Down
GSK-3β (GSK3B)	Down	Down	Down	Down
β-catenin (CTNNB1)	Up	Up	Un-change	Up
Myc (MYC)	Up	Up	Up	Up
P53 (TP53)	Up	Down	Up	Down
TGF-β (TGFB2)	Up	Up	Up	Up
P21 (CDKN1A)	Un-change	Up	Up	Up

Relative changes of protein's activity from stable state A to stable state B, including up, down or unchanged, obtained from model results were listed in the second column. Relative changes of protein's activity from normal liver to HCC obtained from three independent gene expression data were listed in the last three columns]. Model result has agreement ratios of 64.9%, 78.4% and 67.6% with three independent experimental data respectively)

proteins in each stable state were then highlighted to reveal sub-networks that are expected to play key roles in the establishment and maintenance of the stable state. Since the activity of each protein was normalized to range from 0 (minimal activation) to 1 (full activation), the threshold can be selected within a reasonable range (from 0.4 to 0.6). We found that these thresholds will not affect the main conclusions. Figure 2 showed the sub-networks of the normal hepatocyte and cancerous hepatocyte when we set the threshold as 0.4.

These sub-networks then provided means to directly identify a few key features of the genetic mutation patterns in HCC. Biologically, genetic mutations can have varying effects on the function of protein. Mutations that confer enhanced activity were defined as gain-of-function mutations, while those that reduce or abolish protein function were defined as loss-of-function mutations. It was known that some random mutations in cancers were selected and accumulated in response to phenotypic consequences [57–59]. We reasoned that proteins that were inactive in the normal hepatocyte stable state and activate in the cancerous hepatocyte

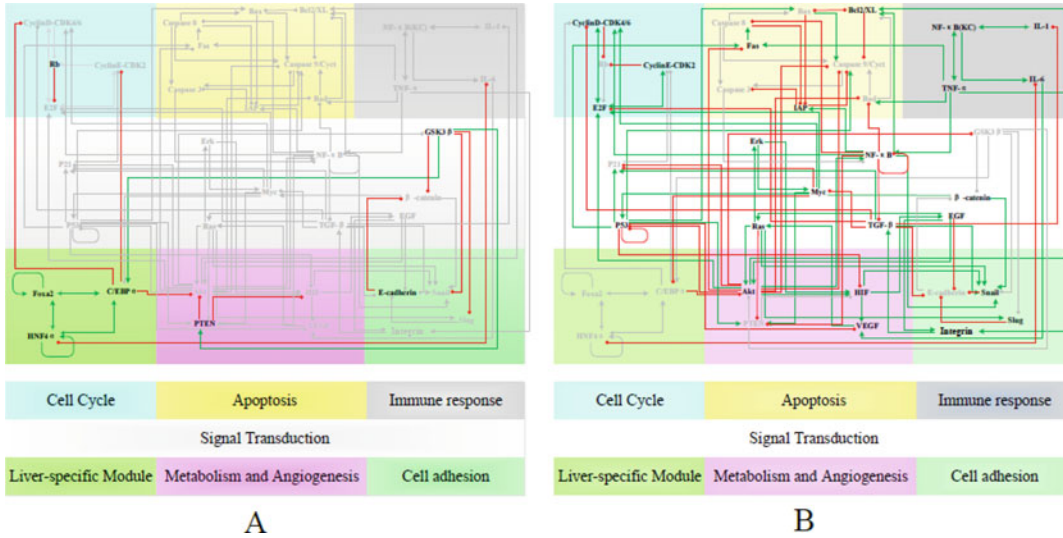


Fig. 2 Sub-networks of normal hepatocyte and cancerous hepatocyte stable states. Activated proteins and interactions are highlighted in *bold* in normal hepatocyte stable state (**a**) and cancerous hepatocyte stable state (**b**) to form different sub-networks

stable state could have a higher probability of undergoing gain-of-function mutations in cancers, as gain-of-function of this gene can adapt to the cancerous hepatocyte stable state and confer selective advantages to establish and maintain the cancerous hepatocyte stable state. For example, as shown in Table 2. Cyclin D-CDK4 was inactive in the normal hepatocyte stable state and active in the cancerous hepatocyte stable state, and so Cyclin D-CDK4 was identified to have a higher probability of undergoing a gain-of-function mutation in HCC. Similarly, proteins that were active in the normal hepatocyte stable state and inactivate in the cancerous hepatocyte stable state were expected to have a higher probability of undergoing loss-of-function mutations in HCC. For example, Rb was active in the normal hepatocyte stable state and inactive in the cancerous hepatocyte stable state, and so Rb was identified to have a higher probability of undergoing a loss-of-function mutation in HCC. In this way, we identified probable genetic mutations in HCC (Table 3). We should note that there were also six proteins whose activity did not significantly differ between the normal hepatocyte stable states and cancerous hepatocyte stable states. For example, Bax was inactive in both the normal hepatocyte stable state and the cancerous hepatocyte stable state. In this approach, we could not decide the probable mutation of this kind of genes.

We next compared these predicted mutated genes with the well-documented genetic mutation data from Catalogue of Somatic Mutations in Cancer (COSMIC) [60]. Given the heterogeneity of mutations in different HCC patients, and even different regions or cells of the same HCC patient, a set of 20 top mutated

Table 3
Prediction of top mutated genes in cancerous hepatocyte from model

Gene name	Model result
Cyclin D-CDK4/6	Gain-of-function
Cyclin E-CDK2	Gain-of-function
Rb	Loss-of-function
E2F	Gain-of-function
C/EBPα	Loss-of-function
Foxa2	Loss-of-function
HNF4α	Loss-of-function
Fas	Gain-of-function
Bcl-2/xL	Gain-of-function
IAP	Gain-of-function
Bax	–
Bad	–
Casp 9/Cytc	–
Casp 8	–
Casp 3	–
E-cadherin	Loss-of-function
Snail	Gain-of-function
Slug	Gain-of-function
Integrin	Gain-of-function
Akt	Gain-of-function
PTEN	Loss-of-function
HIF	Gain-of-function
TNFα	Gain-of-function
IKK-NF-κB	Gain-of-function
IKK-NF-κB (KC)	Gain-of-function
IL-1	Gain-of-function
IL-6/stat	Gain-of-function
EGF	Gain-of-function
VEGF	Gain-of-function
Ras	Gain-of-function
ERK	Gain-of-function

(continued)

Table 3
(continued)

Gene name	Model result
GSK3β	Loss-of-function
β-Catenin	Gain-of-function
Myc	Gain-of-function
P53	Gain-of-function
TGF-β	Gain-of-function
P21	–

There are 24 proteins that were predicted to have gain-of-function mutation and 7 proteins that have loss-of-function mutation. Present approach could not decide the mutation type of the reminding 6 genes (–)

genes in a large number of HCC samples was chosen to reflect key features of genetic mutation data. Four genes (TP53 (P53), CTNNB1 (β-catenin), RB1, PTEN) can be linked to proteins in the core network directly, whereas seven genes (AXIN1, CDKN2A, PIK3CA, HNF1A, ATM, CREBBP, and IL6ST) can be linked to proteins in the core network indirectly, as they have well-defined relationships with proteins in the core network according to Kyoto encyclopedia of genes and genomes (KEGG). These 11 proteins that can be linked to proteins in the core network directly or indirectly were classified into category 1 in Table 4; the remaining 9 proteins that have not been considered in the present network were classified into category 2 in Table 4. Overall, we found that 11 of the top 20 mutated genes in HCC were included in our analysis.

As most of the proteins and interactions in the core network of hepatocyte are conserved in other cell types, the results obtained in the present analysis of HCC are expected to be applicable in other cancers. As such, we also investigated the top 20 mutated genes in biliary tract cancer, bone cancer, breast cancer, central nervous cancer, eye cancer, prostate cancer, skin cancer, small intestine cancer, soft tissue cancer, stomach cancer [61]. The genes that can be linked to proteins in the core network directly or indirectly were classified into category 1 in Table 4. These proteins that have not been considered in the present network were classified into category 2 in Table 4. Strikingly, we found that 10, 13, 9, 10, 13, 12, 11, 13, 13, and 13 of the top 20 mutated genes in these different cancers can be explained by the present core network of the hepatocyte, respectively. Thus, the inherent core network structure, including the relative activity of the constituent proteins, appears to capture common features relevant to the development of cancer in other cell types as well. It should be emphasized that

Table 4
Classification of the top 20 mutated genes in different cancers

Cancer type	Category 1	Category 2	
HCC	TP53, CTNNB1, AXIN1, CDKN2A, PIK3CA, HNF1A, ATM, CREBBP, RB1, IL6ST, PTEN	TERT, ARID1A, ARID2, KMT2C, NFE2L2, KMT2D, PTPRB, TSC2, SMARCA4	11/20
Biliary tract	TP53, KRAS, CDKN2A, SMAD4, AXIN1, CTNNB1, PIK3CA, BRAF, CDH1, PTEN	MLL3, BAP1, IDH1, AR1D1A, PBRM1, TERT, FBXW7, RNF43, IDH2, GNAS	10/20
Bone	TP53,CDKN2A, RB1, CTNNB1, BRAF, AKT1, APC, KIT, NRAS, FGFR2, KRAS, HRAS, EGFR	GNAS, IDH1, PTCH1, IDH2,, SMARCB1, CDC73, SMO	13/20
Breast	PIK3CA, TP53, CDH1, PTEN, AKT1, RB1, ATM, NF1, APC	MLL3, GATA3, ARID1A, MED12, KMT2D, RUNX1, AKAP9, MAP2K4, UBR5,MYH9, BRCA1	9/20
Central nervous	TP53, PTEN, CDKN2A, CTNNB1, EGFR, BRAF, PIK3CA, NF1, RB1, PIK3R1	IDH1, TERT, SMARCB1, H3F3A, ATRX, CIC, CHEK2, PTCH1, KMT2D, SMARCA4	10/20
Eye	GNA11, RB1, BRAF, TP53, PTEN, KIT, NRAS, CDKN2A KRAS, EGFR, PDGFRA, MET, CTNNB1	GNAD, BAP1, SF3B1, TERT, BCOR, FBXW7, DICER1	13/20
Prostate	TP53, PTEN, KRAS, EGFR, CTNNb1, HRAS, ATM, APC, RB1, TRRAP, PIK3CA, BRAF	MLL3, FOXA1, KMT2D, MLL, MED12, AKAP9, MLLT3, KDM6A	12/20
Skin	BRAF, TP53, CDKN2A, FGFR3, NRAS, HRAS, PTEN, PIK3CA, KIT, CTNNB1, MAP2K1	TERT, PTCH1, CYLD, ARID2, ROS1, MLL3, NF, RAC1, GNAQ	11/20
Intestine	KRAS, TP53, APC, SMAD4, CTNNB1, PIK3CA, BRAF, EGFR, KDR, NRAS, ATM, CDKN2A, PDGFRA	GNAS, MEN1, FBXW7, PTPN11, STK11, SMARCB1, ERBB2	13/20
Soft tissue	KIT, CTNNB1, VHL, PDGFRA, TP53, CDKN2A, APC, KRAS, HRAS, NRAS, PTEN, PIK3CA, BRAF	SMARCB1, NF2, MED12, NF, TERT, MEN1, GNAS	13/20
Stomach	TP53, CDH1, APC, PIK3CA, CTNNB1, KRAS, TRRAP, CDKN2A, AXIN1, PTEN, EGFR, PDGFRA, KDR	ARID1A, MSH6, FBXW7, RNF43, NSD1, ERBB2, GNAS	13/20

The first column denotes different cancer types. Genes in category 1 can be linked to proteins in the core network directly or indirectly. Genes in category 2 are not considered in the present network. The last column denotes the coverage rate

these proteins were selected without any prior knowledge of their mutation propensity in cancer.

Moreover, the present model generated the types of aberration, gain-of-function, or loss-of-function, of these probable mutational genes (Table 3). We noted that experiments have identified the type

of aberration of some mutations in HCC [62] and of these, six proteins were in the core network and five were agreed with model prediction. A similar summary of the type of aberration of some mutations in other cancer types [63] reveals that eight proteins were in core network and seven were agreed with model prediction (Table 4). This overall agreement further supports the significant potential of this analysis to predict genetic mutations in cancer. However, it should be mentioned that there is one disagreement between the model and the literature. It was well known that p53 has a loss-of-function mutation in many cancers, while p53 was predicted to have a gain-of-function mutation in our model. This disagreement may be owing to one of two sources. First, given the heterogeneous nature of cancer mutations and the current incomplete network, experimental results of certain genes which do not appear to fit this model are expected. Second, the aforementioned results were obtained with a threshold as 0.4. We found that p53 was one of three genes whose results are sensitive to threshold values. Thus, the behavior of p53 may be more complex than presently believed, as some recent studies have suggested [64].

Our analysis also affords two additional intriguing and testable predictions. First, our model suggests that there are mutations that can confer selective advantages to establish and maintain the normal hepatocytes phenotype: a preferred mutation spectrum in the normal hepatocyte (Table 5). There is some evidence showing that cells and tissues can maintain their normal phenotype in the face of myriad mutated genes [65]. It should be biologically interesting to determine whether there would be such a mutation spectrum in normal liver. Second, while we have used this model to predict mutations in cancer successfully as shown by the above analysis, it showed that normal hepatocyte and cancerous hepatocyte are endogenous stable states of one single endogenous network. This indicated that there are cancers, especially at the early stage, which can take place without major genetic alterations such as these well-documented oncogenes and tumor suppressor genes. Indeed, there is evidence supporting the existence of mutation-free cancer from different standpoints [9, 66]. If firmly established, the cancer genesis and procession is mechanistically completely different from that of the cancer mutation theory.

2.6 New Cancer Therapy Strategies from Endogenous Network Model and Theory

One of the most characteristic properties of dynamical systems is each attractor is maintained by some key molecular-cellular agents and their interactions. Positive feedback loops provide a simple general strategy for the establishment and maintenance of heritable phenotype [67]. The present working model reveals that the agents and interactions maintain normal liver and HCC by forming distinct positive feedback loops in Fig. 3 [68].

The model reveals that positive feedback loopHNF4α-C/EBP-α-Foxa2 is responsible for the establishment and maintenance of

Table 5
Prediction of top mutated genes in normal hepatocyte

Gene name	Model result
Cyclin D-CDK4/6	Loss-of-function
Cyclin E-CDK2	Loss-of-function
Rb	Gain-of-function
E2F	Loss-of-function
C/EBPα	Gain-of-function
Foxa2	Gain-of-function
HNF4α	Gain-of-function
Fas	Loss-of-function
Bcl-2/xL	Loss-of-function
IAP	Loss-of-function
Bax	–
Bad	–
Casp 9/Cytc	–
Casp 8	–
Casp 3	–
E-cadherin	Gain-of-function
Snail	Loss-of-function
Slug	Loss-of-function
Integrin	Loss-of-function
Akt	Loss-of-function
PTEN	Gain-of-function
HIF	Loss-of-function
TNFα	Loss-of-function
IKK-NF-κB	Loss-of-function
IKK-NF-κB (KC)	Loss-of-function
IL-1	Loss-of-function
IL-6/stat	Loss-of-function
EGF	Loss-of-function
VEGF	Loss-of-function
Ras	Loss-of-function
ERK	Loss-of-function

(continued)

Table 5
(continued)

Gene name	Model result
GSK3β	Gain-of-function
β-catenin	Loss-of-function
Myc	Loss-of-function
P53	Loss-of-function
TGF-β	Loss-of-function
P21	–

Twenty-four proteins in endogenous network were predicted to have a higher probability as loss-of-function in normal liver. Seven proteins in endogenous network were predicted to have a higher probability as gain-of-function in normal liver. This approach cannot predict the probable mutation of the reminding six genes (–)

normal liver attractor. Biologically, HNF4α, C/EBPα, and Foxa2 are liver-enriched transcription factors which regulate the synergistic transcriptional activation of hepatocyte specific genes [47, 69]. These liver-enriched transcription factors have mutual activations which form positive feedback loop and maintain normal liver tissue state [47, 69].Experiments also show that fibroblasts can be induced to functional hepatocyte-like cells by overexpressing these liver-specific transcription factors [32].The activated HNF4α-C/EBPα-Foxa2loop also maintains normal liver attractor by affecting the status of other functional modules, such as suppressing hepatocyte cell cycle and related metabolism [70, 71], suppressing inflammation [70], inducing differentiation of hepatocyte [47].

The model reveals that two positive feedback loops, RTKs/Ras, Akt, ERK, β-catenin/Myc, HIF and TNF-α, IL-1, IL-6/NF-κB (Kuppfer cell), are responsible for the establishment and maintenance of HCC attractor. Biologically, RTKs/Ras, Akt, ERK, β-catenin /Myc, HIF is a conserved growth-related positive feedback loop. Receptor tyrosine kinases (RTKs) are the high-affinity cell surface receptors including EGFR, VEGFR etc. Once activated, RTKS will lead to downstream activation of a number of common signaling molecules [72]. The activation of signaling pathways will change the gene expression profiles, the changes in gene expression in turn activate the RTKs ligands, in this way forms a positive feedback loop to support cell proliferation [73, 74].TNF-α, IL-1, IL-6/NF-κB (Kuppfer cell) is an inflammation-related positive feedback loop [75], pro-inflammatory stimuli, TNF and IL-1β, will activate the IκB kinase (IKK), resulting in IκB phosphorylation and leading to the nuclear entry of freed NF-κB dimmers in Kuppfer cell. Activated Kupffer cells produce a panel of inflammatory cytokines and growth factors including IL-1, IL-6, TNF-α, thus forming an

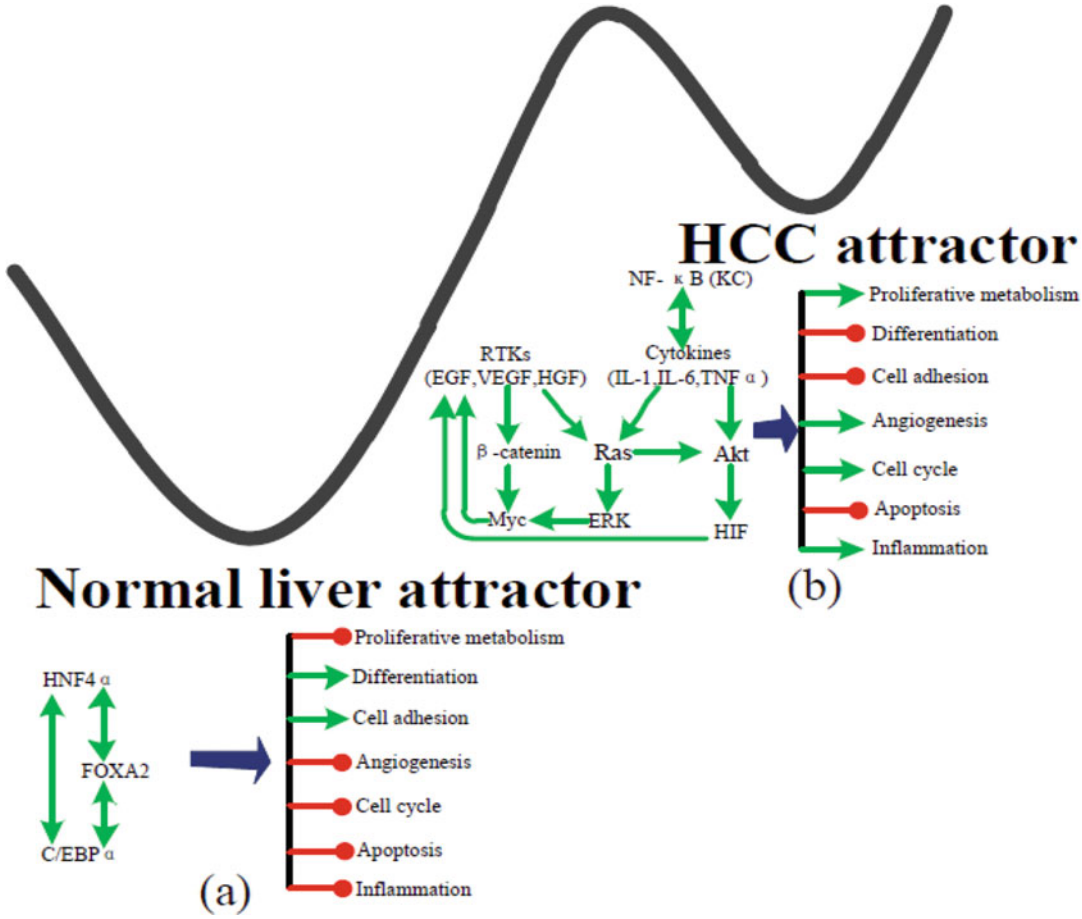


Fig. 3 Maintenance of normal liver and HCC by distinct positive feedback loops. (*a*) Liver-specific positive feedback loop HNF4α–C/EBPα–Foxa2 is responsible for the maintenance of normal liver attractor. (*b*) proliferation related positive feedback loops, RTKs–Ras, Akt, ERK, β-catenin–Myc, HIF and inflammation related positive feedback TNF-α, IL-1, IL-6–NF-κB (in Kupffer cell), are responsible for the maintenance of HCC attractor. HCC may be cured or relieved by inhibiting proliferation and inflammation, and simultaneously inducing liver differentiation

inflammation-related positive feedback loop [76]. The two activated positive feedback loops maintain HCC attractor by affecting the status of other functional modules, such as activating cell proliferation and growth [73, 74], activating angiogenesis [74], inhibiting cell apoptosis [73], inducing dedifferentiation [71].

One straightforward predication is that HCC may be induced to normal liver by inhibiting the proliferative-related feedback loops RTKs/Ras, Akt, ERK, β-catenin/Myc, HIF and inflammation-related feedback loops TNF-α, IL-1/NF-κB (Kuppfer cell)/TNF-α, IL-1, IL-6, and simultaneously activating differentiated liver-specific related loop, HNF4α /C/EBPα /Foxa2 (Fig. 3).The implication of this strategy means that we may induce HCC to normal liver by

inhibiting proliferation and inflammation, and inducing "cancer cell" differentiate into hepatocytes at the same time. There are several reasons for recommending this combined therapy to treat HCC. Firstl, each strategy has been successfully used to cure cancer respectively [77–81]. Proliferation and inflammation-related proteins have been approved by the US Food and Drug Administration (FDA) to treat certain human cancer [77, 81]. Liver-specific transcription factors are overexpressed to induce HCC differentiation has shown an anti-tumor effect [79, 80]. However, the separate treatment strategies may result in dilemma such as drug resistance [78], or just cannot obtain favorable results in clinical practice [80]. Our predication here reveals that we should use combined therapies, in particular we need to induce HCC differentiation. Second, similar strategies have been successfully used in a subtype of APL to make it from highly fatal to highly curable [82], our predication suggests that these concepts should be used in solid tumor. Third, clinical cases have independently reported the spontaneous regression of HCC, some spontaneous regression even reported without specific treatment and evidence of recurrence [83]; however, the spontaneous regression is a rare event. The predicated combined strategies here may make rare HCC spontaneous regression to a more probable phenomenon.

A second characteristic qualitative property of dynamical systems is hysteresis, in the present context which means that the genesis and regression is asymmetric. Expectedly, the model reveals that inducing a normal liver attractor to a HCC attractor may be achieved by activating the proliferative related feedback loops, RTKs/Ras, Akt, ERK, β-catenin/Myc, HIF and inflammation related feedback loops, TNF-α, IL-1, IL-6/NF-κB (Kuppfer cell). However, we cannot induce HCC to normal liver by inhibiting the two positive feedback loops, we still need to activate liver-specific positive feedbacks loop HNF4α/C/EBPα/Foxa2. The asymmetric genesis and regression of HCC is also due to the positive feedbacks loops which may be activated by consistent activation of one agent in the loop; however, inactivation of one agent in the loops often cannot inactivate the positive feedback loops. This mechanism may explain why many tumors will relapse and acquire drug resistance [20, 84].

3 Discussion

3.1 Endogenous Molecular-Cellular Network Hypothesis and Genetic Mutation Hypothesis

The proposed cancer endogenous molecular-cellular network hypothesis incorporates both genetic information and the biochemical reactions among endogenous molecular-cellular agents beyond the genomic information [26, 27]. On the genetic information level, the present hypothesis incorporates the hypothesis that cancer is a genetic disease. The gain and loss of functions via genetic mutations can be simply represented by removing and adding molecular-cellular agents or interactions in the endogenous

network [3, 85], such manipulation can be handled by standard procedures which pose no additional conceptual issues [27, 49]. Beyond the genomic information, the nonlinear dynamical biochemical interactions among the endogenous agents generate normal tissue attractors and cancer attractors. Due to the nonlinear biochemical interactions beyond genomic information, the present hypothesis holds that there is no simple one-to-one relationship between genotype and phenotype in complex diseases such as cancer. Instead, genotype is in general related to phenotypes by a very complex network of biochemical reactions [86, 87]. The genesis and progression of cancer is assumed as the transition from the intrinsic normal state to the intrinsic cancer state. A suitable quantity to describe such a process is the adaptive landscape [11].

Moreover, without any a priori knowledge of genetic mutation propensity in HCC, our results show that such network-level analyses are indeed a powerful approach to enable the prediction and a better understanding of genetic mutations in HCC. This illustrates the usefulness of network-level analyses as a means to predict and understand genetic mutations in cancers.

3.2 Working Endogenous Network and Typical Bioinformatics Network

The aim to establish a liver endogenous network is to reveal the core regulatory mechanisms of HCC genesis and progression at the systemic level. Many other high-throughput-based frameworks, such as ENCODE project [88], also have been proposed with similar aims. Theoretically, the regulatory mechanism can be deduced from high-throughput data if we have enough data [89]. Nevertheless, in reality the present genome-wide gene expression and protein interactions information are far from achieving this goal [18, 89]. Currently, analysis of high throughput is based more on statistics which can deduce the network topology and correlation between these molecular-cellular agents [88]. In the endogenous network construction, we solve this issue by making full use of the well-documented gene regulatory network and signaling transduction pathway which reflect our accumulated knowledge in molecular biology of liver.

We have realized the hypothesis by establishing working endogenous network for liver. Many other high-throughput-based frameworks also have been proposed to grasp the regulatory mechanism of biological systems, such as transcriptomics, ENCODE project, etc. [88]. Those frameworks play an important role in accelerating our understanding of biological systems. Nevertheless, it is unlikely that we can deduce the endogenous molecular-cellular network or other higher-level descriptions of a tissue solely from genome wide information about gene expression and physical interactions between proteins [18, 90, 91]. The quantitative analysis of high throughput is based on statistics, less on biological mechanism, they are often used to deduce the network topology or correlation between these molecular-cellular agents [88]. Compared with the network obtained

from high-throughput data by statistics, the endogenous molecular-cellular network, quantified by dynamical system, is a more ideal framework that reveals the regulatory mechanism of cancer genesis and progression.

Theories and hypotheses similar to the endogenous network hypothesis have been also proposed by others. The cracks in the cancer genetic mutation theory were also noticed and a similar theory predicting the intrinsic inevitability of cancer was proposed [92]. They have also been pushing along a similar qualification road [93]. In a different study it was proposed that ionizing radiation can give rise to similar effects through two distinct and independent routes, genetic, and epigenetic and that phenotype is represented by a stable attractor [94]. From the biological perspective all those theories share the same set of considerations. One major difference may be in quantitative formulations: To the best knowledge of the present authors a complete and consistent framework has been developed for the endogenous network theory. In addition, along our quantitative development the outstanding controversy on the adaptive landscape in evolutionary biology has been resolved [41].

More specific and related theories have also been proposed. It was proposed that cancer is an atavistic condition that occurs when genetic or epigenetic malfunction unlocks an ancient 'toolkit' of pre-existing adaptations [95, 96]. Such a theory is clearly consistent with the evolutionary dynamics structure embedded in the present endogenous network theory. During last few decades there has been a consistent effort to reveal the regulatory mechanism of developmental process in sea urchin [36]. Such a study is another important support to the construction of endogenous network. It should be pointed out after many years of theoretical and experimental studies, the key concepts of the endogenous network theory, such as landscape and states, start to get into mainstream biological and medical research [40, 97–100].

3.3 Quantitative Analysis of Endogenous Molecular-Cellular Network

The endogenous molecular-cellular network should be quantified by a stochastic nonlinear dynamical system [54, 101], recent progress allows us to ignore the stochastic effects at the first step [102, 103]. Characteristic properties of nonlinear dynamical systems declare that a distinct positive feedback loop must be existed to maintain normal liver and HCC, in light of the assumption that normal liver and HCC are distinct attractors of the endogenous molecular-cellular network of liver. A vivid and graphical description of the dynamical system is adaptive landscape which can depict the robustness of these attractors and the transition between the attractors intuitively [11]. Recent progress allows us to construct the adaptive landscape based on the endogenous molecular-cellular network [103, 104]. A quantitative description of the endogenous molecular-cellular network consists of a set of coupled stochastic differential equations [54, 101], despite that we have not explicitly

discussed the stochastic effect here. We have shown that the structure of the endogenous molecular-cellular network model is similar to the Morse-Smale dynamical system [53]. Adaptive landscape is a suitable quantity to describe the transition between the attractors, the robustness of these attractors, the trajectory and escape time from one attractor to another can be obtained from the landscape intuitively. And recent progress on stochastic dynamical systems [105–109] allows us to construct the adaptive landscape based on the endogenous molecular-cellular network [110–112].

3.4 Mutation Theory and Endogenous Network Theory: Further Remarks

Cancer mutation theory is essentially a statistical approach based on associations. Logically, it is not surprising that such a theory leads to a conclusion prevailing in the current literature: Every cancer is a complete different disease. There are enough mutations to do such classification: there are 10^9 base pairs in our genome. As a comparison we know that a little more than 100 chemical elements is already enough to make up the known material world. On the other hand, endogenous network is a biologically mechanistic dynamical theory. It assumes one network beyond all the observed phenomena: normal and abnormal. It is a grand unification theory in biology. We are not there yet: after more than 10 years' effort we have not reached the core network stage. The results already convincingly suggest a common core for all cancers, just the opposite of naively expected from the cancer mutation theory. In our view we have now the technology to get this one endogenous network.

Though the two theories are apparently quite different, we have already found that the endogenous network can predict mutation patterns. The predictions at the core network level are validated by experiments. We would like to point out another important prediction validated by the recent progress in cancer research, so far completely overlooked. In 2011 the next generation of cancer hallmarks was proposed [45]. An examination of the new cancer hallmarks reveals that they were already anticipated by the endogenous network theory. "Deregulating cellular energetics," "avoiding immune destruction," and "tumor promoting inflammation" were directly anticipated in the 2008 endogenous network proposal [26]. This counts for three out of four new hallmarks. The remaining hallmark, "genetic instability and mutations," was discussed earlier, even quantitatively [110]. It was not regarded as the essential part of endogenous network theory, but as one of its consequences. The recent prediction from the endogenous network theory further supports this conclusion [56].

On the other hand, sticking to cancer mutation theory can lead to confusions. The recent much circulated work based on such an associated study is a good example of such confusion: they are lost when anticipated correlation was not found between cancers and mutations [113]: In our view those experts do not really know what they are talking about, while such findings are easy to understand

within endogenous network theory. This and other recent developments strongly suggest that we need to look for new directions, to diversify our resource away from mutation theory.

3.5 Major Possible Caveats in Endogenous Molecular-Cellular Network Approach

Ten years ago there were two real and major problems standing on the way to the development of research program on endogenous network approach to cancer genesis and progression: first, whether or not the endogenous network would really exist; and second, whether or not there would be a functional landscape corresponding to the nonlinear stochastic dynamics of the network. After more than 10 years' effort, the second problem has been essentially solved: The corresponding functional landscape does exist, at least in the mathematical and/or theoretical sense. This second problem has now been replaced by a more practical and potentially important question: For a given endogenous network, how can we know for sure that all the major stable functional states (and all major transition states) would have been found? If not, some critical information on biology might have been missed. This will be a challenging problem on the computational side for years to come.

The first problem has not been solved very satisfactorily. We still do not know the endogenous network for a given biological system for sure. This will be a major effort for both experimental (cancer) biologists and computational modelers now and in the future. Nevertheless, it is clear that there does appear to have a common core network for all cancers. This finding forms a workable base for the further development. Based on the current evident, it is unlikely that the hypothesized endogenous network would not eventually be discovered. It may turn out even simpler than we would have imagined: In biology we have witnessed such solutions several times. One of the most famous is the discovery of double helix for heredity processes: Much simpler than biologists would have been anticipated.

4 Conclusion

Using HCC as an example we have realized the recently proposed cancer endogenous molecular-cellular network hypothesis. The working network of the liver was quantified by a set of nonlinear differentiation equations. We have then demonstrated that the working network reproduces the main features of the normal liver and HCC at both the modular and molecular levels, as a first set of evidence of the validity of the hypothesis. We explicitly obtain two additional testable predications for the further validation. Specifically, (1) the potential strategies to cure or relieve HCC may be inhibiting proliferation and inflammation, and simultaneously inducing liver differentiation; (2) the genesis and regression of

HCC is asymmetric. We explicitly discuss the gaps between the working endogenous network of liver and real liver at each step of realization, including the basic assumptions, working network construction, quantification, validations, and feasibility of the predications. In summary, these predications should be taken seriously when designing the strategies for HCC prevention, cure, and care. We would also suggest that the endogenous molecular-cellular network hypothesis may provide a suitable candidate, both qualitatively and quantitatively, to understand cancer genesis and progression.

We have discussed the basic biological considerations and the key mathematical features behind the cancer endogenous network theory. It is an effort to understand cancer genesis and progression in detail. The hypothesized theory's capturing of the underlying autonomous regulatory machinery is reviewed against two biological systems. By using one of the simplest and most important organisms, the Phage lambda, we have validated hierarchical nature for the endogenous molecular-cellular network of Phage lambda genetic switch. The core network can quantitatively describe the general regulatory machinery for Phage lambda's two important modes. Those validated insights may be used in the more complex systems.

One of the most important applications of this hypothesis is to understand the regulatory machinery underlying cancer genesis and progression. Workflow to construct the endogenous molecular-cellular network of HCC and quantitative analysis of the network were provided. The endogenous molecular-cellular network hypothesis suggests that cancer is an intrinsic state shaped by evolution, the genesis and progression of cancer is the transition from the intrinsic normal state to the intrinsic cancer state and the progression of cancer is not arbitrary, transition from normal to cancer needs to pass through the critical saddle points; this may provide us with quantitative standard for early state cancer detection. With its capacity to take both the genetic and environmental effects into consideration the endogenous network theory may provide a best candidate, both qualitatively and quantitatively, to understand cancer genesis and progression.

References

1. Hajdu SI (2011) A note from history: landmarks in history of cancer, part 1. Cancer 117 (5):1097–1102
2. Nowell PC (1976) The clonal evolution of tumor cell populations. Science (New York, NY) 194(4260):23–28
3. Greaves M, Maley CC (2012) Clonal evolution in cancer. Nature 481(7381):306–313
4. Hou Y et al (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell 148(5):873–885
5. Land H, Parada LF, Weinberg RA (1983) Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. Nature 304(5927):596–602

6. Weinberg RA (1995) The retinoblastoma protein and cell cycle control. Cell 81:323–330
7. Paget S (1989) The distribution of secondary growths in cancer of the breast. 1889. Cancer Metastasis Rev 8(2):98–101
8. Fidler IJ, Poste G (2008) The "seed and soil" hypothesis revisited. Lancet Oncol 9(8):808
9. Wang X et al (2011) Residual embryonic cells as precursors of a Barrett's-like metaplasia. Cell 145(7):1023–1035
10. Koshland DE, Goldbeter A, Stock JB (1982) Amplification and adaptation in regulatory and sensory systems. Science (New York, NY) 217(4556):220–225
11. Ao P (2009) Global view of bionetwork dynamics: adaptive landscape. J Genet Genomics 36(2):63–73
12. Zhu XM et al (2004) Robustness, stability and efficiency of phage lambda genetic switch: dynamical structure analysis. J Bioinform Comput Biol 2(4):785–817
13. Li F et al (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci USA 101(14):4781–4786
14. Zhu X et al (2007) Efficiency, robustness, and stochasticity of gene regulatory networks in systems biology: λ switch as a working example. In: Choi S (ed) Introduction to systems biology. Humana, New York, pp 336–371
15. Bizzarri M et al (2011) Fractal analysis in a systems biology approach to cancer. Semin Cancer Biol 21(3):175–182
16. Pastan I, Gottesman M (1987) Multiple-drug resistance in human cancer. N Engl J Med 316 (22):1388–1393
17. Gimbrone MA et al (1972) Tumor dormancy in vivo by prevention of neovascularization. J Exp Med 136(2):261–276
18. Hartwell LH et al (1999) From molecular to modular cell biology. Nature 402(6761 Suppl):C47–C52
19. Akhurst RJ, Derynck R (2001) TGF-β signaling in cancer—a double-edged sword. Trends Cell Biol 11(Supplement 1):S44–S51
20. Feng G-S (2012) Conflicting roles of molecules in hepatocarcinogenesis: paradigm or paradox. Cancer Cell 21(2):150–154
21. Kauffman S (2008) Control circuits for determination and transdetermination: interpreting positional information in a binary epigenetic code. In: Ciba foundation symposium 29 - cell patterning. Wiley, Chichester, pp 201–221
22. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med 1(1):2
23. Hood L, Flores M (2012) A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. N Biotechnol 29(6):613–624
24. Chalancon G et al (2012) Interplay between gene expression noise and regulatory network architecture. Trends Genet 28(5):221–232
25. Vital-Lopez FG, Memišević V, Dutta B (2012) Tutorial on biological networks. Wiley Interdiscip Rev Data Min Knowl Discov 2(4):298–325
26. Ao P et al (2008) Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. Med Hypotheses 70(3):678–684
27. Wang G et al (2013) From Phage lambda to human cancer: endogenous molecular-cellular network hypothesis. Quant Biol 1 (1):32–49
28. Yuan R et al (2017) Cancer as robust intrinsic state shaped by evolution: a key issues review. Rep Prog Phys 80(4):042701
29. Garber K (2001) Beyond the Nobel prize: cell cycle research offers new view of cancer. J Natl Cancer Inst 93(23):1766–1768
30. Nurse P (2000) A long twentieth century of the cell cycle and beyond. Cell 100(1):71–78
31. Takahashi K et al (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131(5):861–872
32. Huang P et al (2011) Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. Nature 475(7356):386–389
33. Ferrell JE Jr, Tsai TY-C, Yang Q (2011) Modeling the cell cycle: why do certain circuits oscillate? Cell 144(6):874–885
34. Spencer SL, Sorger PK (2011) Measuring and modeling apoptosis in single cells. Cell 144 (6):926–939
35. Meyer BJ, Maurer R, Ptashne M (1980) Gene regulation at the right operator (OR) of bacteriophage λ: II. OR1, OR2, and OR3: their roles in mediating the effects of repressor and cro. J Mol Biol 139(2):163–194
36. Yuh C-H, Bolouri H, Davidson EH (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279(5358):1896–1902
37. Baker SG, Kramer BS (2011) Systems biology and cancer: promises and perils. Prog Biophys Mol Biol 106(2):410–413

38. Alberts B et al (2007) Molecular biology of the cell. Garland Science, New York
39. Smale S, Hirsch MW, Devaney RL (2003) Differential equations, dynamical systems, and an introduction to chaos. Elsevier Science, Amsterdam
40. Bar-Yam Y, Harmon D, de Bivort B (2009) Attractors and democratic dynamics. Science 323(5917):1016–1017
41. Ao P (2005) Laws in Darwinian evolutionary theory. Phys Life Rev 2(2):117–156
42. Matthias S, Sabine W (2008) Cancer as an overhealing wound: an old hypothesis revisited. Nat Rev Mol Cell Biol 9:628–638
43. Williams CS, Mann M, DuBois RN (1999) The role of cyclooxygenases in inflammation, cancer, and development. Oncogene 18 (55):7908–7916
44. El-Serag HB, Rudolph KL (2007) Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. Gastroenterology 132 (7):2557–2576
45. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144 (5):646–674
46. Alberts B et al (2002) Molecular biology of the cell. Garland Science Taylor & Francis Group, New York
47. Monga SPS (2010) Molecular pathology of liver diseases. Springer, New York
48. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15(2):221–231
49. Ao P et al (2010) Towards predictive stochastic dynamical modeling of cancer genesis and progression. Interdiscip Sci Comput Life Sci 2 (2):140–144
50. Shea MA, Ackers GK (1985) The OR control system of bacteriophage lambda: a physical-chemical model for gene regulation. J Mol Biol 181(2):211–230
51. Ao P et al (2007) Generic enzymatic rate equation under living conditions. J Biol Syst 15(04):495–514
52. Huang Y et al (2012) Identification of a two-layer regulatory network of proliferation-related microRNAs in hepatoma cells. Nucleic Acids Res 40(20):10478–10493
53. Holmes P (2005) Ninety plus thirty years of nonlinear dynamics: less is more and more is different. Int J Bifurcat Chaos 15 (09):2703–2716
54. Zhu X-M et al (2004) Calculating biological behaviors of epigenetic states in the phage λ life cycle. Funct Integr Genomics 4 (3):188–195
55. Lei X et al (2015) Biological sources of intrinsic and extrinsic noise in ci expression of lysogenic phage lambda. Sci Rep 5:13597
56. Wang G et al (2016) Endogenous network states predict gain or loss of functions for genetic mutations in hepatocellular carcinoma. J R Soc Interface 13(115):20151115
57. Rosenberg SM (2001) Evolving responsively: adaptive mutation. Nat Rev Genet 2 (7):504–515
58. Cairns J, Overbaugh J, Miller S (1988) The origin of mutants. Nature 335 (6186):142–145
59. Cairns J (1998) Mutation and cancer: the antecedents to our studies of adaptive mutation. Genetics 148(4):1433–1440
60. Bamford S et al (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91(2):355–358
61. Forbes SA et al (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res 38(Database issue):D652–D657
62. Nault J-C, Zucman-Rossi J (2014) Genetics of hepatocellular carcinoma: the next generation. J Hepatol 60(1):224–226
63. Rahman N (2014) Realizing the promise of cancer predisposition genes. Nature 505 (7483):302–308
64. Prives CL (2014) Abstract PL01-02: The two faces of p53: tumor suppressor and oncogene. Cancer Res 74(19 Supplement):PL01-02–PL01-02
65. Rubin H (2006) What keeps cells in tissues behaving normally in the face of myriad mutations? Bioessays 28(5):515–524
66. Versteeg R (2014) Cancer: tumours outside the mutation box. Nature 506 (7489):438–439
67. Ingolia NT, Murray AW (2007) Positive-feedback loops as a flexible biological Module. Curr Biol 17(8):668–677
68. Wang G et al (2014) Quantitative implementation of the endogenous molecular–cellular network hypothesis in hepatocellular carcinoma. Interface Focus 4(3):20130064
69. Odom DT et al (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science 303(5662):1378–1381
70. Wang H et al (2001) C/EBPα arrests cell proliferation through direct inhibition of Cdk2 and Cdk4. Mol Cell 8(4):817–828

71. Dang CV (2012) MYC on the path to cancer. Cell 149(1):22–35
72. Lemmon MA, Schlessinger J (2010) Cell signaling by receptor tyrosine kinases. Cell 141 (7):1117–1134
73. Manning BD, Cantley LC (2007) AKT/PKB signaling: navigating downstream. Cell 129 (7):1261–1274
74. Semenza GL (2010) HIF-1: upstream and downstream of cancer metabolism. Curr Opin Genet Dev 20(1):51–56
75. Karin M (2009) NF-κB as a critical link between inflammation and cancer. Cold Spring Harb Perspect Biol 1(5):a000141
76. Grivennikov SI, Greten FR, Karin M (2010) Immunity, inflammation, and cancer. Cell 140 (6):883–899
77. Bukowski RM, Yasothan U, Kirkpatrick P (2010) Pazopanib. Nat Rev Drug Discov 9 (1):17–18
78. Flaherty KT, Yasothan U, Kirkpatrick P (2011) Vemurafenib. Nat Rev Drug Discov 10(11):811–812
79. Yin C et al (2008) Differentiation therapy of hepatocellular carcinoma in mice with recombinant adenovirus carrying hepatocyte nuclear factor-4α gene. Hepatology 48 (5):1528–1539
80. Zeng X et al (2011) Recombinant adenovirus carrying the hepatocyte nuclear factor-1alpha gene inhibits hepatocellular carcinoma xenograft growth in mice. Hepatology 54 (6):2036–2047
81. Younes A et al (2010) Brentuximab vedotin (SGN-35) for relapsed CD30-positive lymphomas. N Engl J Med 363(19):1812–1821
82. Wang Z-Y, Chen Z (2008) Acute promyelocytic leukemia: from highly fatal to highly curable. Blood 111(5):2505–2515
83. Stoelben E et al (1998) Spontaneous regression of hepatocellular carcinoma confirmed by surgical specimen: report of two cases and review of the literature. Langenbecks Arch Surg 383(6):447–452
84. Pratilas CA, Solit DB (2010) Targeting the mitogen-activated protein kinase pathway: physiological feedback and drug response. Clin Cancer Res 16(13):3329–3334
85. Nowell P (1976) The clonal evolution of tumor cell populations. Science 194 (4260):23–28
86. Wright S (1932The roles of mutation, inbreeding, crossbreeding and selection in evolution. In Proceedings of the sixth international congress on genetics
87. Waddington CH (1957) The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. Allen & Unwin, London, pp ix–262
88. Gerstein M et al (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489(7414):91–100
89. Glass L (1975) Combinatorial and topological methods in nonlinear chemical kinetics. J Chem Phys 63(4):1325
90. Hoenerhoff MJ et al (2011) Global gene profiling of spontaneous hepatocellular carcinoma in B6C3F1 mice: similarities in the molecular landscape with human liver cancer. Toxicol Pathol 39(4):678–699
91. Lovén J et al (2012) Revisiting global gene expression analysis. Cell 151(3):476–482
92. Huang S (2011) On the intrinsic inevitability of cancer: from foetal to fatal attraction. Semin Cancer Biol 21(3):183–199
93. Zhou JX et al (2012) Quasi-potential landscape in complex multi-stable systems. J R Soc Interface 9(77):3539–3553
94. Baverstock K, Karotki AV (2011) Towards a unifying theory of late stochastic effects of ionizing radiation. Mutat Res 718(1–2):1–9
95. Davies PCW, Lineweaver CH (2011) Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. Phys Biol 8(1):015001
96. Vincent MD (2011) Cancer: beyond speciation. Adv Cancer Res 112:283–350
97. Chaffer CL et al (2011) Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. Proc Natl Acad Sci USA 108(19):7950–7955
98. Li S et al (2015) Endogenous molecular network reveals two mechanisms of heterogeneity within gastric cancer. Oncotarget 6 (15):13607–13627
99. Yuan R et al (2016) Core level regulatory network of osteoblast as molecular mechanism for osteoporosis and treatment. Oncotarget 7(4):3692
100. Yuan R et al (2016) From molecular interaction to acute promyelocytic leukemia: calculating leukemogenesis and remission from endogenous molecular-cellular network. Sci Rep 6:24307
101. Qian H (2013) Stochastic physics, complex systems and biology. Quant Biol 1(1):50–53
102. Ao P (2004) Potential in stochastic differential equations: novel construction. J Phys A Math Gen 37(3):L25

103. Ao P, Kwon C, Qian H (2007) On the existence of potential landscape in the evolution of complex systems. Complexity 12(4):19–27
104. Wang J et al (2011) Quantifying the Waddington landscape and biological paths for development and differentiation. Proc Natl Acad Sci USA 108(20):8257–8262
105. Tang Y, Yuan R, Ma Y (2013) Dynamical behaviors determined by the Lyapunov function in competitive Lotka-Volterra systems. Phys Rev E Stat Nonlin Soft Matter Phys 87 (1):012708
106. Tang Y, Yuan R, Ao P (2014) Summing over trajectories of stochastic dynamics with multiplicative noise. J Chem Phys 141(4):044125
107. Tang Y, Yuan R, Ao P (2014) Nonequilibrium work relation beyond the Boltzmann-Gibbs distribution. Phys Rev E Stat Nonlinear Soft Matter Phys 89(6):062112
108. Tang Y et al (2015) Work relations connecting nonequilibrium steady states without detailed balance. Phys Rev E Stat Nonlinear Soft Matter Phys 91(4):042108
109. Tang Y, Yuan R, Ao P (2015) Anomalous free energy changes induced by topology. Phys Rev E Stat Nonlin Soft Matter Phys 92 (6):062129
110. Ao P (2007) Orders of magnitude change in phenotype rate caused by mutation. Cell Oncol 29(1):67–69. author reply 71–2
111. Tang Y et al (2016) Potential landscape of high dimensional nonlinear stochastic dynamics and rare transitions with large noise. arXiv preprint arXiv:1611.07140
112. Tang Y, Yuan R, Ao P (2014) Nonequilibrium work relation beyond the Boltzmann-Gibbs distribution. Phys Rev E 89(6):062112
113. Tomasetti C, Li L, Vogelstein B (2017) Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science 355 (6331):1330–1334

Chapter 12

A Network-Based Integrative Workflow to Unravel Mechanisms Underlying Disease Progression

Faiz M. Khan, Mehdi Sadeghi, Shailendra K. Gupta, and Olaf Wolkenhauer

Abstract

Unraveling mechanisms underlying diseases has motivated the development of systems biology approaches. The key challenges for the development of mathematical models and computational tool are (1) the size of molecular networks, (2) the nonlinear nature of spatio-temporal interactions, and (3) feedback loops in the structure of interaction networks. We here propose an integrative workflow that combines structural analyses of networks, high-throughput data, and mechanistic modeling. As an illustration of the workflow, we use prostate cancer as a case study with the aim of identifying key functional components associated with primary to metastasis transitions. Analysis carried out by the workflow revealed that HOXD10, BCL2, and PGR are the most important factors affected in primary prostate samples, whereas, in the metastatic state, STAT3, JUN, and JUNB are playing a central role. The identified key elements of each network are validated using patient survival analysis. The workflow presented here allows experimentalists to use heterogeneous data sources for the identification of diagnostic and prognostic signatures.

Key words Integrative workflow, Network-based analysis, Large-scale networks, Disease signatures, Mathematical models

1 Introduction

To understand various processes associated with the progression of complex diseases, systems biology-based methods usually begin with the gathering of information from the literature and databases, summarizing components and their interactions relevant for the process under consideration. The information gathered is summarized in interaction maps, which serve as a knowledge-base and being machine readable is amenable to computational analysis. Tumor is one of the complex diseases where mutated and epigenetically modified genes are highly patient and tumor type dependent, and more importantly these genes are integrated in a small set of regulatory pathways [1–3].

Faiz M. Khan, Mehdi Sadeghi and Shailendra K. Gupta contributed equally to this work.

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_12,

Analyzing the structure of biochemical disease networks provides useful information including network hubs, regulatory motifs, and possibly global features like small world organization of the system. Regulatory motifs, feedback, and feedforward loops are a source for nonlinear regulatory behavior [4, 5], which not only challenges human intuition but also limits the application of conventional data analysis tools [6–8]. Moreover, for large-scale biochemical networks a dynamical analysis is particularly difficult with mechanistic (e.g., ODE-based) approaches from the theory of dynamical systems. In order to exploit the advantages of large-scale biochemical networks, in combination with mechanistic modeling, we need integrative approaches and computational workflows to identify disease-specific small regulatory/function modules that can be subjected to a more detailed analysis, followed by the prediction of molecular signatures. Exploring large-scale nonlinear dynamical networks will remain an art form. What we are aiming for here is a rational approach to what is effectively guesswork, forced upon us by the wonderful complexity found in living systems.

In this chapter, we highlight and discuss an integrative workflow (Fig. 1) to study large-scale biochemical disease networks by combining techniques from bioinformatics and systems biology. Integrating experimental and clinical data with the workflow, process-specific hypotheses can be generated and validated. In particular, we present here a flexible and extendible workflow that combines network structural properties with high-throughput and other biomedical data to identify smaller modules/molecular signatures for tumor-specific disease phenotypes. A mathematical model of the identified smaller modules/molecular signatures can be constructed to give mechanistic understanding of the disease and propose new hypotheses, which are subject to experimental validation. For the illustration of the workflow, we used prostate cancer as a case study with the aim of identifying key functional components involved in regulation and progression of primary to metastasis transitions [9]. We construct a network for each of the clinical states of prostate cancer based on differentially expressed

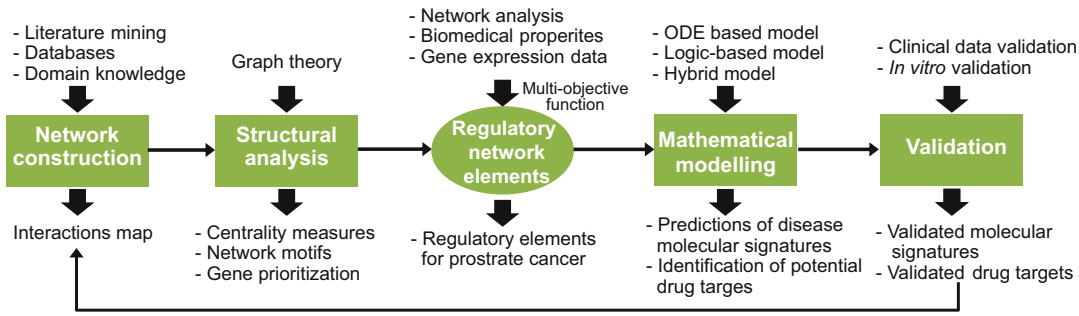


Fig. 1 An integrative workflow to analyze large-scale biochemical disease networks

and significantly correlated gene, miRNA and TF pairs from the patient data. Using the workflow as shown in Fig. 1, we first generated gene/miRNA/transcription factor regulatory networks for primary and metastatic stages of prostate cancer and identified key regulatory interactions responsible for the transitions from primary to metastatic tumor stage by integrating patient-derived gene and microRNA expression data. Analysis carried out by the workflow revealed that *HOXD10*, *BCL2,* and *PGR* are the most important factors affected in primary prostate samples, whereas, in the metastatic state, *STAT3*, *JUN,* and *JUNB* are playing a central role. The identified key elements of each network are validated using patient survival analysis. Our integrative analyses on the disease network also suggest that some of these molecules are targeted by differentially expressed miRNAs which may have a major effect on the dysregulation that led to the disease progression. We observed that in metastatic prostate tumor, five miRNAs (*miR-671-5p*, *miR-665*, *miR-663*, *miR-512-3p,* and *miR-371-5p*) are mainly responsible for the dysregulation of *STAT3,* an important player in the tumor metastasis. These observations provide an opportunity for early detection of metastasis and development of alternative therapeutic approaches.

Ultimately, the integrative workflow discussed in this chapter supports deciphering mechanisms underlying complex diseases. As such, it cannot provide an exact representation of cellular events but nevertheless guides the formation of hypotheses and their validation in experiments.

The specific objectives of this chapter are:

- Review of the existing tools and approaches for the analyses of large-scale biochemical networks.
- Construction of prostate cancer network.
- Integrative workflow to identify tumor-specific *signatures.*
- Validation of the identified signatures through experiments/clinical data.

2 Material and Methods

2.1 The Systems Biology Approach

Biological processes are complex, involving a large variety of components that interact in a nonlinear fashion in space and time. The systems biology approach combines experiments with computational tools and methods to understand such complex processes [10, 11]. We consider the systems biology approach as an interdisciplinary collaboration that realizes an iterative cycle of data-driven modeling and model-driven experimentation (Fig. 2). A research project taking a systems biology approach often starts by gathering information about a process from the literature and databases. This

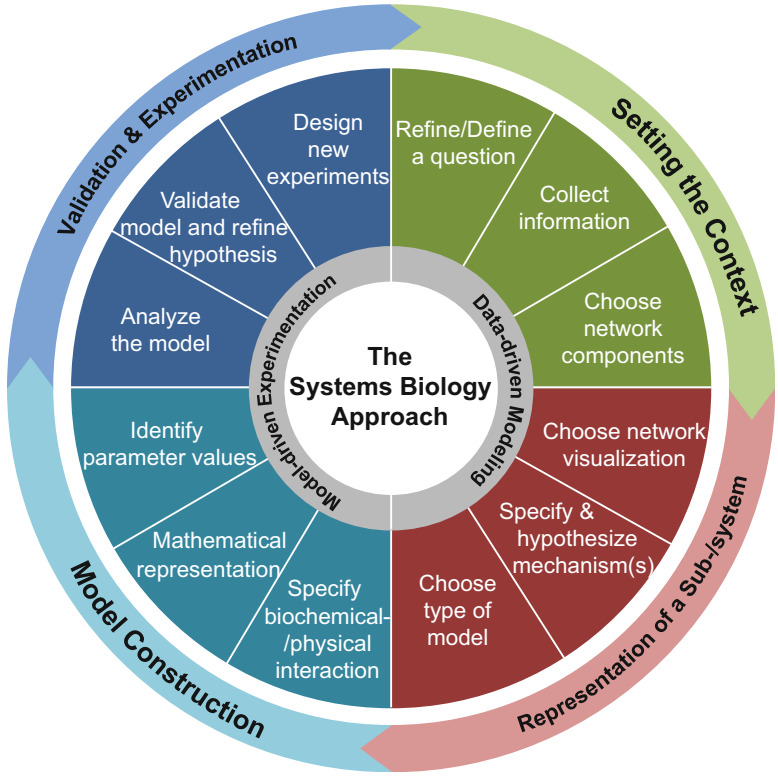


Fig. 2 The systems biology approach: an iterative process of data-driven modeling and model-driven experimentation

is organized and represented in a form of machine readable network, which is the formalized representation of a large number of individual experimental results. The network is then modeled with suitable modeling formalisms, which after calibrating with experimental data should recapitulate the biological phenomena under consideration, is used to formulate or validate hypotheses, and is used to support the design of new experiments. In this way, the systems biology approach cycles into data-driven modeling and model-driven experimentation. More specifically, we can divide this approach into four main stages:

2.1.1 Setting the Context This step starts with the formulation of a biological question that is investigated. For example, a general question could be what the regulatory mechanism(s) underlying prostate tumor metastasis or drug resistance are? This defines the project boundaries and gives directions to collect information from the literature and databases. This information is then converted into a machine readable format (*i.e.*, in the form of network) for computational analysis. With the help of domain experts or using computational methods, project-specific network components or modules are then chosen.

2.1.2 Representation of a Sub-/system

Mapping out the interactions among biochemical entities as a network provides a platform for structural and dynamical analysis of the system. A large set of network visualization tools (e.g., CellDesigner [12], Cytoscape [13], VANTED [14]) are used to construct the networks, representing biological processes from abstract to more detailed level, depending on the requirements of biological question and available knowledge. Based on available knowledge and the domain expert's opinion, the interaction mechanisms (e.g., activation or inactivation) among molecular entities are defined or hypothesized. Depending on network size and kinetic details, a suitable modeling formalism is chosen to analyze the dynamics of the systems for input stimuli and different perturbations.

2.1.3 Model Construction

This step starts with the detailed description of interaction mechanisms providing the biochemical and biophysical information. For example, biochemical interactions characterize the activation/inactivation in terms of phosphorylation/dephosphorylation, and biophysical interactions describe about what enzyme or catalyst regulates the reaction. For dynamical analyses, the interactions are represented by a system of mathematical equations, which we refer to as the mathematical model. Model parameter values are identified and characterized from available biological information [15] and databases like SioABIO-RK [16] and BioModels [17].

2.1.4 Validation and Experimentation

After the identification of model parameter values, analytical tools (e.g., bifurcation and sensitivity analysis) are used to analyze the model's dynamical behavior, stability, and robustness. Then the model is calibrated with biological data by refining its parameter values to recapitulate the biological reality. After successful calibration of the model, new hypotheses are made using predictive simulations that unravel regulatory mechanisms underlying complex processes. Hypotheses made by model simulations need to be validated by designing new experiments. If the model predictions are validated by experiments, it will provide a reasonable explanation of the biological phenotypes and sharpen our understanding of the complex processes that generate them.

2.2 Network Construction

Processes in living cells are carried out by complex interactions among biological elements such as genes, proteins, RNAs, mRNAs, enzymes, transcription factors, and other molecules. To understand the mechanisms behind normal and malfunctional execution of processes (linked to diseases), it is necessary to have a blueprint of these interactions in the form of network (i.e., vertices connected by edges). The construction of networks is a painstaking exercise of manual validation and encoding. Various approaches are used to map out molecular interactions underlying certain biological processes. A number of computational techniques have

Table 1
Important databases for retrieving interactions for various biochemical interactions

Type of network	Databases
Gene regulatory networks (GRN)	KEGG, TRANSFAC, TRED, TransPath
Metabolic networks	KEGG, BioCyc, MetaCyc, BRENDA, BiGG, metaTIGER
Protein-protein interaction (PPI)	BioGrid, HPRD, STRING, IntAct, DIP, MIPS
Signal transduction networks (STN)	PID, BioCarta, SPIKE, WikiPathways, CST Signaling Pathways, The Cell Collective, iHOP, SignaLink, NetPath
MicroRNA interaction network	miRecords, TarBase, miRTarBase, miRWalk, miRGen, TransmiR, UCSC browse

been developed to infer biochemical networks from high-throughput data [18, 19]. Despite their importance, these techniques face many challenges, for example false positive and false negative interactions are one of the main challenges that influence the results of inference. On the other hand, machine learning algorithms for inference are complex; but simple methods, like Naive Bayes, may not work well for complex situations, they are slow to train and prone to overfit [19]. A more detailed and highly focused network, centered around particular disease [20, 21] or process [22–24], can be constructed by expert domain knowledge (functional and structural information), diligent manual search for published literature, and publically available databases (*see* Table 1 for some of available databases for retrieving interaction for different types of biochemical networks). To avoid the laborious manual curation for network construction, some methods are developed to automatically reconstruct networks by retrieving interactions or sub-networks from existing maps and models [25, 26]. Combining automatic reconstruction with domain knowledge, manual search of literature and databases would provide a reasonable strategy to construct detailed and fully annotated large-scale biochemical networks (*see* Fig. 3).

All these maps are a formalized representation of information that can subsequently be analyzed with computational algorithms. They are organized interactions knowledge-bases which help in: (1) Gathering disperse information about complex biological systems at one place. (2) Managing and organizing information in a standard pathway diagram format that is helpful to conceptually analyze and intuitively visualize the network components. (3) Provide information about the interactions to develop hypothesis that can experimentally be tested. (4) Provide a foundation to derive simulation models to analyze the dynamics of interacting components. Their structural analysis allows the identification of functional modules [1–3], regulatory motifs [4, 5], and hub nodes [5, 27–29] along

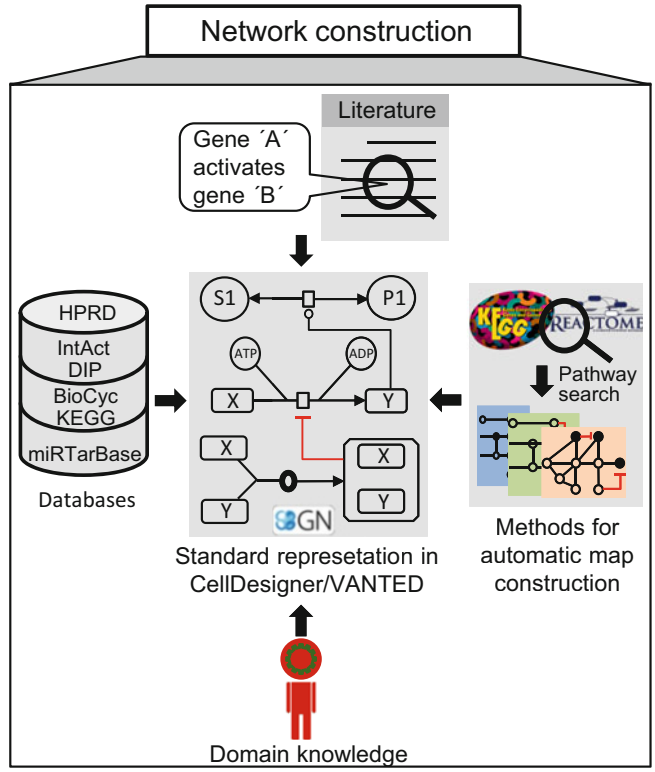


Fig. 3 Scheme for biochemical network construction. More detailed and highly focused networks, centered around a particular disease or process, can be constructed by expert domain knowledge (functional and structural information), diligent manual search of published literature, and publically available databases (e.g., HPRD, IntAct, DIP, BioCyc, KEGG, REACTOME, and miRTarBase)

with critical interactions that might be critical for specific phenotype. The network-based approach with the aim of getting insights into the mechanisms underlying processes dysregulated in diseases is almost the "state of the art" now. For example, Matsuoka et al. created a large detailed pathway map of influenza A virus replication cycle [20]. The map is annotated with around 500 scientific articles and includes information from already developed influenza maps and incorporate pathways information from KEGG, PANTHER, and Reactom databases. This study was intended to develop a broader picture of functional mechanism of influenza A virus and its associated host response. Further, the map was used for in silico *analysis* to identify several critical targets for influenza A virus life cycle. Calzone et al. constructed a comprehensive map of RB/E2F pathway interactions in the regulation of cell cycle [23]. It contains more detailed and systematic information than any general purpose databases about the study and serves as a knowledge-base. They identified different structural modules in the map based on clusters of relevant cycles in the reaction graph.

2.3 Network Analysis

After gathering and managing information regarding biological systems in the form of interactions networks, structural and dynamical analysis provides useful information about the network architecture and the dynamics of regulatory pathways. The structural analysis of networks allows the identification of functional modules, regulatory motifs (including feedback and feed-forward loops), and node properties (including node degree (ND) and betweenness centrality (BC)). In biochemical networks, nodes can be protein, gene, miRNA, etc. Modules are the aggregations of the densely interconnected neighboring nodes. It is observed that the functionally related nodes are located in close proximity and thus form functional modules [30, 31]. These functionally related genes could be associated with the same biological pathway and have similar effects on certain disease phenotype (which fits to the proverb "guilt by association") and may be targeted by structurally similar drugs [32, 33].

Biological networks are enriched in recurring structural patterns called motifs. Network motifs are the interacting patterns that recur significantly more often than in random networks [4]. These motifs are sort of small molecular circuitry that the cell uses to process information and governing dynamic response to external or internal fluctuations [34, 35]. Feedback and feedforward loops (FBL and FFL) are the important regulatory network motifs. Feedback loops are characterized by direct/indirect inhibition/activation of a node by its own target, e.g., Fig. 4a, b show the indirect activation/inhibition of node "X" by its own target. FBLs can be either positive or negative depending on the parity of negative links in the loop. If parity is odd the FBL is negative (Fig. 4a) and if it is even then the FBL is positive (Fig. 4b). Feedforward loops are characterized by interactions in which a node is a mutual target of a node and its target, e.g., in Fig. 5a, b where "*X*" regulates "*Y*," and then "*X*" and "*Y*" mutually target "*Z*." FFLs can either be

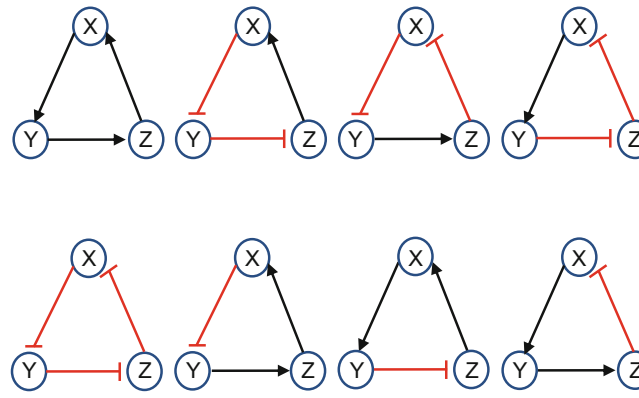


Fig. 4 Representation of all possible feedback loops (either positive or negative) in a three-node network

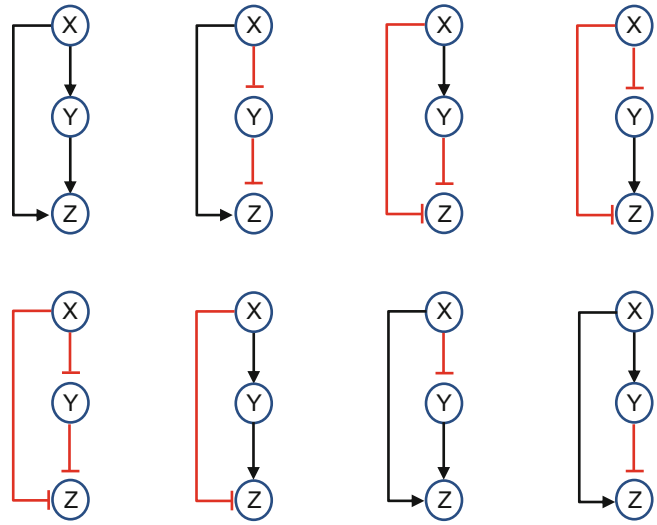


Fig. 5 Representation of all possible feedforward loops (either coherent or in-coherent) in a three-node network

coherent or in-coherent depending on the parity of negative links in the loop. FFL is coherent (Fig. 5a) if the parity is even and it is in-coherent (Fig. 5b) when the parity is odd.

As sets of genes involved in certain phenotypes are highly interconnected and regulate each other through coherent and incoherent regulatory loops (motifs) from different pathways, the analysis of these can provide insights into the structure and dynamics of the network [5, 35] followed by the identification of disease biomarkers [36, 37].

The other important network topological parameter of a node is betweenness centrality, which indicates the sum of the number of shortest paths from all vertices to all others pass through that node. Node with high BC serves as a gate keeper in the communication between different components in a network, for example in Fig. 6 node "A" connects the left and right parts of a network, so it gets a highest BC value [38] which is a non-intuitive behavior and plays a crucial role in controlling the dynamics of a system [4, 34]. Node properties have a significant role in the network topology [5, 27–29]. For example, networks with degree (number of edges connected to a node) distribution follow a power law $P(ND) \sim (ND)^{-r}$, where r is an approximated parameter whose value ranges in $2 \leq r < 3$, called scale-free networks. There are two important characteristics of scale-free networks; first they contain "hubs," nodes comprising many more connections than others; second due to hubs, these networks are heterogeneous in terms of node degree and considered to be robust against single-random perturbation.

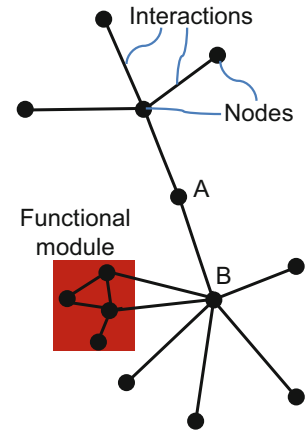


Fig. 6 Illustration of network topological properties: The circles represent 'Nodes' and the lines/edges between them portray the 'Interactions' of the network. The node 'A' act as a bridge between upper and lower part of the network therefore it stands out as a high BC node. The node 'B' has large number of edges so it gets high value of node degree. The sub-network highlighted by red rectangle termed as 'Functional module' means the constituent nodes take part in a similar cellular function

The network structure analysis revealed that the functioning and regulation of a network are governed by a certain set of organizing principles [5]. To understand the mechanisms of these organizing principles, the mechanistic dynamical models are used to analyze the systems. The dynamical properties of a network characterize the temporal behavior of a network under certain conditions, so as to understand the nature of regulation of interacting components in response to perturbations that affect the functionality of the cells, and ultimately their consequences on the cellular phenotype.

2.4 Selection of a Mathematical Modeling Formalism

Recent advances in omics technology and the availability of databases with information about interactions among proteins, genes, and miRNA, make it possible to model cellular processes as biochemical networks [20, 21, 23, 24, 39]. These networks provide useful information, such as the identification regulatory motifs and hub nodes, but their utility is confined to a static analysis that does not help explain "cause and effect" relationships, which are ubiquitous in biological systems [40]. Moreover, it is impossible to characterize complex network structures, like feedback/feedforward loops and cross-talk in network modules, which induce counter-intuitive behavior in systems dynamics, from the network connectivity only. Nevertheless, networks are an important step toward understanding system and provide a foundation for the development of dynamical models [6]. The dynamical analysis helps in understanding the functioning of a system and supports the

formulation of new hypotheses about the effect of specific internal or external perturbations in a system [40]. Using a systems biology approach (Fig. 2), the iterative cycle of data-driven modeling and model-driven experimentation refines formulated hypotheses until they are validated.

"Models" are an abstract representation of reality which provide a reliable sense (i.e., understanding the behavior) of the original system depending on available information and the purpose of modeling [6, 40, 41]. "Modeling" is the process of creation and usage of a model [41]. Mathematical models describe the reality (i.e., processes in a cell) in terms of functions or equations which contain variables and parameters: $f(X_1,\ldots,X_N; k_1,\ldots,k_N)$, where f is the function that evaluates, e.g., the temporal behavior of a system depending on the variables $X_1,\ldots,X_N$ and parameters $k_1,\ldots,k_N$. Variables are the quantity of interest in model analysis which typically change over time, e.g., concentration of protein in a cell. Parameters are quantities which are fixed for a given computational experiment to characterize specific quantitative behavior of a model. Parameter values are typically characterized from the literature, databases, for example BioModels and SABIO-RK, and can be estimated from experimental data by calibrating the model to recapitulate the real biological process [15, 42]. After calibrating the model with certain experimental data, it can perform a large set of repetitive in silico experiments for many different conditions that may be quite time-consuming and expensive with wet-lab experiments. Models can be created at different levels of abstraction, ranging from coarse grained qualitative models of (large) subcellular processes to a detailed quantitative model of a (small) functional module (Fig. 7).

2.4.1 ODE-Based Modeling

If the relevant components in a network are largely known and sufficient quantitative data available, for a small-scale network, ODE-based models are widely used to analyze the functional role of nonlinear biochemical networks [43–45]. In such models the reactions are represented by a set of differential equations (Fig. 7) describing change in quantity of reactants to products and vice versa, in case of reversible reactions. Based on reaction rate and kinetic parameters such models usually yield high-quality predictions of the system's dynamics with quantitative information about molecular concentrations [46]. These models, however, require accurate kinetic parameters which is often infeasible for large networks; therefore, the ODE-based model of large biochemical systems is very difficult if not possible. In such cases the Boolean/logic-based model is a suitable option [47, 48].

2.4.2 Logic-Based Modeling

Logic-based modeling is a popular approach to describe the qualitative temporal behavior of a large system of interactions where experimental data are frequently sparse (not all can be measured,

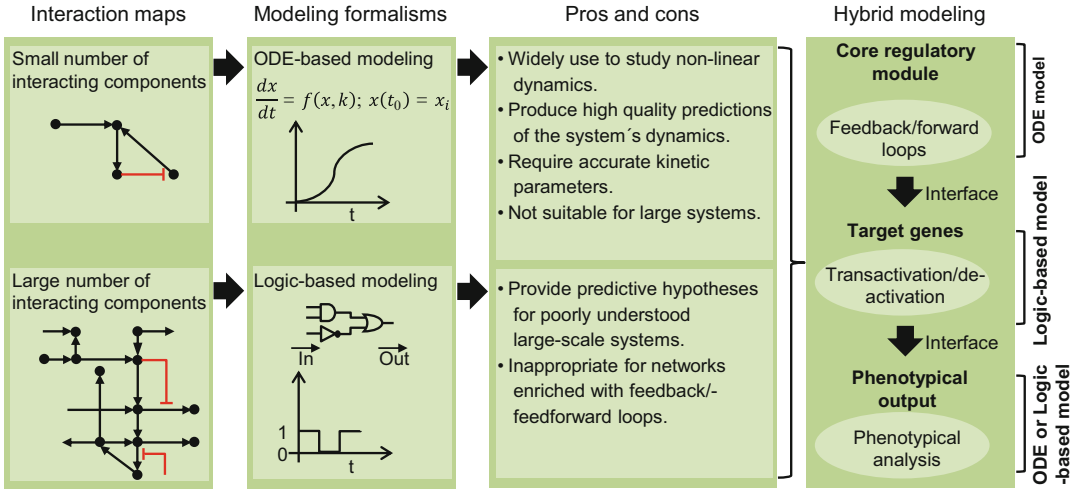


Fig. 7 Selection of appropriate mathematical modeling formalism: The choice of suitable modeling formalism is based on available information of a system, the structure and the size of a network. ODE-based models are widely used for small non-linear system with detailed information available. For a large network size and little available knowledge, logic-based models are preferably used for their dynamics understanding. Hybrid modeling is a strategy to model large-scale non-linear systems by combing ODEs and logic-based models

few time points) and uncertain (lake of replicates and precision) [49]. Logic-based models are qualitative and do not require detail quantitative parameters which make it suitable for large-scale biochemical networks [50–52]. In such a modeling formalism, a network is represented as a graph with nodes and edges, where a node represents any molecular specie and edges depict the type of effect that one specie exerts on the state of other in terms of activation and inactivation (Fig. 7). The state of each species is determined by a logic-based function that links the incoming effect to a state. Logic-based models can provide predictive testable hypotheses, which are especially valuable in poorly understood large-scale systems [53, 54].

The basic/simplest logic-based model is the Boolean model popularized by Kauffmann [55] where the components of a network can be in one of two states: (1) "on" state (also denoted by 1 or "true" state) which represents the "active" or "expressed" state of the component; (2) "off" state (also denoted by 0 state or "false" state) which represents the "inactive" or "not expressed" state of the component. In the Boolean models, nodes ($X_{1,\ldots,n}$) of a network correspond to the Boolean variables that can have values either 1 or 0, and edges define the type of interactions (e.g., activation or inhibition). The future state $X(t + 1)$ of a node is a Boolean function (BF) of the current state $X(t)$ of the nodes regulating it, i.e., $X_i(t + 1) = \mathrm{BF}(X_1(t), X_2(t),\ldots,X_n(t))$ (Fig. 8). Boolean functions determine the states of the node using Boolean gates (NOT, ACTIVE, OR, and AND), where the NOT gate is

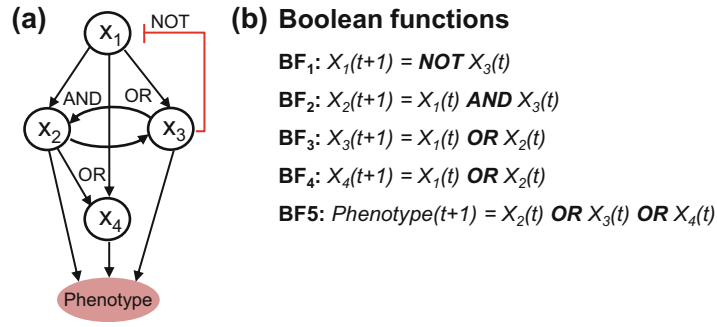


Fig. 8 Logic-based representation of biochemical network. *Left*: A model of biological network consisting of four nodes (X_{1-4}), and regulatory interactions among, linked to a certain phenotype. *Right*: Derivation of Boolean functions (BF) based on regulatory interactions among nodes

used when a molecule inhibits/suppresses other molecules (e.g., interaction from X_3 to X_1; BF_1). In case of independent regulation of more than one node on downstream species, we connected them using OR gate (e.g., interactions from X_1 and X_2 to X_3; BF_3). An ACTIVE gate is used when one molecule activates another (e.g., interaction from X_1 to X_3; BF_3). Finally, an AND gate represents the interaction where more than one molecule together regulate the expression level of the downstream molecule (e.g., interactions from X_1 and X_3 to X_2; BF_2). Boolean functions NOT and ACTIVE are derived based on the network structure but the rules for OR and AND gate determined by training/calibrating the model with sort of qualitative data [56]. Numerous simulation tools are available to simulate logic-bases models, for example, the CellCollective [57], CellNetAnalyzer [58], CellNOpt [56], GINSim [59], BoolSim, BoolNet [60], BooleanNet [61], SimBoolNet [62], SQUAD [63], and ADAM [64].

Model construction: We constructed a small toy logic-based model of signaling and transcriptional pathway in cancer shown in Fig. 9. The upper central part of the model captures the mechanism how the extra-cellular ligands, i.e., epidermal growth factor (EGF) binds to the epidermal growth factor receptor (EGFR) kinase to regulate the cell cycle progression by accumulating the Cyclin dependent kinase (CDK) via downstream activation of protein kinase like Ras, Raf, and ERK. The Cyclins and CDKs complex plays a critical role in cell cycle. It phosphorylates the retinoblastoma (RB) and E2F1 complex causing a transition of the cell through the check point G1/S and enters in S-phase [65]. The top right part of the model represents the survival signaling pathway (PI3K/AKT), which usually has oncogenic behavior in cancer [66]. The survival signaling pathway inhibits the pro-apoptotic genes regulated by E2F1 and leads the cancer cell to an uncontrolled proliferation [67]. The

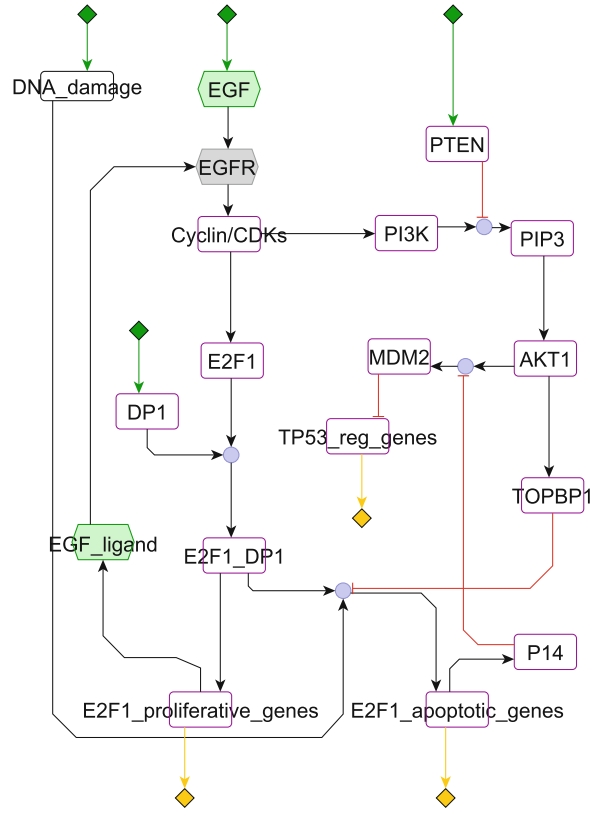


Fig. 9 Logic-based model of signaling and transcriptional pathway in cancer: The *black* and *red lines* represent the type of interactions (i.e., activation and inhibition respectively) among the interacting components. Nodes attached by *green box* and *arrow* are the input, while the nodes attached with *yellow box* and *arrow* are the output nodes of the system. The model is constructed in ProMoT [69] and layout of the graphic is produced in yEd

model is constructed in CellNetAnalyzer, a MATLAB tool to carry out structural and dynamical analysis (here we used a logical steady state) of a large system of interactions. In the logical steady state, the signal drives from input to output of system for all species and reactions until no further state update. All the Boolean functions are derived from the network structure (*see* Table 2). The ligand EGF, DNA damage, PTEN, and DP1 are input node while TP53 and E2F1 regulated apoptotic genes [68] are the output of the systems. In Table 2 reactions, which are derived from the pathway shown in Fig. 9, Boolean functions are shown in the CellNetAnalyzer format. Species on the left-hand side of the equation represent "Reactants" and the species and the right-hand side represents "Products." The "Coefficient" indicates the level of expression of reactants and products. The "+" sign represent "AND" logic and the "!" sign is used for "NOT."

Table 2
CellNetAnalyzer representation of the Boolean functions

1	1 EGF_LIGAND = 1 EGFR
2	1 EGF = 1 EGFR
3	1 EGFR = 1 CYCLIN_CDKS
4	1 CYCLIN_CDKS = 1 PI3K
5	1 PIP3 = 1 AKT1
6	1 E2F1 + 1 DP1 = 1 E2F1_DP1
7	1 CYCLIN_CDKS = 1 E2F1
8	1 E2F1_APOPTOTIC_GENES = 1 P14
9	1 AKT1 = 1 TOPBP1
10	1 E2F1_PROLIFERATIVE_GENES = 1 EGF_LIGAND
11	1! P14 + 1 AKT1 = 1 MDM2
12	1! PTEN +1 PI3K = 1 PIP3
13	1 E2F1_DP1 = 1 E2F1_proliferative_genes
14	1! MDM2 = 1 PT53_reg_genes
15	1! TOPBP1 + 1 E2F1_DP1 + 1 DNA_damage = 1 E2F1_apoptotic_genes

Model simulation: Model is simulated to analyze the input-output response for normal execution of the system. The model recapitulates a "healthy system" for normal pathway function and "cancer system" when the pathway is mutated. First, we initialized the input nodes by 1 (i.e., activated). The model simulations show that both the apoptotic and proliferative genes are 1 (i.e., expressed), which is a necessary condition for healthy cell systems. Second, we mutate the pathway by deactivating the PTEN which results in the constitutive activation of the AKT/PI3K pathway. The activated AKT/-PI3K pathway inhibits the pro-apoptotic genes regulated by TP53 and E2F1 [70] and leads to an uncontrolled proliferation [71], which recapitulates the cancer system. In Table 3, the active and in-active states of the nodes are represented by "1" and "0" respectively. In the table "healthy system" represents normal pathway execution where both the proliferative and apoptotic genes are active when there is no mutation in PTEN and other input nodes are active. The last "Perturbed system" represents the mutation scenario in PTEN (expression state indicated by "0") which results in the activation of proliferative genes and inactivation of apoptotic genes.

The choice for modeling formalism with the aim of applying on biochemical networks which are large and have complex structure,

Table 3
Logical steady-state analysis of the model

	Input				Output		
	PTEN	EGF	DP1	DNA damage	E2F1_proliferative_genes	PT53_reg_genes	E2F1_apoptotic_genes
Healthy system	1	1	1	1	1	1	1
Perturbed system	0	1	1	1	1	0	0

composed of multiple crosstalk pathways that even contain overlapping regulatory loops, is highly challenging. We proposed a modeling strategy that combines ODE-based and logic-based models to accommodate a large-scale, nonlinear system of interactions that we here refer to as "hybrid models" [72].

2.4.3 Hybrid Model

Hybrid models combine different modeling formalisms to handle systems that contain multiple aspects; for example discrete and continuous, linear and nonlinear dynamics. The biological system of cell cycle is an appealing example that contains discrete and continuous aspects [73, 74]. Alfieri et al. modeled the cell cycle as a hybrid system using hybrid automata where the R-point transition was modeled as a discrete event while the mitogenic stimulation of the system was realized as a continuous state by ODEs [74]. In Khan et al. we proposed hybrid modeling formalism that combines the feature of ODEs and logic-based frameworks provide an efficient solution to model large-scale, non-linear biochemical networks [72]. The network is organized and divided into different parts with distinctive regulatory features (Fig. 7) and each part is modeled with suitable modeling formalism. For instance, sub-networks that enriched with feedback and feed-forward loops, and which are therefore expected to display a highly nonlinear behavior are modeled using ODEs, whereas the target gene module that contains activation or inactivation regulation of dozens to hundreds of genes is modeled using logic-based formalism. Hybrid model provides good compromise between quantitative/qualitative accuracy and scalability when considering large networks.

2.5 Integration of Omics Data

Recent advances in the high-throughput techniques made it possible to measure spatio-temporal genomics, transcriptomics, proteomics, and metabolomics data in the context of complex diseases. Most omics technologies are already at impressive level regarding data quality, robustness, time, and cost efficiency. Integration of Omics data with biochemical disease networks has been shown to acquire better insights of system-wide impact of perturbation and therapy in the progression and management of the diseases [75]. Among various Omics datasets, analysis of gene expression data and its integration with biochemical disease networks is the most fundamental process to answer questions like to which degree a gene is active in the process under investigation, what environmental changes alter its expression, which cellular processes are associated with it, and in case of deregulation, which diseases can be caused or mediated by this particular gene. Thus, integration of expression data on the network lets us develop a clearer picture of the role of the genes in the disease progression. In case of cancer, genes that function as tumor suppressors can cause tumorigenesis if their production (expression) is reduced; on the other hand, the increase of production of oncogenes may have similar effects.

Microarray and RNAseq are the most widely used techniques to observe the expression of gene and miRNAs in the context of various diseases and several analytical methods/tools were developed to identify differentially expression genes/miRNAs. Most of these tools use statistical methods such as student's t-test and its variants [76], ANOVA [77], Bayesian method [78], and/or Mann-Whitney test [79] to rank genes for differential expression. Once the –omics data is integrated on the network, this helps in prioritizing regulatory motifs along with other network topological and biomedical parameters to derive a small sub-network responsible for disease progression. This small network is then subjected to a suitable mathematical formalism as described above to find context-specific suitable molecular signatures.

2.6 Validation of Model-Derived Hypotheses Using Experimental and/or Clinical Data

The most significant step in any of the systems biology project is to validate molecular signatures derived after modeling simulations. Validation of predictive results from any of the computational approaches is not only necessary to identify operational/technical errors but is also important to justify the need of new analysis procedures. The molecular signatures, for example gene signatures, could be validated by using more precise gene expression measure along with larger sample size. However, this strategy is generally not used because of the cost of array experiments. Many researchers validate significant results by extracting fresh mRNA from the same specimens and measuring the expressing level using different mRNA-measurement techniques such as RT-PCR (real-time polymerase chain reaction) or by using targeted gene overexpression/silencing experiments. The availability of larger number of clinical studies in the public domain also made it possible to analyze model-derived results quickly before planning time consuming and expensive experiments. One of such resources is "The Cancer Genome Atlas" (TCGA) project which host multi-dimensional maps of key genomic changes in 33 various types of cancer from thousands of independent studies [80]. Predicted cancer related molecular signatures can be easily validated for their relevance using the Kaplan-Meier survival analysis tool available on the UCSC xena browser (http://xena.ucsc.edu) which provides a window to access a large collection of UCSC-hosted public databases such as TCGA [80], International Cancer Genome Consortium (ICGC) [81], Therapeutically Available Research to Generate Effective Treatments (TARGET), Genotype-Tissue Expression (GTEx), and others.

3 Results/Discussion

In order to validate the presented workflow (Fig. 1), we used prostate cancer as a case study and performed various analysis steps to predict molecular signatures associated with primary to metastatic tumor transition.

3.1 Data Collection, Preprocessing, and Analysis

Prostate cancer is a highly heterogeneous cancer and a leading cause of cancer related death worldwide [82]. A large number of genes and miRNAs which are associated with various signaling cascades were found to be dysregulated in several independent studies [83–87]. Several large-scale gene expression datasets were deposited and publically available for research purposes. For the construction of networks involving important factors associated with primary and metastatic prostate cancer phenotypes, we search published literatures along with gene expression datasets. In particular, we used GSE21032 microarray dataset available on Gene Expression Omnibus (GEO) which contains prostate cancer expression data in primary and metastatic states [88]. This dataset contains 218 patient-derived samples, 98 primary tumors, 13 metastatic tumors, and 28 normal prostate tissue samples ($N = 139$) with mRNA and miRNA expression profiles. Microarray data preprocessing was implemented by aroma.affymetrix R package [89]. Data preprocessing consists of three stages including background correction, normalization, and summarization. The RMA method that is the most confident approach for Exon Array data normalization was used to gene expression data normalization. To further analyze the normalized expression, values were transformed to log2 scale. Differential expression analyses were conducted using the popular limma R package [90]. In order to explore differentially expressed genes (DEGs) and differentially expressed miRNAs (DEMs), primary prostate tumor samples were compared to normal prostate tissue samples and metastases prostate cancer samples were compared to primary prostate cancer samples. Absolute log fold change greater than 1 and p-value less than 0.05 were considered as cutoff to explore differentially expressed genes and miRNAs. p-Values were calculated and were adjusted for multiple testing by applying the Benjamini-Hochberg (BH) correction. In total, we found 549 DEGs (179 upregulated and 370 downregulated) in primary and 1008 DEGs (254 upregulated and 754 downregulated) in metastatic stages at p-value <0.05 and absolute log fold change >1. In case of DEMs, we found 55 miRNAs upregulated and 43 miRNAs downregulated in primary state and 88 miRNAs upregulated and 89 downregulated in metastatic state at the same cutoff selected for the analysis of DEGs.

3.2 Construction of Gene-Transcription Factor-miRNA Interaction Network for Primary and Metastatic Prostate Cancer

From the datasets of identified DEGs in primary and metastatic prostate tumors, we first predicted genes that can regulate other DEGs as transcription factors (TF). This was predicted using TRANSFAC database [91], which is a comprehensive and unique database on eukaryotic TFs. Thus, we constructed a dataset of DEGs and differentially expressed TFs in the primary and metastatic phenotypes of prostate cancer. We then constructed a co-expression network of DETF and corresponding target genes. Furthermore, we connected miRNAs (DEMs) with the DETF and

DEGs network using the information available on MirTarbase [92], a database for experimentally validated miRNA target genes, miRanda [93] and TargetScan [94], two databases for predicted miRNA target genes. For each of the DETF-DEG pairs, we calculated the Pearson correlation coefficient (PCC) of expression values and selected the significant pairs with an absolute PCC value more than 0.4 (as TF can either trans-activate or trans-suppress their target genes). The main regulatory effect of miRNAs is induced through binding and degradation of their target mRNA. Therefore, we considered only negative correlation between miRNA and the target mRNA expression [95–97] and set the PCC cutoff to <-0.4. This filtering significantly reduced number of pairs for the construction of regulatory network. Finally, from the selected DEM-DEG and DETF-DEG pairs, we constructed and visualized a composite highly differentially regulated co-expression network for each clinical stage of prostate cancer progression using Cytoscape [13] as shown in Fig. 10. The number of nodes in the

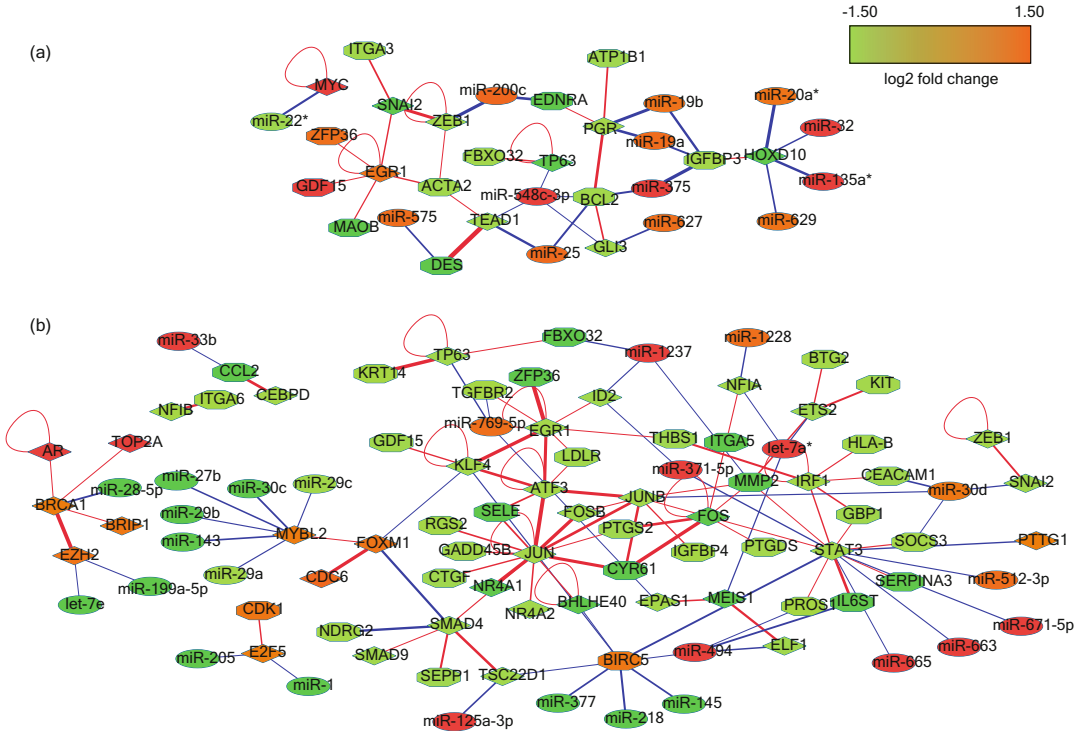


Fig. 10 Integrative network constructed using significant pairs of differentially expressed gene, transcription factors, and miRNAs based on the co-expression correlation of connected nodes for (**a**) primary; and (**b**) metastatic prostate cancer phenotype. MiRNA, genes, and TFs are represented as oval, hexagonal, and diamond shape nodes respectively. Nodes are colored based on their log2 fold change expression values (*green*: down-regulated; *red*: up-regulated). Edge color indicates type of regulation (*red*: activation; *blue*: suppression) and edge width is proportional to the absolute correlation coefficient for the expression values of connected pairs

metastatic integrative co-expression network is more than that of the primary integrative network which indicates that a large number of biological processes and pathways are dysregulated in the transition from the primary to metastatic state in prostate cancer.

3.3 TF and miRNA Co-expression Network Analysis

In order to identify key molecular signatures responsible for the transition from healthy to primary and from primary to metastatic phenotypes in prostate cancer, we analyzed two integrative networks shown in Fig. 10. In particular, we considered two important network topological parameters: (1) degree centrality and (2) betweenness centrality of nodes in the network. The degree of a node is the number of edges connected to the node. Degree centrality of a node is a local centrality index that indicates hub nodes which are involved in a large number of interactions with other nodes in network. Betweenness centrality differs from the other centrality measures. A node can have quite low degree, and still with high betweenness centrality depending on the position of node in the flow of information. In other words, betweenness centrality shows to what extent a node can serve as a bridge in the network. This measurement also indicates how much a given node in a network has control on the interactions of other nodes [38]. Betweenness centrality for node $v \epsilon G$ is the fraction of shortest paths between pairs of nodes $i, j \epsilon G$ that pass through node v. The betweenness centrality $BC(v)$ of a node v is computed as follows:

$$BC(v) = \sum_{i,j \epsilon G} \frac{\sigma_{ij}(v)}{\sigma_{ij}}, \textit{where } i \neq j \neq v.$$

In this equation, $\sigma_{ij}(v)$ denotes the total number of shortest paths between i and j that pass through node v and σ_{ij} denotes the total number of shortest paths between i and j.

Z-Scores for these two parameters were computed for all the nodes in the integrative networks of primary and metastatic prostate cancer. Nodes with z-score > 2.0 for degree and betweenness centrality were considered network key elements and may have a critical effect in prostate cancer initiation and progression. Based on the node degree and betweenness centrality of highly differentially expressed and significantly correlated nodes, we selected top three candidates from each of the networks. These are BCL2, PGR, HOXD10 in case of primary stage and STAT3, JUN and JUNB in case of metastatic state of the prostate cancer.

HOXD10 is a nuclear sequence-specific transcription factor. Its downregulation has been reported in prostate cancer [73]. IGFBP3 that also present in the integrated co-expression network for primary cancer is one of the HOXD10 downstream target genes. IGFBP3 is a pro-apoptotic and anti-angiogenic protein, and induces cell apoptosis [87, 98]. PGR (progesterone receptor) and

BCL2 role in the prostate cancer progression has not been clarified completely. There are also reports that imply association between PGR and BCL2 overexpression and aggressive phenotypes of prostate cancer [99–102]. It is clear from the integrative co-expression network of primary tumor that there is a positive correlation between PGR and BCL2. Regulation of BCL2 through direct binding of PGR has also been reported previously [102, 103]. Taking advantage of the miRNA regulatory interactions in the integrative co-expression networks, we can track the effects of miRNAs on these key regulatory signatures. In the integrative co-expression network for primary tumor PGR and its downstream target, PGR and BCL2 are targeted by multiple overexpressed miRNAs (*miR-20a*, *miR-32*, *miR-135a* and *miR-629* for *HOXD10*, *miR-19a*, *miR-19b* and *miR-375* for *IGFBP3, miR-25, miR-375 and miR-548c-3p for BCL2*, and *miR-19a* and *miR-19b for PGR)*, these inhibitory regulations might be an important reason for the downregulation of the mentioned genes in the primary prostate tumor.

STAT3, JUN, and JUNB are three key signatures that have been identified in the metastatic prostate cancer integrative network. STAT3 is a member of STAT (Signal Transducers and Activators of Transcription) family. It acts in response to cytokines and growth factors, particularly IL6 [104]. STAT3 overexpression has been observed in prostate cancer and based on this finding, inactivation of STAT3/IL6 was examined for prostate cancer treatment; however, this inhibition resulted in the progress of prostate cancer to metastatic phase [86, 105]. Despite the indispensable role of STAT3 and IL6 in prostate cancer, their role in prostate cancer progression must be stage and condition dependent. IL6ST and IRF1 that are present in the metastatic state integrative co-expression network are STAT3 important regulators [106]. Based on the network analysis, their downregulation may have an important effect on the STAT3 inhibition. The integrative co-expression network also introduces five overexpressed miRNAs (*miR-671-5p*, *miR-665*, *miR-663*, *miR 512-3p,* and *miR-371-5p*) which suppress STAT3 expression. The synergistic inhibitory effect of these mentioned factors has a substantial effect on suppression of STAT3 in metastatic prostate cancer.

JUN and JUNB are other key molecular signatures in the metastatic state integrative co-expression network. These genes are members of the AP1 transcription factor family [107]. There are other members of the AP1 transcription factor family in the network as well (*FOS*, *FOSB*, and *ATF3*) which are downregulated. These genes are the mediators of TGFβ signaling pathway and its misfunction leads to progressive phenotypes of prostate cancer [108, 109]. Downregulation of AP1 transcription factors is an important factor for TGFβ signaling pathway aberration in metastatic prostate cancer based on our integrative co-expression

network. Downregulation of JUNB was associated with aggressive metastatic prostate cancer. Based on the previous reports the JUNB expression level can be used to assess aggressiveness of a prostate tumor and to see its potential to progress to metastatic states [110, 111].

3.4 Validation of Molecular Signatures Using Kaplan-Meier Survival Analysis

The Kaplan-Meier survival analysis was conducted for two purposes: (1) Assessment of the methodology developed for creating integrative co-expression networks; and (2) Validation of key molecular signatures. It is important to see how a given biomarker has a critical role in situ based on patient data. The Kaplan-Meier survival analysis is a nonparametric method to estimate the probability that a patient survives beyond a given time; the idea is to construct a series of probability tables for groups (patients with different expression values) versus survival status at each time point at which a failure (biochemical recurrence) occurs, assessing the significance of difference between the survival plots for groups by the log-rank test. In a sense, survival analysis was used to verify the effectiveness of the co-expression network analysis. In fact, better separation in the Kaplan-Meier curve for the identified key molecular signatures has been used as a standard to judge the quality of our methodology for the prediction. The Kaplan-Meier analysis was conducted using biochemical recurrence (BCR) data which demonstrates tumor relapse in patients for identified molecular signatures. Samples were divided into two groups of high and low expression values based on median expression values of each key molecular signature. BCR data was used as a representative index for prostate tumors progression. Among key molecular signatures for primary prostate tumors the survival curves for HOXD10 and PGR show significant difference between the high and low expression groups (p-value <0.05). We can expect long life without BCR event in patients with high expression values for these two genes in contrast to patients with low expression values. These results confirm HOXD10 and PGR as prognosis molecular signatures for primary prostate tumor (Fig. 11, left panel).

In the metastatic state key signatures (STAT3, JUN, and JUNB) the survival curves for all of the key genes show significant difference between high and low expression groups. Taking the advantage of survival analysis in R we could take a look on mean survival time and number of BCR events in the different groups and key signatures (Fig. 11, right panel). Based on this information, among all key signatures the highest and lowest mean survival times were 65.71 and 42.54 months in the high and low expression groups of STAT3, respectively. Furthermore, the highest and lowest BCR events were observed for STAT3 which were 10 and 25 BCR events for high and low expression groups of STAT3, respectively (Fig. 11).

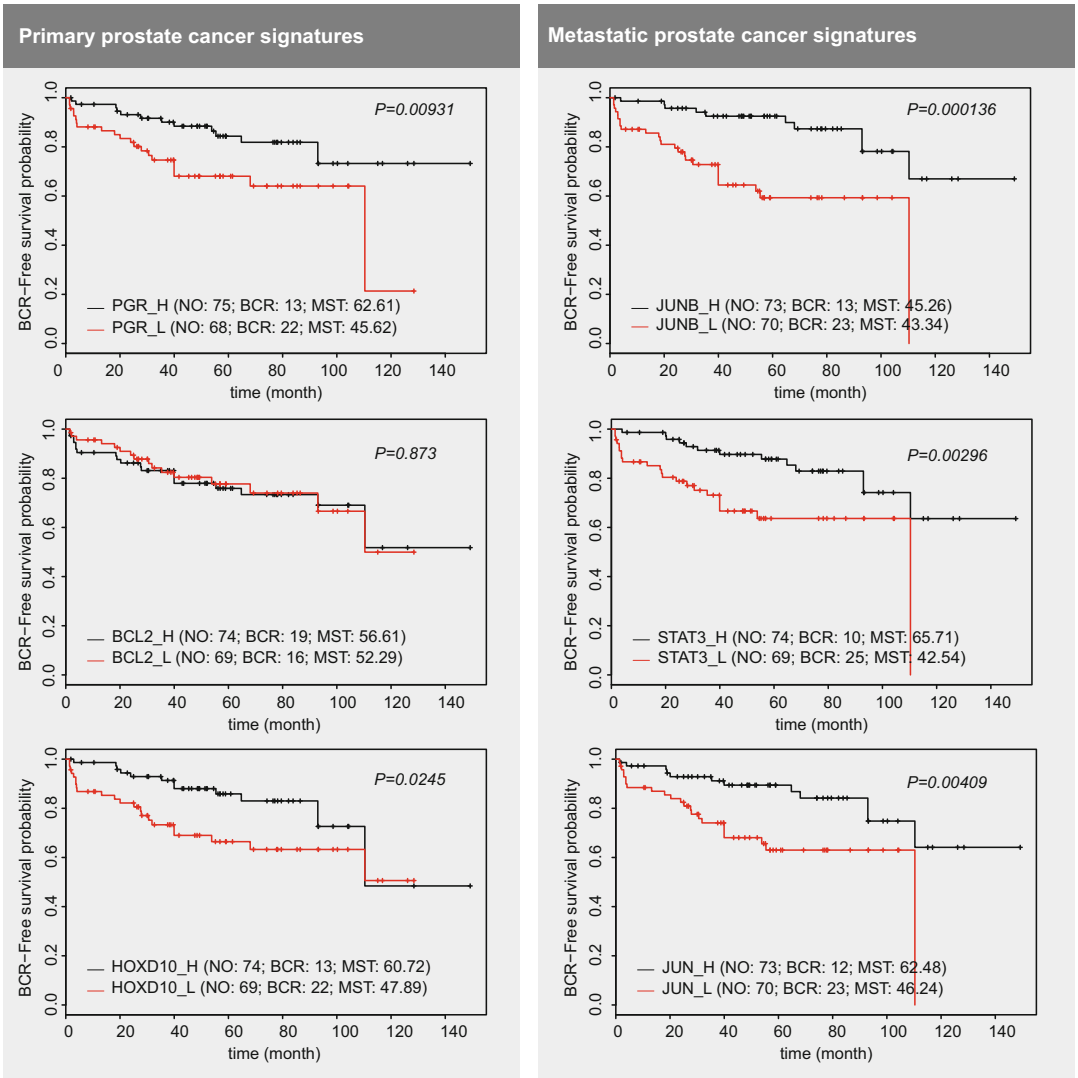


Fig. 11 Kaplan-Meier survival curves of BCR-free survival probability for key molecular signatures in primary (left panel) and metastatic (right panel) states of prostate cancer. High expression of *STAT3* is associated with the highest mean survival time (MST = 65.71 months) and lowest BCR event (10 BCR events) among all of genes. NO: number of samples in the group, BCR: biochemical recurrence, *MST* mean survival time

Based on the integrative co-expression network analysis, survival analysis and literature review STAT3 may be the most significant key signature in the metastatic prostate cancer. The metastatic integrative co-expression analysis also reveals many genes that involve in TGFβ signaling pathway (*SMAD4*, *SMAD9*, *TGFBR2*, *GDF15*, *JUN*, *FOS*, *JUNB*, *FOSB*, and *ATF3*). All of these genes were downregulated in metastatic prostate tumor. As mentioned before, malfunction of TGFβ signaling pathway results in aggressive phenotypes of prostate cancer [109]. JUN and JUNB are the other

explored key signatures for metastatic prostate cancer. These genes are members of the AP1 transcription factor family. These transcription factors were participated in a variety of pathways which implicated in cancer beginning and progression (cell differentiation, proliferation, apoptosis, and oncogenic transformation) [107]. AP transcription factors are mediators of TGFβ signaling pathway [108]. Based on the integrative co-expression network for metastatic prostate cancer and the large number of AP1 transcription factors (JUN, JUNB, FOS, FOSB, and ATF3) and TGFβ signaling pathway members, AP1 transcription factors and TGFβ signaling pathway interrelationship dysregulation may have a critical role in metastatic prostate cancer which needs more studies to clarify.

4 Notes

To improve the treatment outcomes of patients who develop metastatic cancer, a mechanistic understanding of the determinants of the progression of disease is indispensable. From the last few decades, the systems biology and bioinformatics approaches produce successful results to study complex disease with the aim of unraveling their regulatory mechanisms but still it is a long way to go. To understand complex processes and gain new insights into a complex disease like cancer, the interdisciplinary collaborations in systems biology usually begin with the gathering of information from the literature and databases, summarizing components and their interactions relevant for the process under investigation. The information gathered is summarized in interaction maps, which serve as a knowledge-base and being machine readable are amenable to computational analysis. Interaction maps are the summary of a large number of components interacting through feedback and feedforward loops and overlapping pathways.

Recently, numerous large-scale biochemical networks have been constructed with the intention to broaden the understanding of the regulatory events behind the normal and dysregulated function of the pathways involve in certain processes [20, 21, 23, 24, 39]. Using structural analysis, these networks provide useful information about the organization of the network by identifying hub nodes, regulatory motifs, small interconnected modules, and factors that might be used as therapeutic targets [20]. Networks can also be used to analyze large-scale data to identify expression patterns by using the mapping function of the CellDesigner, Cytoscape, or other visualization tools. So far, most of the large-scale network-based studies are confined to static analysis only; nevertheless, they can provide foundation for mechanistic understanding of complex processes that are dysregulated in disease.

We promote here an integrative workflow that combines network structural and dynamical analysis with high-throughput –omics data and other biomedical information to gain mechanistic insights into the causes of differentiated expression patterns from normal to disease state. To this end, we constructed the transcription factor-miRNA regulatory network to understand the mechanisms of metastatic phenotype in prostate cancer.

To further substantiate the analysis, one can use the dynamical systems theory to construct a mathematical model using suitable modeling formalism. The dynamical analysis, for example in silico stimulus response or perturbation analyses, of biochemical networks helps in understanding the functioning of a system and provides an opportunity to formulate new hypotheses about the effect of specific internal or external perturbations in a system. In the systems biology approach, the iterative cycle of data-driven modeling and model-driven experimentations refine the formulated hypotheses until they are validated, which help to understand complex mechanisms in certain biological traits.

References

1. Barabasi A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genetics 12:56–68
2. Chuang H-Y, Lee E, Liu Y-T et al (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140
3. Csermely P, Korcsmáros T, Kiss HJM et al (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther 138:333–408
4. Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8:450–461
5. Kitano H (2007) A robustness-based approach to systems-oriented drug design. Nat Rev Drug Discov 6:202–210
6. Wolkenhauer O (2014) Why model? Front Physiol 5:21
7. Le Novère N (2015) Quantitative and logic modelling of molecular and gene networks. Nat Rev Genet 16:146–158
8. Voit EO (2009) A {systems-theoretical} framework for health and disease. Math Biosci 217:11–18
9. Sadeghi M, Ranjbar B, Ganjalikhany MR et al (2016) MicroRNA and transcription factor gene regulatory network analysis reveals key regulatory elements associated with prostate cancer progression. PLoS One 11:e0168760
10. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372
11. Kitano H (2002) Systems biology: a brief overview. Science (New York, NY) 295:1662–1664
12. Funahashi A, Morohashi M, Kitano H et al (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Biosilico 1:159–162
13. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
14. Junker BH, Klukas C, Schreiber F (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. BMC Bioinformatics 7:109
15. Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp JA et al (2009) Systems biology: parameter estimation for biochemical models. FEBS J 276(4):886–902
16. Wittig U, Kania R, Golebiewski M et al (2012) SABIO-RK—database for biochemical reaction kinetics. Nucleic Acids Res 40: D790–D796
17. Li C, Donizelli M, Rodriguez N et al (2010) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol 4:92

18. Lee W-P, Tzou W-S (2009) Computational methods for discovering gene networks from expression data. Brief Bioinform 10:408–423
19. Zahiri J, Bozorgmehr JH, Masoudi-Nejad A (2013) Computational prediction of protein–protein interaction networks: algorithms and resources. Curr Genomics 14:397–414
20. Matsuoka Y, Matsumae H, Katoh M et al (2013) A comprehensive map of the influenza A virus replication cycle. BMC Syst Biol 7:97
21. Wu G, Zhu L, Dent JE et al (2010) A comprehensive molecular interaction map for rheumatoid arthritis. PLoS One 5:e10137
22. Caron E, Ghosh S, Matsuoka Y et al (2010) A comprehensive map of the mTOR signaling network. Mol Syst Biol 6:453
23. Calzone L, Gelay A, Zinovyev A et al (2008) A comprehensive modular map of molecular interactions in RB/E2F pathway. Mol Syst Biol 4:173
24. Oda K, Matsuoka Y, Funahashi A et al (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1:2005.0010
25. Ritz A, Poirel CL, Tegge AN et al (2016) Pathways on demand: automated reconstruction of human signaling networks. Syst Biol Appl 2:16002
26. Supper J, Spangenberg L, Planatscher H et al (2009) BowTieBuilder: modeling signal transduction pathways. BMC Syst Biol 3:67
27. Gursoy A, Keskin O, Nussinov R (2008) Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans 36:1398–1403
28. Zhang Z, Zhang J (2009) A big world inside small-world networks. PLoS One 4:e5686
29. Jeong H, Mason SP, Barabási A-L et al (2001) Lethality and centrality in protein networks. Nature 411:41–42
30. Kotlyar M, Fortney K, Jurisica I (2012) Network-based characterization of drug-regulated genes, drug targets, and toxicity. Methods 57:499–507
31. Mitra K, Carvunis A-R, Ramesh SK et al (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14:719–732
32. Wang J, Lu M, Qiu C et al (2010) TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res 38:D119–D122
33. Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clin Genet 71:1–11
34. Yeger-Lotem E, Sattath S, Kashtan N et al (2004) Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. Proc Natl Acad Sci U S A 101:5934–5939
35. Tyson JJ, Novák B (2010) Functional motifs in biochemical reaction networks. Annu Rev Phys Chem 61:219–240
36. Zhang Y, Xuan J, de Los Reyes BG et al (2008) Network motif-based identification of breast cancer susceptibility genes. Conference proceedings: annual international conference of the IEEE engineering in medicine and biology society. IEEE engineering in medicine and biology society annual conference, 2008, pp 5696–5699
37. Wang X, Gulbahce N, Yu H (2011) Network-based methods for human disease gene prediction. Brief Funct Genomics 10:280–293
38. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
39. Guebel DV, Schmitz U, Wolkenhauer O et al (2012) Analysis of cell adhesion during early stages of colon cancer based on an extended multi-valued logic approach. Mol BioSyst 8:1230–1242
40. Voit EO (2016) The inner workings of life: vignettes in systems biology. Cambridge University Press, Cambridge, NY
41. Bezručko BP, Smirnov DA (2010) Extracting knowledge from time series: an introduction to nonlinear empirical modeling. Springer, New York, NY
42. Vera J, González-Alcón C, Marín-Sanguino A et al (2010) Optimization of biochemical systems through mathematical programming: methods and applications. Comput Oper Res 37:1427–1438
43. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
44. Zi Z, Klipp E (2007) Constraint-based modeling and kinetic analysis of the smad dependent TGF-β signaling pathway. PLoS One 2: e936
45. Raia V, Schilling M, Böhm M et al (2011) Dynamic mathematical modeling of IL13-induced signaling in Hodgkin and primary mediastinal B-cell lymphoma allows prediction of therapeutic targets. Cancer Res 71:693–704
46. Vera J, Schmitz U, Lai X et al (2013) Kinetic modeling-based detection of genetic signatures that provide chemoresistance via the E2F1-p73/DNp73-miR-205 network. Cancer Res 73:3511–3524

47. Samaga R, Saez-Rodriguez J, Alexopoulos LG et al (2009) The logic of EGFR/ErbB signaling: theoretical properties and analysis of high-throughput data. PLoS Comput Biol 5:e1000438
48. Schlatter R, Philippi N, Wangorsch G et al (2012) Integration of Boolean models exemplified on hepatocyte signal transduction. Brief Bioinform 13:365–376
49. Bornholdt S (2005) Less is more in modeling large genetic networks. Science 310:449–451
50. Saez-Rodriguez J, Simeoni L, Lindquist JA et al (2007) A logical model provides insights into T cell receptor signaling. PLoS Comput Biol 3:e163
51. Schlatter R, Schmich K, Avalos Vizcarra I et al (2009) ON/OFF and beyond—a boolean model of apoptosis. PLoS Comput Biol 5: e1000595
52. Saadatpour A, Wang R-S, Liao A et al (2011) Dynamical and structural analysis of a T cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia. PLoS Comput Biol 7: e1002267
53. Chowdhury S, Pradhan RN, Sarkar RR (2013) Structural and logical analysis of a comprehensive Hedgehog signaling pathway to identify alternative drug targets for glioma, colon and pancreatic cancer. PLoS One 8: e69132
54. Assmann SM, Albert R (2009) Discrete dynamic modeling with asynchronous update, or how to model complex systems in the absence of quantitative information. Methods Mol Biol (Clifton, NJ) 553:207–225
55. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467
56. Terfve C, Cokelaer T, Henriques D et al (2012) CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. BMC Syst Biol 6:133
57. Helikar T, Kowal B, McClenathan S et al (2012) The cell collective: toward an open and collaborative approach to systems biology. BMC Syst Biol 6:96
58. Klamt S, Saez-Rodriguez J, Gilles ED (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst Biol 1:2
59. Chaouiya C, Naldi A, Thieffry D (2012) Logical modelling of gene regulatory networks with GINsim. Methods Mol Biol (Clifton, NJ) 804:463–479
60. Müssel C, Hopfensitz M, Kestler HA (2010) BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics (Oxford) 26:1378–1380
61. Albert I, Thakar J, Li S et al (2008) Boolean network simulations for life scientists. Source Code Biol Med 3:16
62. Zheng J, Zhang D, Przytycki PF et al (2010) SimBoolNet—a cytoscape plugin for dynamic simulation of signaling networks. Bioinformatics (Oxford) 26:141–142
63. Di Cara A, Garg A, De Micheli G et al (2007) Dynamic simulation of regulatory networks using SQUAD. BMC Bioinformatics. 8:462
64. Hinkelmann F, Brandon M, Guang B et al (2011) ADAM: analysis of discrete models of biological systems using computer algebra. BMC Bioinformatics 12:295
65. Swat M, Kel A, Herzel H (2004) Bifurcation analysis of the regulatory modules of the mammalian G1/S transition. Bioinformatics (Oxford) 20:1506–1511
66. Saal LH, Johansson P, Holm K et al (2007) Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proc Natl Acad Sci U S A 104:7564–7569
67. Pützer BM, Engelmann D (2013) E2F1 apoptosis counterattacked: evil strikes back. Trends Mol Med 19:89–98
68. Polager S, Ginsberg D (2009) p53 and E2f: partners in life and death, Nature Reviews. Cancer 9:738–748
69. Mirschel S, Steinmetz K, Rempel M et al (2009) ProMoT: modular modeling for systems biology. Bioinformatics 25:687–689
70. Hennessy BT, Smith DL, Ram PT et al (2005) Exploiting the PI3K/AKT pathway for cancer drug discovery. Nat Rev Drug Discov 4:988–1004
71. Hallstrom TC, Mori S, Nevins JR (2008) An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. Cancer Cell 13:11–22
72. Khan FM, Schmitz U, Nikolov S et al (2014) Hybrid modeling of the crosstalk between signaling and transcriptional networks using ordinary differential equations and multivalued logic. Biochim Biophys Acta 1844:289–298
73. Ramachandran S, Liu P, Young AN et al (2005) Loss of HOXC6 expression induces apoptosis in prostate cancer cells. Oncogene 24:188–198
74. Alfieri R, Bartocci E, Merelli E et al (2011) Modeling the cell cycle: from deterministic models to hybrid systems. Biosystems 105:34–40

75. Kristensen VN, Lingjaerde OC, Russnes HG et al (2014) Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 14:299–313
76. Storey JD, Tibshirani R (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA Microarrays. In: Irizarry RA (ed) The analysis of gene expression data: methods and software. Springer, New York, NY, pp 272–290
77. Kerr MK, Martin M, Churchill G (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7:819–837
78. Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17:509–519
79. Wu TD (2001) Analysing gene expression data from DNA microarrays to identify candidate genes. J Pathol 195(1):53–65
80. Tomczak K, Czerwińska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 19(1A):A68–A77
81. Hudson TJ, Anderson W, Aretz A et al (2010) International network of cancer genome projects. Nature 464:993–998
82. Siegel RL, Miller KD, Jemal A (2016) Cancer statistics. CA Cancer J Clin 66:7–30
83. Varambally S, Dhanasekaran SM, Zhou M et al (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419:624–629
84. Barton BE, Karras JG, Murphy TF et al (2004) Signal transducer and activator of transcription 3 (STAT3) activation in prostate cancer: Direct STAT3 inhibition induces apoptosis in prostate cancer lines. Mol Cancer Ther 3:11–20
85. Abdulghani J, Gu L, Dagvadorj A et al (2008) Stat3 promotes metastatic progression of prostate cancer. Am J Pathol 172:1717–1728
86. Nair S, Barve A, Khor T-O et al (2010) Regulation of Nrf2- and AP-1-mediated gene expression by epigallocatechin-3-gallate and sulforaphane in prostate of Nrf2-knockout or C57BL/6J mice and PC-3 AP-1 human prostate cancer cells. Acta Pharmacol Sin 31:1223–1240
87. Mehta HH, Gao Q, Galet C et al (2011) IGFBP-3 is a metastasis suppression gene in prostate cancer. Cancer Res 71:5154–5163
88. Taylor BS, Schultz N, Hieronymus H et al (2011) Integrative genomic profiling of human prostate cancer. Cancer Cell 18:11–22
89. Bengtsson H, Wirapati P, Speed TP (2009) A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics 25:2149–2156
90. Ritchie ME, Phipson B, Wu D et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47
91. Matys V, Fricke E, Geffers R et al (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374–378
92. Chou C-H, Chang N-W, Shrestha S et al (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res 44(D1): D239–D247
93. Betel D, Koppal A, Agius P et al (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 11:R90
94. Agarwal V, Bell GW, Nam J-W et al (2015) Predicting effective microRNA target sites in mammalian mRNAs. eLife 4:PMC4532895
95. Cai Y, Yu X, Hu S et al (2009) A brief review on the mechanisms of miRNA regulation. Genomics Proteomics Bioinformatics 7:147–154
96. Lorio MV, Croce CM (2012) MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO Mol Med 4:143–159
97. Lai X, Schmitz U, Gupta SK et al (2012) Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. Nucleic Acids Res 40:8818–8834
98. Wang L, Chen S, Xue M et al (2012) Homeobox D10 gene, a candidate tumor suppressor, is downregulated through promoter hypermethylation and associated with gastric carcinogenesis. Mol Med 18:389–400
99. Yu Y, Liu L, Xie N et al (2013) Expression and function of the progesterone receptor in human prostate stroma provide novel insights to cell proliferation control. J Clin Endocrinol Metab 98:2887–2896
100. Bonkhoff H, Fixemer T, Hunsicker I et al (2001) Progesterone receptor expression in human prostate cancer: correlation with tumor progression. Prostate 48:285–291
101. Li M, Ma H, Yang L et al (2016) Mangiferin inhibition of proliferation and induction of apoptosis in human prostate cancer cells is

correlated with downregulation of B-cell lymphoma-2 and upregulation of microRNA-182. Oncol Lett 11:817–822

102. Esber N, Le Billan F, Resche-Rigon M et al (2015) Ulipristal acetate inhibits progesterone receptor isoform A-mediated human breast cancer proliferation and BCl2-L1 expression. PLoS One 10:e0140795
103. Yin P, Lin Z, Cheng Y-H et al (2007) Progesterone receptor regulates Bcl-2 gene expression through direct binding to its promoter region in uterine leiomyoma cells. J Clin Endocrinol Metab 92:4459–4466
104. Horvath CM (2000) STAT proteins and transcriptional responses to extracellular signals. Trends Biochem Sci 25:496–502
105. Abell K, Watson CJ (2005) The Jak/Stat pathway: a novel way to regulate PI3K activity. Cell Cycle (Georgetown, TX) 4:897–900
106. Ho HH, Ivashkiv LB (2006) Role of STAT3 in type I interferon responses: negative regulation of stat1-dependent inflammatory gene activation. J Biol Chem 281:14111–14118
107. Jochum W, Passegue E, Wagner EF (2001) AP-1 in mouse development and tumorigenesis. Oncogene 20:2401–2412
108. Karin M, Liu Z, Zandi E (1997) AP-1 function and regulation. Curr Opin Cell Biol 9:240–246
109. Tu WH, Thomas TZ, Masumori N et al (2003) The loss of TGF-β signaling promotes prostate cancer metastasis. Neoplasia (New York, NY) 5:267–277
110. Thomsen MK, Bakiri L, Hasenfuss SC et al (2015) Loss of JUNB/AP-1 promotes invasive prostate cancer. Cell Death Differ 22:574–582
111. Birner P, Egger G, Merkel O et al (2015) JunB and PTEN in prostate cancer: "loss is nothing else than change". Cell Death Differ (4):22, 522–523

Chapter 13

Spatiotemporal Fluctuation Analysis of Molecular Diffusion Laws in Live-Cell Membranes

Francesco Cardarelli

Abstract

A present challenge of membrane biophysics is deciphering the dynamic behavior of molecules, such as lipids and proteins, within the natural environment of a living-cell membrane. Here, a fluorescence fluctuation-based approach will be described, which makes it possible to probe the "diffusion law" of molecules directly from imaging, in the form of a mean square displacement vs time-delay plot (*i*MSD), with no need for interpretative models. Of note, the presented approach does not require extraction of the molecular trajectories nor the use of bright fluorophores. Conversely, it can be used at high fluorophore density and with relatively dim fluorophores, such as GFP-tagged molecules transiently expressed within cells. The ability of this approach to resolve average molecular dynamic properties well below the diffraction limit will be discussed. Overall, this novel approach is proposed as a powerful tool for the determination of kinetic and thermodynamic parameters over wide spatial and temporal scales.

Key words Fluorescence correlation spectroscopy, Fluctuation analysis, Protein dynamics, Diffusion law, Membrane heterogeneity, Transient confinement, Dynamic partitioning, GFP

1 Introduction

A major challenge of present (and future) membrane biophysics is to quantitatively study how molecular ensembles dynamically interact and exert their functional role in live-cell membranes. In this context, high-speed single-particle tracking (SPT) techniques play a crucial role: several individual molecular components (proteins, lipids, etc.) can be typically produced (e.g., by cloning or synthesis), purified, fluorescently labeled, and re-introduced within the living cell for SPT analysis. Based on SPT, for instance, Kusumi and coworkers quantitatively addressed the compartmentalization of the fluid plasma memb into submicron domains by extracting the diffusion law (in the form of the classical mean squared displacement, or MSD) of many relevant membrane components (for a review *see* [1, 2]). Yet, the SPT approach is inherently endowed with challenging experimental requirements such as: (1) production,

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_13,

purification, labeling of the molecule of interest with a suitable marker and its re-introduction into the living system; (2) the use of relatively large, bulky labels that can induce cross-linking of target molecules or steric hindrance effects, (3) the need for a large number of single-molecule trajectories to obtain trustable statistics.

In this regard, fluorescence correlation spectroscopy (FCS) is rapidly emerging as a very attractive experimental platform. In fact, thanks to its intrinsic single-molecule "sensitivity" in the presence of many similarly labeled molecules, it can easily afford the required statistics in a limited amount of time. In addition, FCS works well with genetically encoded fluorescent proteins and, in general, with relatively dim fluorophores. The basic principle of fluctuation analysis is that the fluorescent molecules stochastically crossing the open detection volume defined by the laser spot lead to a fluctuating occupation number that follows the Poisson statistics (i.e., the variance is proportional to the average number of molecules). The underlying molecular dynamics is extracted as a characteristic decay time through fluctuation correlation analysis. In its classic view, FCS is commonly used as a *local* measurement of the concentration and characteristic transit time of molecules across the laser beam. Many efforts targeted the extension of the FCS principle to the spatial dimension. For instance, the focal area was duplicated [3], moved in space in laser *scanning* microscopes [4–8], or combined with fast cameras [9–11]. Using these "spatio-temporal" approaches, heterogeneity of diffusion constants and concentrations across space was addressed for several molecules on both the model and actual biological membranes [12, 13]. In the effort to fill the gap between the FCS and SPT approaches, it was recently demonstrated that the molecular FCS-based diffusion laws can be recovered by performing fluctuation analysis at various spatial scales larger (by spot-variation FCS [14, 15]) or smaller (by STimulated Emission Depletion, STED [16–19]) than the laser focal area and then by extrapolating the dynamic behavior of molecules below the diffraction limit. In all the FCS experiments described, however, the size of the laser beam is a limit that cannot be overcome. Also, accurate modeling of the dynamics under study is typically required. To tackle these issues an alternative method based on spatio-temporal image correlation spectroscopy (STICS [20]) will be presented here, which is suitable for the study of the dynamics of fluorescently tagged molecules on live-cell membranes with high spatiotemporal resolution (schematic representation in Fig. 1). In particular, TIRF microcopy is exploited to provide accurate optical sectioning of the plasma membrane, while wide-field imaging by an EMCCD camera is applied to reach sub-millisecond resolution (Fig. 1a). The spatiotemporal fluctuation analysis proposed here converts a stack of fluorescence intensity images (Fig. 1b) into a stack of images representing the spatiotemporal evolution of

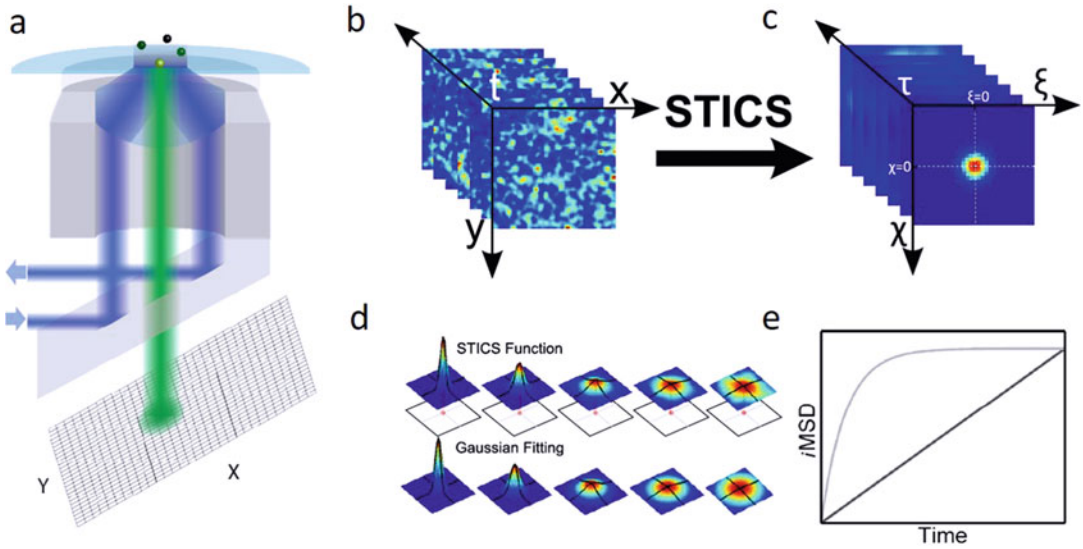


Fig. 1 (**a**) Image acquisition can be performed in wide-field excitation. In particular, a convenient wide-field strategy takes advantage of the total internal reflection (TIR) of the excitation light at the interface between the glass and the sample, limiting the excitation to the close proximity of the interface. In this case, fluorescence light produced in the sample is measured concomitantly in all the positions in space using an array of photodetectors such as an EMCCD. (**b**) The spatiotemporal fluctuation analysis proposed here converts a stack of fluorescence intensity images $I(x, y, t)$, into (**c**) a stack of images representing the spatiotemporal evolution of correlation $G(\xi, \chi, \tau)$ (Eq. 1). (**d**) If the dynamic behavior of molecules is governed exclusively by diffusion the peak of correlation will remain fixed in the origin of the Cartesian axes. If the correlation function is interpolated by a Gaussian function (Eq. 3), the variance of the Gaussian will provide a measurement of the average molecular displacement (*i*MSD). (**e**) The *i*MSD plot can be used to distinguish among the different diffusion modes, such as free and confined diffusion

correlation (Fig. 1c). In the presence of diffusion only, the correlation function can be interpolated by a Gaussian function (Fig. 1d), whose variance provides a measurement of the average molecular displacement, directly from imaging (hereafter, *i*MSD). Thanks to the *i*MSD vs time plot, protein diffusion modes can be directly identified with no need for an interpretative model or assumptions about the spatial organization of the membrane (Fig. 1e). The method is put to test in live cells with two benchmark molecules diffusing on the plasma membrane: a GFP-tagged variant of the trans-membrane transferrin receptor (TfR), a well-known benchmark of confined diffusion [21, 22] and a GFP-tagged variant of glycosylphosphatidyl inositol (GPI), a benchmark of dynamic partitioning into cholesterol-enriched membrane compartments, also known as "lipid rafts." It will be shown how characteristic diffusion constants, confinement areas, and partitioning coefficients can be quantitatively extracted over many microns in the sample. From a theoretical point of view, as previously shown, the spatio-temporal auto-correlation function of the acquired image series critically

depends on the dynamics of the molecules moving in the collected image series. In particular, the correlation function is defined as

$$G(\xi,\chi,\tau) = \frac{\langle I(x,y,t)I(x+\xi,y+\chi,t+\tau)\rangle}{\langle I(x,y,t)\rangle} - 1 \qquad (1)$$

where $I(x, y, t)$ represents the measured fluorescence intensity in the position x, y and at time t, ξ and χ represent the distance in the x and y directions respectively, τ represents the time lag, and $\langle \ldots \rangle$ represents the average. This function can be transformed into

$$G(\xi,\chi,\tau) = \frac{1}{N} p(\xi,\chi,\tau) \bigotimes W(\xi,\chi), \qquad (2)$$

where "N" represents the average number of molecules in the observation area, $\bigotimes$ represents the convolution operation in space, and $W(\xi, \chi)$ represents the autocorrelation of the instrumental waist, the so-called Point Spread Function (PSF) generally well approximated by a Gaussian function. Finally, $p(\xi, \chi, \tau)$ represents the probability to find a molecule at a distance ξ and χ after a time delay τ. If we consider a diffusive dynamics, in which molecules move randomly in all directions and net fluxes are not present, $p(\xi, \chi, \tau)$ is also well approximated by a Gaussian function

$$G(\chi,\xi,\tau) = g(\tau) \cdot \exp\left(-\frac{\chi^2+\xi^2}{\sigma^2(\tau)}\right), \qquad (3)$$

where the variance ($\sigma^2(\tau)$) can be identified as the average molecular Mean Square Displacement directly derived from imaging, with no need for molecular trajectories (iMSD).

2 Materials

2.1 Solutions and Gel

1. Agarose gel at 3% density (prepared from Agar) in TBE buffer.
2. Latex beads, yellow-green fluorescent, 30 nm size dissolved 1:10 (v:v) in distilled water prior to the measurement.

2.2 Cell Manipulation

1. Cell line: Chinese Hamster Ovary (CHO)-K1 cells.
2. Humidified and thermostated incubator for cell maintenance.
3. Medium for cell maintenance: DMEM-F12 supplemented with 10% Fetal Bovine Serum (FBS).
4. Medium for cell transfection: DMEM-F12 with no FBS.
5. 100 mm Petri dish for cell maintenance and 35 mm Glass-bottom Petri Dishes for optical microscopy.
6. Lipofectamine 2000 Transfection Reagent. Store at 4 °C.
7. Trypsin from bovine pancreas.

8. Phosphate Buffer Solution (PBS) for cell washing procedures.
9. Latrunculin B solution for actin depolymerization. Store at −20 °C.
10. pcDNA3 plasmid for eukaryotic cell transfection, encoding for a GFP-tagged variant of the trans-membrane transferrin receptor (GFP-TfR) or a GFP-tagged variant of glycosylphosphatidyl inositol (GPI-GFP).

2.3 Imaging Equipment and Software

1. DMI6000 microscope equipped with TIRF modulus and iXon Ultra 897 camera (*see* **Note 1**).
2. LAS AF image acquisition and processing software.
3. MatLab software for data analysis.

3 Methods

3.1 Sample Preparation

1. 48 h before the experiment, wash three times a 100 mm Petri dish of confluent cells (CHO-K1, in the example reported here) with PBS, add 1 ml of trypsin, and store the Petri in the incubator (37 °C, 5% CO_2) for 5 min. Resuspend detached cells by adding 9 ml of DMEM-F12 medium supplemented with 10% of FBS and seed 100 μl of cell solution to a Petri dish containing 900 μl of the same medium (*see* **Note 2**).
2. Incubate the cells for 24 h at 37 °C and 5% CO_2.
3. 24 h before the experiment transfect cells accordingly by using Lipofectamine 2000 (following the manufacturer's instructions) using the desired plasmid and incubate the cells for 24 h at 37 °C and 5% CO_2 before imaging.

3.2 Camera and PSF Calibration

1. Turn on the camera and wait for it to cool down. Set the proper camera acquisition parameters (i.e., for the experimental system proposed here, the exposure time is typically set to 0.5 ms, the EMgain to 1000, the acquisition mode to "Cropped Mode" (*see* **Note 3**), the ROI size to 32 × 128 (*see* **Note 4**), and the total number of repetitions to 10^4). More technical details can be found elsewhere [23]
2. Start the acquisition of the camera background signal.
3. Import acquired frame series to a data processing program (e.g., Matlab). Calculate and inspect the average intensity in each pixel in order to verify that the camera background is approximately flat in the selected region of the chip. Create a histogram of the values (also defined Digital Levels, DLs) in acquired images stack (e.g., by using the "hist" command in Matlab) and plot the logarithm of the resulting frequency (e.g., by using the "semilogy" command in Matlab) (Fig. 2, *see* **Note 5**).

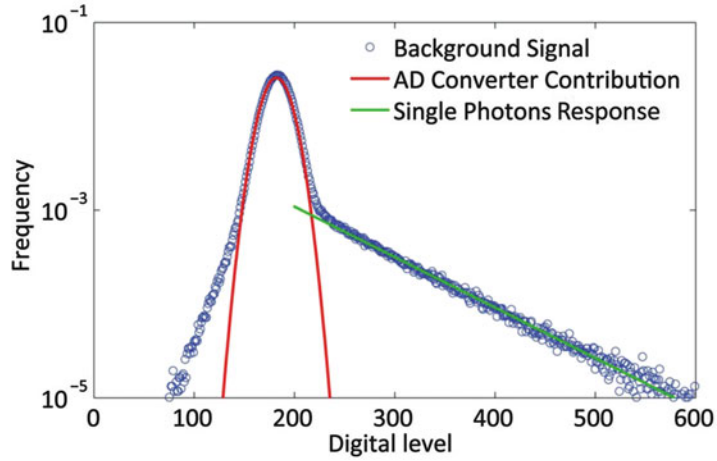


Fig. 2 Calibration of camera response to single photons. The figure shows the Digital Level (DL) distribution for camera background in a 32X128 ROI, exposure 0.5 ms, in Cropped Sensor Mode. The peak at about 180 DL represents the camera response to no photons. Particularly, it represents the contribution of the Analog Digital (AD) converter and can be approximated by a Gaussian function to estimate the offset and the variance introduced by the signal recording. Above 200 DL the distribution of digital levels becomes exponential and represents the average camera response to a single photon. The measurement of these parameters allows estimating the density of photons recorded during the acquisition. (Reproduced from ref. 24 with permission)

4. Dilute 10 μl of 30 nm fluorescent beads solution (about 5 μM) in 90 μl of distilled water (*see* **Note 6**). Cut a squared piece (1 cm × 1 cm) of agarose gel (3%) and deposit 10 μl of the solution on the top of the gel. Turn over the piece of gel on the bottom glass of a 35 mm Glass-bottom Patri dish and squeeze the drop on the glass.
5. Put the sample in the microscope holder, set the camera exposure and EMgain (100 milliseconds and 1000, respectively, are appropriate parameter values according to the system proposed here), and eventually wait for the camera to cool down. Find a field of view with isolated beads, accurately focus on a single bead and acquire 100 frames (Fig. 3a). Five to six repetitions of the same measurement are recommended, in order to acquire enough statistics.
6. Import the acquired image series to a data processing program (e.g., MatLab) and average the stack in time to identify easily isolated fluorescent spots. Take care of selecting the smallest ones to avoid aggregates (Fig. 3b). Fit the selected intensity distribution with a Gaussian function (e.g., by Matlab "gaussfit" function) and verify the goodness of the fit by inspecting the associated fit residuals (Fig. 3c). The half width at half maximum retrieved from fitting is a good estimate of the characteristic PSF size.

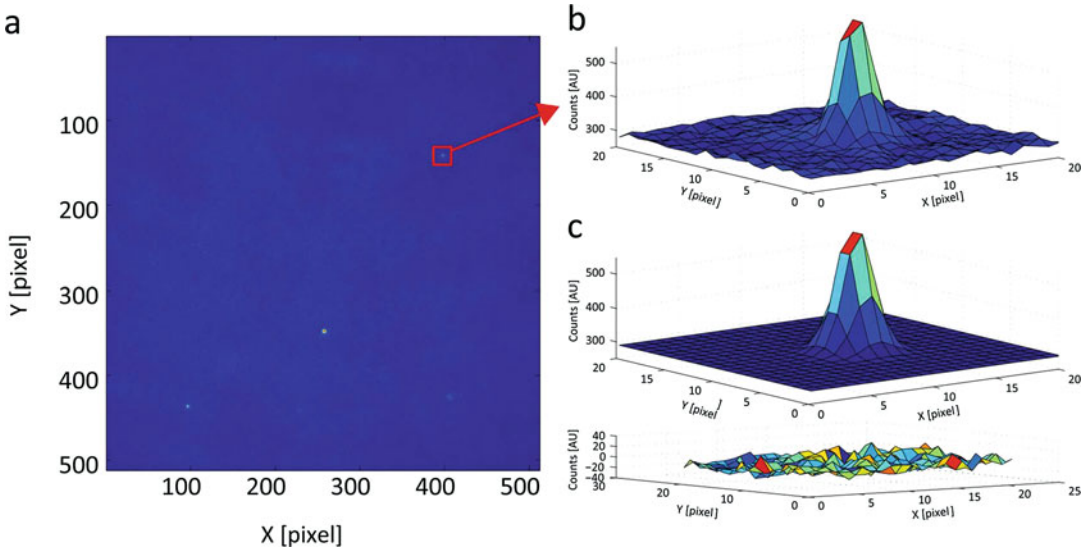


Fig. 3 Calibration of point spread function. (**a**) Pseudocolor image of an isolated bead and beads aggregates. (**b**) 3D plot of the intensity profile of an isolated bead shows a well-defined Gaussian profile. (**c**) Fit of the intensity distribution by a Gaussian function (*upper panel*) with the corresponding residuals (*lower panel*). The good agreement between the fitted distribution and the measured intensity profile is also a proof that the instrumental PSF can be approximated by a Gaussian function (Reproduced from ref. 24 with permission)

3.3 Data Acquisition

1. Align the TIRF laser excitation according to the procedure of your setup. In the setup proposed here, an auto-alignment procedure is available. Once the laser is aligned, an optimal penetration depth of about 70 nm can be selected. Put the sample in the holder.
2. Focus on a fluorescent cell by using the eyepiece (*see* **Note 7**), send the light to the camera, and gently push the slits allowing only the light from the selected ROI to pass (here a 32 × 32 pixels ROI).
3. Set the appropriate exposure time and EMGain on the camera, select the desired number of image repetitions, then start the acquisition. Reset the EMGain and exit from the Cropped Mode to allow temperature stabilization before acquiring a new cell. Repeat the last two steps in order to acquire the desired number of cells.

3.4 Calculation of the Mean Square Displacement from Imaging (iMSD)

1. Import the acquired image series into a data processing program (e.g., Matlab, appropriate scripts can be found elsewhere [23]).
2. Calculate the average intensity of each image and plot it in time. If more than 10% of photobleaching is present, it is suggested to discard the image series. If it is lower, try to correct the effect on the correlation function by subtracting

from each image its average intensity, as previously demonstrated [24].

3. Calculate the spatiotemporal correlation (Eq. 1, $G(\xi,\chi,\tau)$). Remove $G(\xi,\chi,0)$ because the correlation due to the shot noise in low-light regime dominates $G(0,0,0)$. The correlation due to the detector dominates the $G(\pm1,0,0)$, and molecular movement during the exposure time could deform $G(\xi,\chi,\tau)$ for $\tau = 0$ by producing an apparent increase in the measured PSF waist (this effect disappears for $\tau > 0$). Interpolate $G(\xi,\chi,\tau)$ by a Gaussian fit (Eq. 3) (e.g., by the Matlab "gaussfit" function) to recover the *i*MSD.
4. Plot the obtained waist $\sigma(\tau)^2$ (*i*MSD) as a function of time (*see* **Note 8**). The first few points can be fitted to extrapolate the intercept at zero time delay ($\sigma_0{}^2$) (5 points are usually enough but more points can be fitted if they show a linear behavior) and compare the obtained value with the previously calibrated value of the PSF (*see* **Note 9**).
5. Transient confinement: the case of TfR-GFP. Many studies showed that the cytoplasmic tail of this receptor interacts with the membrane-associated F-actin skeleton, which in turn acts as a fence for the receptor mobility [15, 25] (Fig. 4a). A representative TIRF image of a CHO-K1 cell expressing TfR-GFP is presented in Fig. 2b (*see* **Note 10**). The temporal evolution of the correlation function with the corresponding Gaussian fit and residues shows the expected decrease in height and increase in width of the correlation, due to molecular movement (Fig. 2c). Also, the autocorrelation plot shows that the characteristic time of the fluctuations is shorter than the total length of the measurement (arrow) (Fig. 2d). Thus, immobile fraction removal is a safe operation. As expected, the measured diffusion law (Fig. 2e, red curve) for TfR-GFP shows a first flat behavior below 100 nm, with an average D_{app} of about 0.7 $\mu m^2/s$, followed by consequent rapid decrease in apparent diffusivity down to 0.2 $\mu m^2/s$ (the value typically measured by diffraction-limited FCS [15]). This result shows that our approach can easily measure the average displacement of GFP labeled proteins with a resolution of few tens of nanometers (*see* **Note 11** and ref. 26). Moreover, the spatial scale at which the *i*MSD starts to decrease its slope sets the characteristic spatial scale of protein (partial) spatial confinement induced by the membrane skeleton at around 120 nm, in keeping with previous estimates [1]. F-actin digestion by Latrunculin B (Fig. 2f) produces the expected change in the TfR-GFP diffusion law (Fig. 2e, green curve).
6. Dynamic partitioning: the case of GFP-GPI. The *i*MSD approach has the ability to discriminate between transiently

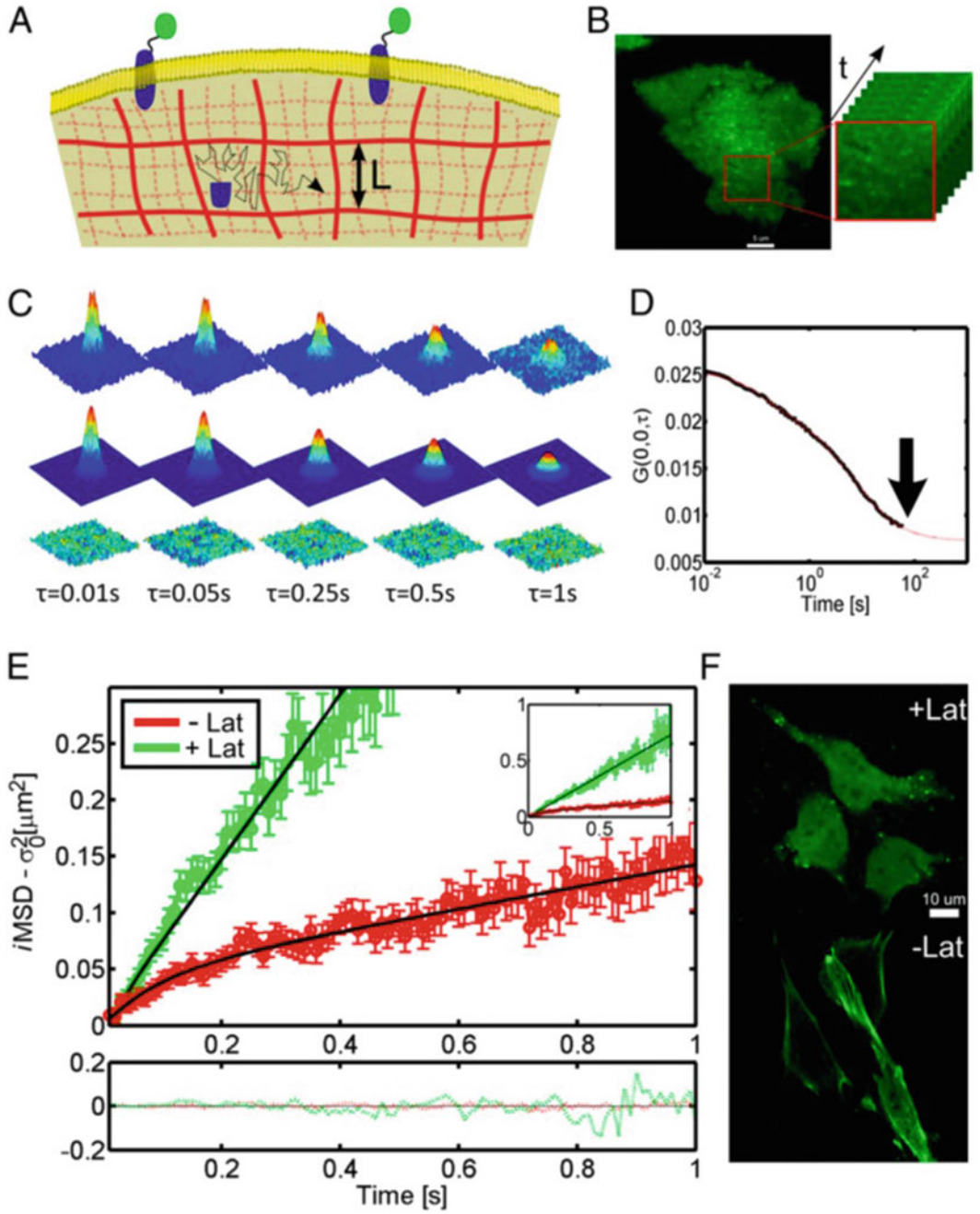


Fig. 4 Analysis of TfR-GFP dynamics in living cells. (**a**) Schematic representation of a GFP-tagged TfR diffusing within the cytoskeleton meshwork, with particular emphasis on the F-actin mesh close to the plasmamembrane. (**b**) TIRF microscopy image of a cell expressing GFP-tagged TfR (*left*) and the detail of the ROI on the membrane selected for imaging (*right*). (**c**) Temporal evolution of the correlation function with the corresponding Gaussian fit and residues. (**d**) The autocorrelation plot shows that the characteristic time of the fluctuations is shorter than the total length of the measurement (*arrow*). Thus, immobile fraction removal is a safe operation. (**e**) *i*MSD vs time plot for GFP-TfR in physiological conditions (*red curve*) and after 30 min of

confined diffusion in a meshwork and dynamic partitioning into membrane nano-domains (i.e., lipid rafts). In fact, this latter is expected to need two Gaussian components for a satisfactory fitting [11]. This reflects the presence of two segregated diffusive behaviors: (a) pure isotropic diffusion outside of the nanodomains and (b) confined diffusion within the nanodomains. Figure 3a shows a representative TIRF image of GFP-GPI transiently transfected into a living cell. The temporal evolution of GFP-GPI correlation function (Fig. 3b, first row and Fig. 5) and the corresponding Gaussian fitting and residuals (Fig. 3b, second and third rows) show that a one-component Gaussian model does not provide a satisfactory description of the system spatiotemporal evolution. By contrast, fitting to a two-Gaussian model (Fig. 3b, fourth row) well describes the experimental function, producing random residuals (Fig. 3b, fifth row). The experimental *i*MSD curves are qualitatively in keeping with theoretical predictions [11], in the sense that they show an almost constant *i*MSD component (trapped molecules) and a linearly increasing one (diffusing molecules) (Fig. 3c).

4 Notes

1. To properly collect the fluorescence from membranes a combination of high-magnification, high-numerical-aperture objective (100×, NA: 1.47) with selective membrane illumination by TIRF, and an EMCCD camera (physical size of the pixel on the chip 16 μm) for detection are used. To reach a pixel size of 100 nm an additional magnification lens of 1.6× is used. As discussed above, a time resolution below 1 ms is desirable to properly describe the dynamics of fast membrane lipids/proteins. To this end a region of interest (ROI) smaller than the whole chip of the camera (512 × 512 pixels) is required, although the frame time would be limited by the time (typically milliseconds for 512 × 512 pixel EMCCD) required to shift the charges from the "exposure" to the "readout" chip on the camera.
2. The dilution proposed typically corresponds to approximately 10^5 cells, but this number may vary depending, for instance, on the cell line and the degree of actual confluence in the

Fig. 4 (continued) Latrunculin-B treatment (*green curve*). The *inset* shows the *i*MSD trend at a short timescale. (**f**) Fluorescence images of cells transfected with actin-GFP that show the effect of Latrunculin-B on the integrity of actin filaments after 30 min of treatment (Reproduced from ref. 11 with permission)

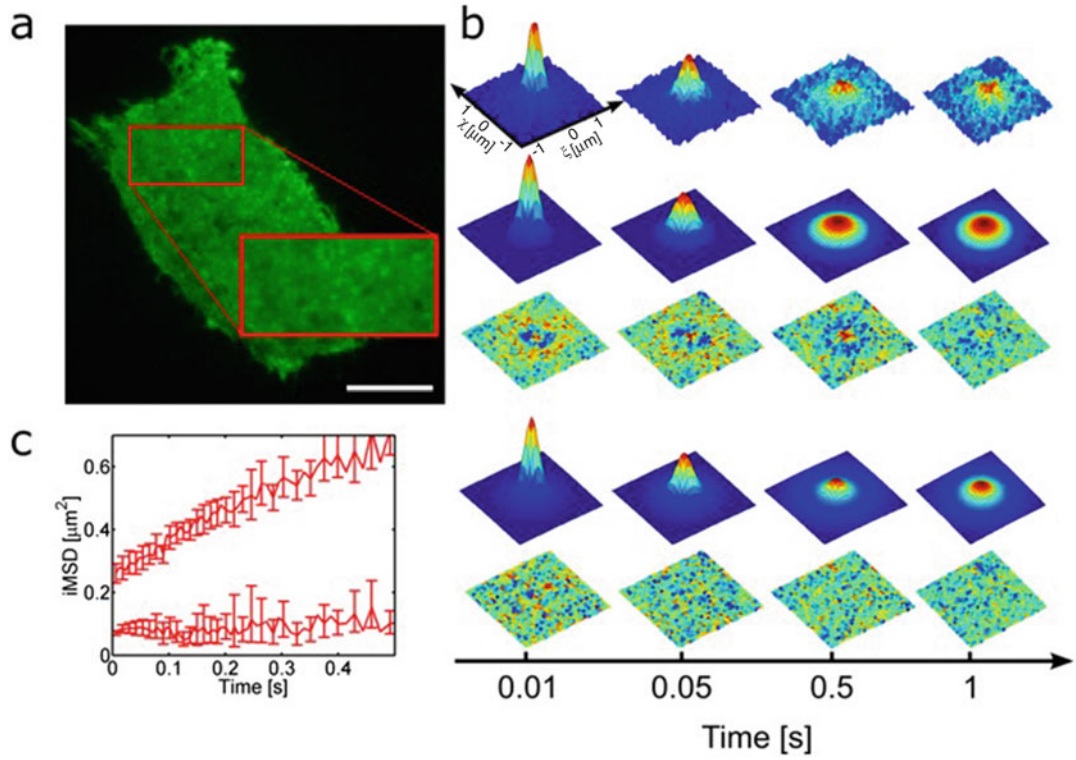


Fig. 5 Analysis of GFP-GPI dynamics in living cells. (**a**) A representative TIRF image of GFP-GPI. (**b**) Time evolution of the correlation function (*first row*) and the corresponding fitting (*second row*). Residuals (*third row*) show that a one-component Gaussian model does not afford a satisfactory description of the system. Fitting with a two-Gaussian model (*fourth row*) better describes the experimental function, producing random residuals (*fifth row*). (**c**) the two calculated iMSD curves are qualitatively in keeping with the theoretical predictions, in the sense that they show an almost constant iMSD component (trapped molecules) and a linearly increasing one (diffusing molecules) (Reproduced from ref. 11 with permission)

maintenance Petri dish. Higher accuracy can be reached by counting the detached cells before plating them.

3. The "Cropped Mode" technology allows shifting lines from the ROI only instead of the whole frame, with a practical effective reduction of the exposed chip size (called Cropped Sensor Mode in our EMCCD). For this configuration to be effective, a couple of slits mounted in the optical path must cover the chip outside of the ROI. Thanks to this setup, a time resolution down to 10^{-4} s can be achieved. When working in Cropped Mode, please remember to remove the first and the last few horizontal lines (3–10 depending on the size of the ROI) for each frame as the camera background is usually biased close to borders.
4. To accurately fit the correlation function a spatial sampling (pixel size) lower than the waist of the PSF is suggested. The PSF waist is typically in the 200–500 nm range (depending on

the Objective numerical aperture and the excitation wavelength). Thus, a pixel size of 70–150 nm (three times lower than the instrumental waist) can be usually set. Furthermore, please consider that the minimum size of the image should be at least 3-times larger than the maximum displacement of interest (plus the instrumental waist). This is needed to reach a good convergence of the fitting algorithm and a statistically significant sampling of molecular displacements. Concerning the camera-based system used here, please note that the pixel physical size on the chip is fixed. Therefore, decreasing the pixel size inevitably decreases the signal in each pixel (that depends on the square of the pixel size), makes the field of view smaller, and demands for a higher magnification power.

5. A typical frequency distribution of camera background yields a peak at a certain value of Digital Levels (DL) (due to the camera response to no photon) that represents the contribution of Analog Digital (AD) converter. This contribution can be approximated as a Gaussian distribution to estimate the offset and the variance introduced by the signal recording. At higher DL values, the distribution becomes exponential and represents the average camera response to a single photon. The center and the variance of the Gaussian function represent the camera offset and error, respectively, while the decay constant of the exponential part affords an estimate of the DL assigned by the camera to each single photon. The higher is the ratio between the average DL of a single photon and the AD converter error, the lower will be the noise in the calculated correlation function. For more details please refer to ref. 23.
6. It is suggested to extensively sonicate the solution for 20 min to avoid the presence of beads aggregates.
7. Particular attention is required in order to avoid art factual correlations. In fact, as previously shown for similar techniques [27], cell borders as well as out-of-focus vesicles could introduce a strong correlation. If the inspection of the average image reveals cell borders or out of focus vesicles, try to exclude the region involved, otherwise discard the acquisition. To correct the effect of this immobile structures subtract the average temporal intensity from each pixel [28].
8. If the data are too noisy, try to increase the number of acquired frames, increase the laser power, average more $G(\xi,\chi,\tau)$ together.
9. If the ${\sigma_0}^2$ and PSF values are comparable, it means that the dynamics of isolated fluorophores (not large aggregates) is extracted. By contrast, if ${\sigma_0}^2 >>$ PSF, it means that either large aggregates are present or hidden dynamics are present (i.e., a faster time resolution is needed) [11].

10. Low fluorescence intensity cells should be preferred, as the membrane is closer to the native condition and the probability of artifacts related to the over-expression is minimized. In addition, the central part of the cell should be avoided, as the effects of out-of-focus fluorescence (from cytoplasm, for instance) may be present.
11. The actual resolution in the measurement of particle displacements is technically not limited by diffraction as it depends on how accurately the correlation function is measured [26, 29]. Few photons per molecule (usually below 10) in each frame are typically enough. In fact, the contributions of all the observed molecules are averaged together when the correlation function is calculated, even if molecules are not isolated (as in the case dim and dense labels, such as fluorescent proteins transfected in live cells). It thus appears clear that the minimum measurable displacement depends on the diffusivity of the molecule and on the time resolution of the imaging setup. The same principle applies to 3D applications of spatio-temporal fluctuation analysis [30].

Acknowledgments

This work is dedicated to the memory of Dr. Carmine Di Rienzo. Carmine left us with a precious heritage of knowledge. Dissemination of his ideas is the greatest tribute to his science and memory.

References

1. Kusumi A, Nakada C, Ritchie K et al (2005) Paradigm shift of the plasma membrane concept from the two-dimensional continuum fluid to the partitioned fluid: high-speed single-molecule tracking of membrane molecules. Annu Rev Biophys Biomol Struct 34:351–378
2. Kusumi A, Shirai YM, Koyama-Honda I et al (2010) Hierarchical organization of the plasma membrane: investigations by single-molecule tracking vs. fluorescence correlation spectroscopy. FEBS Lett 584:1814–1823
3. Ries J, Schwille P (2006) Studying slow membrane dynamics with continuous wave scanning fluorescence correlation spectroscopy. Biophys J 91:1915–1924
4. Berland KM, So PT, Chen Y et al (1996) Scanning two-photon fluctuation correlation spectroscopy: particle counting measurements for detection of molecular aggregation. Biophys J 71:410–420
5. Ruan Q, Cheng MA, Levi M et al (2004) Spatial-temporal studies of membrane dynamics: scanning fluorescence correlation spectroscopy (SFCS). Biophys J 87:1260–1267
6. Heinemann F, Betaneli V, Thomas FA et al (2012) Quantifying lipid diffusion by fluorescence correlation spectroscopy: a critical treatise. Langmuir 28:13395–13404
7. Cardarelli F, Lanzano L, Gratton E (2012) Capturing directed molecular motion in the nuclear pore complex of live cells. Proc Natl Acad Sci U S A 109:9863–9868
8. Cardarelli F, Lanzano L, Gratton E (2011) Fluorescence correlation spectroscopy of intact nuclear pore complexes. Biophys J 101: L27–L29
9. Kannan B, Har JY, Liu P et al (2006) Electron multiplying charge-coupled device camera based fluorescence correlation spectroscopy. Anal Chem 78:3444–3451
10. Unruh JR, Gratton E (2008) Analysis of molecular concentration and brightness from fluorescence fluctuation data with an electron

multiplied CCD camera. Biophys J 95:5385–5398

11. Di Rienzo C, Gratton E, Beltram F et al (2013) Fast spatiotemporal correlation spectroscopy to determine protein lateral diffusion laws in live cell membranes. Proc Natl Acad Sci U S A 110:12307–12312
12. Schwille P, Haupts U, Maiti S et al (1999) Molecular dynamics in living cells observed by fluorescence correlation spectroscopy with one- and two-photon excitation. Biophys J 77:2251–2265
13. Weiss M, Hashimoto H, Nilsson T (2003) Anomalous protein diffusion in living cells as seen by fluorescence correlation spectroscopy. Biophys J 84:4043–4052
14. Wawrezinieck L, Rigneault H, Marguet D et al (2005) Fluorescence correlation spectroscopy diffusion laws to probe the submicron cell membrane organization. Biophys J 89:4029–4042
15. Lenne PF, Wawrezinieck L, Conchonaud F et al (2006) Dynamic molecular confinement in the plasma membrane by microdomains and the cytoskeleton meshwork. EMBO J 25:3245–3256
16. Hell SW, Wichmann J (1994) Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Opt Lett 19:780–782
17. Eggeling C, Ringemann C, Medda R et al (2009) Direct observation of the nanoscale dynamics of membrane lipids in a living cell. Nature 457:1159–1162
18. Mueller V, Ringemann C, Honigmann A et al (2011) STED nanoscopy reveals molecular details of cholesterol- and cytoskeleton-modulated lipid interactions in living cells. Biophys J 101:1651–1660
19. Sezgin E, Levental I, Grzybek M et al (2012) Partitioning, diffusion, and ligand binding of raft lipid analogs in model and cellular plasma membranes. Biochim Biophys Acta 1818:1777–1784
20. Hebert B, Costantino S, Wiseman PW (2005) Spatiotemporal image correlation spectroscopy (STICS) theory, verification, and application to protein velocity mapping in living CHO cells. Biophys J 88:3601–3614
21. Sako Y, Kusumi A (1995) Barriers for lateral diffusion of transferrin receptor in the plasma membrane as characterized by receptor dragging by laser tweezers: fence versus tether. J Cell Biol 129:1559–1574
22. Sako Y, Kusumi A (1994) Compartmentalized structure of the plasma membrane for receptor movements as revealed by a nanometer-level motion analysis. J Cell Biol 125:1251–1264
23. Di Rienzo C, Gratton E, Beltram F et al (2014) From fast fluorescence imaging to molecular diffusion law on live cell membranes in a commercial microscope. J Vis Exp 9:e51994
24. Ries J, Chiantia S, Schwille P (2009) Accurate determination of membrane dynamics with line-scan FCS. Biophys J 96:1999–2008
25. Ritchie K, Shan XY, Kondo J et al (2005) Detection of non-Brownian diffusion in the cell membrane in single molecule tracking. Biophys J 88:2266–2277
26. Di Rienzo C, Gratton E, Beltram F et al (2016) Spatiotemporal fluctuation analysis: a powerful tool for the future nanoscopy of molecular processes. Biophys J 111:679–685
27. Kolin DL, Wiseman PW (2007) Advances in image correlation spectroscopy: measuring number densities, aggregation states, and dynamics of fluorescently labeled macromolecules in cells. Cell Biochem Biophys 49:141–164
28. Digman MA, Brown CM, Sengupta P et al (2005) Measuring fast dynamics in solutions and cells with a laser scanning microscope. Biophys J 89:1317–1327
29. Di Rienzo C, Gratton E, Beltram F et al (2016) Super-resolution in a standard microscope: from fast fluorescence imaging to molecular diffusion laws in live cells. In: Alberto D, AMJ v Z (eds) Super-resolution imaging in biomedicine. Taylor & Francis Group, Abingdon, pp 19–47
30. Di Rienzo C, Piazza V, Gratton E et al (2014) Probing short-range protein Brownian motion in the cytoplasm of living cells. Nat Commun 5:5891

Chapter 14

A Method for Cross-Species Visualization and Analysis of RNA-Sequence Data

Stephen A. Ramsey

Abstract

In this methods article, I describe a computational workflow for cross-species visualization and comparison of mRNA-seq transcriptome profiling data. The workflow is based on gene set variation analysis (GSVA) and is illustrated using commands in the R programming language. I provide a complete step-by-step procedure for the workflow using mRNA-seq data sets from dog and human bladder cancer as an example.

Key words mRNA-seq, Cross-species, Transcriptome, Bioinformatics, Gene function

1 Introduction

Transcriptome profiling by high-throughput short-read end-sequencing of cDNA fragments, i.e., mRNA-seq, has become a standard systems biology research modality due to its specificity for transcript detection, large dynamic range, and ability to quantify the entire polyadenylated transcriptome in one assay [1, 2]. However, applications such as comparative oncology, comparative immunology, or evolutionary functional genomics involve comparing and contrasting transcriptome responses across tissues in *different species*, for example between cancer and normal tissue for neoplasms that occur in humans and domestic dogs [3–5], between human and mouse in various immune cell types [6, 7], or in various organs or tissues in various species in the context of evolutionary studies [8–12]. While for simple two-sample mRNA-seq study designs it is possible to use scatter plots of gene expression ratios between pairs of orthologs [3], such an approach does not naturally extend to more complicated (e.g., multi-factor or time-course) study designs and it does not enable cross-species analysis that is unsupervised by the sample types. Alternatively, normalized measures of gene expression such as fragments per kilobase of transcript per million reads (FPKM) can be used to compare expression for

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_14,

ortholog pairs across species [10], but that approach has been criticized as insufficiently accounting for species-specific effects [11, 12]. Another challenge that arises in cross-species mRNA-seq analysis is that the baseline count level for a gene to be considered above-background can vary from data set to data set, such that a pan-species expression level cutoff to eliminate low-expressed genes [10, 11] may not be optimal. In this chapter, I describe a novel approach to cross-species RNA-seq analysis that circumvents the above problems by (1) using a kernel density estimation approach to select the normalized count cutoffs for low-expressed genes on a per-species basis, (2) reducing mRNA-seq data from gene-level transcript abundances to *gene-function*-level indices of transcriptional activity (using mappings of human genes to gene functional annotations and the Gene Set Variation Analysis technique [13]), and (3) comparing the gene function-level indices across species. Using mRNA-seq data from a comparative oncology study of human and dog bladder cancer [3], I illustrate in a step-by-step fashion (using the example code in the R programming language) how this method can enable unsupervised cross-species mRNA-seq analysis as well as enable supervised cross-species mRNA-seq comparisons.

2 Materials

The example mRNA-seq data sets shown in this section are from a cross-species (human and dog) study of bladder cancer, specifically transitional cell carcinoma (TCC) of the bladder [3]. The raw sequence data sets for the dog samples (TCC and normal bladder) are publicly available in the National Center for Biotechnology Information Gene Expression Omnibus database under accession number PRJNA339175. The human bladder cancer mRNA-seq data are publicly available from the Cancer Genome Atlas (TCGA) project through the Data Portal of the Genomic Data Commons website [14] at the National Cancer Institute (TCGA-BLCA mRNA-seq data set). The human bladder cancer mRNA-seq data files were obtained from under an approved Data Use Request for General Research Use (dbGaP project # 8059, approval # 34645) and processed to produce a matrix of per-gene/per-sample raw counts as described in [3]. In this section, I list the software tools and loaded data tables that would be required to carry out a cross-species analysis of mRNA-seq data through this workflow, using data from the bladder cancer cross-species study for illustration purposes. Completing this analysis workflow will require the following:

1. An installation of the R computing environment [15], which is a free and open-source implementation of the R statistical computing language [16] with an integrated software package

management system based on the Comprehensive R Archive Network (CRAN) [17]. The R installation should have the following software packages installed (including all of their dependencies): reshape2 [18], GSEABase {GSEABaseGeneSete:tz}, GSVA [13], ggplot2 [19], and DESeq2 [20], using the built-in "install.packages" function in R.

2. A data set of ortholog-based mappings of Ensembl gene identifiers for the non-human species (in the example vignette shown here, dog) to Human Gene Nomenclature Committee (HGNC) Gene Symbols. Such a mapping can be obtained using the BioMart tool [21] through the Ensembl genome portal [22]. This information should be contained in a single-column data frame "dog_ensgene_to_symbol" in which the Ensembl gene identifiers are the row names, as shown here (in this chapter, blue text indicates screen output from an R session):

```
>head(dog_ensgene_to_symbol)

                  Associated.Gene.Name
ENSCAFG00000022708
ENSCAFG00000022709
ENSCAFG00000022710
ENSCAFG00000022711
ENSCAFG00000022712
ENSCAFG00000022713                 ND1
```

(*see* **Note 1**). In the above, the R function "head" is used; this function prints out the first n lines (the default is $n = 6$ lines) of whatever object is the function argument.

3. A data set of human gene annotations mapping Ensembl gene identifiers to HGNC gene symbols. Such a mapping can be obtained using Ensembl BioMart. This information should be contained in a single-column data frame "human_ensgene_to_symbol" in which the Ensembl gene identifiers are the row names:

```
>head(human_ensgene_to_symbol)

                  Associated.Gene.Name
ENSG00000210049                 MT-TF
ENSG00000211459               MT-RNR1
ENSG00000210077                 MT-TV
ENSG00000210082               MT-RNR2
ENSG00000209082                MT-TL1
ENSG00000198888                MT-ND1
```

4. A file containing mappings of HGNC gene symbols to gene functional annotation categories, in Gene Matrix Transposed (GMT) format {GSEATeam:wt}. A comprehensive file of human gene annotations (for the Gene Ontology functional annotation categories [23]) can be obtained from the Molecular Signatures Database (MSigDB [24, 25]) web site [26] via a downloadable file "c5.all.v5.2.symbols.gmt."

5. For each species, raw counts and normalized counts of aligned mRNA-seq reads for each gene, based on Ensembl gene identifiers. Normalized counts can be obtained from the raw counts using the mRNA-seq analysis R software packages edgeR [27] (using the "cpm" function) or DESeq2 [20] (using the "counts" function with the option "normalized=TRUE"). The mRNA-seq counts for the first species (in the example analysis vignette for this article, dog) should be contained in a data frame "rsc_dog" for which the row names are Ensembl gene identifiers and the column names are sample identifiers. A portion of the data frame (along with the dimensions of the data frame) are shown here:

```
>head(rsc_dog)
                   TCC.1 TCC.2 TCC.3 TCC.4 TCC.5 TCC.6 TCC.7 normal.1 normal.2 normal.3
ENSCAFG00000014413     4    20   221    11    51     8    18        7       42       27
ENSCAFG00000014412    94    29    64    91   603   271   234       26      126      143
ENSCAFG00000014410    66    15    26    14   200    78    16       14       80      118
ENSCAFG00000014417     0     0     0     0     0     0     0        0        0        0
ENSCAFG00000014416   440   196   202   373   411   766   629      208      605      225
ENSCAFG00000014415    37    27    30    39   244   102   111       25      162       99

>dim(rsc_dog)

[1] 24580    10
```

In the above example, the R function "dim" gives the row and column dimensions of its argument (the data frame "rsc_dog").

6. Normalized mRNA-seq counts for the first species should be contained in a data frame "rsc_norm_dog" with the same row and column names as "rsc_dog." A portion of the data frame (along with the dimensions of the data frame) are shown here:

```
>head(rsc_norm_dog)
                    TCC.1  TCC.2  TCC.3  TCC.4  TCC.5  TCC.6  TCC.7 normal.1 normal.2
normal.3
ENSCAFG00000014413   4.15  38.30 386.92  16.99  15.42   5.00  13.44    20.73    21.81
19.48
ENSCAFG00000014412  97.44  55.53 112.05 140.53 182.27 169.37 174.69    77.00    65.44
103.18
ENSCAFG00000014410  68.42  28.72  45.52  21.62  60.46  48.75  11.94    41.46    41.55
85.14
ENSCAFG00000014417   0.00   0.00   0.00   0.00   0.00   0.00   0.00     0.00     0.00
0.00
ENSCAFG00000014416 456.10 375.33 353.65 576.01 124.24 478.74 469.57   616.00   314.24
162.35
ENSCAFG00000014415  38.35  51.70  52.52  60.23  73.76  63.75  82.87    74.04    84.14
71.43

>dim(rsc_norm_dog)

[1] 24580    10
```

Raw and normalized mRNA-seq counts for the second species (in this vignette, human) should be stored in data frames named "rsc_human" and "rsc_norm_human," respectively.

7. Sample group information for the mRNA-seq datasets for each species. For each species, the sample group information should be contained in a single-column data frame in which the row names are unique sample names. A portion of the data frame for the example human-dog analysis vignette (along with the dimensions of the data frame) is shown here:

```
>head(dog_sample_info)

    external_name
s01         TCC.1
s02         TCC.2
s03         TCC.3
s05      normal.1
s06         TCC.4
s34         TCC.5

>dim(dog_sample_info)

[1] 10  1
```

   Sample information for the other species (in this vignette, human) should be stored in a similar data frame (in this vignette, we will assume the data frame is named "human_sample_info").

8. Ortholog mappings between the two species, in the form of a two-column data frame whose first column contains Ensembl gene identifiers for the second species (in this example vignette, human) and whose second column contains the Ensembl gene identifier of an ortholog (if any) for the gene in the first species (in this example vignette, dog). Such a mapping can be obtained using Ensembl BioMart. A portion of the data frame for the example human-dog analysis vignette (along with the dimensions of the data frame) is shown here (*see* **Note 2**).

```
>head(human_dog_ensg)

  Ensembl.Gene.ID Dog.Ensembl.Gene.ID
1 ENSG00000261657
2 ENSG00000223116
3 ENSG00000233440
4 ENSG00000207157
5 ENSG00000229483
6 ENSG00000252952  ENSCAFG00000025776

>dim(human_dog_ensg)
[1] 65999     2
```

3 Methods

Below, I outline the steps required to carry out an unsupervised and a supervised comparison of mRNA-seq data sets from two species, using as an example mRNA-seq data sets from a cross-species (dog and human) study of bladder cancer. The first five steps of the

workflow involve performing identical transformations on the data sets for individual species, and thus, for those steps I show the example R commands only for the dog mRNA-seq data set (the R commands for the human data set are identical except for replacing "dog" with "human" in the commands).

1. For each species, compute (for each gene and each sample group) the geometric mean of the normalized $\log_2$ expression levels of the gene for all the samples in the sample group. Then, for each gene, compute the maximum sample-group-averaged expression level. Using the reshape2 R package, these steps can be performed for a given species using a single R command. For the dog data set, the command would be (example output also shown):

```
library(reshape2)
rsc_maxexp_dog <- setNames(aggregate(value~gene,
                                     data=aggregate(value~gene+condition,

data=merge(setNames(melt(log2(1+as.matrix(rsc_norm_dog))),

c("gene","sample_name","value")),
                                                          col_data_dog, by.x="sample_name",
by.y=0),
                                        FUN=mean),
                                    FUN=max),
                           c("gene","max_exp"))


>head(rsc_maxexp_dog)

                gene       maxexp
1 ENSCAFG00000014413  68.60036224
2 ENSCAFG00000014412 133.12610293
3 ENSCAFG00000014410  56.05202922
4 ENSCAFG00000014417   0.00000000
5 ENSCAFG00000014416 404.80579062
6 ENSCAFG00000014415  76.53825143

>dim(rsc_maxexp_dog)

[1] 24580     2
```

In the above R command, the function "melt" converts a data frame from a wide format to a narrow "melted" format [28], the function "merge" combines sample information with the melted data frame of normalized $\log_2$ counts, and the outer and inner calls to the function "aggregate" compute the inter-sample-group maximum and intra-sample-group geometric mean (respectively) of the $\log_2$ counts on a gene-level basis. Usually, the distribution (over genes) of the sample-group-maximum normalized expression levels is bimodal, with the lower mode corresponding to genes with either zero or extremely low transcript abundances in any sample group; however, the distribution can differ between experimental data sets and between species (Fig. 1). It is usually convenient to filter out low-expressed genes under the premise that their mRNA-seq counts are likely noise-dominated (this step also

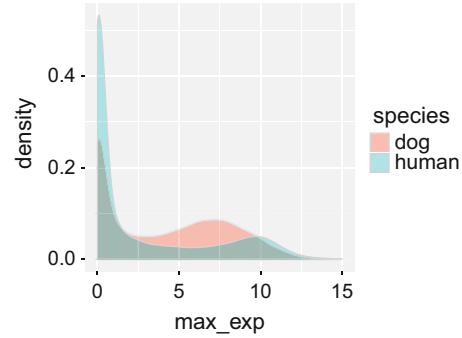


Fig. 1 Density distributions of the per-gene maximum (across sample groups) of the within-sample-group-average $\log_2$ expression level, for the dog and human mRNA-seq data sets. It is clear that different expression level thresholds apply for the two mRNA-seq data sets

has a benefit of reducing the number of hypothesis tests performed for individual-gene-level differential expression analysis) [20, 29]. Therefore, it is convenient to select (on a per-species basis) the local intermodal minimum as the cutoff for defining a gene that is expressed in at least one sample group [30], as shown in the next step.

2. For each species, compute the minimum-expression-level cutoff using kernel density estimation. For the dog dataset, the R commands and expected output would be

```
maxexp_density_dog <- density(rsc_maxexp_dog$max_exp)
cutoff_dog <- optimize(approxfun(maxexp_density_dog$x, maxexp_density_dog$y),
interval=c(1,10))$minimum

>cutoff_dog

[1] 2.96
```

(indicating an expression level cutoff of 2.96 for the dog mRNA-seq dataset). In the above statement, the R function "density" returns an R object representing the kernel density-estimated distribution of the function argument, the R function "approxfun" performs linear interpolation, and the R function "optimize" finds the point at which a given function of one or more variables attains a minimum value over a specified region of the domain of the given function.

3. For each species, filter the data matrices to remove any genes that are not expressed based on the minimum-expression-level cutoff that was defined above. For the case of the dog dataset, the R commands and example output would be

```
genes_exp_dog <- rsc_maxexp_dog$gene[which(rsc_maxexp_dog$max_exp >= cutoff_dog)]
rsc_exp_dog <- rsc_dog[genes_exp_dog, ]
rsc_norm_exp_dog <- rsc_norm_dog[genes_exp_dog, ]
```

```
>head(rsc_norm_exp_dog)
                    TCC.1  TCC.2   TCC.3  TCC.4  TCC.5  TCC.6  TCC.7 normal.1 normal.2
normal.3
ENSCAFG00000014413   4.146  38.30  386.92  16.99  15.42   5.00  13.44    20.73    21.81
19.48
ENSCAFG00000014412  97.440  55.53  112.05 140.53 182.27 169.37 174.69    77.00    65.44
103.18
ENSCAFG00000014410  68.416  28.72   45.52  21.62  60.46  48.75  11.94    41.46    41.55
85.14
ENSCAFG00000014416 456.104 375.33  353.65 576.01 124.24 478.74 469.57   616.00   314.24
162.35
ENSCAFG00000014415  38.354  51.70   52.52  60.23  73.76  63.75  82.87    74.04    84.14
71.43
ENSCAFG00000020948 572.204 712.36 1442.62 237.82  29.93 204.37 431.50   796.66   382.80
324.69

>dim(rsc_norm_exp_dog)
[1] 13572    10
```

4. For each species, select only the genes that are expressed (as in **step 3**) and whose Ensembl gene identifier maps to a HGNC gene symbol, including normalized expression data in $\log_2$ scale. For cases where multiple Ensembl gene identifiers map to a single HGNC gene symbol, average the gene expression data in $\log_2$ scale. For the case of the dog dataset, the R code and example output would be

```
rsc_norm_exp_mapped_dog <- data.frame(
     aggregate(. ~ Associated.Gene.Name,

data=merge(dog_ensgene_to_symbol[which(dog_ensgene_to_symbol$Associated.Gene.Name != ""),
                                      , drop=FALSE],
                   rsc_norm_exp_dog, by.x=0, by.y=0)[,-1],
        FUN=function(expvals) {mean(log2(expvals+1))}),
     row.names=1)
```

```
 >head(rsc_norm_exp_mapped_dog)
        TCC.1 TCC.2  TCC.3 TCC.4  TCC.5  TCC.6 TCC.7 normal.1 normal.2 normal.3
5S_rRNA 1.050 4.302  3.509 3.793 0.7252 1.4334 1.436    3.957    4.023    2.957
7SK     4.045 4.216  4.017 4.733 4.1991 4.1246 4.722    4.090    4.226    3.943
A2M     9.674 9.387 10.456 5.063 8.7988 4.1598 4.453    8.441    9.698   10.338
A4GALT  5.779 4.333  4.673 5.128 4.5833 4.3923 4.297    5.982    4.096    5.001
AAAS    4.908 5.148  5.171 4.768 5.8385 5.2668 4.762    3.983    4.261    5.155
AADAC   0.000 3.114  2.644 0.000 0.3810 0.7004 0.000    3.683    3.751    3.879

>dim(rsc_norm_exp_mapped_dog)
[1] 11703    10
```

In the above statement, the R function "merge" is used to combine data frames containing the mRNA-seq data and containing mappings between Ensembl gene identifiers and HGNC gene symbols; the R function "aggregate" is used to compute the per-sample average expression level for all Ensembl genes that map to a given HGNC gene symbol; and the function "data.frame" is here used to construct a new data frame from a given data frame, taking the first column of the given data frame as row names for the new data frame.

5. For each species, map the gene expression data to gene function-based expression index levels (specifically, enrichment scores based on the Gene Set Enrichment Analysis test statistic [24]) using the Gene Set Variation Analysis (GSVA) algorithm [13]. This can be done in two R commands:

```
library(GSEABase)

library(GSVA)

gsc5 <- geneIds(getGmt("c5.all.v5.1.symbols.gmt",
                       collectionType=BroadCollection(category="c5"),
                       geneIdType=SymbolIdentifier()))

gsva_dog_c5 <- gsva(data.matrix(rsc_norm_exp_mapped_dog), gset.idx.list=gsc5,
                    rnaseq=TRUE, method=c("gsva"), verbose=TRUE)

>head(gsva_dog_c5$es.obs)

                           TCC.1  TCC.2 TCC.3 TCC.4 TCC.5 TCC.6 TCC.7 normal.1 normal.2
normal.3
NUCLEOPLASM               -0.193 -0.205  0.02  0.13 -0.17  0.12  0.16   -2e-01   -0.033
0.09
EXTRINSIC_TO_PLASMA_MEMB. -0.022  0.049 -0.11  0.24  0.18  0.14  0.31    2e-01   -0.234
-0.25
ORGANELLE_PART            -0.206 -0.083  0.04  0.08 -0.05  0.13  0.07   -2e-01   -0.006
0.02
CELL_PROJECTION_PART       0.287 -0.124 -0.07  0.05  0.13  0.08 -0.23    2e-01    0.101
0.28
CYTOPLASMIC_VESICLE_MEMB. -0.008 -0.001 -0.22  0.34  0.19  0.45  0.39   -4e-01   -0.311
-0.37
GOLGI_MEMBRANE             0.122 -0.063  0.01  0.02  0.15  0.14 -0.04   -3e-04   -0.109
-0.17

>dim(gsva_dog_c5$es.obs)

[1] 1454   10
```

In the above example R code, the function "getGmd" reads in the gene set information from a file in a GMT format; the function "geneIds" returns the gene set information as a list; the "data.matrix" function constructs a numeric matrix from the contents of a data frame; and the function "gsva" transforms the mRNA-seq normalized $\log_2$ count data to gene function-level, per-sample enrichment scores using the Gene Set Variation Analysis method.

6. Merge the GSVA-transformed expression data matrices from the two species together, and merge that data with the sample metadata for the datasets for the two species together:

```
gsva_merged <- merge(gsva_dog_c5$es.obs, gsva_human_c5$es.obs, by.x=0, by.y=0)

>dim(gsva_merged)

[1] 1452  443

sample_data_merged <- data.frame(rbind(col_data_dog, col_data_human),
                                 species=c(rep("dog", nrow(col_data_dog)),
                                           rep("human", nrow(col_data_human))))

>head(sample_data_merged, n=15)

         condition species
TCC.1       cancer     dog
TCC.2       cancer     dog
TCC.3       cancer     dog
TCC.4       cancer     dog
TCC.5       cancer     dog
TCC.6       cancer     dog
TCC.7       cancer     dog
normal.1    normal     dog
normal.2    normal     dog
normal.3    normal     dog
UCCB_1      cancer   human
UCCB_2      cancer   human
UCCB_3      cancer   human
UCCB_4      cancer   human
UCCB_5      cancer   human

>dim(sample_data_merged)

[1] 443   2
```

In the above example R code, the "rep" function creates a vector by replicating its function argument, and "rbind" function combines two data frames by stacking them vertically.

7. Analyze the merged, GSVA-transformed data using Multidimensional Scaling [31] in two dimensions (also known as Principal Coordinates Analysis or PCoA). PCoA gives a planar coordinate location for each mRNA-seq sample such that the distance between each pair of samples in the plane corresponds (as closely as possible) to a quantitative dissimilarity measure between their respective mRNA-seq samples. The R command for this step, and example output, are as follows:

```
pcoa_results <- cmdscale(dist(t(gsva_merged)))$points

>head(pcoa_results, n=15)

          [,1] [,2]
TCC.1    -5.2   -1
TCC.2    -4.0   -1
TCC.3    -2.3    6
TCC.4     3.7    3
TCC.5    -2.7    2
TCC.6     0.3    5
TCC.7     3.7    2
normal.1 -1.1   -4
normal.2  1.7   -5
normal.3  3.9   -5
UCCB_1    2.2    2
UCCB_2   -6.0    6
UCCB_3   -7.7    6
UCCB_4   -3.0    6
UCCB_5   -9.3   -5

>dim(pcoa_results)

[1] 443   2
```

In the above R command, the "t" function returns the transpose of its argument, and the "cmdscale" function returns a list object whose "points" element contains the PCoA coordinates.

8. Merge the PCoA results with the sample metadata and visualize PCoA results as a scatter plot. Each of these steps can be performed in a single R command as shown below (with example output):

```
merged_pcoa_data <- data.frame(sample_data_merged, pcoa_results, )
```

```
>head(merged_pcoa_data, n=15)
```

```
         condition species    X1    X2
TCC.1       cancer     dog -5.20 -1.01
TCC.2       cancer     dog -3.97 -1.38
TCC.3       cancer     dog -2.31  5.53
TCC.4       cancer     dog  3.65  3.43
TCC.5       cancer     dog -2.70  2.01
TCC.6       cancer     dog  0.27  4.89
TCC.7       cancer     dog  3.66  1.68
normal.1    normal     dog -1.11 -4.41
normal.2    normal     dog  1.73 -5.33
normal.3    normal     dog  3.94 -5.44
UCCB_1      cancer   human  2.23  2.44
UCCB_2      cancer   human -5.97  5.65
UCCB_3      cancer   human -7.65  5.80
UCCB_4      cancer   human -2.98  6.34
UCCB_5      cancer   human -9.29 -5.45
```

```
>dim(merged_pcoa_data)
```

```
[1] 443   4
```

```
library(ggplot2)
```

```
ggplot_res <- ggplot(data=merged_pcoa_data) +
geom_point(aes(x=X1, y=X2, colour=condition, shape=species, alpha=species), size=4) +
guides(alpha=FALSE, size=FALSE) + theme(text=element_text(size=30)) +
scale_alpha_manual(values = c(1.0, 0.3))
```

```
>print(ggplot_res)
```

The PCoA co-visualization of the human and dog bladder datasets is shown in Fig. 2 (*see* **Note 3**). In the above R command, the function "ggplot" sets the data source for the plot, the function "geom_point" creates a scatter plot; the "theme" function is used to set the point size for the text labels in the plot; and the function "scale_alpha_manual" defines the two alpha transparency levels to be assigned to the two values for the "species" column in the data frame.

9. For each species, and for each gene that is expressed above background, compute an intraspecies $\log_2$ expression ratio and p-value (for the test of equal means of the $\log_2$ expression levels of the two sample groups) based on a negative binomial distribution-based count model:

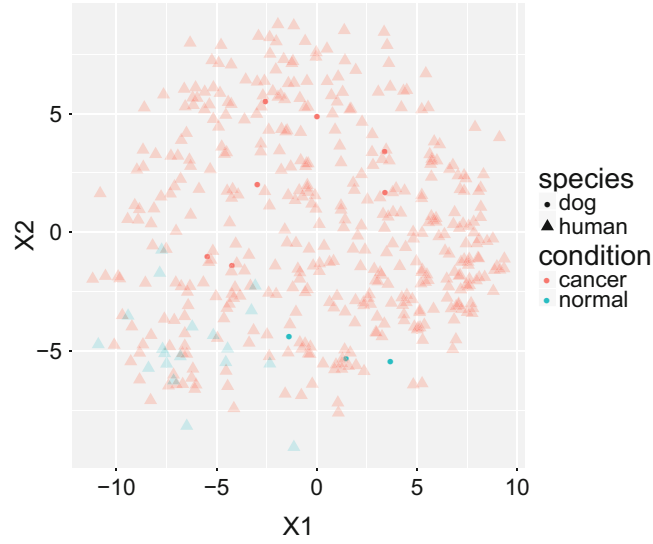


Fig. 2 Principal coordinates analysis (PCoA) showing combined transcriptome profiling data sets from human bladder samples and dog bladder samples. Even in this unbiased (i.e., blinded to sample group type) analysis of the data, there is a similar line of separation between normal and cancer samples in both dogs and humans, indicating a high degree of similarity between the human and dog bladder cancer transcriptomes at the levels of gene functions

```
library(DESeq2)

deg_dog_df <- results(DESeq(DESeqDataSetFromMatrix(countData=rsc_exp_dog,
                            colData=data.frame(condition=factor(col_data_dog$condition,
                                                               levels=c("normal","cancer")),

row.names=rownames(col_data_dog)),
                                                    design=~condition)))


>head(deg_dog_df)

log2 fold change (MAP): condition cancer vs normal
Wald test p-value: condition cancer vs normal
DataFrame with 6 rows and 6 columns

                   baseMean log2FoldChange     lfcSE      stat    pvalue      padj
                  <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
ENSCAFG00000014413      17.3        -0.4098     0.613   -0.6684    0.5039     0.655
ENSCAFG00000014412     117.8         0.6927     0.399    1.7362    0.0825     0.185
ENSCAFG00000014410      45.4        -0.4414     0.538   -0.8211    0.4116     0.575
ENSCAFG00000014416     392.7         0.1547     0.489    0.3162    0.7518     0.846
ENSCAFG00000014415      65.3        -0.3228     0.314   -1.0291    0.3034     0.468
ENSCAFG00000020948     513.7         0.0463     0.738    0.0628    0.9499     0.969

>dim(deg_dog_df)

[1] 13572     6
```

In the above output table, "padj" (i.e., p_{adj}) is the p-value that has been adjusted for multiple hypothesis tests (i.e., a q-value) using the Benjamini-Hochberg correction [32]. The "DESeqDataSetFromMatrix" function constructs an object containing the sample information and mRNA-seq count data for each (expressed) gene; the "DESeq" function performs

per-gene differential expression testing; and the "results" function returns the results of the differential expression testing as a data frame. The R command to compute $\log_2$ expression ratios and p-values for the second species (in this vignette, human) would be identical to the above command, but with "human" replacing "dog."

10. Select ortholog pairs of genes for which the genes are differentially expressed ($p_{adj} < 0.05$ for each of the genes in the ortholog pair), and organize $\log_2$ ratios for each of the ortholog pairs (in the two respective species) in a two-column data frame.

```
deg_dog_human_comb_df <- merge(as.data.frame(deg_dog_df),
                               merge(as.data.frame(deg_human_df),
                                     human_dog_ensg, by.x=0, by.y=1), by.x=0, by.y=8)

deg_dog_human_comb_df_filt <- deg_dog_human_comb_df[which(deg_dog_human_comb_df$padj.x <
0.05 &
deg_dog_human_comb_df$padj.y < 0.05),]

>dim(deg_dog_human_comb_df_filt)

[1] 1788   14
```

Thus, 1788 ortholog pairs of genes are differentially expressed in both dog and human bladder cancer (vs. normal bladder) with a false discovery rate of 0.05.

11. Scatter plot of $\log_2$ ratios for ortholog gene pairs, across mRNA-seq studies in two species.

```
scatter_plot <- ggplot(deg_dog_human_comb_df_filt) +
      geom_point(aes(x=log2FoldChange.x, y=log2FoldChange.y)) +
      theme(text=element_text(size=20)) +
      xlab(expression(log[2](cancer/normal)~"in dog")) +
      ylab(expression(log[2](cancer/normal)~"in human"))

>print(scatter_plot)
```

This analysis shows that there is a significant correlation between upregulation or downregulation of a gene in human bladder cancer, and upregulation or downregulation of its dog ortholog in canine bladder cancer (Fig. 3).

4 Notes

1. Not all non-human genes will have a known human ortholog with a HGNC symbol.
2. A single-human gene can have multiple dog orthologs (or vice-versa), and thus, a given Ensembl human or dog gene identifier can appear multiple times in the above data frame (thus explaining the large number of rows in the data frame).
3. Even in the unsupervised analysis shown in Fig. 2, a general separation of normal from cancer samples is evident across the

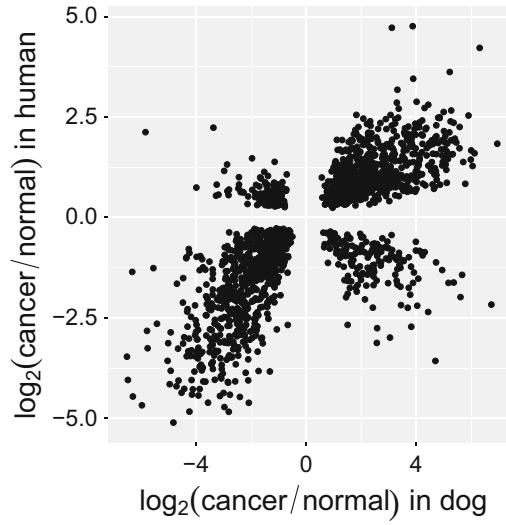


Fig. 3 Scatter plot of gene expression ratios of 1788 pairs of gene orthologs (human-dog) in bladder cancer vs. normal bladder. Up- or downregulation of a gene in human bladder cancer is highly predictive of up- or downregulation of its dog ortholog in bladder cancer (odds ratio = 29.8, $p < 10^{-15}$, Fisher's exact test)

two species (although given the very large TCGA cohort size, some overlap between cancer and normal is seen for the TCGA dataset).

Acknowledgments

This work was supported by the National Science Foundation (award 1553728-DBI), the PhRMA Foundation (Research Starter Grant in Informatics), the Medical Research Foundation of Oregon (New Investigator Grant), and the Animal Cancer Foundation (Comparative Oncology Award). S.A.R. thanks Shay Bracha and Cheri Goodall for kindly providing the dog bladder RNA samples that were used in the transcriptome profiling study [3], Tanjin Xu for assistance with the mRNA-seq data processing, Brent Kronmiller for help with designing the dog mRNA-seq study, and Ilya Shmulevich, Sheila Reynolds, and Matti Nykter for advice.

References

1. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226
2. Lister R, O'Malley RC, Tonti-Filippini J et al (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523–536. https://doi.org/10.1016/j.cell.2008.03.029
3. Ramsey SA, Xu T, Goodall C et al (2017) Cross-species analysis of the canine and human bladder cancer transcriptome and exome. Genes Chrom Cancer (4):56, 328–343. https://doi.org/10.1002/gcc.22441

4. Fowles JS, Brown KC, Hess AM et al (2016) Intra- and interspecies gene expression models for predicting drug response in canine osteosarcoma. BMC Bioinformatics 17:93. https://doi.org/10.1186/s12859-016-0942-8
5. Dhawan D, Paoloni M, Shukradas S et al (2015) Comparative gene expression analyses identify luminal and basal subtypes of canine invasive urothelial carcinoma that mimic patterns in human invasive bladder cancer. PLoS One 10:e0136688. https://doi.org/10.1371/journal.pone.0136688
6. Seok J, Warren HS, Cuenca AG et al (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci 110:3507–3512. https://doi.org/10.1073/pnas.1222878110
7. Shay T, Jojic V, Zuk O et al (2013) Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc Natl Acad Sci 110:2946–2951. https://doi.org/10.1073/pnas.1222738110
8. Chan ET, Quon GT, Chua G et al (2009) Conservation of core gene expression in vertebrate tissues. J Biol 8:33. https://doi.org/10.1186/jbiol130
9. Brawand D, Soumillon M, Necsulea A et al (2011) The evolution of gene expression levels in mammalian organs. Nature 478:343–348. https://doi.org/10.1038/nature10532
10. Lin S, Lin Y, Nery JR et al (2014) Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci 111:17224–17229. https://doi.org/10.1073/pnas.1413624111
11. Gilad Y, Mizrahi-Man O (2015) A reanalysis of mouse ENCODE comparative gene expression data. F1000Research 4:121. 10.12688/f1000research.6536.1
12. Sudmant PH, Alexis MS, Burge CB (2015) Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol 16:287. https://doi.org/10.1186/s13059-015-0853-4
13. Hänzelmann S, Castelo R, Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 14:7. https://doi.org/10.1186/1471-2105-14-7
14. NIH Genomic Data Commons Data Portal (2016) v. 4.0. https://gdc-portal.nci.nih.gov
15. Ripley BD (2001) The R project in statistical computing (2001). MSOR Connections. Newsl LTSN Maths Stat OR Network 1:23–25
16. Ihaka R, Gentleman R (1995) R: a language for data analysis and graphics. J Comp Graph Stat 5:299–314
17. Hornik K (2012) The comprehensive R archive network. Comput Stat 4:394–398. https://doi.org/10.1002/wics.1212
18. Wickham H (2007) Reshaping data with the {reshape} package. J Stat Software 21:1–20
19. Wickham H (2009) ggplot2: elegant graphics for dataanalysis. Springer, New York, NY
20. Love MI, Huber W, Anders S (2013) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/PREACCEPT-8897612761307401
21. Smedley D, Haider S, Ballester B et al (2009) BioMart—biological queries made easy. BMC Genomics 10:22. https://doi.org/10.1186/1471-2164-10-22
22. Cunningham F, Amode MR, Barrell D et al (2015) Ensembl 2015. Nucleic Acids Res 43:D662–D669. https://doi.org/10.1093/nar/gku1010
23. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29. https://doi.org/10.1038/75556
24. Subramanian A, Tamayo P, Mootha VK et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 102:15545–15550. https://doi.org/10.1073/pnas.0506580102
25. Liberzon A (2014) A description of the Molecular Signatures Database (MSigDB) Web site. Methods Mol Biol 1150:153–160. https://doi.org/10.1007/978-1-4939-0512-6_9
26. Molecular Signatures Database (MSigDB) (2016) v. 5.2. http://software.broadinstitute.org/gsea/msigdb
27. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616
28. Wickham H (2014) Tidy data. J Stat Software 59:10. 10.18637/jss.v059.i10
29. Lin Y, Golovnina K, Chen Z-X et al (2016) Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. BMC Genomics 17:28. https://doi.org/10.1186/s12864-015-2353-z
30. George NI, Chang C-W (2014) DAFS: a data-adaptive flag method for RNA-sequencing data to differentiate genes with low and high expression. BMC Bioinformatics 15:92. https://doi.org/10.1186/1471-2105-15-92
31. Cox MAA, Cox TF (2001) Multidimensional scaling, 2nd edn. Chapman and Hall, Boca Raton, FL
32. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Chapter 15

Multi-agent Simulations of Population Behavior: A Promising Tool for Systems Biology

Alfredo Colosimo

Abstract

This contribution reports on the simulation of some dynamical events observed in the collective behavior of different kinds of populations, ranging from shape-changing cells in a Petri dish to functionally correlated brain areas in vivo. The unifying methodological approach, based upon a Multi-Agent Simulation (MAS) paradigm as incorporated in the NetLogo™ interpreter, is a direct consequence of the cornerstone that simple, individual actions within a population of interacting agents often give rise to complex, collective behavior.

The discussion will mainly focus on the emergence and spreading of synchronous activities within the population, as well as on the modulation of the collective behavior exerted by environmental force-fields. A relevant section of this contribution is dedicated to the extension of the MAS paradigm to Brain Network models. In such a general framework some recent applications taken from the direct experience of the author, and exploring the activation patterns characteristic of specific brain functional states, are described, and their impact on the Systems-Biology universe underlined.

Key words Multi agent systems, Complex adaptive systems, Dynamic simulations, Ising models, Brain networks

1 Introduction

1.1 Generalities

Little doubt remains that the emergence of complexity in most natural phenomena [1, 2] is not an exception but a rule imposing, as a corollary, a severe limit to their quantitative prediction on the long term: hence, a *systemic* perspective relying upon statistical [3] and simulation studies [4, 5] could open new avenues in the field. Moreover, any reliable approach to the simulation of population dynamics stemming from the interactions among individuals as well as produced by environmental factors, in the last fifty years should be most welcome.

The left panel of Fig. 1 (modified from [6]) shows, in the evolving landscape of modern informatics, the relatively recent

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_15,

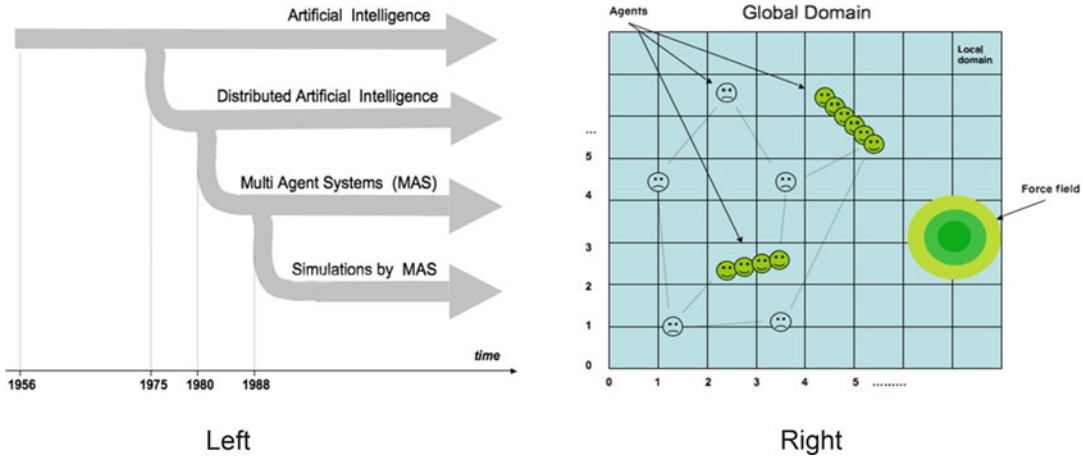


Fig. 1 *Left*: Development of Multi-Agent Systems (MAS) from other similar disciplines (Modified from [6]). *Right*: The space (global domain) in which agents move is a grid of variable size where each single location (local domain) can be endowed with specific features, like attractive/repulsive ability caused by force-fields of defined intensity and shape

appearing of computational strategies appropriate to reproduce complex functions and collective behavior of interacting agents.

The general philosophy inspiring Multi Agent Systems (MAS) is reminiscent of cellular-automata and adaptive systems [7] whose basic idea [8] is generalized including the possible influence of agents in any cell of the grid on the behavior of agents localized in any other cell.

As a matter of fact, MAS are most often implemented in studies concerning numerous populations of interacting agents capable of simple, autonomous actions producing complex, collective behavior,. It is worth noting that single agents lack a full global view of the system on their own, but show self-organization as well as self-steering and sophisticated individual behaviors. In other words, the utility of MAS remains outstanding when one looks for:

- A qualitative account of the global population's behavior from the known features of the single agents and of their spatial location.
- A quantitative refinement of the model parameters describing the collective behavior with the aim of fitting at best the dynamics observed in experimental data.

Among the several software tools available in this framework NetLogo™ [9] has the advantage of being: (1) frequently updated by an active users-community, (2) endowed with a relatively smooth learning curve, and, (3) last but not least, freely available. NetLogo™ is an interpreted programming environment based upon LOGO™, one of the classical Artificial Intelligence (A.I.) languages, and is particularly well suited for modeling complex

systems' behavior under the influence of environmental factors [10, 11]. In particular, it allows representing such factors in the form of space-dependent force fields, and making the agents responsive to them by associating the force-fields with the properties of the spatial grid where the agents live and move. Some obvious applications include temperature/pressure gradients, chemical diffusion, electromagnetic/gravitational interactions, etc.

1.2 MAS and System Biology

Multi-Agent Systems (MAS) and Systems Biology (SB) in their relatively recent history share a number of deep connections, both from a theoretical perspective and in a more practical, applicative dimension. It could not be different from that, since Systems Biology too is based on the understanding that the whole is greater than the sum of the parts and actually shares with MAS a holistic approach to explore the biological complexity and to design predictive, multiscale models. Thus, once again Complexity is the most significant, pregnant, and meaningful keyword characterizing the problems of interest for both MAS and System Biology.

Having stressed that, however, it is difficult to overemphasize the tenet that both are concerned with networks. As a matter of fact, the mutual relationships between agents in MAS find an immediate operational representation in the network's links. The semantics of words like *network* and *system*, on the other hand, directly refers to an ensemble of (even nonlinearly) interacting elements.

Finally, it is quite interesting that the following citations from ISB [12]:

> "*.... our bodies are made up of many networks that are integrated and communicating on multiple scales .....*" and "*... Systems biology looks at these (biological) networks across scales to integrate behaviors at different levels, to formulate hypotheses for biological function and to provide spatial and temporal insights into dynamical biological changes....*"

seem to reflect exactly the most demanding and exciting ambition of MAS.

2 The Influence of External Force-Fields

In our own experience we could appreciate the advantage of using MAS-based simulations particularly in the study of:

1. Shape changes induced by force fields in cell populations in vitro.
2. Synchronous activation of brain neurons.

In all the cases reproducing the observed time-dependent trends provided qualitative, precious hints to work out solid mechanistic hypotheses.

The unifying element in the above cases was the presence of force-fields affecting the dynamic behavior of the agents. The fields were time-invariant and their spatial dependence was of the form

$$FF_d = FF_o e^{-d}.$$

where FF_d is the force-field strength at distance d from the source FF_o.

The numerical resolution of such space dependence was directly proportional to the granularity of the agents' world (the grid).

In the detailed description of each case the research problem in the background will be briefly sketched, as well as the simple rules defining the behavior of the single agents and the typical results provided by the simulations.

2.1 Cell Shapes and Fields

2.1.1 The Problem

The morphofunctional changes within a population of initially similar elements can be the result of coexisting endogeneous (genetic) and/or exogeneous (environmental) factors. The quantification of their relative importance remains of crucial interest, due to the obvious functional—pathological modifications linked to the morphological changes.

According to a crude but quite useful approximation, switching from one conformation ($\mathbf{C}_1$) to another ($\mathbf{C}_2$), namely $\mathbf{C}_1 \rightarrow \mathbf{C}_2$, can be described by a simple first-order process. The rate of the process, V_1, is ruled by a kinetic constant $\mathbf{k}_1$, namely $V_1 = \mathbf{k}_1 \times [\mathbf{C}_1]$, where the square brackets indicate molar concentration. If the switch is reversible, the same reasoning applies to the reverse process, $\mathbf{C}_1 \mathrm{J} \leftarrow \mathbf{C}_2$, leading to an equilibrium condition when $V_1 = \mathbf{k}_1 \times [\mathbf{C}_1] = V_2 = \mathbf{k}_2 \times [\mathbf{C}_2]$. Thus, under equilibrium conditions the $\mathbf{C}_1/\mathbf{C}_2$ ratio is given by $\mathbf{k}_2/\mathbf{k}_1$.

2.1.2 The Model

The interconversion from a "regular," yellow (Y) to a "spiky," blue (B) shape, within a population of up to several thousand cells is modeled as a standard biochemical equilibrium between two conformations [13] influenced by:

- Chemical factors ("apparent" kinetic constants for the B→Y and Y→B conversions).
- Physical factors (external force-fields).

The "apparent" kinetic constants include an "intrinsic" factor as well as two "environmental" factors regulated by two specific force-fields. Each of the two force-fields influences only one of the two kinetic constants.

Figure 2 shows how a population of agents all in the same yellow state at time = $\mathbf{t}_0$ (panel A), undergoes a reversible shift to another (blue) state, reaching different equilibria ruled by different $\mathbf{k}_1$:$\mathbf{k}_2$ ratios. Such ratios are 1:1; 0.7:0.3 and 0.3:0.7 in panels B, C

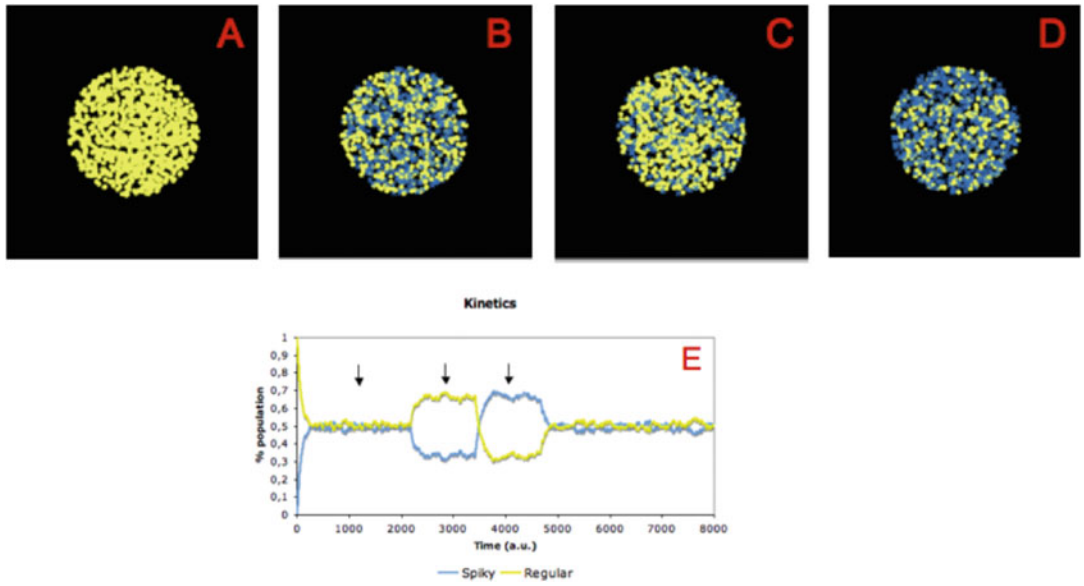


Fig. 2 Dynamics of conformational changes. A 100% "regular, yellow" population of agents (pretty similar to a cell population in a Petri dish, panel **A**) at $\mathbf{t}_0$ undergoes a reversible transition to a "spiky, blue" state. The yellow/blue ratio at equilibrium, as defined by the rates of the direct and inverse change, is about 1, 2, and 0.5 in panels **B**, **C,** and **D**, respectively (see also the text). Panel **E** shows the time course of an overall process including the transitions A→B→C→D→A, and the *arrows* indicate, from left to right, the equilibrium state in **B**, **C,** and **D**

and D, respectively, where $\mathbf{k}_1$ rules the yellow → blue rate. In panel D the global time course of the considered process is reported and the arrows indicate the three equilibrium states. Such a simulation exercise can be obviously approached through the corresponding set of differential equations, available in this case. However, the appealing feature of the MAS simulator remains the straightforward, intuitive connection with the statistical nature underlying any first-order kinetic process: for example, when $\mathbf{k}_1 = 0.7$ the switch from $\mathbf{C}_1$ to $\mathbf{C}_2$ of each member of the agents population is provided by the following statement:

ask agents [*if color* = *yellow and random 100* < *70* [*set color* = *blue*]]

where random <100 is a random integer between 0 and 99.

2.1.3 Some Results

Connecting structural and functional properties has always been the most useful way to describe and predict the behavior of living systems, independent of size and specific features. This entitles the use of the above modeling approach at different dimensional levels and, in particular, in the case of cell populations where phenotypical (structural) changes may reflect events of huge physiological and pathological relevance. More specifically, having in mind the changes in cellular and tissue architecture associated with neoplastic

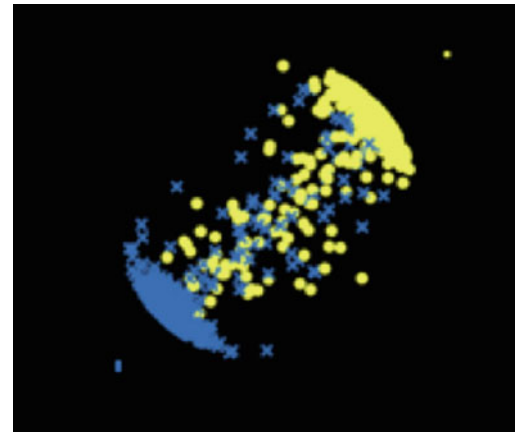


Fig. 3 Clustering of "regular" and "spiky" agents induced by specific force-fields. The agents simulate a mixture of two (*yellow* and *blue*) cell types in a Petri dish and the sources of the force-fields are indicated by *small* dots of the corresponding *color*, outside the Petri dish

transformations [14] the two types of agents in the shown simulations have been characterized by both different color (yellow/blue) and shape (regular/spiky). Due to the crucial influence on the cells behavior attributed to a manifold of exogenous (environmental) factors, it also appears of special interest the easy reproduction of that influence by means of external force-fields, as exemplified in Fig. 3. The aggregation induced by different force-fields on a specific cell type is just one of the many relevant effects which could be made more and more realistic as far as their type and specificity is concerned. As a matter of fact, some relatively slight improvements in the simulator would produce relevant predictions, amenable to experimental tests. Such improvements mainly deal with a precise modeling the influence of gravitational-fields, cell density, and various chemical factors on the conformational switch.

2.2 Synchronized Neurons

2.2.1 The Problem

A depolarizing wave moving at a 3 mm/min speed in the rabbits cortex has been observed since a long time by Leao [15] and named Cortical Spreading Depression (CSD), since after its passage the cortex remained inactive for some time. In 1994, however, Lauritzen [16] showed that associated with the visual aura in human migraine was a high-activity wave moving in the anterior direction from the occipital region at speed from 2 to 6 mm/min. The synchronous activity of large neuron patterns is triggered by the synchronous activity in a relatively restricted area, which represents the "epileptic focus." From that area, the synchronous activity spreads throughout the whole brain, so that an increasing number of neurons become active at the same time. CSD is not limited to the occipital area: its starting point may be observed most

frequently in a specific (CA1) hyppocampal area, followed by the neo-cortex, and it remains a most interesting phenomenon of neural synchronization.

2.2.2 The Model

The facilitated tuning of the reciprocal interactions between agents in a MAS has always produced a most impressive output whenever such interactions are endowed with a rythmic trend reflected by a set of regular/periodic oscillations.

Brain neurons may be represented by agents having an activity-cycle of the same duration and able to switch from a resting to an active state at a given instant within its activity-cycle. Activation of each single neuron occurs at a randomically chosen instant within its activity-cycle; it often occurs, however, that synchronous activation patterns emerge as a result of an autocatalytic process in which more and more neurons become active at the same time. The initial random activation becomes synchronized since each neuron tries and rearranges its own activity cycle in order to match the activity cycle of other neurons in the neighborhood. It is allowed to do that, however, only if the neuron senses a minimum number of active neurons within defined spatial and temporal windows. In order to reproduce periodic synchronization we made the activity status of each neuron depending upon its metabolic energy so that, below a given energy threshold, the neuron is unable to synchronize with other neurons. In addition, the metabolic energy level influences the maximal distance at which the neuron may sense other neurons. Under resting conditions (random activation), the metabolic energy increases at a relatively slow rate; during synchronous activity, however, the decrease in energy proceeds at a much higher rate. In other words, the synchronous activation regime is characterized by a high metabolic cost, which cannot be sustained for a long time. Whenever a lower energy threshold is reached, the synchronization mechanism is stopped and the random activation (resting condition) restored. Thus, the energy supplies can be replenished making a further synchronization event possible.

2.2.3 Some Results

Figure 4 shows the activity patterns observed in the area representing a coronal section of the human brain, by means of a simulation device described elsewhere [17]. The six panels in the figure refer to the neurons activity distribution at various times (proportional to the ticks number in each panel) from an initial fully randomic distribution (see the inset) to the synchronous ring of neurons clustering in different and alternating regions of the brain.

As compared to other programmable tools specialized for neuronal systems, like Neuron™ [18], the MAS approach appears much more flexible, although probably less powerful in terms of manageable models size. As an example, by a minimum amount of programming effort in the Netlogo environment, it was relatively straightforward to work out simulations based upon completely

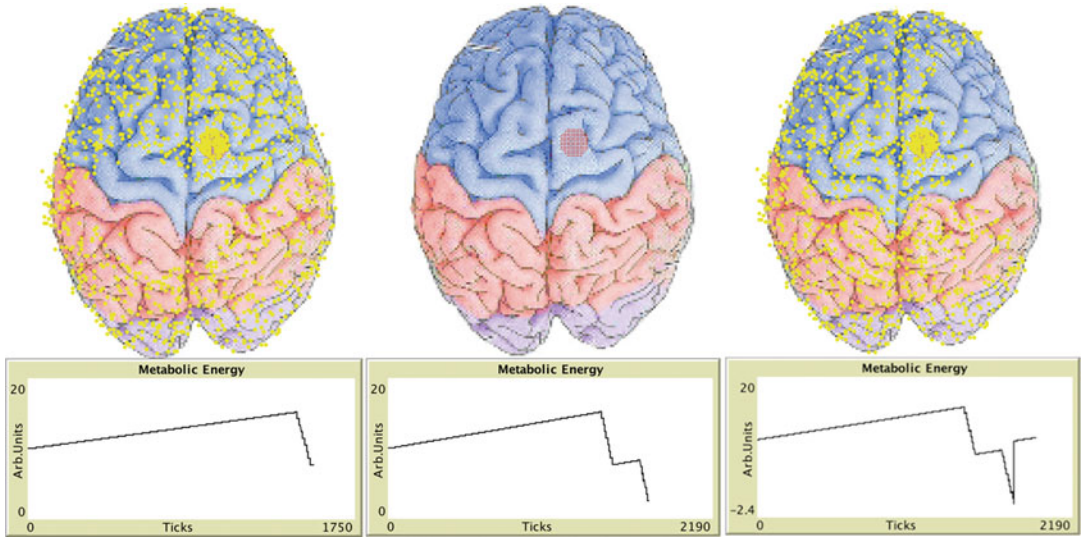


Fig. 4 Synchronized activity in brain neurons. The *top left*, *middle,* and *right panels* show the active neurons at the maximum, minimum, and intermediate synchronization levels, respectively. The *bottom panels* show the corresponding trend of metabolic energy. Notice, close to the central sulcus in the right hemisphere, the "epileptic focus" triggering the synchronization wave

different mechanisms, as those shown in Figs. 4 and 5. On the basis of the latter one, in particular, it looks very promising the possible synergic combination of a multi-agent environment with other general-purpose programmable simulation algorithms (like genetic algorithms) to underpin the basic mechanisms at the root of highly complex functions in the human Central Nervous System.

In any case, a common mechanism for migraine and epilepsy based upon the synchronization of specific neuronal regions was confirmed and satisfactorily simulated also by a multi-agents programming environment, which paves the way for further, more ambitious extensions [19].

3 Functional Connections and Correlations in Brain Networks

Using a MAS-based simulation engine to study the functional properties of a neuronal network is simplified by the straightforward identification of MAS agents and links with, respectively, network nodes and ties. In a MAS representation of neuronal networks, in fact, an agent can be assimilated to a single neuron, to a set of functionally correlated neurons, or even to a given brain region. Moreover, functional relationships among sub-networks are not necessarily associated with well-defined anatomical connections.

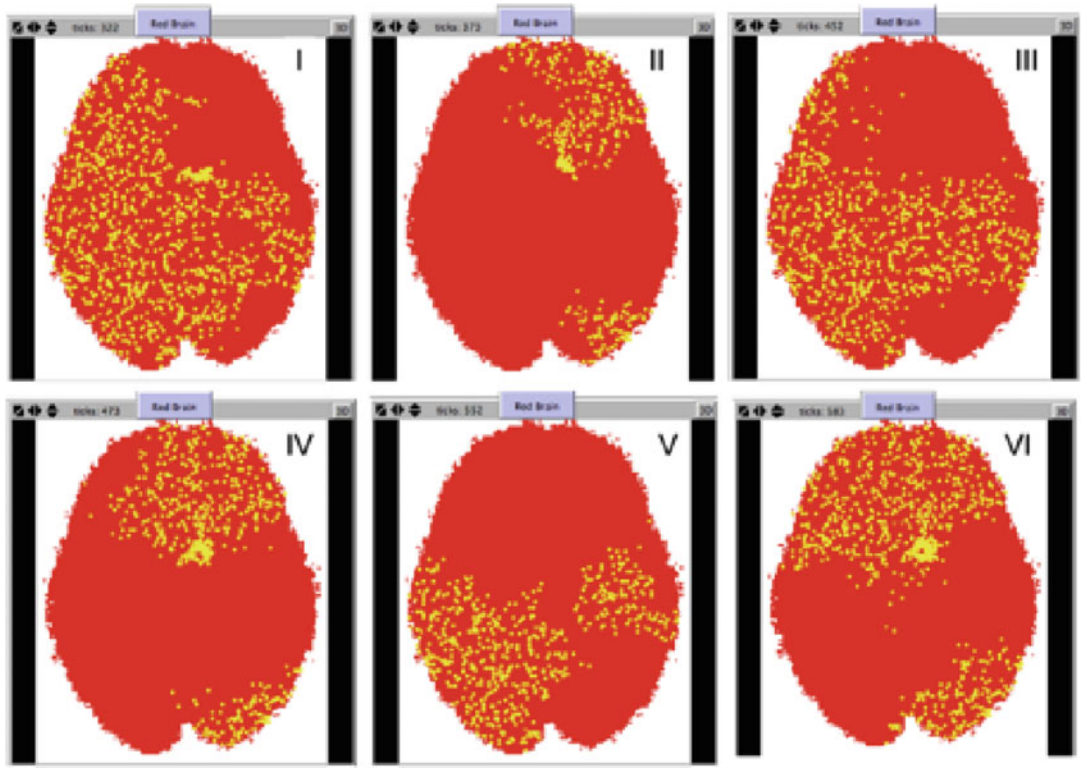


Fig. 5 Oscillating activity waves of brain neurons simulated by a MAS. Panels from *I* to *VI* have been recorded in sequence at about the same time interval (10 ± 5 s) from each other. The program used in the simulation is described in [17]

The scarcity of detailed experimental information is often a severe limit to designing accurate and realistic mechanistic models; the intrinsic flexibility of a MAS-based simulation strategy, however, may often smooth the problem. At our knowledge, such a strategy has not been systematically attempted as yet to simulation of regulatory phenomena where local interactions between couples of nodes (neurons) induce the emergence of a global behavior pointing to a homeostatic equilibrium.

A direct comparison with an alternative strategy based on differential equations would probably be unfavorable to our approach in terms of speed and efficiency, but not in terms of straightforward, intuitive correspondence between in vivo and in silico events.

The basic experimental reference typical of any simulation of functionally correlated brain areas is depicted in Fig. 6: from such information a functional connectivity matrix [20] is derived and henceforth used as a compact reference to the number of nodes in the network and to their reciprocal interaction.

The simulation strategies designed in the two following sections deal with the so-called *default* condition of brain activity, which is in a resting state characterized by the absence of any

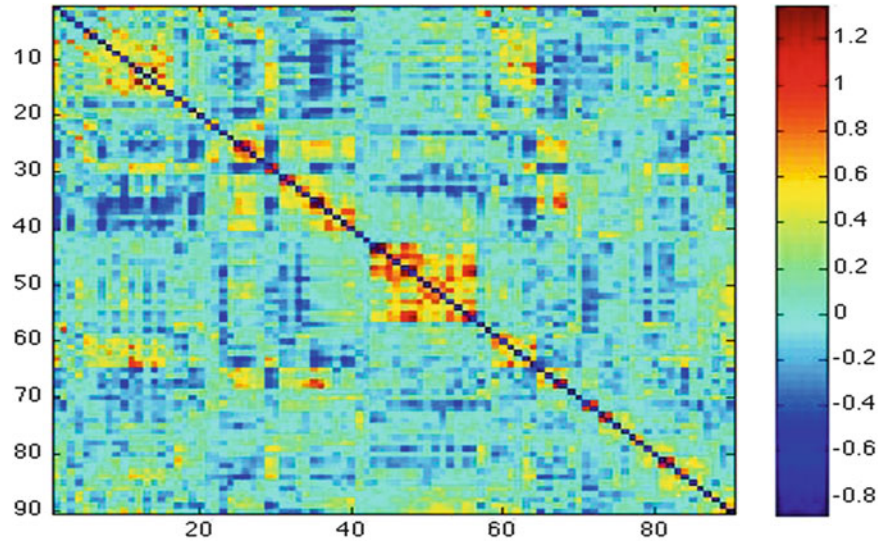


Fig. 6 Correlation matrix between Regions of Interest (ROIs) of a single subject from functional Magnetic Resonance Imaging (fMRI) data. The raw activation (BOLD effect) data are plotted in a normalized, false color scale. The original fMRI images of the subject are divided into 90 ROIs and from each ROI the time series are extracted (acquisition time 450 s at 3 Hz frequency, in order to obtain discrete series of 150 elements). The correlation matrix was calculated for all possible couples of the 90 ROIs (modified from [32])

external stimulation. The focus was on the type of interactions occurring between different brain areas in terms of positive/negative correlations in one case, and in terms of synchronous activation in the other.

3.1 The Role of Negative Correlations

3.1.1 The Problem

Up to now, studies on functional brain networks mainly focused on positive correlations between cerebral areas due to the still not well-defined nature of the negative ones. Negative correlations have been often interpreted as a preprocessing artifact, namely as a global signal regression [21], although significant relations have been found between negative correlations and caffeine intake, meditation, brain development and aging, schizophrenia, different social and cognitive tasks [22]

3.1.2 The Model

A possible topology of negative functional networks has been described by several recent papers by means of Network Theory [23] and of Balance Theory [24]. While the former theory is particularly useful in providing quantitative parameters of network topologies, the latter one helps in defining the conditions for the networks' functional stability. This is of special interest, since it may suggest possible mechanisms accounting for the influence of negative correlations on the functional equilibria between brain areas.

A well-known feature of the MAS is the flexible rearrangement of connection patterns according to simple rules. Such rules, in the case of networks, smooth the difference in the activity levels of

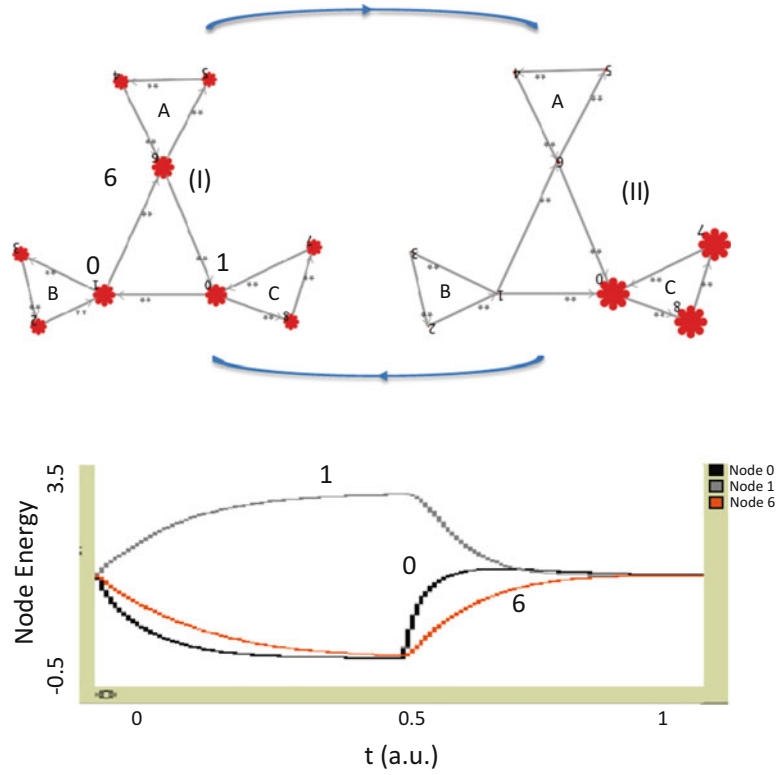
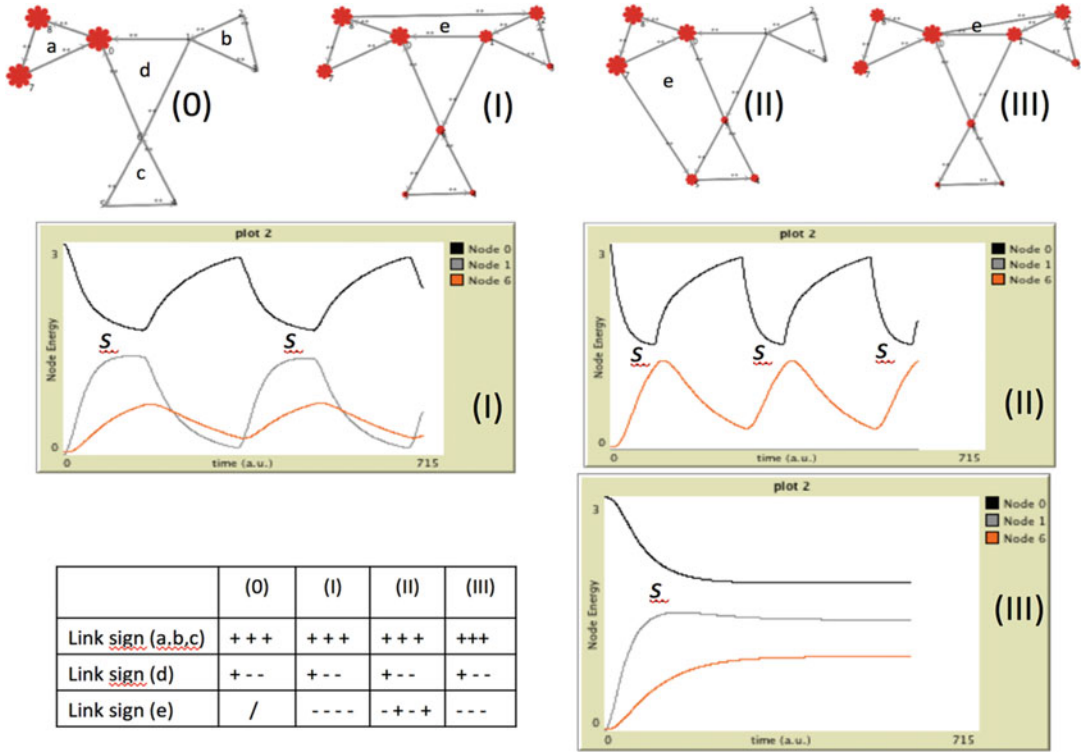


Fig. 7 *Top*: Dynamic features of a simple 9-node network. The direction of the link between nodes 0, 1 is 0 → 1 in (*II*) and 1 → 0 in (*I*). The latter condition is the only one compatible with a global stability. *Bottom:* The (*I*) → (*II*) and (*II*) → (*I*) transitions start at $t = 0$ and $t = 0.5$, respectively. The graph describes the consequential time-dependent changes in activity (size) of nodes 0, 1, and 6 (representative of the three sub-networks) (modified from [32])

different sub-networks by acting on the sign of links or even forming new links aiming to reach a more stable global condition.

Figure 7 (top panel) shows a simple 9-node network implemented in a Multi Agents System accounting for the arrangement of both positive and negative correlations as well as for the homeostatic regulation between regions (sub-networks A, B, C) through a negative feedback mechanism. Thanks to the graphical facilities of Netlogo™ [9, 10] environment, the (de)activation signals traveling from input node to output nodes are easily reproduced and visualized in terms of time-dependent changes in the nodes' size.

The mechanism in Fig. 7, coupled to the underlying MAS programming environment, reproduced some relatively sophisticated homeostatic phenomena (*see* below), and its possible extension to the study of stability alterations of pathological significance is foreseeable.



	(0)	(I)	(II)	(III)
Link sign (a,b,c)	+ + +	+ + +	+ + +	+++
Link sign (d)	+ - -	+ - -	+ - -	+ - -
Link sign (e)	/	- - - -	- + - +	- - -

Fig. 8 Simulating homeostatic equilibria of sub-networks by MAS. The architecture of the network is the same as in Fig. 3. From an initial state characterized by the hyperactivity of the sub-network in (0), a more balanced global configuration can be reached by activating some extra links between the nodes in sub-network a and nodes in sub-networks b and c. This gives rise to a new sub-network, e, located in different possible locations, three of which are represented in (*I*), (*II*), and (*III*). The table in the lower left corner contains the sign of the links in the sub-networks of the more balanced (*III*), less balanced (0), and intermediate (*I*, *II*), networks. The "S" indicates the energy level corresponding to an equilibrium state (modified from [32])

3.1.3 Some Results

Figure 8 shows an intuitive dynamical representation of the network global activation states as obtained making the nodes' size proportional to the activation level which, in turn, depends upon the energy(information) flowing through the links. Thus, in order to overcome the unbalanced state (0) in the figure, a new link of appropriate sign can in principle be formed involving one of three different couples of nodes, as shown in (I), (II), and (III). In full agreement with the second stability theorem of the Balance Theory [24], only in the last case the global pattern of signs guarantees a global stability state. In the (I) and (II) conditions, the global instability produces an oscillatory regime of activation, as shown by the time courses in the bottom panels of the figure. It should be noted that the oscillations in the time courses reflect the autonomous generation in the network of an extra link whenever a given threshold in the energy of some node is trespassed (as in the three hypertrophic nodes in the sub-network "a" of the (0) state). The

new links generated in (I) and (II), however, can only relieve the system from the surplus of energy in the "a" sub-network and cannot guarantee as table equilibrium (S): soon after brushing against the equilibrium, in fact, a new cycle is immediately initiated. Solely in state (III) the global link pattern is stabilizing the equilibrium.

Thus, the role of restoring a global stability in the network is assigned to the emergence of a new sub-network in a specific location with a specific link architecture (pattern(III) in Fig. 8). It is interesting to note that the above picture is strongly reminding the "negative feedback" mechanism so often invoked to describe the modulation of physiological equilibria in metabolic cycles. However, at odds with metabolic cycles, where the chemical nature (or concentration) of substance(s) flowing through the direct and inverse pathways is different, in the present case the activation level of the target node is the only essential trigger of the negative feedback.

3.2 Ising Models of Brain Areas Activity

3.2.1 The Problem

The areas of human brains active in the "default" mode [25] elicit peculiar interest since their role under that condition is not fully understood [26]. Due to the huge complexity of brain physiology [27] even a crude modeling approach to the problem is welcome, particularly if—as in the case of the Ising model—it looks able to describe a phase transition (flipping) of an ensemble based upon the dynamical behavior of its single components. Such an abstract model can be extended from the original context of spin alignment of microscopic magnets to other transition-like phenomena, as the liquid/gas transition in a fluid or—as in the present case—to the active/resting transition in the functional states of brain networks.

From experimental evidence of the type in Fig. 6, the problem can be tackled considering the following sequential steps:

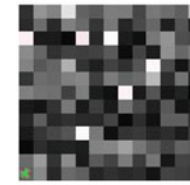
- Extracting from fMRI records the time series which describe the activity levels of the Regions Of Interest (ROIs). In the present case, for each of 90 ROIs the time series included 150 elements.
- Calculating from the time series a correlation matrix that is a functional connectivity matrix (Fig. 6).
- Filtering the values in the above matrix by some threshold (including the sign) in order to work out a symmetric Adjacency Matrix [AM] like that in Fig. 9 (A) of 19 selected ROIs in the present case, whose generic element $[AM]_{i,j}$ represents the weight of the link connecting agents i and j. The first three rows in panel (A) contain the XYZ coordinates of the corresponding ROI, according to the MRI Atlas of the Human Brain-Harvard Medical School [28].

(A)

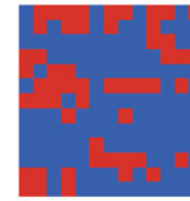
X	-12.14	-24.33	-28.33	-36.06	13.05	0.89	25.59	28.44	43.24	14.64	-3.17	-47.57	19.31	0.65	2.30	49.95	-0.54	-23.52	26.83
Y	-57.80	11.66	-37.15	-81.40	-53.43	-57.24	26.48	-33.47	-74.44	-45.78	49.37	-67.88	38.32	-53.25	-14.88	-63.53	-8.46	-28.52	-22.56
Z	15.33	54.93	-15.07	32.22	14.27	53.98	45.02	-19.27	31.67	-53.21	13.70	34.56	47.40	28.23	36.12	32.05	4.19	-12.79	-17.11
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0.00	0.17	0.54	0.47	0.88	0.35	0.32	0.45	0.44	0.25	0.43	0.39	0.34	0.69	0.15	0.39	0.15	0.46	0.38
2	0.17	0.00	0.19	0.43	0.13	0.26	0.37	0.11	0.30	0.12	-0.06	0.09	0.09	0.03	-0.09	0.14	-0.02	0.10	0.05
3	0.54	0.19	0.00	0.43	0.54	0.28	0.26	0.50	0.36	0.17	0.23	0.22	0.23	0.32	0.04	0.24	0.10	0.53	0.37
4	0.47	0.43	0.43	0.00	0.44	0.47	0.31	0.36	0.66	0.18	0.11	0.23	0.17	0.30	-0.03	0.27	0.07	0.24	0.19
5	0.88	0.13	0.54	0.44	0.00	0.39	0.37	0.46	0.48	0.22	0.44	0.35	0.38	0.71	0.16	0.41	0.13	0.46	0.41
6	0.35	0.26	0.28	0.47	0.39	0.00	0.40	0.23	0.45	0.10	0.06	-0.05	0.09	0.30	0.08	0.12	0.02	0.09	0.07
7	0.32	0.37	0.26	0.31	0.37	0.40	0.00	0.18	0.43	0.09	0.22	0.09	0.41	0.26	0.11	0.31	0.07	0.10	0.13
8	0.45	0.11	0.50	0.36	0.46	0.23	0.18	0.00	0.34	0.14	0.17	0.15	0.14	0.24	0.03	0.16	0.06	0.34	0.34
9	0.44	0.30	0.36	0.66	0.48	0.45	0.43	0.34	0.00	0.17	0.09	0.19	0.25	0.30	-0.02	0.43	0.05	0.18	0.20
10	0.25	0.12	0.17	0.18	0.22	0.10	0.09	0.14	0.17	0.00	0.13	0.16	0.10	0.21	0.03	0.16	0.11	0.12	0.10
11	0.43	-0.06	0.23	0.11	0.44	0.06	0.22	0.17	0.09	0.13	0.00	0.47	0.48	0.60	0.43	0.33	0.24	0.35	0.33
12	0.39	0.09	0.22	0.23	0.35	-0.05	0.09	0.15	0.19	0.16	0.47	0.00	0.38	0.57	0.18	0.50	0.09	0.32	0.24
13	0.34	0.09	0.23	0.17	0.38	0.09	0.41	0.14	0.25	0.10	0.48	0.38	0.00	0.44	0.21	0.42	0.08	0.24	0.23
14	0.69	0.03	0.32	0.30	0.71	0.30	0.26	0.24	0.30	0.21	0.60	0.57	0.44	0.00	0.35	0.49	0.14	0.38	0.34
15	0.15	-0.09	0.04	-0.03	0.16	0.08	0.11	0.03	-0.02	0.03	0.43	0.18	0.21	0.35	0.00	0.13	0.17	0.14	0.16
16	0.39	0.14	0.24	0.27	0.41	0.12	0.31	0.16	0.43	0.16	0.33	0.50	0.42	0.49	0.13	0.00	0.07	0.24	0.22
17	0.15	-0.02	0.10	0.07	0.13	0.02	0.07	0.06	0.05	0.11	0.24	0.09	0.08	0.14	0.17	0.07	0.00	0.20	0.20
18	0.46	0.10	0.53	0.24	0.46	0.09	0.10	0.34	0.18	0.12	0.35	0.32	0.24	0.38	0.14	0.24	0.20	0.00	0.60
19	0.38	0.05	0.37	0.19	0.41	0.07	0.13	0.34	0.20	0.10	0.33	0.24	0.23	0.34	0.16	0.22	0.20	0.60	0.00

(B)

```
[[ 0.17  0.54  0.47  0.88  0.35  0.32  0.45  0.44  0.25  0.43  0.39  0.34  0.69]
 [ 0.15  0.39  0.15  0.46  0.38  0.19  0.43  0.13  0.26  0.37  0.11  0.3   0.12]
 [-0.06  0.09  0.09  0.03 -0.09  0.14 -0.02  0.1   0.05  0.43  0.54  0.28  0.26]
 [ 0.5   0.36  0.17  0.23  0.22  0.23  0.32  0.04  0.24  0.1   0.53  0.37  0.44]
 [ 0.47  0.31  0.36  0.66  0.18  0.11  0.23  0.17  0.3  -0.03  0.27  0.07  0.24]
 [ 0.19  0.39  0.37  0.46  0.48  0.22  0.44  0.35  0.38  0.71  0.16  0.41  0.13]
 [ 0.46  0.41  0.4   0.23  0.45  0.1   0.06 -0.05  0.09  0.3   0.08  0.12  0.02]
 [ 0.09  0.07  0.18  0.43  0.09  0.22  0.09  0.41  0.26  0.11  0.31  0.07  0.1 ]
 [ 0.13  0.34  0.14  0.17  0.15  0.14  0.24  0.03  0.16  0.06  0.34  0.34  0.17]
 [ 0.09  0.19  0.25  0.3  -0.02  0.43  0.05  0.18  0.2   0.13  0.16  0.1   0.21]
 [ 0.03  0.16  0.11  0.12  0.1   0.47  0.48  0.6   0.43  0.33  0.24  0.35  0.33]
 [ 0.38  0.57  0.18  0.5   0.09  0.32  0.24  0.44  0.21  0.42  0.08  0.24  0.23]
 [ 0.35  0.49  0.14  0.38  0.34  0.13  0.17  0.14  0.16  0.07  0.24  0.22  0.2 ]]
```



(C)



(D)

Fig. 9 Brain activity patterns represented as matrices. (**A**) Adjacency matrix (AM) derived from a correlation matrix of the type in Fig. 6. The 19 ROIs can be identified by the 3D coordinates in the first three rows [44]; (**B**) *Upper Triangular Submatrix* of the AM in the form of Square Lattice (SL); (**C**) Recoding of (SL) on a discrete Gray Scale (GS) from 0 to 1, and white for negative values; (**D**) binary and colored rendering of (**C**)

- Recoding one of the triangular, specular, halves (above or below the main diagonal) in the symmetric Adjacency Matrix as a 2D squared lattice (Fig. 9(B)).
- Discretizing first (Fig. 9(C)) and then binarizing into $+1$ or -1 the values in the squared lattice (Fig. 9(D)) according to a given threshold (in the present case $= 0.35$).

The whole procedure is shown in Fig. 9, with reference to the time series of the selected 19 ROIs which, arranged in any possible binary combination produce, from the correlation matrix reported in panel (a), the final output reported in panel (d). The final output is a binary squared lattice to be submitted to the Ising model implemented in the MAS environment. Notice that the lattice's side length of 13 is an approximation of the correlations' number in one of the triangular submatrices in panel (a), given by

$$\sqrt{((n \times (n-1)/2)},$$

where $n = 19$, number of rows or columns of the Adjacency Matrix in panel (a).

3.2.2 The Model

An abstract Ising model [29, 30] considers each site i in a lattice as associated with a discrete variable, s_i, with value $+1$ or -1 standing for up or down spin, respectively. In the absence of external forces the total energy of the lattice is given by

$$H(s) = -J \sum_{<i,j>} s_i, s_j,$$

where the notation $<i,j>$ indicates that sites i and j are nearest neighbors and J is a coupling constant. A competition arises between thermal fluctuations (reflecting the interaction with the environment), which induce the system to get disordered, and the opposite tendency to get organized in some specific way depending on the interaction or coupling (J) between the sites. The spins flip if flipping decreases their energy, but sometimes also flip into a higher energy state. The flipping probability is calculated by a Metropolis algorithm, based on the formula

$$e^{-\text{Ediff}/T},$$

where Ediff is the potential gain in energy and T is the temperature. Thus, flipping to a higher energy state directly depends upon temperature and inversely upon Ediff. The Ising model is lying in the calculation of the energy at each site as the negative of the sum of the products of its spin with each of its neighbors' spins.

We took advantage of the 2D Ising algorithm in the software library of the Netlogo environment [31] which uses a Metropolis/Monte Carlo method for the probabilistic time evolution of the spins in conjunction with space periodic boundary conditions and four nearest neighbors for each node. First, we traced, according to the Ising model, each element of the square lattice (SL) depicted in Fig. 9d in its evolution along time windows of different lengths: thanks to the association to brain areas in the (AM) of Fig. 9a, this allows following at each time step the evolution dynamics of each ROI in the original (AM) through its correlations with all other ROIs. Thus, by summing up for each row i (or column j) the values of the corresponding column j (or row i), the global activation dynamics of each (ROI) can be reconstructed in the considered time window.

The implementation of the above procedure was tested through the temperature dependence of activity trends of the type shown in Fig. 10 concerning three randomly selected areas. The activity trend observed at a reference temperature of 2.27 a.u., (Fig. 10, upper panel) disappears by about doubling the value of

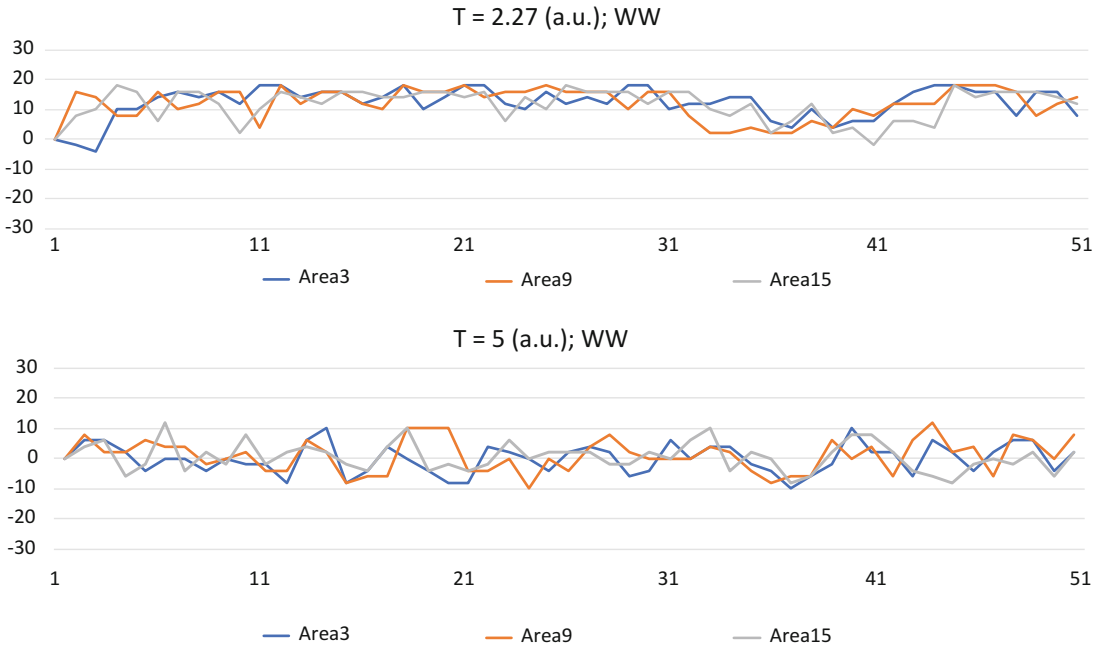


Fig. 10 MAS simulation of activity trends in brain areas by an Ising model—I. *Vertical* and *horizontal axes* indicate, respectively, activity levels and time, both in a.u

the temperature (Fig. 10, lower panel). This is expected, due to temperature-induced disorder overcoming the ordering effect of the spin alignment.

3.2.3 *Some Results*

The trend appearing in Fig. 10a can be explored over a much longer time window, as in Fig. 11, where an oscillatory behavior is clearly shared by the three areas at hand, namely 3, 9, and 15.

Once again the role of temperature in defining the conditions under which the phase transition in the spin alignment can be observed, is clearly emerging: Table 1 summarizes the Pearson correlation of the trends reported in Figs. 10 and 11 for the three brain areas under consideration. Pretty similar results have been observed in other areas. Lower correlation values are associated with the higher temperature particularly in the longer time window (1000 a.u.), indicating essentially random trends.

In addition, the information reported in Fig. 12 points to a well-synchronized and self-sustaining order appearing after a relatively short time (Fig. 12, middle panel) and keeping stable from there on if no change occurs in the experimental conditions. Needless to say, at higher temperature (5 a.u.), the corresponding trends in Figs. 10 and 11 are constantly reproduced.

Whether or not the above observations are amenable to physiological interpretation remains open to discussion. Despite the simple data acquisition setup, any Ising-inspired model necessarily

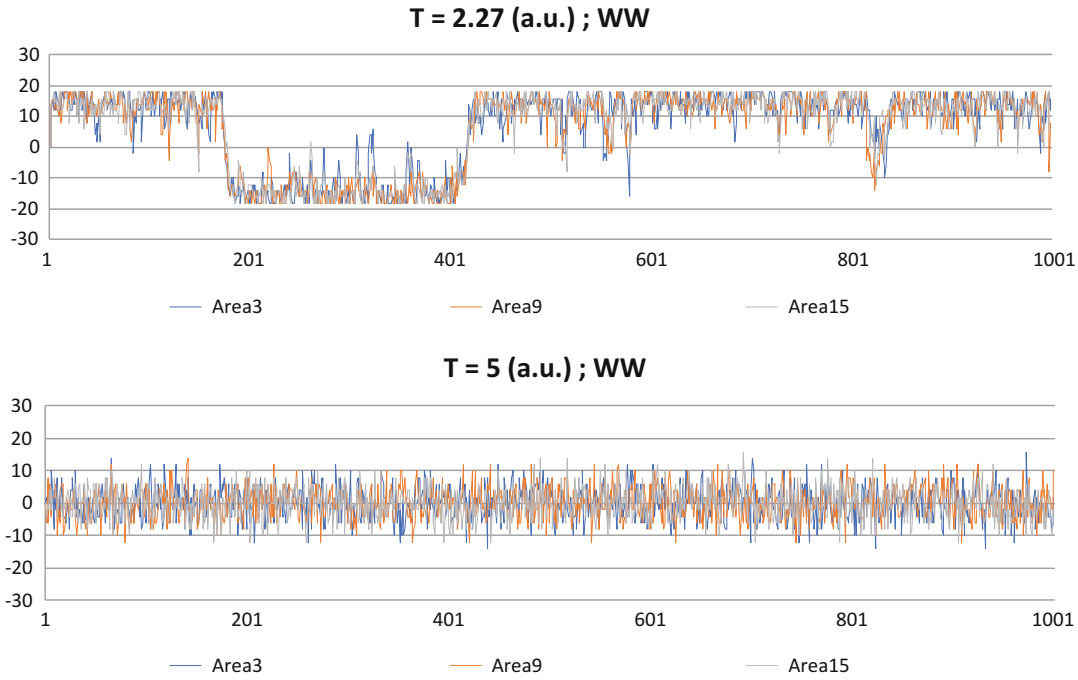
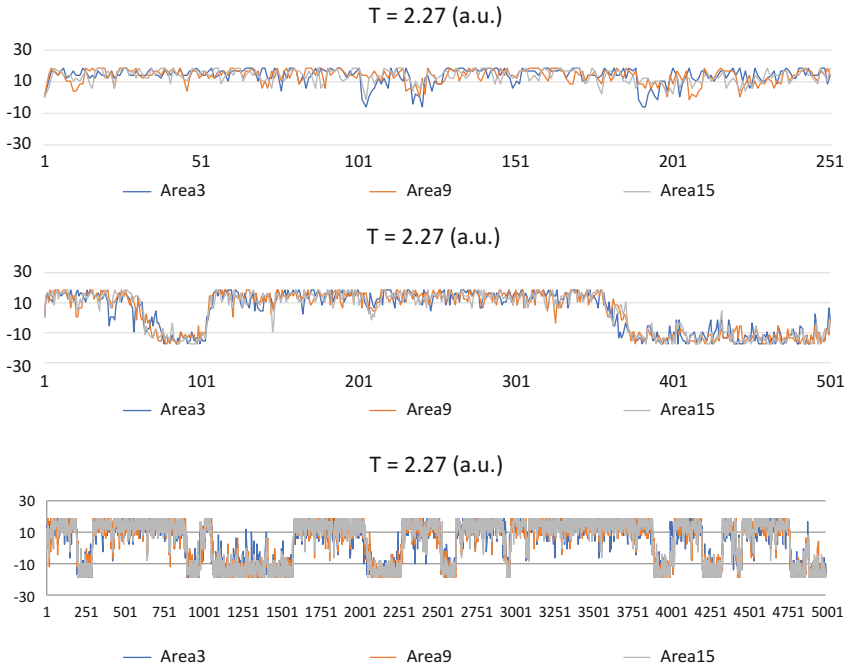


Fig. 11 MAS simulation of activity trends in brain areas by an Ising model—II. But for the much wider time window, all other conditions are identical to those in Fig. 10. Notice that 1000 a.u., namely machine time units, correspond to about 35 s on a MacBook Pro equipped with 2.6 GHz Intel Core i7 processor

Table 1
Pearson correlation coefficient between numerical series of time-dependent activity in some brain areas

		$T = 2.27$ a.u.		$T = 5$ a.u.	
		Area3	**Area9**	**Area3**	**Area9**
t window = 50 a.u.	Area9	0.34		0.28	
	Area15	0.45	0.43	0.37	0.12
t window = 1000 a.u.	Area9	0.91		0.15	
	Area15	0.92	0.94	0.12	0.12

requires a careful analysis of the temperature effects and of the probabilistic threshold for spontaneous spin flipping in the reproduction of brain areas activation dynamics, which is in due course now in our lab.



PEARSON CORRELATION COEFF.		
	Area3	*Area9*
T = 2.27; t = 250		
Area9	0.33	
Area15	0.39	0.46
T = 2.27; t = 500		
Area9	0.91	
Area15	0.92	0.94
T = 2.27; t = 5000		
Area9	0.92	
Area15	0.93	0.94

Fig. 12 Arising of stable and highly correlated activity trends in brain areas under default state. All conditions are the same as in Figs. 10 and 11. The table on the *right* contains the correlation coefficient at the considered temperature (*T*) and time span (*t*)

4 Conclusion

Following the Newtonian track, originated in the realm of celestial mechanics and immediately generalized to any kind of time-dependent events, dynamical simulations are traditionally obtained by mathematical models based upon numerical integration of differential equations. The chaotic behavior of complex systems, however, provides severe limitations to any long-term prediction by deterministic models and opens the door to an alternative approach based upon qualitative, stochastic models. It is worth stressing that the latter remains the only possible approach whenever the phenomena of interest concern the collective behavior of large populations of relatively simple elements.

On these premises, Multi Agent Systems (MAS) perform reliable simulations of population-dynamics phenomena depending upon the interactions of the population members among each other and with the environment. Thus, the utility of MAS is to be primarily appreciated in the design of mechanistic models accounting for some collective behavior. Another, more ambitious, context deals with the quantitative refinement of the model's parameters, in the aim to fit the experimentally observed dynamics. All in all, as compared to the traditional approach based on approximated

solutions of differential equations, Multi Agent Systems appear as a more intuitive and rewarding approach, particularly in the initial exploration phase of complex phenomena.

The applications presented here provided a description of the model's dynamic response to various combinations of endogenous and exogenous factors, useful in the development of new ideas and mechanisms. In some cases, however, whenever the MAS showed able to faithfully reproduce the observed data trends, its aim exceeded the purely descriptive dimension and a quantitative answer to some specific mechanistic questions could be proposed and empirically tested.

Finally, it is fair to stress that, due to the prevailing interest of the author, a major emphasis has been given to only a few among the many characteristic features of MAS, namely: 1) the straightforward account of the environmental force-fields influence on the agents behavior, and 2) the synchronous and/or correlated activation exerted by agents upon each other.

Dealing with a much wider range of phenomena the euristic power of MAS could also emerge [10] as an outstanding tool to face many relevant issues typical of System Biology.

References

1. Holland J (1996) Hidden order: how adaptation builds complexity. Addison-Wesley, Reading
2. Laughlin R (2005) Un universo diverso: reinventare la Fisica da cima a fondo. Codice, Torino
3. Rodgers J, Nicewander W (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42(1):59–66
4. Bar-Yam Y (1998) Dynamics of complex systems. Addison-Wesley, Boston, MA
5. Borschev A (2016) The big book of simulation modeling. www.anylogic.com
6. DeToni F, Bernardi E (2009) "Il pianeta degli agenti" UTET-Torino. UTET-Torino, Turin
7. Holland J (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence. The MIT Press, Cambridge MA
8. Wolfram S (2002) A new kind of science. Wolfram Media, Inc, Champaign, IL
9. Wilenski U (1999) http://ccl.northwestern.edu/netlogo/
10. Wilensky U, Rand W (2015) Introduction to agent-based modeling: modeling natural, social and engineered complex systems. The MIT Press, Cambridge MA
11. Colosimo A (2011) Multi Agent Simulators: flexible tools to reproduce collective behaviors, Biophys Bioeng Lett 4(2):34–41
12. Institute for Systems Biology. https://www.systemsbiology.org
13. Wyman J, Gill S (1990) Binding and linkage: functional chemistry of biological macromolecules. University Science Books, Sausalito, CA
14. Bizzarri M, Cucina A, Biava P, Proietti S, D'Anselmi F, Dinicola S, Pasqualato A, Lisi E (2011) Embriyonic morphogenetic field induces phenotypic reversion in cancer cells: review article. Curr Pharm Biotechnol 12:243–253
15. Leao A (1944) Spreading depression of activity in the cerebral cortex. J Neurophysiol 7:359–390
16. Lauritzen M (1994) Pathophysiology of the migraineaura. the spreading depression theory. Brain 117:199–210
17. Colosimo A (2008) Biological simulations by autonomous agents: two examples using the netlogo environment. Biophys Bioeng Lett 1 (3):40–50
18. Hines M, Carnevale N (2001) Neuron: A tool for neuroscientists. Neuroscientist 7:123–135
19. Prinz A (2008) Understanding epilepsy through network modeling. Proc Natl Acad Sci U S A 105(16):5953–5954

20. Friston K (2011) functional and effective connectivity: a review. Brain Connect 1(1):13–36
21. Fox M, Snyder A, Vincent J, Corbetta M, Essen DV, Raichle M (2005) The human brain is intrinsicallyorganized into dynamic, anticorrelated functional networks. Proc Natl Acad Sci U S A 102:9673–9678
22. Schwarz AJ, McGonigle J (2011) Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. NeuroImage 55(3):113246
23. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10(3):186–198
24. Rubinov M, Sporns O (2009) Complex network measures of brain connectivity: uses and interpretations. NeuroImage 53 (3):1059–1069
25. Deco G, Jirsa VK, McIntosh AR (2013) Resting brains never rest: computational insights into potential cognitive architectures. Trends Neurosci 36(268–274):3
26. Deco G, Jirsa VK (2012) Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. J Neurosci 23:3366–3375
27. Nakagawa TT, Jirsa VK, Spiegler A, McIntosh AR, Deco G (2013) Bottom up modeling of the connectome: linking structure and function in the resting brain and their changes in aging. NeuroImage 80:318–329
28. Caviness V, Meyer J, Makris N, Kennedy D (1996) Mri-based topographicparcellation of human neocortex: ananatomically specified-method with estimate of reliability. J Cogn Neurosci 8:566–587
29. Binder K (2001) Ising model. In: Michiel H (ed) Encyclopedia of mathematics. Springer, New York, NY
30. Marinazzo D, Pellicoro M, Wu G, Angelini L, Cortes JM, Stramaglia S (2014) Information transfer and criticality in the Ising model on the human connecrtome. PLoS One 9:1–7
31. Wilensky U (2003) NetLogo ising model. Evanston, IL, Center for Connected Learning and Computer-Based Modeling, Northwestern University. http://ccl.northwestern.edu/netlogo/models/Ising
32. Parente F, Colosimo A (2016) The role of negative links in brain networks. Biophys Bioeng Lett 9(1):1–13

Chapter 16

Metabolomics: Challenges and Opportunities in Systems Biology Studies

Luca Casadei, Mariacristina Valerio, and Cesare Manetti

Abstract

Metabolomics has the capability of providing predisposition, diagnostic, prognostic, and therapeutic biomarker profiles of individual patients, since a large number of metabolites can be measured in an unbiased manner from biological samples. In this setting, ^{1}H-Nuclear Magnetic Resonance (NMR) spectroscopy of biofluids such as plasma, urine, and fecal water offers the opportunity to identify patterns of biomarker changes that reflects the physiological or pathological status of an individual patient.

In this chapter, we show as a metabolomics study can be used to diagnose a disease, classifying patients as healthy or as pathological taking into account individual variability.

Key words Precision medicine, Metabolomics, NMR spectroscopy, Principal component analysis, Linear discriminant analysis, Covariance analysis

Abbreviation

λ	g-Log transformation parameter
CF	Cystic fibrosis
FID	Free induction decay
g-log	Generalized log
JRES	2D ^{1}H J-resolved
LDA	Linear discriminant analysis
NaN3	Sodium azide
NMR	Nuclear magnetic resonance
PC	Principal component
PI	Pancreatic insufficiency
p-JRES	Proton-decoupled skyline projections
PQN	Probabilistic quotient normalization
TSP	Sodium salt of 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid, 98 atom % D

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_16,

1 Introduction

Systems biology, acquiring a holistic perspective to explain the complexity of a biological system as a whole, is having and will have an ever-greater impact on the current health care. We are shifting from diagnosis and treatment of diseases by symptoms to precision medicine in which each patient is treated taking into account environmental factors, lifestyle habits, genetic information, and molecular phenotype [1, 2].

One of the most challenging goals in the precision medicine era is to develop an experimental approach to establish the phenotypic properties of an individual in an objective and repeatable way as well as the phenotypic changes related to genetic and environmental factors [3]. In this effort, metabolomics, i.e., the study of all low molecular weight (~50–1500 Da) molecules or metabolites within biological samples, plays an important role. By providing a comprehensive biochemical fingerprint of a biological system at the cell, tissue, or organism level, metabolomics offers a systemic approach to the study of various diseases without first having to identify markers of the disease [4]. Consequently, metabolomics can help clinicians in the screening, diagnosis, treatment, and monitoring of many diseases.

The main complementary high-throughput platforms for metabolomics are Mass Spectrometry and Nuclear Magnetic Resonance (NMR). NMR gives a direct fingerprint of the system, thus providing a unified picture of whole metabolome across the identification of all major metabolite classes simultaneously. The subsequent use of statistical and mathematical tools plays a key role in extracting meaning from this "big data." The metabolic descriptors so identified become the coordinates of a new system of reference represented by metabolomic maps on which patients and their response to therapy are located.

NMR-based metabolomics has already been applied in many disease studies [5].

Moreover, metabolomic analysis showed the potential to predict the prognosis or the response to treatments from baseline metabolic profiles [6, 7], highlighting the potential of metabolomics analysis in patient stratification and personalized medicine.

We propose here as a case study the analysis of fecal samples from young patients with fibrosis cystic (CF) and healthy children (controls) to characterize the metabolic impact of variations of the gut microbiota. In fact, cystic fibrosis is a lethal hereditary disorder involving respiratory infections, chronic inflammation, repeated antibiotic treatments and hence influencing the gut microbiota profiles.

2 Materials

2.1 Fecal Sample Collection

1. Collect at least 50 mg of feces for sample in a 1.5 ml Eppendorf tube (*see* **Note 1**).

2.2 Reagents for Sample Preparation

1. D_2O, 99.9 atom % D.
2. Phosphate-buffered saline (PBS) tablet.
3. Sodium azide (NaN3).
4. Sodium salt of 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid, 98 atom % D (TSP).

2.3 Equipment

1. 500 MHz spectrometer (Bruker BioSpin Corp., Billerica, MA, USA or Agilent Technologies, Santa Clara, CA, USA or Jeol Ltd., Akishima, Tokyo, Japan). However, 400 or 600 MHz NMR instruments are commonly also used in metabolomic studies.
2. Analytical balance.
3. 1.5 ml Eppendorf tubes.
4. Spatula.
5. Micropipettes and pipette tips.

2.4 Software for Data Analysis

Several commercial and free licensed software packages are available for NMR data processing, postprocessing, and statistical analysis. Only the most widely used software is reported.

1. Software for processing NMR data: TopSpin (Bruker BioSpin Corp., Billerica, MA, USA), VNMRJ (Agilent Technologies, Santa Clara, CA, USA), MetaboLab [5] in the MATLAB programming environment (MathWorks, Inc., Natick, MA).
2. Software for postprocessing NMR data: ACD/NMR processor (Advanced Chemistry Development Inc. (ACD/Labs), Toronto, ON, Canada), MetaboLab [8] in the MATLAB programming environment (MathWorks, Inc., Natick, MA).
3. Software for multivariate data analysis: SIMCA-P+ (Umetrics, Umeå, Sweden), PLS-Toolbox (Eigenvector Research, Manson, WA) in MATLAB.

3 Methods

3.1 Phosphate-Buffered Saline for NMR Sample Preparation

Prepare the 10 mM D_2O (99.9 atom % D) phosphate-buffered saline solution at pH 7.4 and 25 °C by mixing PBS tablets (as indicated by the manufacturer), TSP 1 mM and NaN_3 10 mM.

3.2 Sample Preparation for NMR Spectroscopy

1. Dissolve each sample in 1 ml of 10 mM D_2O phosphate-buffered saline solution at pH = 7.4.
2. Homogenize samples by using a vortex mixer for 1 min.
3. Centrifuge samples at 14,000 × *g* for 10 min at 20 °C to obtain fecal water. After centrifugation, transfer 600 μl of each resulting supernatant into a 5 mm NMR tube.

3.3 Acquisition, Processing, and Postprocessing of NMR Data

3.3.1 NMR Setup

1. Set temperature to 298 K.
2. Properly position a representative sample inside the probe and leave 5 min to equilibrate the sample temperature.
3. For obtaining a good signal-to-noise ratio: adjust the probe-head tuning and matching; lock and shim the sample on D_2O; calibrate the 90° pulse length; determine the power, length, and frequency offset for HDO signal suppression by using the presaturation pulse.
4. Once an optimal signal is obtained, transfer the setting parameters to the other samples (*see* **Notes 2–4**).

3.3.2 Two-Dimensional 2D ^{1}H J-Resolved

For the analysis of fecal water samples, J-resolved pulse sequence is used to observe resonances better, as they are partially or completely buried in a typical 1D spectrum. This sequence improves the quality of the metabolic information extracted.

1. Acquire 2D ^{1}H J-resolved (JRES) NMR spectra using a double spin echo sequence [9], suppressing the residual water signal with the presaturation technique.
2. Use the following parameters to acquire the JRES spectra: transients per increment, 16; total increments, 32; dummy scans, 16; data points, 16k; spectral width for direct (F2 or chemical shift) dimension, 6 kHz; spectral width for indirect (F1 or J-coupling) dimension, 40 Hz; relaxation delay, 2 s.
3. Processing the NMR data carrying out the following operations: zero-fill the F1 data to 256 data points; multiply each Free Induction Decay (FID) with a combined sine-bell/exponential function in the F2 dimension and a sine-bell function in the F1 dimension; apply Fourier Transform to each dimension; tilt the spectra by 45°; symmetrize the spectra about F1 dimension; calibrate chemical shifts to the TSP methyl protons at 0.00 ppm; apply a zero-order baseline correction of spectrum.
4. Exporting the proton-decoupled skyline projections (p-JRES) in a suitable format (arrange the exported 1D-skyline projections into a matrix of N samples (rows) by M variables (columns)) for subsequent postprocessing treatment.

3.3.3 NMR Data Postprocessing

NMR data postprocessing is a necessary step of metabolomics pipeline to extract useful information related to the state of biological system. This step helps to avoid sources of variation in the data, such as dilution effect, subtle changes in chemical shifts, line-widths, and baseline across series of spectra, which can interfere with the outcome of the statistical analysis, leading to false deductions.

NMR data postprocessing usually includes exclusion of non-informative regions, binning, normalization, scaling, and data export for subsequent multivariate statistical analysis.

1. Remove the regions in the spectra that contain only noise and/or exogenous peaks. Therefore, exclude the spectral regions outside the window 0.5 (including TSP signal) and 9.0 ppm and those containing the residual water (δ 4.7–5.0 ppm) and drug peaks.
2. Reduce the dimensionality of data splitting the p-JRES spectra into small segments (bins or buckets) with variable widths ranging from 0.01 to 0.04 ppm to ensure that each bin contains the same signals throughout all the spectra. If local peak shifts across series of spectra are still observed, compress groups of bins into single bins or alignment of the spectra. Then, integrate the signal within each bin (*see* **Note 5**).
3. Normalize the binned spectra by applying the Probabilistic Quotient Normalization (PQN) [10, 11] method to make spectra comparable:
   (a) Set the total spectral area of every spectrum to 100.
   (b) Calculate as a reference spectrum the median spectrum (median of each variable/bin area) of healthy group samples.
   (c) Calculate the quotient between the area of each spectral bin of the considered spectrum and that of the corresponding bin in the reference spectrum.
   (d) Calculate the median of all the quotients.
   (e) Divide all the variables of the considered spectrum by the median quotient.
   (f) Repeat **steps c–e** for all spectra.
4. Scaling the data by applying the generalized log (g-log) transformation [12, 13] to make the variables within spectra comparable:
   (a) Estimate the g-log transformation parameter (λ) by the maximum likelihood method using a set of five replicate measurements.
   (b) Obtain these five replicates from a single homogeneous pool of fecal water samples from healthy and pathological patients. Process the replicate spectra as described above

(i.e., selection of exclusion regions, binning, and normalization).

5. Mean centered the data: subtract the mean value of each variable from the original data of that bin.

3.4 Statistical Analysis of NMR Data

1. Reduce the data by using Principal Component Analysis. This process assigns to each sample a score relative to each extracted component (Principal Component, PC). The extracted components are independent of each other by construction, thus they are non-overlapping features of the studied system. Use the component scores to plot PC maps of the samples which best provide an indication of the differences between the classes (healthy or disease groups) in terms of metabolic similarity.
2. Carry out separate inferential statistics (t-test) on the different component scores, so as to check for the statistical significance of the between groups differences.
3. Compare the metabolic profiles and the clinical features of each patient by Pearson's correlation and/or ANOVA test, having as dependent variables the components and as regressors (sources of variation) potential modulating or confounding factors.
4. After having verified the absence of potentially confounding factors on the PCs, apply a linear discriminant analysis (LDA) to the components so as to develop a predictive model for the classification of patients in healthy or disease groups.
5. In the case of statistically significant effects of confounding factors on discriminant components, correct (covariance analysis or partial correlation analysis) for the effect of the above-mentioned factors. This procedure will allow estimating the actual degree of association between DA-based membership class probability and clinical status.

We investigated the NMR data by using Principal Component Analysis (PCA) carried out on samples from young patients with cystic fibrosis (CF) and healthy children. Five components are sufficient to explain the 40% of the variance in the metabolic data. The score plot in Fig. 1-NMR shows a clear separation between the CF and healthy children on the PC1 ($p = 0.001$ by t-test) and PC4 ($p < 0.0001$ by t-test).

In this study, since the metabolic status of the CF patients could be influenced by several variables such as age, gender, and antibiotic and probiotic assumption, we decided to assess whether any of these factors could influence the separation between CF patients and healthy children. To address age and gender as potential confounding factors, the metabolic profiles and the clinical features of each child were compared by Pearson's correlation, while for assessing antibiotic and probiotic assumption variables, the metabolic

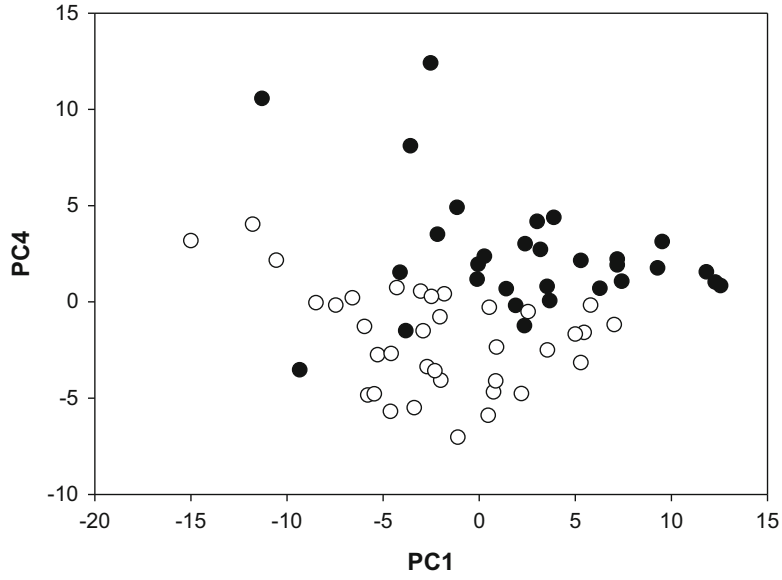


Fig. 1 PCA score plot (PC1 vs. PC4) of the ^{1}H-NMR profiles of fecal samples obtained from 30 patients with cystic fibrosis and 36 healthy children

Table 1
Spearman's correlations between age and metabolic profiles in CF and healthy patients

	Principal components				
Patients	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
All	−0.213	−0.197	−0.241	0.381	0.117
CF	−0.204	−0.027	−0.440	0.200	0.318
Healthy	**−0.426**	−0.372	−0.164	**0.456**	−0.148

Significant values ($p < 0.05$) are in bold

profiles were analyzed by the ANOVA test for each CF patient (Tables 1 and 2 NMR).

Age and gender as well as antibiotic and probiotic assumption were no confounding factors. Only for healthy children, a negative correlation with PC1 ($\rho = -0.426$; t-test 2006 vs. 2011 $p = 0.03$; 2007 vs. 2011 $p = 0.04$) as well as a positive correlation with PC4 were found ($\rho = 0.456$; t-test 2006 vs. 2009 $p = 0.02$; 2006 vs. 2010 $p = 0.03$; 2006 vs. 2011 $p = 0.003$; 2007 vs. 2011 $p = 0.02$).

After having verified the lack of effects of potentially confounding factors on PC1 and PC4, a linear discriminant analysis (LDA) was applied to the PC1 and PC4 components to develop a predictive model for the classification of children in healthy or CF groups. The model discriminated the two groups with a sensitivity, specificity, and accuracy of 86%.

Table 2
Sources of variability in metabolomic data of CF patients and healthy children as measured by analysis of variance (ANOVA)

		Principal components				
Factors	**Patients**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Gender	All	0.001 (0.976)	0.199 (0.657)	0.032 (0.858)	0.303 (0.584)	1.367 (0.246)
	CF	1.077 (0.309)	3.309 (0.080)	0.090 (0.767)	1.296 (0.265)	1.75 (0.197)
	Healthy	0.010 (0.923)	1.239 (0.274)	0.216 (0.645)	0.229 (0.635)	0.078 (0.781)
Antibiotic consumption	CF	0.018 (0.894)	1.768 (0.195)	10.922 (**0.003**)	0.121 (0.731)	5.362 (**0.029**)
Probiotic consumption	CF	3.116 (0.089)	0.015 (0.904)	2.909 (0.100)	0.199 (0.660)	0.590 (0.449)

The table reports the *F*-values for each factor and, in parenthesis, the corresponding *p*-values. Significant values ($p < 0.05$) are in bold

Table 3
Pearson's correlation analysis among DA-based membership class probability, clinical status (healthy or sick), and pancreatic insufficiency factor (PI)

	DA-based membership class probability	**Pancreatic insufficiency factor**	**Clinical status (healthy or sick)**
DA-based membership class probability	1.0	0.762	0.768 (0.383)
Pancreatic insufficiency factor	0.762	1.0	0.822
Clinical status (healthy or sick)	0.768 (0.383)	0.822	1.0

In parenthesis, partial correlation coefficient between DA-based membership class probability and clinical status adjusted for the PI effect is reported

Interestingly, from the results of Pearson's simple and partial correlation analysis among DA-based membership class probability, clinical status (healthy or sick), and pancreatic insufficiency (PI) factor, we inferred that the metabolic variations in cystic fibrosis are mainly associated with pancreatic insufficiency. In fact, the correlation coefficient between DA-based membership class probability and clinical status ($r = 0.77$) adjusted for the PI effect dropped to 0.38 (Tables 3 and 4 NMR), thus pointing to pancreatic insufficiency as the main driver of metabolism-based classification.

In particular, PC4 was highly correlated with PI ($r = 0.60$), while PC1 carries information partially independent of PI.

Table 4
Pearson's correlation analysis among PC1, PC4, clinical status (healthy or sick), and pancreatic insufficiency factor (PI)

	PC1	PC4	Pancreatic insufficiency factor	Clinical status (healthy or sick)
PC1	1.000	−0.012 (−0.334)	0.388	0.382 (0.120)
PC4	−0.012 (−0.334)	1.000	0.604	0.592 (0.210)
Pancreatic insufficiency factor	0.388	0.603	1.000	0.822
Clinical status (healthy or sick)	0.382 (0.120)	0.592 (0.210)	0.822	1.000

In parenthesis, partial correlation coefficients among PC1, PC4, and clinical status adjusted for the PI effect are reported

4 Notes

1. If you cannot prepare the fecal samples immediately after collection, store them at −80 °C.
2. For each experiment, the magnetic field homogeneity must be optimized through an accurate shimming. To check if a sample is properly shimmed, you can observe the full-width at half-maximum of lactate peak that should be less than 1.7 Hz, before applying apodization and symmetric shape. TSP peak is not a reliable signal to check the quality of shimming due to the huge amount of proteins present in the sample, which influence its line-width because TSP binds to proteins.
3. Samples must be acquired in randomized order.
4. It is useful to run a standard solution to identify potential impurities arising from reagents and preparation procedures.
5. The most common method of spectral binning is the so-called equidistant binning, i.e., each spectrum is divided into bins with fixed width, typically 0.04 ppm. The weakness of this method is that, under certain experimental conditions, single peaks can be divided into two neighboring bins, generating artifacts. To avoid this problem, several mathematical algorithms [14–17] have been developed to vary the individual size bin. For example, ACD intelligent bucketing method (ACD/NMR processor, Advanced Chemistry Development Inc. (ACD/Labs), Toronto, ON, Canada), a combination of equidistant binning and non-equidistant binning, sets the bucket divisions at local minima (within the spectra) to ensure that each resonance is in the same bin throughout all spectra.

Acknowledgments

We thank all the members of Dr. Lorenza Putignani's laboratories, Unit of Human Microbiome, Genetic and Rare Diseases Area, Bambino Gesù Children's Hospital, IRCCSRome, Italy and Unit of Parasitology, Bambino Gesù Children's Hospital, IRCCSRome, Italy, for providing the samples and the clinical data for the metabolomics study here described.We are grateful to Dr. Alessandro Giuliani, Department of Environment and Primary Prevention, Istituto Superiore di Sanità, Rome, Italy for his useful comments and suggestions on the data analysis.

References

1. Chen R, Snyder M (2012) Systems biology: personalized medicine for the future? Curr Opin Pharmacol 12:623–628
2. Saqi M, Pellet J, Roznovat I et al (2016) Systems medicine: the future of medical genomics, healthcare, and wellness. Methods Mol Biol 1386:43–60
3. Wishart DS (2016) Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov 15:473–484
4. Nicholson JK, Lindon JC (2008) Metabonomics. Nature 445:1054–1056
5. Nicholson JK, Holmes E, Kinross JM et al (2012) Metabolic phenotyping in clinical and surgical environments. Nature 491:384–392
6. Everett JR (2015) Pharmacometabonomics in humans: A new tool for personalized medicine. Pharmacogenomics 16:737–754
7. Priori R, Casadei L, Valerio M et al (2015) ^{1}H-NMR-based metabolomic study for identifying serum profiles associated with the response to etanercept in patients with rheumatoid arthritis. PLoS One 10(11):e0138537
8. Ludwig C, Günther UL (2011) MetaboLab – advanced NMR data processing and analysis for metabolomics. BMC Bioinformatics 12:366
9. Thrippleton MJ, Edden RA, Keeler J (2005) Suppression of strong coupling artefacts in J-spectra. J Magn Reson 174:97–109
10. Dieterle F, Ross A, Schlotterbeck G et al (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. Anal Chem 78:4281–4290
11. Dieterle F, Riefke B, Schlotterbeck G et al (2011) NMR and MS methods for metabolomics. Methods Mol Biol 691:385–415
12. Purohit PV, Rocke DM, Viant MR et al (2004) Discrimination models using variance-stabilizing transformation of metabolomic NMR data. OMICS J Integr Biol 8:118–130
13. Parsons HM, Ludwig C, Gunther UL et al (2007) Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. BMC Bioinformatics 8:234–250
14. Dieterle F, Ross A, Schlotterbeck G et al (2006) Metabolite projection analysis for fast identification of metabolites in metabonomics. Application in an amiodarone study. Anal Chem 78:3551–3561
15. Davis RA, Charlton AJ, Godward J et al (2007) Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform. Chemometr Intell Lab 85:144–154
16. De Meyer T, Sinnaeve D, Van Gasse B et al (2008) NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. Anal Chem 80:3783–3790
17. Anderson PE, Mahle DA, Doom TE et al (2011) Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data. Metabolomics 7:179–190

Chapter 17

Systems Biology-Driven Hypotheses Tested In Vivo: The Need to Advancing Molecular Imaging Tools

Garima Verma, Alessandro Palombo, Mauro Grigioni, Morena La Monaca, and Giuseppe D'Avenio

Abstract

Processing and interpretation of biological images may provide invaluable insights on complex, living systems because images capture the overall dynamics as a "whole." Therefore, "extraction" of key, quantitative morphological parameters could be, at least in principle, helpful in building a reliable systems biology approach in understanding living objects. Molecular imaging tools for system biology models have attained widespread usage in modern experimental laboratories. Here, we provide an overview on advances in the computational technology and different instrumentations focused on molecular image processing and analysis. Quantitative data analysis through various open source software and algorithmic protocols will provide a novel approach for modeling the experimental research program. Besides this, we also highlight the predictable future trends regarding methods for automatically analyzing biological data. Such tools will be very useful to understand the detailed biological and mathematical expressions under in-silico system biology processes with modeling properties.

Key words Molecular imaging, Software, System biology, Omics, Modeling, Laboratory data management

1 Introduction

System biology is a growing field of research, stimulated by the combination of multiple factors: primarily, the advent of high-performance computing enables the simulation of the behavior of increasingly complex dynamic systems, described by the large number of equations. Moreover, the abundance of data from experimental biology is amenable to providing the appropriate input to sophisticated models of biological domains, as well as data for the validation of such models. System biology aim is centered on the description by quantitative experimental methodologies and the calculation of the behavior of the mathematical models representing the biological processes in order to discover the novel

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6_17,

phenomena and to identify biochemical events, including cell signaling, cell cycle, and different biological pathways [1, 2].

In agreement with this perspective, new opportunities are disclosed to understand biological systems and functions of cellular tissues and components, leading to interesting predictions of biological interest, even for an evidence-based, personalized approach to treatment [3]. Nowadays, system biology is becoming the borderline of modern biological research with a great amount of data stemming from the new omics approaches. This complex situation can be very challenging to be understood without a network or a systems point of view, with the associated computational analyses [3, 4].

Furthermore, a systems biology perspective will contribute to reshaping our view regarding the theory of biological phenomena [4]. To understand the system biology approach in modern biological research we need the powerful computational tools currently available, which are used to manage the large-scale data sets of information on cellular structures, genetics, proteins, cytoskeleton [5]. These types of tools are able to build the dynamic system models to perform the simulation and interpretation of the mechanisms of some cellular behavior from a system viewpoint [6, 7]. There are several mathematical techniques that were created on the basis of systems biology and have been developed to study the methodological properties of the complex living networks [8]. At present, mathematical modeling is gaining increasing importance to explain and predict the behavior of biological systems.

New system biology applications are continuously appearing, spurred by the development of tools and techniques [9–11]. Such instruments leverage on the platforms for the local and global analysis, e.g., high-throughput genomics and proteomics equipment. The important point here is to define the quantitative models that are able to decipher the biological information [12].

We believe that quantitative models of System biology are helpful in identifying key dynamical features, which would have a fruitful impact on pharmaceutical and medicinal practice. According to Hiroaki Kitano the technical and system level analysis not only permits the visualization for molecular interactions but also speeds up the measurement with various developed methodologies [13].

The borders of System biology are not yet clearly defined. From the basic biological knowledge, System biology can be brought to consider also predictive capabilities for patients, in case of a possible healthcare service supply. Therefore, this field of investigation can have far-reaching consequences, in the patients' standpoint.

Moreover, the wealth of data available about a particular biological process is usually not cast in a harmonized form, since many different sources can be used, typically. Indeed, data structured by the System biology approach are heterogeneous:

they span from textual information (such as that retrieved from repositories of biomedical papers, e.g., Medline) to quantitative results (from biological, physical, chemical, or bioengineering studies). The former is usually affected by semantic ambiguities, so that a proper processing phase is required to filter textual information. The case of semantic heterogeneity is then frequently occurring.

Nature of semantic heterogeneity—It is important to clarify what semantic heterogeneity is. Sheth and Larson [14] suggest that heterogeneity occurs "... when there is a disagreement about the meaning, interpretation or intended use of the same or related data [in different databases]." But they noted that "... this problem is poorly understood, and there is not even an agreement regarding a clear definition of the problem." Ontology's role is to help unbundle the objects and make clear the relation between them.

For the purposes of database integration, the traditional philosophical (metaphysical) notion of ontology is useful—where this is "the set of things whose existence is acknowledged by a particular theory or system of thought." From the perspective of database integration, each database can be regarded as a "theory" that acknowledges the existence of a set of objects—its ontology. Some care needs to be taken to distinguish this traditional metaphysical use of the word ontology from one that was recently developed in Computer Science. Here an ontology is regarded as a "specification of a conceptualization" and has been applied to a wide range of things, including dictionaries. This sense of the word does not give us a fine-grained enough tool for our needs: it regards a database simply as an ontology—and so it cannot make sense of talking about the ontology underlying it, let alone underlying a set of databases.

Along with the traditional philosophical sense of ontology there is a related notion of semantics—where this is the relationship between words (data) and the world—the word (data) describe. This needs to be distinguished from the different, but related, sense of the word in linguistics where it means the study of meaning. These notions of ontology and semantics can then be used to describe two other useful notions—that of an ontological model and semantic divergence.

An ontological model is a model that directly reflects the ontology. There is a simple semantics where each object in the ontology has a direct relationship with the corresponding representation in the model. One of the characteristics of an ontological model is that the representations in it can be regarded as the names of the objects in the ontology.

Semantic divergence occurs where an item in the representation does not map directly onto an object in the ontology. Semantic heterogeneity occurs when apparently similar items in two different representations have different semantics. The notion of semantic divergence and semantic heterogeneity overlap—but do not

coincide. By itself, semantic divergence does not necessarily lead to semantic heterogeneity. If two databases that need to be integrated have identical semantic divergences, then they are not semantically heterogeneous, they work semantically in the same way. In practice, much of the semantic heterogeneity in databases has its sources in differing semantic divergences and most database integration projects have to deal with significant semantic divergence.

The distinction between semantic heterogeneity and divergence can be used to characterize the way in which the ontological matching strategy proposed here differs from that typically adopted. Currently, many integration projects view the semantic matching process as a mechanism for dealing with semantic heterogeneity—focusing on resolving the semantic differences between the databases. And they analyze these differences using "real world semantics." The unified database is then a combination of the homogenous and resolved heterogeneous data, both of which may or may not be semantically divergent. The ontological strategy focuses on purging the semantic divergence from each of the databases, and in doing so, mapping the underlying ontology. This ontology then provides a basis for designing the "single unified database" that is the output of the integration.

The preceding terms can be used to characterize what ontological analysis for semantic integration is. Ontology provides a framework and suggests a process for the analysis needed for semantic matching. This process focuses on the semantics of the database, identifying semantic divergence. It aims to purge this divergence to produce an ontological model. One key aspect of this model is that it explicitly contains at its top level the categories that inform the ontological paradigm.

One of the most important concepts of system biology is robustness [13], which can be defined as constancy of behavior regardless of unsteady situations (e.g., environmental changes). Actually, a striking feature of living organisms is their remarkable resilience to external stimuli: understanding the mechanisms underlying robustness could certainly enhance system biology's clinical translability.

The system biology approach is also considered the constraint-based elucidation for the regulatory mechanisms in metabolic linkages [15]. At the lowest levels, some behaviors of the systems are limited by constraints, at the same time the latter allow other behaviors to emerge [16].

2 Materials and Methods

In order to be able to obtain data that can be input to system biology models, or serve as validation for the latter, the traditional biology lab must be equipped with analytical capabilities, easily

usable by personnel who is not necessarily trained in, e.g., information engineering.

The traditional biology laboratory of the past was mostly concerned with the application of experimental techniques in different areas: e.g., microbiology, biochemistry, cell and developmental biology. The data generated by the experiments were easily managed with a laboratory notebook, possibly together with a collection of images from, e.g., microscopy or electrophoresis. Then, in the past data management was not systematically aimed at sharing of knowledge between researches, leaving this aspect to the cooperation between the individual researchers, given the task assignments by the lab's head.

The organization of the traditional biology lab reflected the view of biological research at the time, in which particular mechanisms were singled out and studied accurately by just a few researchers. In this framework, the originality of ideas could work very well even in the absence of a tight organizational structure.

Nowadays, after the advent of high-throughput techniques, in particular those related to molecular biology, the view of scientific discovery in biology is much less linked to the single researcher's ingenuity and initiative, and is regarded as a complex process, involving the analysis of massive amounts of data (possibly from diverse investigational standpoints), to be interpreted in a system-level perspective. As underlined by Kitano, "System-level understanding requires a shift in our notion of "what to look for" in biology. While an understanding of genes and proteins continues to be important, the focus is on understanding a system's structure and dynamics. Because a system is not just an assembly of genes and proteins, its properties cannot be fully understood merely by drawing diagrams of their interconnections" [13].

With such considerations, it is evident that advanced biological research is only possible today with remarkable resources of data management and analysis. In the next section, we will illustrate the experience of our laboratory on the transition to such a new paradigm for biological research.

2.1 Quantitative Data in Biology: An Infrastructure for Data Analysis and Exchange

In the everyday activity of the up-to-date experimental biology laboratory, different techniques and methods are used, relative to molecular sciences, statistical techniques, and System biology constraints. Experimental biology consists of the integration of typical laboratory activities with scientific theories, drawn from several domains: chemistry, physics, information engineering, and of course biology, just to name a few.

It is widely recognized that available laboratory data (especially from optical microscopy) are not fully exploited. Even though the biologist's insight is always extremely useful in assessing the relevance of cell samples in different experimental conditions, nevertheless the need for more objective and repeatable assessments is

becoming urgent, given the possibility to automatize—at least partially—the analysis of experiments. An obvious example is the necessity to perform statistical analyses on the size distribution of cell or cellular compartments, which has been a traditional, time-consuming chore in biological laboratories, involving manual delineation of contours and calculation of areas/volumes before statistical testing. Image processing techniques (segmentation, either supervised or unsupervised) allow saving considerable amounts of researcher's time, in doing this type of evaluations.

In light of these considerations, at our laboratory, the decision was taken to fit the lab with an informatics infrastructure for data exchange and processing. In doing so, it has been considered that the lab personnel had generally no particular skill or previous experience in information engineering, so one key requirement for managing a quantitative, computer-based approach in image analysis was to make available easily usable software tools, preferably from open source developers.

In the following, we briefly describe the design and deployment at our lab of a suitable platform (COSYSBI, for COoperative SYStems Biology) for System biology. It is a tool for communication, data collection, sharing and dissemination of scientific results of the various members of the lab.

Through a structured organization, the COSYSBI portal, namely, the COSYSBI Communication Center, provides access via the web to a diverse range of information, document repositories, applications, internal processes, and services, breaking the barriers of space and time in which different researchers' activity usually occurs. This results in efficiency improvement of the individual as well as of the groups referring to the lab. While the advantages of using a portal in the field of dissemination of information are clear, the use of the portal for the development of collaborative and interactive activities is not yet established.

The COSYSBI Communication Center, therefore, in addition to being a repository of information and a channel of communication, is an online collaboration tool that, through the virtual workplace, allows and promotes collaboration between cross-functional workgroups, allowing the real-time evaluation of the best solution.

The relocation of the space/time is achieved through the use of the network and the ability to converge the channels of communication and of these functions/services: Audio/video conferencing, email discussion, and/or web (mailing list, news groups), share and edit documents in real time (co-editing), group browsing (co-browsing), surveys, brainstorming.

The intranet portal was implemented through a multidisciplinary approach that takes into account different aspects of communication, web design, organization, architecture, integration, and application development.

The objective of COSYSBI Communication Center is "managing knowledge." The knowledge is then diffused and transferred within the network in an organized way.

The main criteria that must inspire the methodology of design, planning, and construction of the portal are the web usability: here "usability" stands for the degree of ease of use and the functional efficiency of the product.

The variables on which the correct application of usability must be verified are the following:

Web Interface—is the set of all portal web pages and links, at internal as well as at external level. It is possible to analyze the interface from three standpoints: web design, navigation system, personalization; Content—The type of information available will vary depending on the typical user and the nature of the portal.

The following two tables (Tables 1 and 2) suggest how Web Interface and Content should ideally be built.

After the design of the structure of the entire portal, before proceeding to the realization, a structuring of the information flow is necessary. This operation is preliminary with respect to the construction of the portal, because it has an impact on the tools that will be used.

In order to plan the activities of a portal you need to know:

1. Who are the users that you want to address; which are their needs.
2. What areas of the portal are subject to change and which not?
3. The average time required to make the changes.

The tools that can be used for the creation of a portal are countless. However, it is possible to draw some general requirements that any platform must meet. These requirements are:

Table 1
Criteria to be followed in the web interface design (design, navigation, personalization)

Hierarchical structure of information
Side connections
Multiple usability models
Internal search engine
Recognizable design
Accessible design
Navigable design
Predictable design

Table 2
Favorable properties for content

Up-to-date
Of high value
With appropriate depth and extension
Attractive: informative but understandable
Stimulating

- Scalability of the structure.
- Open structure.
- Transportability on any platform of the functions of the portal.

Scalability—It is important to create a flexible and easily scalable structure. From the hardware point of view, this means ensuring continuity of service regardless of the maintenance operations, data storage, or renewal. From a software point of view, it means using tools, languages, and easily expandable methods. Actually, static tools would involve the impossibility to expand the portal without redesigning it again from scratch.

Open structure—The Internet is now a huge archive, with the desired information obtainable by several different devices. Because of this, it is important to build a multichannel access to portal solutions or at least to use open solutions, to be enriched in the future without having to re-implement a new project.

Portability—The portability is the ability to access a resource, regardless of the platform on which it resides. In the implementation phase, this means to bear in mind the presence of: different browsers, often not perfectly compatible with each other; different operating systems, obviously not compatible with each other; the necessity of having to change the tools with technological innovation growth.

The analysis of user requirements concerns the acquisition of the specifications regarding the features, the architecture, and the product technologies to be realized.

The definition of user requirements was particularly long and laborious, since it is related to a purely scientific context. At this point, a series of interviews with researchers of the lab was carried out, in order to design a technological tool able to support projects management as well as communication and sharing between partners, and at the same time serving the real needs of the particular context in which the instrument was to be inserted.

Particular attention was paid to the demand for high flexibility, scalability, and adaptability of the systems and applications.

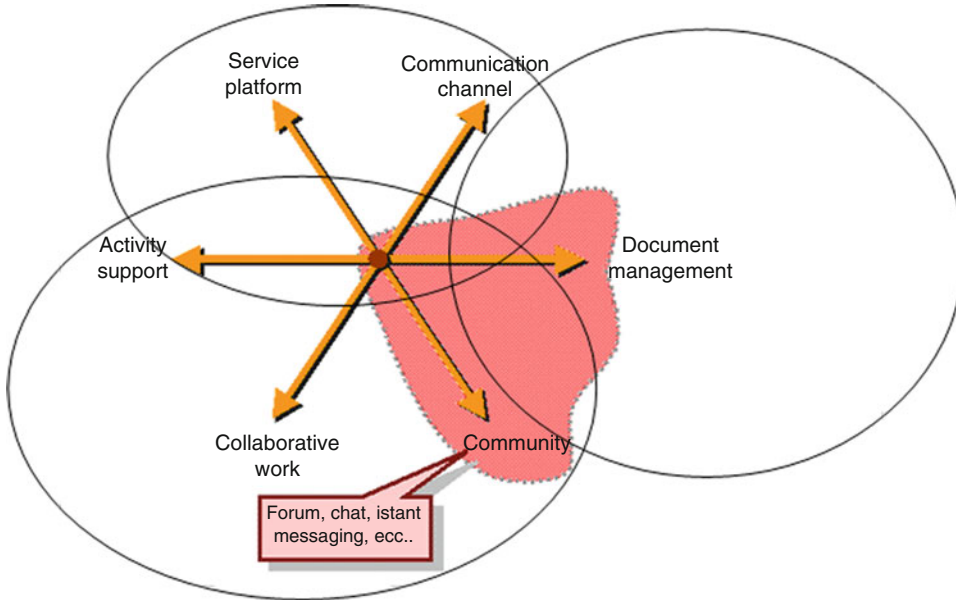


Fig. 1 Logical schematic of COSYSBI Communication Center

The goal was to make individual knowledge understandable and applicable to the entire network, with the aim of collecting all the documents and creating a database through a directory and a search engine, so as to make any relevant piece of information immediately available.

The COSYSBI Communication Center is a tool for knowledge management: then, e.g., if the principal aim is to collect as many documents as possible, to make them easily available, on the other hand you will have to avoid collecting and preserving all those documents that become useless and obsolete.

Hence, referring to a hypothetical ongoing research project, some criteria and some reference group member, with appropriate experience and acquired knowledge will have to be identified, in order to make a sort of review and selection of project documentation. All the details necessary to contact the author of every document will be provided: the possibility of a personal dialogue with whoever is the holder of the special knowledge cannot and should not be excluded from a system of knowledge management.

Thanks to this organizational system, it was found to be easier to maintain consistency and continuity in the analysis of knowledge and documentation collected by the group.

The logical model of COSYSBI Communication Center is shown in Fig. 1

2.2 Current Molecular Imaging Lab Activities

The COSYSBI framework is used not only for data and document management but also for access to computing resources and imaging software (*see* **Note 1**).

In our lab we collect a set of Software tools to approach the quantitative measurement of molecular imaging data. For example, ImageJ [17] is one of the most popular software used for biological imaging analysis (Fractal Dimension and Lacunarity, Color Histogram, Roundness, Cell counter, etc.). Another relevant Software tool is CellProfiler [18]. These softwares are made available, via the COSYSBI architecture already described, to all researchers attending the lab.

2.2.1 An Example of Data Analysis Workflow for Cell Cultures

Innovative research in experimental biology is enabled by the availability of advanced molecular imaging tools. In this review, to characterize cytoskeleton proprieties, we have used the confocal images from in-vitro cells cultures experiments, through multiple analyses and in combination with the morphological information on optical microscopy. The COSYSBI biological system repository is an essential management tool for gathering information on cells (cell lines, morphological characteristics, and information useful for image analysis), to be linked to external databases on cells or previous experiments, performed also by other research groups. The easy accessibility of data stemming from multiple sources is a key factor in deriving characteristic parameters, to be fed into the suitable algorithms.

Confocal microscopy images are extremely useful for the quantitative analysis of experiments and model construction. In particular, visualization of the cytoskeleton structure is useful to understand cell motility, stiffness, and more generally the cell phenotype. Cytoskeleton properties can be investigated, among other methods, by the calculation of the fractal dimension, starting from microscopy data. Such an analysis has been performed on confocal images cells, to observe the changes of cell proprieties due to the experimental protocol. Clear meaningful relationships are identified from the data experimented on different treatment condition and are assessed graphically and statistically interpreted.

Figure 2 shows the protocol used for the quantitative evaluation of the shape parameters.

The first step is to snake the cell membrane. This is a difficult task to do at computational level. Often the biologist needs to manually administer this task. He is able to identify the profile of the cell membrane with all its protrusions, based on his own experience, even when the membrane is not on the microscope's focus plane. This task is particularly decisive for the correct evaluation of cell morphology.

The next step is to subtract the background, due to the nutrient medium and microenvironment. Even this step is particularly crucial for a good fit. The ability to use automatic threshold algorithms is almost always impossible: in this case too, the biologist supervision is required.

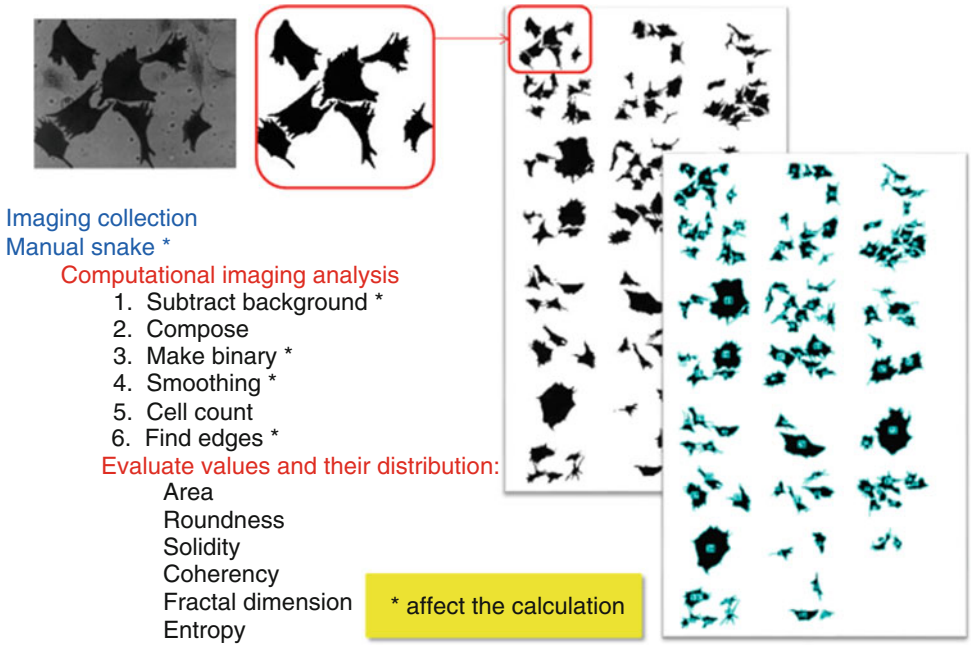


Fig. 2 Computational protocol to quantitative evaluation of cell shape parameters

After finishing the segmentation and background subtraction work, all the cells can be brought back to a single image, representing our virtual slide.

Before proceeding to the quantitative measurement of shape parameters, it is usually necessary to perform other preprocessing operations on the images. Figure 2 shows binarization, smoothing, and vectorization (find edges) as an example. At this point you can start measuring shape parameters for each cell.

Each parameter is linked to a biological meaning, so that collaboration between bioinformatics and the biologist is really strategic. Measurement and interpretation work should be understood as a continuous work cycle. During this cycle, the bioinformatician and the biologist further refine the measurement protocol and decide which tasks can be left to automated calculations and which others do not. It is important to observe that the quantitative measurement of shape parameters can also provide useful indications for the evaluation of cell functions. The results of this measure will be supplemented with those from confocal microscopy to quantitatively characterize the cell phenotype. With the Systems Biology approach, applying fusion between information from various imaging techniques, we can realize the construction of an electronic phenotype that virtually represents our cell culture.

In this part, we are briefly outlining how to extract the information about the cell structures of Cytoskeleton from confocal microscopy images with ImageJ and CellProfiler open source

Table 3
Pipeline for performing quantitative morphological analysis on confocal images with ImageJ

Step 1	• Identifying the cells in ImageJ Software • Calculate the split channels to separate the microtubules and nuclei
Step 2	• Refining and adjusting the threshold • Adjust the top slider • Adjust the bottom slider • Calculate the threshold of each confocal image

Table 4
Pipeline for performing quantitative morphological analysis on confocal images with CellProfiler

Step 1	• Calculate the ratios (area nucleus) • Measuring the image area occupied • Calculating the image intensity (cells) • Measure object intensity (nuclei) • Predict correlation behavior of the cells
Step 2	• Measuring object size shape (nuclei) • Performing the calculation and setting up the parameters for graphical representation to display the data comparison

software. ImageJ is one of the principal and most popular scientific image analysis software. To facilitate more advanced and clear results the program provides many plugins, made by a lively scientific community, to add functionality and visualization tools. This program is used in over 30,000 laboratories. ImageJ is a very popular research tool helpful in imaging processing and run in any web browser. *Cell profiler* enables gathering information regarding number, thickness and length of actin, internal filaments, and microtubules. Thus, it provides comprehensive structural information about the cytoskeleton under investigation.

Details of how to acquire the information about the image processing of the cellular data from ImageJ are descripted in [19, 20]. In our experience, we applied the algorithms for the identification of functional information, analyzing molecular imaging data regarding experiments with cell line. In doing so, we have used the built-in functions of ImageJ to filter and analyze our data. An outline of the algorithmic steps is given in Tables 3 and 4.

We investigate area, roundness, fractal dimensions, solidity, entropy, coherence to evaluate the cytoskeleton morphotype and phenotype properties and to compare the results retained from treated and untreated cells data. *See* in **Note 2** for more details about these steps.

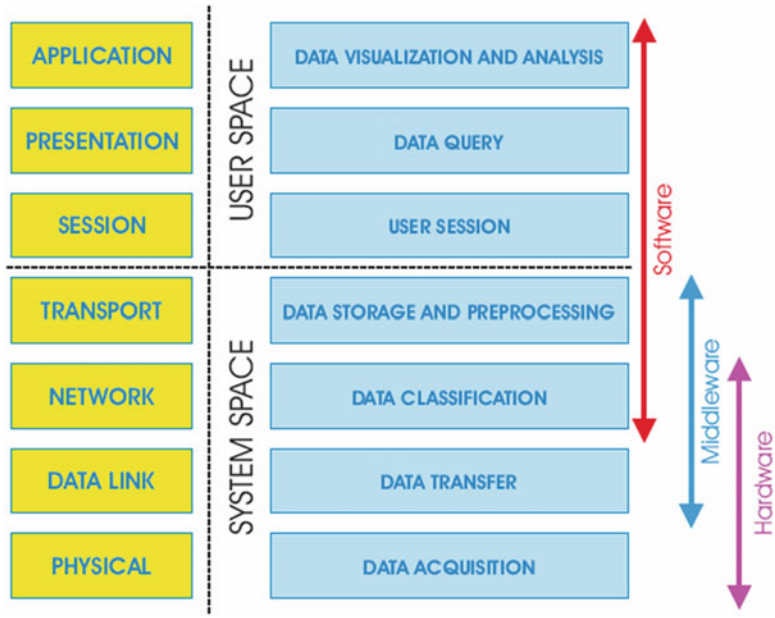


Fig. 3 Conceptual architecture of the COSYSBI integrated platform

2.2.2 An Example of Data Analysis Workflow for MRI Imaging

The diagram in Fig. 3 shows the conceptual architecture of the COSYSBI integrated platform for the MRI imaging management.

At the base of the system space are devices that interfere with data acquisition equipment: MR scanners operating at 1.5, 3.0, and 7.0 T, electroencephalographic equipment (can be used in isolation or in conjunction with MR scanners in the case of simultaneous acquisition of neuroradiological and neurophysiological data). The raw data produced at this first level (data acquisition) is transferred through the DICOM protocol (images) and dedicated protocols (EEG and MRS) (second level, data transfer) to a (classification) third tier structure. In the third level, the data is sorted by two possible addresses:

- Usable data for clinical purposes.
- Usable data for research purposes.

The fourth level (data storage and preprocessing) involves three procedures:

- Convert raw data into interchange format.
- Preprocessing automated raw data (clinical data only).
- Population of repository tables based on the data header.

At this level, appropriate automatic backup procedures will be implemented. The first level of the section in the user space implies the opening of a user session, during which the user (local or remote) authenticates and accesses the system, which grants

privileges corresponding to the different user groups. The three main types of users envisaged are:

1. Medical personnel who are exclusively clinical.
2. Medical staff and researcher carrying out clinical research purposes.
3. Researchers with different profiles.

At the next level (data query), the user queries the repository via a different interface according to the allowed activity by access profile.

The last level (data visualization and analysis) provides:

1. View the output as a result of the queries.
2. Interactive analysis corresponding to automated level 1 and 2 processing (e.g., thresholding, interactive multi-plan view, etc.).
3. Download formatted data for custom analysis (level 3).

Figure 4 represents the operative scenario of the integrated framework.

Through the Acquisition Console, users can select the type of activity to implement:

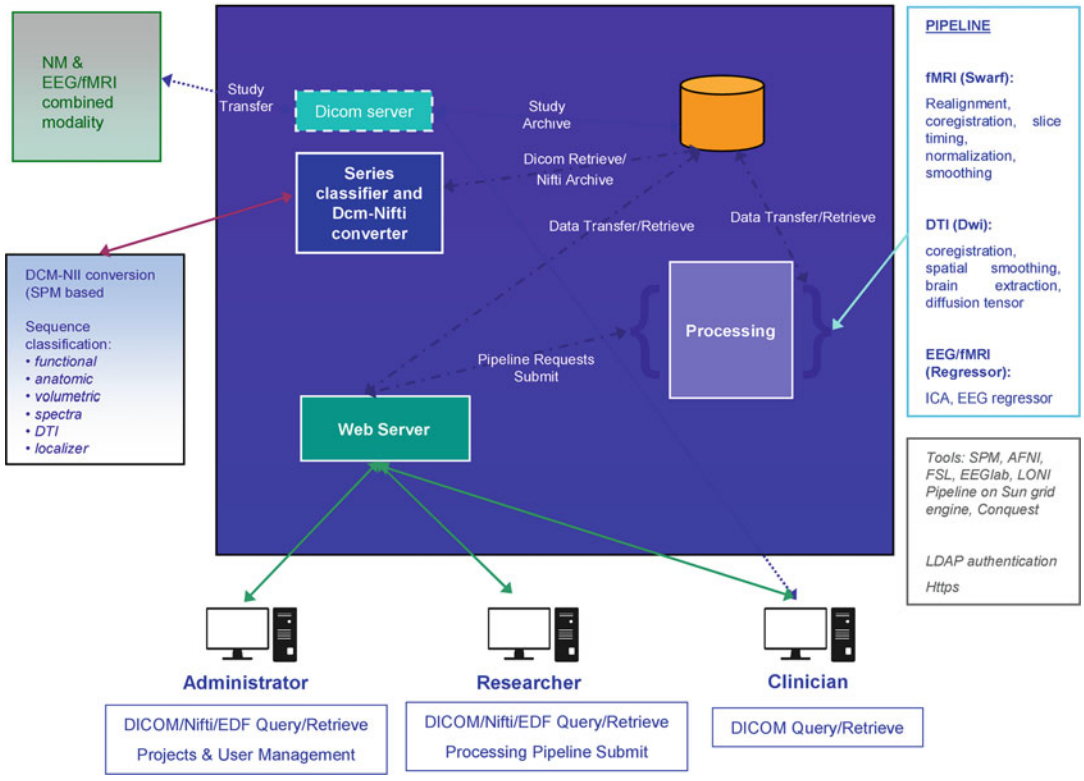


Fig. 4 Operative scenario of the integrated framework

- "Acquisition for Clinical Purpose."
- "Acquisition for Research Purposes."

Then the sequence type to be recorded is selected and the acquisition phase begins.

The acquired data is transferred through the DICOM protocol (images) and dedicated protocols (EEG and MRS) to a data classification structure.

Data is classified into:

- Usable data for clinical purposes.
- Usable data for research purposes.

The Central Server receiving the acquired data identifies the typology of the data (by reading the header); thus, data processing for the storage phase begins.

The sequences of captured data (images and signals from external devices) will be processed and the results will be made available to medical users under a cooperation model that will also enable them to intervene with further processing on the sequences.

The integration of acquired data is to be understood as an integrated view of the various image sources, but above all as image processing and signaling aimed at obtaining the association of the information typologies of the different techniques to the same anatomical structures and, when assimilated, even electrophysiological signals, including the temporal dimension.

In case the data were transferred with the "EEG" protocol, conversion of the raw data into the EEG interchange format is expected.

In case the data were transferred with an MRS protocol, conversion of the raw data into MRS interchange format is expected, and in this case also occurs before storage in the repository, if it is a sequence of Type R (Acquisition for Research Purposes). If this is not the case, preprocessing is performed.

Figure 5 shows how to use the integrated platform to extract from MRI 3D images anatomic 3D structures of interest. Such structures can then be used to create models of computational fluid dynamics or to create real models by 3D printing technique.

2.3 Future Trends for Automated Data Analyses in Biology

Technological breakthroughs and developments continue to give the experimentalists ever more tools, with increasing spatial and temporal resolution, to investigate living tissues and cells. For instance, metabolomics—a post genomic research field concerned with developing methods for analysis of low molecular weight compounds in biological systems (cells, organs, or organisms)—is currently investigated with an impressive array of techniques: no less than 14 different common analytical techniques for metabolomics were already listed by Hollywood [21].

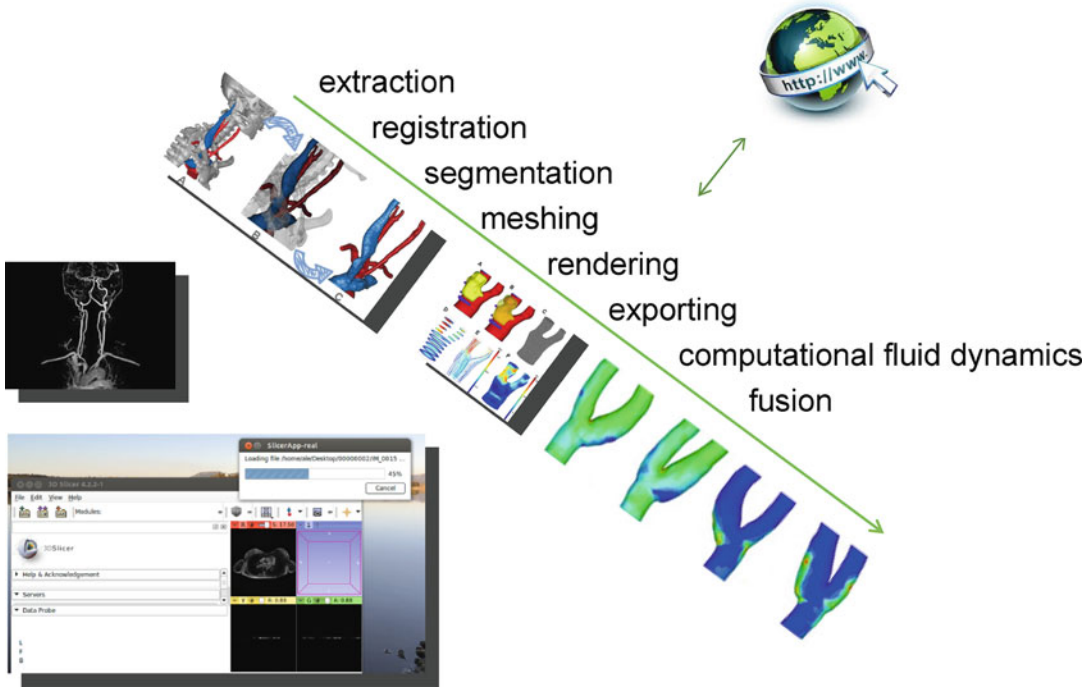


Fig. 5 The typical pipeline to obtain 3D in silico model starting from MRI 3D DICOM image set

Of course, in metabolomics analysis, as in the other omics analyses, large amounts of data are already routinely produced. It is easily predictable that this "data deluge" will not decrease in the future, requiring careful attention in the analysis of the results of each experiment.

Of equal—or even greater—importance will be the consideration of the quality of measurements. For instance, when acquiring continuous data streams for hours or even days, electronic drift can bias the results, so equipment checks must be scheduled periodically.

Statistical methods will be necessary to demonstrate compliance to accuracy and precision requirements. Also reference samples—especially for the elusive metabolomics domain—need to be developed in conjunction with such statistical methods.

In the last 5 years, there has been a great increase in the number of molecular bioimaging tools and bioimaging control software which support the microscopy hardware to perform the analyses very quickly and flawlessly.

To efficiently analyze the ever increasing amounts of data, and extracting information from them, it is essential to have Big Data systems that can quickly find the correct information, process it,

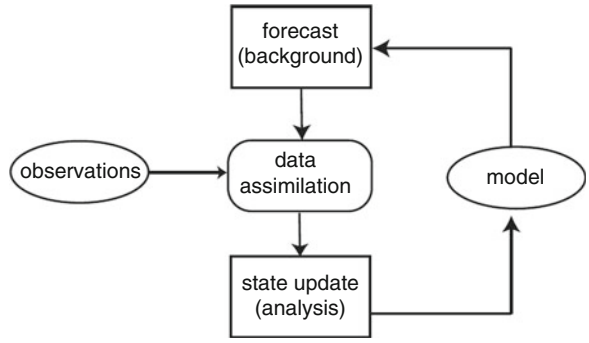


Fig. 6 Schematic illustration of the data assimilation

and produce analytical results. However, although extremely important, the Big Data Analytics function is only one of the functions that can be achieved with the engineering techniques involved in the Big Data area. Even more important are data mining, data assimilation, and machine learning techniques; Machine learning techniques are used for data analysis and pattern recognition and thus they can play a key role in the development of data assimilation procedure (Fig. 6), data mining applications, and, in the near future, artificial intelligence.

Data analysis can be performed after acquisition of biological data. Correlations between different biological factors, clinical and non-clinical data and molecular analysis variants can be performed, giving, i.e., an estimation of the probability of developing a pathology, given the presence of each one of the risk factors; a predictive model can be constructed during the follow-up time to monitor the disease's onset.

3 Notes

1. Here, we will look at some image analysis tools to analyze and visualize our imaging data. Some open source software includes ImageJ, Fiji, Meta Morph, Amira, Imaris; these softwares are able to offer microscopy data analysis with direct linking to imaging instrumentations.

   The open source software is helpful for the image analysis and image informatics workflow to visualize the data in an easier way such as: Bioimage XD [22], Icy, Fiji [23], Cell profiler.

   There are image database in which the public repositories are available for the data associated with the number of articles already published. These include as follows:

The protein subcellular location image database (http://pslid.org/), The human protein data base atlas (http://proteinaltalas.org), LOCATE (http://locate.imb.uq.edu.au/).

Besides, there are many integrative platforms for the image data management. For example OMERO or BISQUE or KNIME is used for bioimage analysis. Interestingly, machine learning powerful microscopy analysis is concerned with powerful machine learning algorithms, which automatically classify the pixels of the images and applied to the experiments for identifying the microscopy images [24]. In the last two decades, several groups have innovated various collections of scientific images for biological experiments. Workflow systems tools have also enabled serving multiple data resources. Open source workflow tools are used as Taverna (http://www.taverna.org.uk/) and Galaxy (http://galaxy.psu.edu/).

2. Let us start with one confocal image of the sample from our experimental data. First, we load the data into ImageJ. For this purpose, we use ImageJ built-it functions: File --- open --- image from folder. Further, we split the channels by using the functions inside the ImageJ session through, file --- image --- Color --- Split channels (Fig. 7).

   It is important to split the channels to separate the nuclei and microtubules and to extract the resulting information. Next, we contour the nuclei and microtubules to measure the cell parameter that we choose to characterize morphological and phenotypic features (Fig. 8).

   We can also use split images from the confocal data to overlay the nuclei over the underlying microtubules (Fig. 9).

   Second, we analyze the nuclei that we have loaded before and saved into the folder with the nuclei Region Of Interest (ROI) manager tool. Once the threshold parameters have been accomplished, it is of interest to analyze the effects and the different structural dynamics onto the biological processes in which the evaluations of the model were hypothesized. We have adjusted the nuclei threshold at 9.20–15.20% (approx. for the sample images) with the size = 100–20,000 pixel for the analyses of the nuclei that has been adjusted. In this workflow, we have calculated the morphological information for nuclei image data.

   The integration of the information about the confocal images is analyzed with the CellProfiler. We need to adjust the image according to a preprocessing calibrating protocol. This is a very important step, at the basis of quantitative measurements. And then, by drag and drop the individual confocal image we load the data into the cell profiler. The resulting

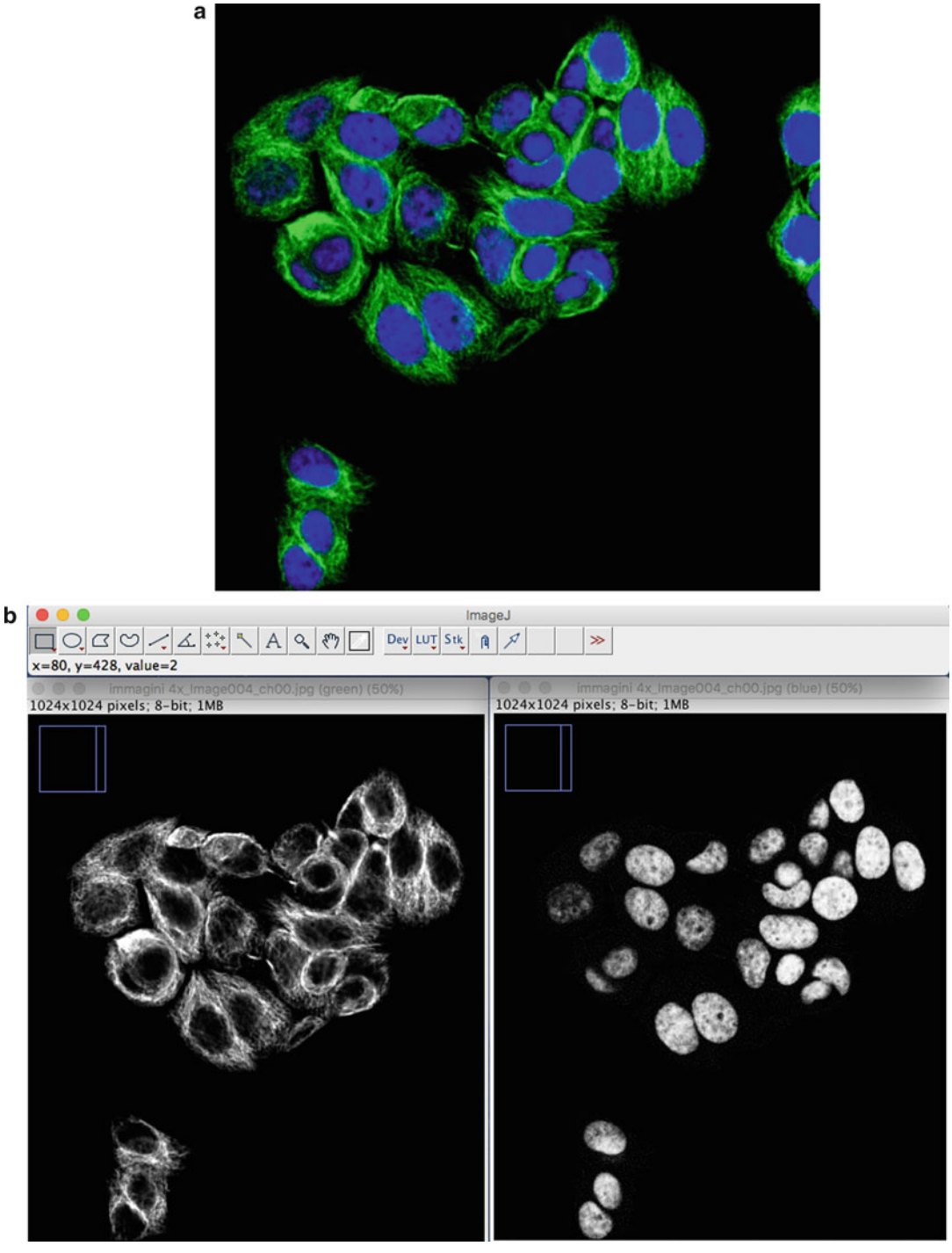


Fig. 7 Split images of the confocal image data (**a**). The enhanced contrast and brightness plug-in has been used to visualize the microtubules and the nuclei (**b**)

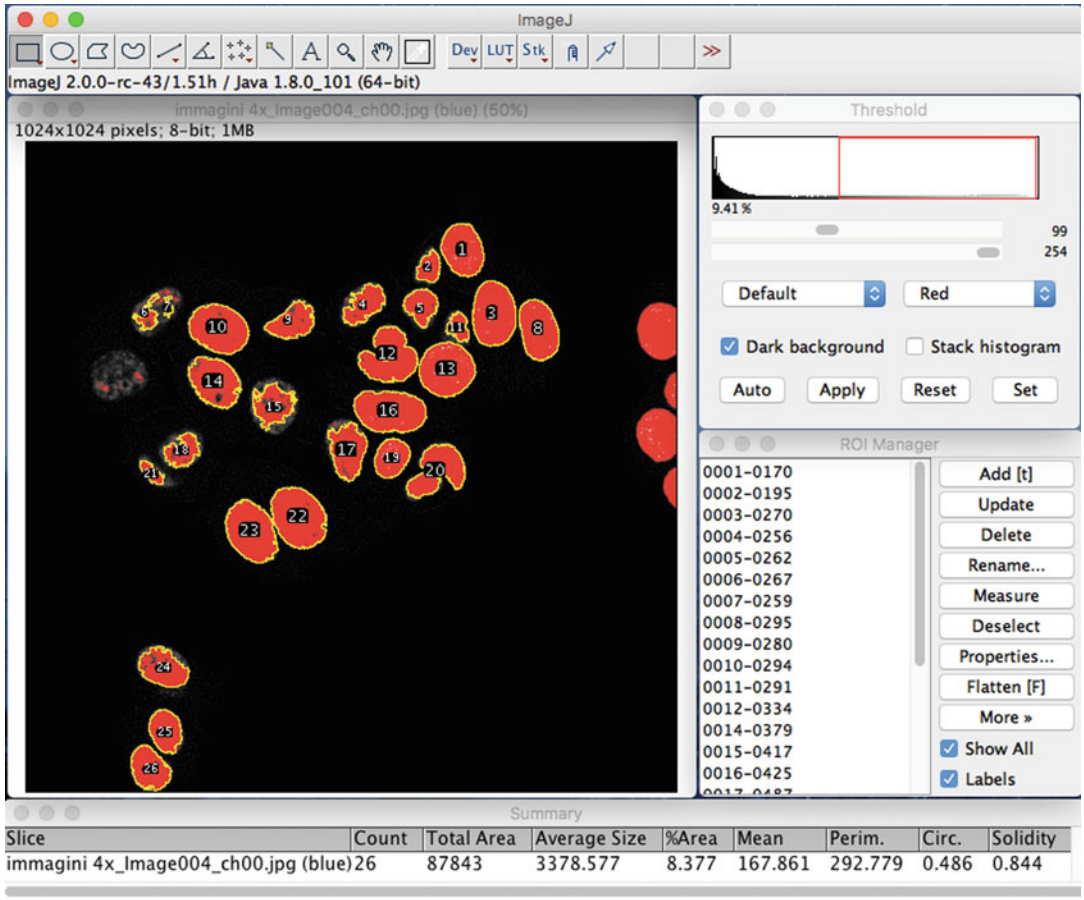


Fig. 8 ImageJ screenshot of the analysis and measurement of the nuclei. The confocal image is selected to adjust the threshold parameters to count the nuclei in the image. The nuclei contain information on cytoskeleton experiment. All nuclei are *red* in *color* has been applied threshold. The particles were analyzed from the in-built functions to allow the calculation and identification of the cells

image will enable extracting the information with modules settings underneath.

Now, we need to add the modules to use our pipeline. When the modules are selected through the settings, we can run the pipeline and model the output at every step through the embedded execution functions (Fig. 10, 11, and 12).

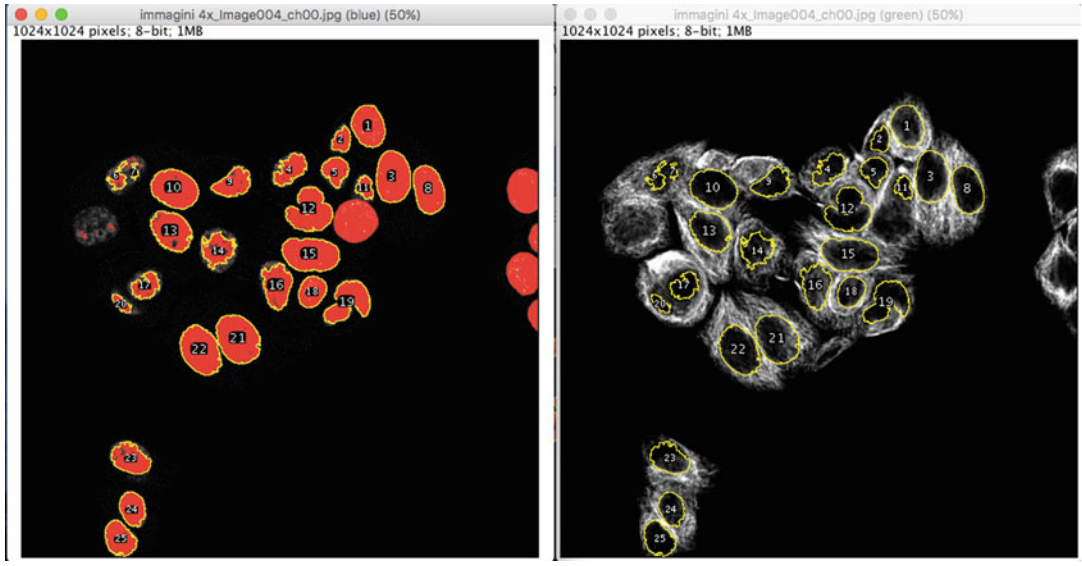


Fig. 9 ImageJ analysis screenshot of selected nuclei by applying the threshold parameter for the selection on the confocal images data

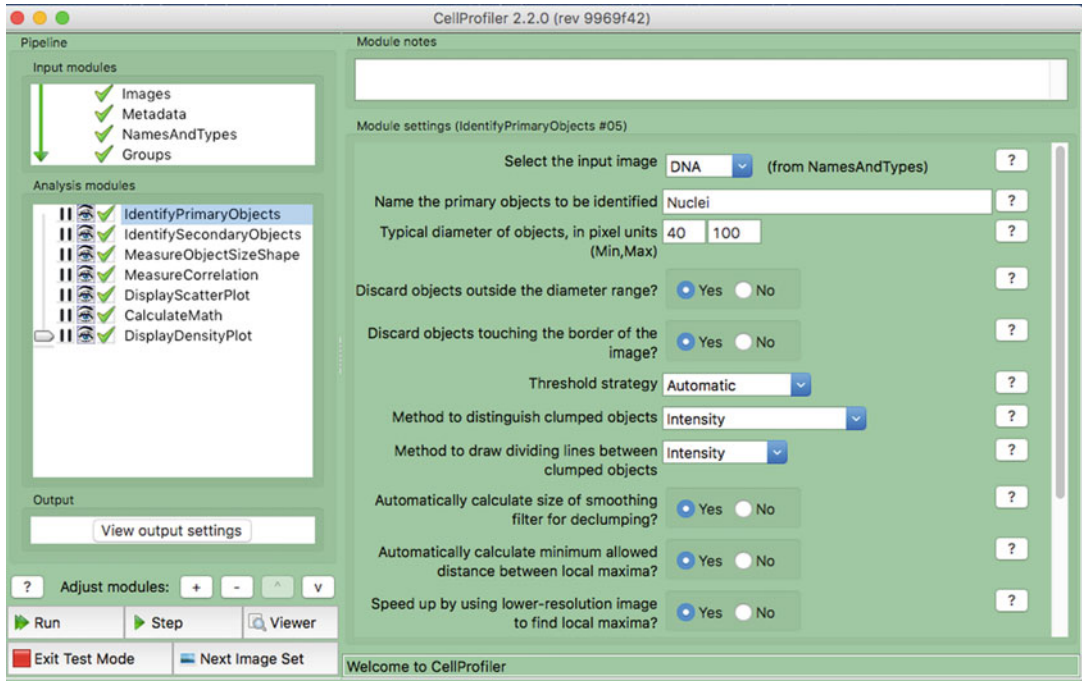


Fig. 10 CellProfiler pipeline

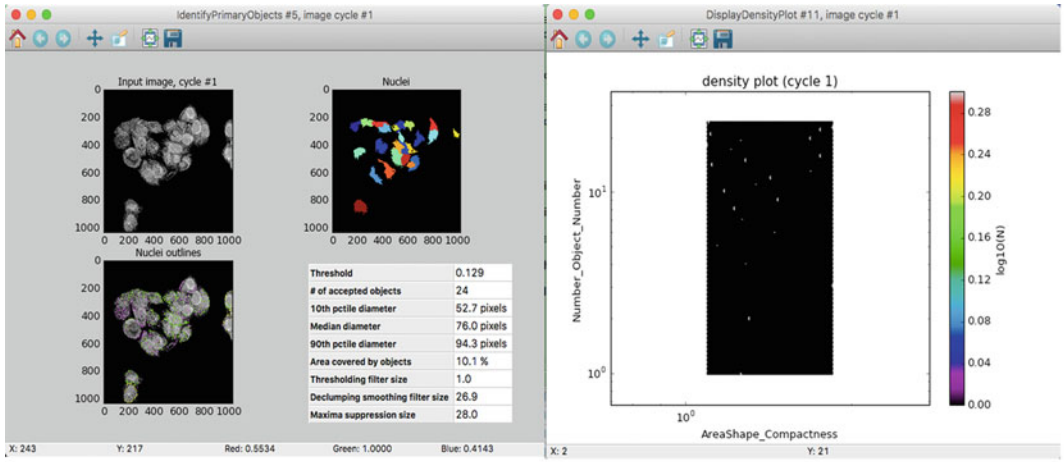
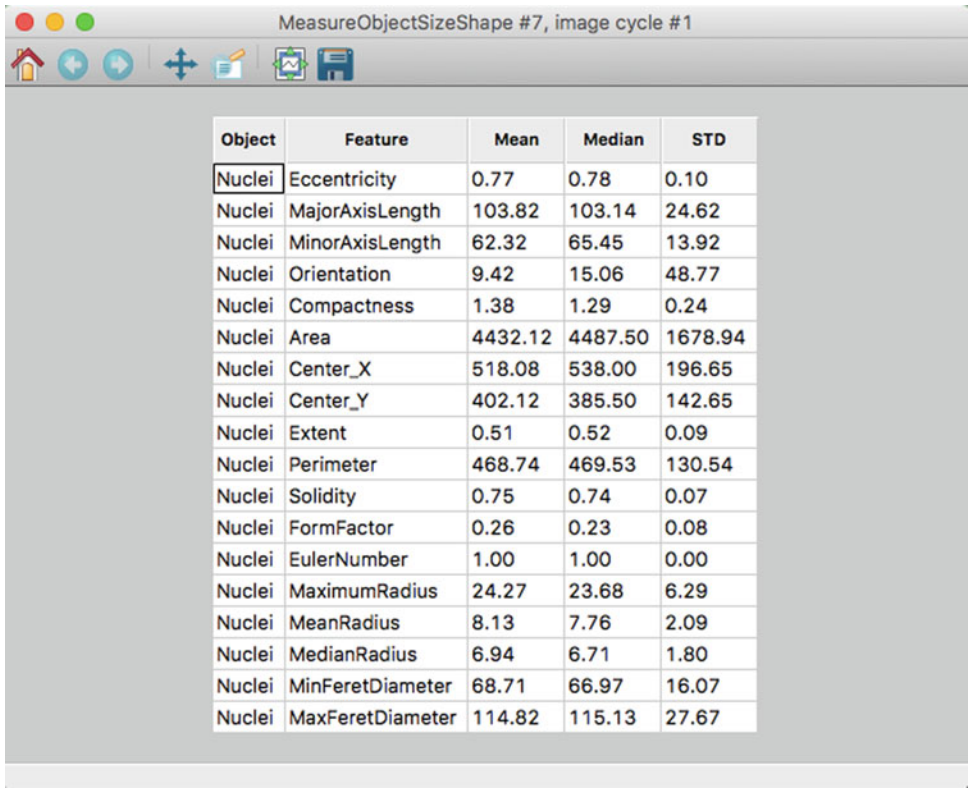


Fig. 11 Cell profiler to import the data from a drag and drop function. It enables improving cell handling through a computerized processing tool cell profiler. The tool is able to count the cells and threshold and on the left is to measure the density plot



Object	Feature	Mean	Median	STD
Nuclei	Eccentricity	0.77	0.78	0.10
Nuclei	MajorAxisLength	103.82	103.14	24.62
Nuclei	MinorAxisLength	62.32	65.45	13.92
Nuclei	Orientation	9.42	15.06	48.77
Nuclei	Compactness	1.38	1.29	0.24
Nuclei	Area	4432.12	4487.50	1678.94
Nuclei	Center_X	518.08	538.00	196.65
Nuclei	Center_Y	402.12	385.50	142.65
Nuclei	Extent	0.51	0.52	0.09
Nuclei	Perimeter	468.74	469.53	130.54
Nuclei	Solidity	0.75	0.74	0.07
Nuclei	FormFactor	0.26	0.23	0.08
Nuclei	EulerNumber	1.00	1.00	0.00
Nuclei	MaximumRadius	24.27	23.68	6.29
Nuclei	MeanRadius	8.13	7.76	2.09
Nuclei	MedianRadius	6.94	6.71	1.80
Nuclei	MinFeretDiameter	68.71	66.97	16.07
Nuclei	MaxFeretDiameter	114.82	115.13	27.67

Fig. 12 An example of result table of nuclei analysis through cell profiler inbuilt pipeline functions

References

1. Schneider MV (2013) Defining systems biology: a brief overview of the term and field. Methods Mol Biol 1021:1–11
2. Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664
3. Bizzarri M, Palombo A (2015) Recognizing the "patient's phenotype" through systems biology. Curr Synth Syst Biol 3:1
4. Bizzarri M, Palombo A, Cucina A (2013) Theoretical aspects of systems biology. Prog Biophys Mol Biol 112(1–2):33–43
5. Bizzarri M et al (2011) Fractal analysis in a systems biology approach to cancer. Semin Cancer Biol 21:175–182
6. Fuchs R, Rice P, Cameron GN (1992) Molecular biological databases—present and future. Trends Biotechnol 10:61–66
7. Gasch AP et al (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4241–4257
8. Kitano H (2004) Cancer as a robust system: implications for anticancer therapy. Nat Rev Cancer 4(3):227–235. https://doi.org/10.1038/nrc1300
9. Kitano H (2002) Computational systems biology. Nature 420:206–210
10. Gibelli L (2015) Stochastic features and strategy of computational methods: comment on "On the interplay between mathematics and biology, hallmarks toward a new systems biology" by N. Bellomo et al. Phys Life Rev 12:74–75
11. You L (2004) Toward computational systems biology. Cell Biochem Biophys 40:167–184
12. Hood L (2003) Systems biology: integrating technology, biology, and computation. Mech Ageing Dev 124:9–16
13. Kitano H (2004) Biological robustness. Nat Rev Genet 5:826–837
14. Sheth AP, Larson JA (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput Surv 22(3):183–236
15. Bertalanffy LV (1968) General system theory: foundations, development, Applications. Springer, New York, NY, p 295
16. Svoboda J (2008) Foundations in cancer research. The turns of life and science. Adv Cancer Res 99:1–32
17. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9:671–675
18. Lamprecht MR, Sabatini DM, Carpenter AE (2007) CellProfiler: free, versatile software for automated biological image analysis. BioTechniques 42:71–75
19. Girish V, Vijayalakshmi A (2004) Affordable image analysis using NIH Image/ImageJ. Indian J Cancer 41:47
20. Collins TJ (2007) ImageJ for microscopy. BioTechniques 43:25–30
21. Hollywood K, Brison DR, Goodacre R (2006) Metabolomics: current technologies and future trends. Proteomics 6:4716–4723
22. Kankaanpaa P et al (2012) BioImageXD: an open, general-purpose and high-throughput image-processing platform. Nat Methods 9:683–689
23. Schindelin J et al (2012) Fiji: an open-source platform for biological-image analysis. Nat Methods 9:676–682
24. Murphy RF (2011) An active role for machine learning in drug development. Nat Chem Biol 7:327–330

INDEX

Mariano Bizzarri (ed.), *Systems Biology*, Methods in Molecular Biology, vol. 1702,
https://doi.org/10.1007/978-1-4939-7456-6, © Springer Science+Business Media LLC 2018

N

O

P

R

S

T

V