An evolutionary multi-agent algorithm to explore the high degree of selectivity in three-dimensional protein structures

Conference Paper · June 2017

DOI: 10.1109/CEC.2017.7969431

3 authors:

Leonardo De Lima Corrêa
Universidade Federal do Rio Grande do Sul
8 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Mario Inostroza-Ponta
University of Santiago, Chile
40 PUBLICATIONS **440** CITATIONS
SEE PROFILE

Marcio Dorn
Universidade Federal do Rio Grande do Sul
58 PUBLICATIONS **188** CITATIONS
SEE PROFILE

An evolutionary multi-agent algorithm to explore the high degree of selectivity in three-dimensional protein structures

Leonardo de Lima Corrêa
Institute of Informatics
Laboratory of Structural Bioinformatics and Computational Biology
Federal University of Rio Grande do Sul
Porto Alegre - RS - Brazil
llcorrea@inf.ufrgs.br

Mario Inostroza-Ponta
Departamento de Ingenieria Informatica
Center for Biotechnology and Bioengineering
Universidad de Santiago de Chile
Santiago - Chile
mario.inostroza@usach.cl

Márcio Dorn
Institute of Informatics
Laboratory of Structural Bioinformatics and Computational Biology
Federal University of Rio Grande do Sul
Porto Alegre - RS - Brazil
mdorn@inf.ufrgs.br

***Abstract*—Tertiary protein structure prediction in silico is one of the most challenging problems in Structural Bioinformatics. The challenge arises due to the combinatorial explosion of plausible shapes, where a long amino acid chain ends up in one out of a vast number of three-dimensional conformations. The rules that govern the biological process are partially known, which difficult the development of robust prediction methods. Many computational methods and strategies were proposed over the last decades. Nevertheless, the problem remains open. The agent-based paradigm has been shown a useful technique for the applications that have repetitive and time-consuming activities, knowledge share and management, such as the integration of different knowledge sources and modeling of complex biological systems. In this paper, we propose a first principle method with database information for the 3-D protein structure prediction problem. We do so by designing a multi-agent approach that uses concepts of evolutionary algorithms to speed up the search phase by improving local candidate solutions from the protein conformational space. To validate our method, we tested our computational strategy on a test bed of eight protein sequences. Predicted structures were analyzed regarding root-mean-square deviation, global distance total score test and secondary structure arrangement. The obtained results were topologically compatible with their correspondent experimental structures, thus corroborating the effectiveness of our proposed method. As observed, the evolutionary multi-agent approach achieved good results in terms of the evaluted measures and was able to efficiently search the roughness of protein energy landscape.**

I. INTRODUCTION

The prediction of the three-dimensional structure of proteins is an important research area in Structural Bioinformatics and could be described as the efforts to predict the unknown 3-D structures of proteins [1]. Proteins are present in all living systems, performing different functions. The function performed by a protein is strictly related to its adopted conformation [2]. This is the primary incentive for researchers in the field, besides the large gap between the number of known protein sequences and known 3-D protein structures.

Proteins or polypeptides are polymers made of 20 different amino acid residues. Each protein is defined by a unique sequence of amino acids that under some physiological conditions fold into a particular 3-D shape, known as the native state of the protein [3]. Determining the 3-D structure of a protein is both experimentally expensive (due to the costs associated with crystallography, electron microscopy or NMR), and time-consuming. Proposing methods that determine quickly and at low-cost, accurate protein structures will contribute to life science fields such as Medicine, Biotechnology, and the Pharmaceutical industry. Tertiary protein structure prediction is currently one of the challenging problems in Structural Bioinformatics [4]. The challenge arises due to the combinatorial explosion of plausible shapes, where a long amino acid chain ends up in one out of a vast number of 3-D conformations. The 3-D protein structure prediction (PSP) problem is classified in computational complexity theory as an NP-hard problem due the high dimensionality and complexity that the search space can assume, even for a small protein [5]. Therefore, currently, there is not any general method capable of achieving the optimal solution.

Over the last years, several computational strategies have been proposed as a solution to the PSP problem. Existing methods can be studied into four classes [1]: (*i*) first principle methods without database information [6]; (*ii*) first principle methods with database information [7]; (*iii*) fold recognition methods [8]; and (*iv*) comparative modeling methods [9]. The first group of methods, which do not rely on sequence similarity of known structures, aim at predicting new folds only through computational simulation of physicochemical properties of the folding process of the proteins in nature. This methodology is guided by the fact that the native structure of a protein corresponds to the global minimum of its free energy [10]. This kind of methods are also known as *ab initio* approaches. Groups *ii*, *iii* and *iv* are classified as knowledge-based methods and are capable of making predictions when template structural information's are available from experimentally determined protein structures. Specifically, the group *ii* represents a hybrid class of methods, that make use of model

978-1-5090-4601-0/17/$31.00

information combined with a classic *ab initio* approach.

Currently, to predict the 3-D structure of proteins only from its amino acids sequence (first principle methods), a wide range of optimization algorithms and metaheuristics are being developed and applied. These methods try to find approximated solutions for this problem. To improve the results and reduce the protein conformational search space, metaheuristics commonly make use of previous knowledge of known protein structures stored in structural databases, such as the popular *Protein Data Bank* (PDB) [11]. Despite the advances in the development of computational methods for the PSP problem, further research remains to be done. The development of new strategies, the adaptation and investigation of new approaches and the combination of existing state-of-the-art computational techniques is clearly needed [1]. In this paper, we explore concepts of autonomous agents and multi-agent systems (MAS) in an attempt to search in a more effective way the protein conformational space to identify native-like protein structures. Commonly, computational search strategies have to handle with the roughness of the protein conformational space where a protein molecule can have multiples conformations. Thereby, we propose a multi-agent system that incorporates concepts of evolutionary algorithms and makes use of the knowledge stored in the PDB to deal with the 3-D PSP problem. Our main contribution in this work is the design and assessment of an efficient and robust method to search the protein conformational space. The multi-agent paradigm can be effective exploited as an important tool to investigate the properties of biological systems that are difficult to study in more traditional ways, for example with in vivo or in vitro experiments. The remainder of the article is structured as follows: Section II describes some fundamental concepts of protein structure representation and fitness function; conformational preferences of amino acids and multi-agent systems. Section III presents an overview of the most recent works developed in the area of the multi-agent systems applied on the PSP problem, focusing on *ab initio* methods with database information. Section IV describes the proposed approach to exploring the three-dimensional protein conformational space. Section V shows the experiments and presents the evaluation of the obtained results. Section VI concludes and points out directions for further research.

II. BACKGROUND

A. Proteins and its representation

Protein structures can adopt a variety of shapes, and the 3-D structure of a protein is defined by its amino acid sequence that folds spontaneously during or after the biosynthesis. The amino acid sequence is not the only responsible for the final protein conformation. The relation between the amino acid sequence of a protein and its structure depends on many factors such as solvent, the concentration of salts, temperature, etc. The computational representation of a 3-D protein structure is a challenging task due to the difficulty in representing the protein structure and simulating the factors that contribute to the inherent structure stability. This representation is related to the level of detail used to describe the 3-D protein structure. The higher the number of features, higher is the capacity of representing the protein in its native state. The geometric representation is one of the most important elements of 3-D protein structure prediction methods and is directly related to the reduction or increase of the protein conformational search space. Using all atoms to represent the protein is computationally expensive, and thus, simplified representations are needed. There are two most common representations of polypeptides structures found in the literature. The first model represents the 3-D protein structure through the Cartesian position of the atoms. The second model represents the protein structure using the set of dihedral torsion angles and is based on the fact that bond lengths are nearly constants in a polypeptide chain (Fig. 1). The use of dihedral angles has the advantage over the Cartesian model for having reduced degrees of freedom.

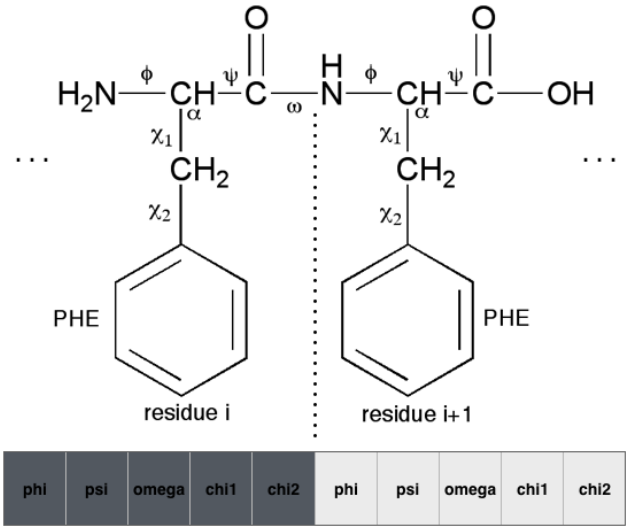


Fig. 1: Schematic representation of a protein molecule and their torsion angles.

A peptide is a molecule composed by two or more amino acid residues chained by a chemical bond known as *peptide bond*. All amino acids found in proteins have the same main structure, identified as main chain or backbone, and differ only in the format of the side chain. Plenty of chained peptides are commonly called polypeptides or proteins. The peptide bond (C−N) has a partial double bond and tends to be planar. The rotation is only permitted around the bonds $\mathrm{N}-\mathrm{C}_\alpha$ (Phi ϕ) and $\mathrm{C}_\alpha-\mathrm{C}$ (Psi ψ). These angles are the most responsible for the conformation adopted by a protein molecule. The stable local arrangement of amino acid residues in the protein forms its secondary structure [2]. Similar to the polypeptide backbone, side-chain also have dihedral angles, and its conformation contributes to the protein structure stabilization and packing. The number of angles Chi (χ) of the side-chain depends on the amino acid type. In this work, protein structures are represented only by the dihedral angles (Phi ϕ), (Psi ψ), and the side-chain Chi (χ) angles, as a way to reduce the complexity of the *all-atom* protein representation. In our method, the algorithm receives as input parameters only the amino acids

sequence of the target protein and its respective secondary structure. For the backbone representation of a polypeptide, this gives rise to $2 \times n$ (main-chain) $+ m \times n$ (side-chain) degrees of freedom, where n is the number of amino acid residues and m the number of χ angles which varies according to the amino acid type. The main disadvantage of the usage of dihedral angles is that a small change in one dihedral angle can cause drastic changes in the polypeptide structure.

B. Fitness Function

Protein structure prediction methods change the orientation of atoms of the protein structure to minimize an energy function [12] since the native structure of a protein corresponds to the global minimum of its free energy. A potential energy function incorporates two types of terms: bonded and non-bonded. The bonded terms (bonds, angles, and torsions) are covalently linked. The bonded terms constrain bond lengths and angles near their equilibrium values. The bonded conditions also include a torsional potential that models the periodic energy barriers encountered during bond rotation. The non-bonded potential comprises ionic bonds, hydrophobic interactions, hydrogen bonds, *van der* Waals forces, and dipole-dipole bonds.

To evaluate the quality of a predicted structure, we employed the Rosetta energy function implemented by the PyRosetta toolkit [13] as the fitness function. We also have included the term of Solvent Accessible Surface Area (SASA) from the PyRosetta, using 1.4Å as atomic radius, to help on the packing of the 3-D structures. To improve the formation of correct secondary structures, we also included a secondary structure term (Eq. 1). The procedure consists of giving a positive reinforcement, adding a negative constant $(-const)$ to the summation of all amino acids of the protein $(Pset)$. This procedure occurs when the corresponding secondary structure (zp_i) of the $i-th$ amino acid (aa_i), is equal to the secondary structure (zi_i) of the same residue provided as input to the algorithm. On the other hand, the technique gives a negative reinforcement to the summation, adding a positive constant $(+const)$, when the secondary structure of the corresponding amino acid residues are not equal. All amino acids of the protein are comparable during the evaluation of the conformation. The DSSP [14] algorithm was used to assign the secondary structures. This strategy is similar to the presented by Correa et. al. [41].

$$SS_{term} = \sum_{aa \in Pset}^{i+1} V_{aa,zp,zi}(aa_i, zp_i, zi_i) \quad (1)$$

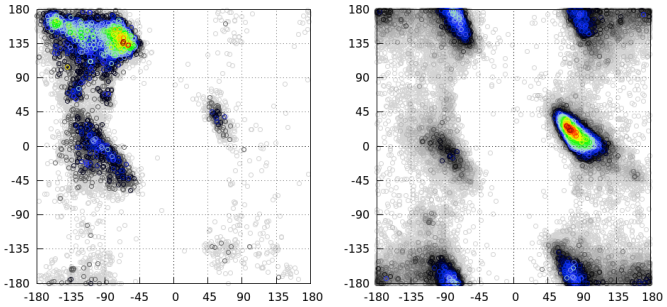
$$V_{aa,zp,zi}(aa, zp, zi) = \begin{cases} \textit{-const}, & zp = zi \\ \textit{+const}, & zp \neq zi \end{cases} \quad (2)$$

Both terms described above are added to the result of the PyRosetta energy function, forming the final scoring function (E_{final}) adopted in this work (Eq. 3).

$$E_{final} = E_{pyrosetta} + SASA_{term} + SS_{term} \quad (3)$$

C. Angle Probability List

Our system takes advantages of the experimental knowledge stored in the PDB. The main benefit of incorporating this information to the proposed method is to reduce the protein search space. We have incorporated the *Angle Probability List* (APL - http://sbcb.inf.ufrgs.br/apl) approach proposed and developed by Borguesan et al. [15]. Such technique analyzes the conformational preferences of amino acids in proteins according to its secondary structure. To incorporate the structural information of protein templates, they have build a histogram $\mathtt{H}_{aa,z}$ of $[-180, 180] \times [-180, 180]$ cells for each amino acid residue (aa) and secondary structure (z) [15]. Each cell (i,j) has the number of times that a given amino acid aa in secondary structure z has a pair of torsion angles ($\mathtt{i} \leq \phi < \mathtt{i+1}$, $\mathtt{j} \leq \psi < \mathtt{j+1}$). For each amino acid residue and secondary structure they have computed the torsion *Angle Probability List* $\mathtt{APL}_{aa,z}$ that represents the normalized frequency of each square. Meanwhile the initialization of solutions, our multi-agent system uses this information to initialize the solutions of each agent in an attempt to reduce the size of the search space and inject high-quality solutions as a starting point. The algorithm uses a weighted random selection to get the angle values from the APL histograms to give more chance to the cells of the histograms which have a higher relative frequency of occurrence, i.e., the most abundant conformational regions of the Ramachandran illustrated in Fig. 2. Section IV describes how this procedure works.



(a) Phe - Coil (b) Gly - Coil

Fig. 2: Angle Probability List for the amino acids Phe and Gly when present in coil secondary structures . APLs were prepared with NIAS Server [40].

D. Multi-Agent Systems

Multi-agent systems (MAS) are being used to face complex problems, devising tasks among the agents of the system and exploring a more distributed approach. An agent is an independent computer process that runs in an independent way, where under some given circumstances, interacts and cooperates with the other agents to solve a major problem [16].

The main particularities of simple agents are the capability of reacting or respond to the environment and change it based on their perceptions. The pro-activity that allow the agent not simply respond to the environment but also follows its own goals when seems an appropriate situation and the

interactivity with the other entities that incorporates the social feeling to the agents allows to solve complex problems [17], [18]. The agent-based paradigm has been shown a useful technique for the applications that have repetitive and time-consuming activities, knowledge share and management, such as integration of different knowledge sources and modeling of complex systems [19]. Thus, a MAS presents a suitable strategy to model and tackle the complex PSP problem [42]. For example, the MAS for the PSP problem can focus on first principle methods, and the agents can divide tasks and goals, interact and compete in an attempt to explore the search space in a more efficient way [43]. Or it can be modeled as a framework integrating different prediction methods or considering different biological datasets [29], [44]. Section III presents and discusses the most important aspects of the use multi-agent systems in the PSP problem.

III. RELATED WORK

Focusing on the critical aspects of biological systems, concerning to structures, activities, and interactions, the use of agent-based strategies can be modeled as abstractions of this, which are kept alive during the whole process from design to simulation. These concepts allow MAS become suitable to simulate biological systems that can be decomposed in several independent but interacting entities, each one represented by an agent. In the literature, we observed that there is a lack of published works related to multi-agent systems applied to the 3-D PSP problem, principally when restricting the search to methods based on concepts of the first principle methods. It is possible to cite the work of Campeotto et al. [20], where a multi-agent system was developed aiming to predict tertiary structures of proteins in a faster platform using a GPU. They have created a set of agents with distinct functions to explore the conformations of different parts of protein. Thus, the protein is broken into smaller parts and distributed between the responsible agents that try to find a good fold for that parts, and at the end of the process, all the best parts are grouped again to create the entire protein. In our approach, the protein is not divided between agents. Each agent work with the complete protein (torsional angles) representation, called solutions, like in an evolutionary algorithm. We have chosen that option, because of the influence caused by the neighbor's residues in the torsional angles values assumed by an arbitrary amino acid residue [21]. Another multi-agent approach was proposed by Lipinski-Paes and Norberto De Souza [22] and was called as MASTERS Framework. This framework uses a different protein representation, which is the AB lattice model [23]. Such model consists in a highly simplified model for protein-folding phenomena and is based only on the interactions between hydrophobic and polar amino acid residues. Lipinski-Paes and Norberto De Souza also allow the use of different energy functions, but these fitness functions are simpler and designed specially to this type of representation. Pérez et al. [24] proposed an iterative multi-agent architecture to be used as a virtual laboratory to explore minimalist models of protein folding. Those models were described by the HP model and coarse-grained 2D-square lattice representation [25]. An interesting thing is that the communication between agents is made through the blackboard technique, providing both data sharing and coordination artifacts. In Muscalagiu et al. [26], [27], was proposed an agent-based framework for the PSP problem, composed of autonomous agents which collaborate to find good conformations. They have employed the Distributed Constraint Programming strategy [28]. On this system, each amino acid residue is viewed as an autonomous agent that communicates with others by transmitting messages. The input protein is represented by lattice models with distributed constraints. The frameworks proposed by Bates et al. [29], Garro et al. [30] and Jin and Kim [31], are not based on first principles methods, but are systems developed to predict the protein folding using and integrating the results of different existing predictors based on comparative modeling and fold recognition techniques, that are out of the scope of this work.

IV. THE PROPOSED METHOD

To deal with the 3-D PSP problem, we designed a multi-agent system that incorporates the knowledge extracted from experimental structures stored in the PDB, and concepts of population-based evolutionary algorithms [32], [33]. The system also uses a Simulated Annealing [34] implementation as local search operator to eliminate stereochemical clashes that arise due to the unnatural overlap of any two non-bonding atoms [2]. In our approach, as illustrated in Figure 3, we have designed different types of agents, where each one performs specific roles and interacts in an attempt to obtain good solutions to the problem. The method uses the information stored in the PDB through the *Angle Probability List* (APL), already explained in a previous section, on the initialization of the solutions to reduce the size of the search space and use high-quality solutions as starting point to find an approximation to the native-like 3-D structure of a target sequence, and also to diversify the population of solutions during the execution of the method. The developed MAS was structured into two main parts called as *Initialization step* and *Optimization step*. These steps will be described in the sections below.

A. Implementation

The proposed multi-agent system was developed in the SPADE (Smart Python multi-Agent Development Environment) platform [35], which is a multi-agent platform coded in Python that provides an easy and fast development of agent-based applications. SPADE also explores a new communication model, based on the Jabber protocol (http://www.jabber.org), and presents a simple interface to create agents using this communication concept. In our system, the communication between agents was done through explicit messages, using the classical commands of *send* and *receive*. All simulations of the system were performed in a local network. Thus we do not have considered fails in the communications, neither lost messages or duplicated ones. The platform works according to the FIPA standard proposal [36] for a multi-agent platform. The communication between agents in the system is made

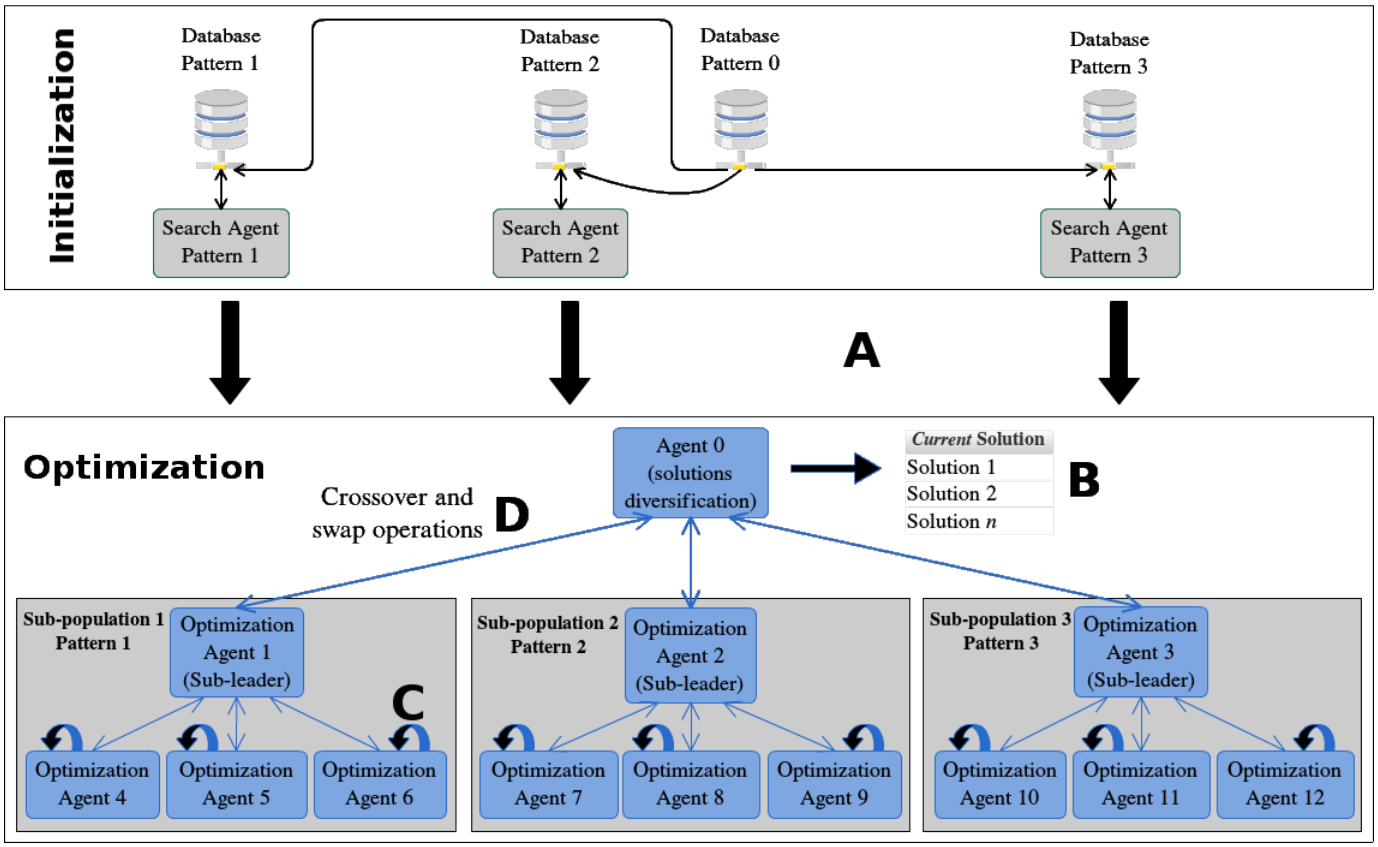


Fig. 3: Schematic representation of the structure, actions and interactions of our multi-agent system.

through the FIPA-ACL protocol, which provides an organized and effective agent communication.

B. Solution Representation

Each residue of the protein is represented by a set of six values: two for ϕ and ψ angles and four for the χ angles (not all the amino acids have the four side-chain angles). The side-chain angles that do not exist in nature were set with a null value. Based on that, a solution for a peptide of n residues is represented by a vector of real values of size $n \times 6$.

C. Initialization Step

The APLs (example in Figure 2) of a target sequence were partitioned into sub-groups, according to the combination of amino acid residues and their respective secondary structures, resulting in different patterns (*Database Patterns* - Fig. 3). The pattern 1 considers a consecutive sequence composed of 3 amino acids to determine the torsion angles of the amino acid located in the middle of the sequence. The pattern 2 represents a consecutive sequence of 2 amino acids where the goal is to determine the torsional angles of the amino acid at the left of the sequence, and the pattern 3 is used to set the torsion angles of the amino acid at the right of the sequence also considering a sequence of 2 residues. The division of data was done to increase the diversity in the space of solutions. Hence, for each database pattern we linked a *Search agent* which is responsible for creating good initial solutions from its respective database, and send these to the *Optimization Agents* (OA) when requested, as represented (Fig. 3 - A).

- *Search agent:* The search agents have two main tasks, and their behavior is quite simple: (*i*) initialize the solutions that will be optimized by the OA in the Optimization step; and (*ii*) keep trying to create better individuals, while waiting for requests from the Optimization Agents.

D. Optimization Step

- *Optimization agent:* this kind of agent aims to optimize the initial solutions received from the Search Agents. We have incorporated, as behaviors of the Optimization Agents, concepts of evolutionary-based algorithms, such as crossover using the roulette wheel selection method and swap operators, and a Simulated Annealing implementation as a local search method. The interactions among the agents through global search operators lead to evolution and progressive improvements of the solutions. All communications are realized by the change of ACL messages between the involved agents. Each agent has its own cycles or generations.

We have employed a population of thirteen OA that were organized in a hierarchical ternary tree, as it is shown in Figure 3. Each OA maintains a set of thirty solutions where one of them are called current solution, and the others are the

pockets (Fig. 3 - B). The population is also organized in four overlapped subpopulations composed of three supporters and one leader agent. An agent can only interacts with the leader agent of the subpopulation that it belongs to. It can see only one database pattern of the APL and communicate just with the corresponding Search Agent of the pattern. The OA have some defined tasks to be done:

- On the initialization, each agent requests solutions to the corresponding Search Agent until fill in the pockets;
- In each generation, the agent makes an *inner* crossover operation with the solutions in its respective pockets. The offspring of the inner crossover is also stored in the current solution of the agent (Fig. 3 - C);
- Every ten generations, the leader agent of a subpopulation makes crossover with agents located in the lower level. The offspring resultant of the crossover operation is stored in the current solution of the lower level agent (Fig. 3 - D);
- Due the roughness of the protein search space, we proposed a local search procedure to speed up the search by improving candidate solutions locally. This strategy explores the neighbors of a solution aiming at finding a solution that is better than the current one. Every five hundred cycles, the Simulated Annealing method is applied to do small fixes in the current solution;
- Each agent keeps the pockets always sorted;
- The agent updates the population in each generation. This task can be divided into small steps. First, the current solution is stored in one of the pockets if it is better than one that is already stored. Second, if the agent is in the lower level of a subpopulation, then it sends the best solution to the leader agent (swap operation). Therefore, the best solutions are kept on the top of the hierarchy in the *Agent 0* pockets, diversifying the solutions from different sub-population patterns (Fig. 3 - D);
- At the end of each cycle, each agent discards the worst solution stored in the pockets and request a new one to the respective Search Agent to avoid premature convergence and escape from local minimal.

V. EXPERIMENT AND RESULTS

A. Target proteins and experiments

The amino acid sequences of eight proteins were obtained from the PDB and used as case studies in our experiments. Table I presents details of the target proteins. The target protein sequences were submitted to the proposed MAS in order to predict their 3-D structures. Structural analysis of the predicted structures are presented in Section V-B. We analyse the root-mean-square deviation (RMSD, minimization measure) [37] and the global distance total score test (GDT_TS, maximization measure) [37] of the predicted 3-D structures when compared with their corresponding experimentally determined structures and stereochemical qualities of the secondary structures of the predicted 3-D protein structures. The proposed method was ran eight times per six hours.

TABLE I: Target amino acid sequences.

Protein	Size	Description
1AB1 (Fig.4a-red)	46	One sheet/Two helices
1ACW (Fig.4b-red)	29	One sheet/One helix
1L2Y (Fig.4c-red)	20	Two helices
1DFN (Fig.4d-red)	30	One sheet
2P81 (Fig.4e-red)	44	Two helices
1K43 (Fig.4f-red)	14	One sheet
2MR9 (Fig.4g-red)	44	Three helices
1WQC (Fig.4h-red)	26	Two helices

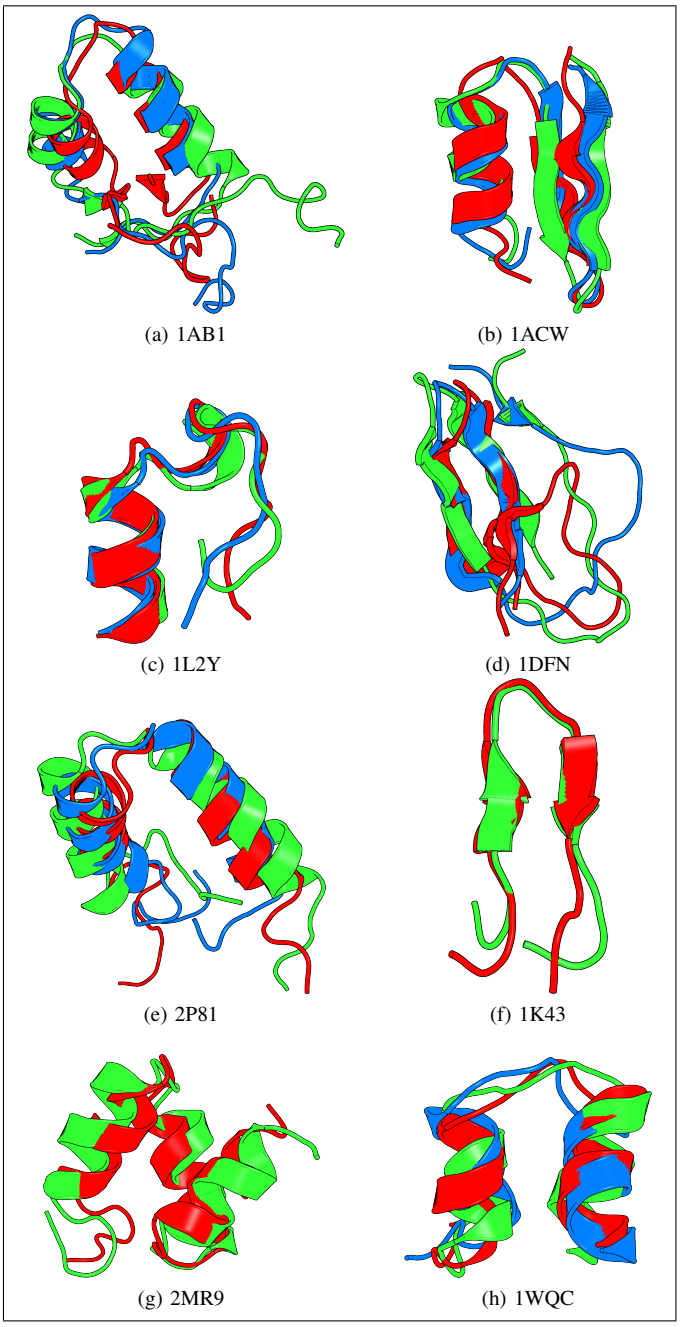


Fig. 4: Cartoon representation of the experimental (red), lowest RMSD (blue) and lowest energy (green) of the predicted structures. Graphic representation was prepared with PYMOL [39].

TABLE II: Algorithm simulation results.

PDB ID	Low. Energy Kcal/mol	RMSD (Å)	GDT_TS	Lowest RMSD (Å)	Energy Kcal/mol	GDT_TS	Avg. Energy Kcal/mol	Avg. RMSD (Å)	Avg. GDT_TS
1AB1	-17,520.18	6.06	46.74	3.84	-17,177.39	55.98	-14,725.32 (±2,679.50)	7.63 (±2.63)	49.70 (±8.25)
1ACW	-15,519.24	2.77	68.10	1.48	-14,636.67	79.31	-12,357.97 (±6,832.25)	2.53 (±0.50)	70.26 (±5.47)
1L2Y	-8,526.28	1.86	76.25	1.62	-8,385.99	81.25	-8,433.43 (±59.66)	1.80 (±0.13)	80.25 (±2.40)
1DFN	-10,769.66	5.02	45.00	4.69	-10,654.82	45.83	-7,255.80 (±2,305.63)	5.37 (±0.52)	42.81 (±3.59)
2P81	-23,103.27	5.37	37.50	3.26	-23,095.96	38.64	-22,999.13 (±91.96)	5.12 (±1.07)	36.27 (±1.70)
1K43	-4,614.26	0.56	83.93	0.56	-4,614.26	83.93	-4,559.54 (±40.87)	0.77 (±0.15)	85.20 (±2.86)
2MR9	-26,986.69	1.90	74.43	1.90	-26,986.69	74.43	-26,889.63 (±91.09)	3.35 (±1.79)	65.18 (±7.82)
1WQC	-13,971.78	2.61	73.08	2.21	-13,882.32	73.08	-13,848.54 (±68.87)	3.33 (±1.11)	66.83 (±7.07)

TABLE III: Structural analysis for the lowest energy solution using the Q-Index measure. (P/E) represents the secondary structure of the predicted (P) and the experimental (E).

PDB	% Q_H(P/E)	% Q_E(P/E)	% Q_T(P/E)	% Q_C(P/E)	% Q4
1AB1	100.0%(20/20)	100.0%(4/4)	80.0%(4/5)	76.4%(13/17)	89.1%
1ACW	100.0%(9/9)	100.0%(10/10)	40.0%(2/5)	20.0%(1/5)	75.8%
1L2Y	100.0%(12/12)	–	–	87.5%(7/8)	95.0%
1DFN	–	87.5%(14/16)	55.5%(5/9)	80.0%(4/5)	76.6%
2P81	100.0%(27/27)	–	60.0%(3/5)	91.6%(11/12)	93.1%
1K43	–	100.0%(6/6)	40.0%(2/5)	100.0%(3/3)	78.5%
2MR9	100.0%(30/30)	–	66.6%(6/9)	60.0%(3/5)	88.6%
1WQC	100.0%(18/18)	–	–	100.0%(8/8)	100.0%
Average	100.0%(116/116)	94.4%(34/36)	57.8%(22/38)	82.6%(50/63)	89.1%

B. Structural analysis

For each target protein, we present biochemical and structural analysis over the solution with the minimum energy function value among the eight runs performed. The quality of the predicted structures were evaluated by similarity comparisons with the structures of the experimental proteins obtained from the PDB. Quality measurements have been made in terms of RMSD and GDT_TS. To compute these measures we have considered only the C_α atoms of the protein structures. Table II shows the achieved results of the eight runs of the MAS on each sequence. The method can reach low energy values while at the same time reaches good solutions in terms of RMSD and GDT_TS. Protein molecules are not static but are in varying degrees of motion. According to the results in Figure 4, our proposal can predict structures with similar fold organization when compared with the experimental ones. This can be also observed when the RMSD values in Table II (column 5) are evaluated.

We run STRIDE [38] to analyze the hydrogen bonds that define the secondary structure (SS) of the predicted structures. We calculate the percentage of correctly classified SS of residues using Q-index (more details can be found in [15]). The SS states from STRIDE was reduced to four states using the following schema: H and G to H; E, B to E; T; and all other states to C. We compare the SS contents of the predicted structures against the secondary structure of the experimentally determined structures. As can be observed in Table III, the SS of the predicted and experimental structures are comparable in terms of SS content. The proposed method achieved an accuracy (Q_4) of about 89%. When we compare the topology (Fig. 4) of the predicted structures against the experimental 3-D ones we observe that the topologies are similar in terms of secondary structure.

VI. CONCLUSIONS AND FURTHER WORK

Actually, there is an increasing need for new computational strategies that make use of template information from experimentally determined protein structures and their use to predict the unknown 3-D structure of proteins. In this paper, we proposed a multi-agent system that incorporates in the search process the information extracted from the PDB and concepts of population-based evolutionary algorithms for the PSP problem. As corroborated by experiments, the results show the proposed approach has a promising ability to predict good three-dimensional conformations in terms of potential energy, RMSD and GDT_TS measurements when compared with the experimental protein structures. Considering protein molecules are not static but are in varying degrees of motion, our proposal could predict structures with similar fold organization when compared with the experimental ones.

There are several research opportunities to be explored in this field, with relevance multidisciplinary applications in Computer Science and Bioinformatics. For instance, one could test the viability to expand the system to a real distributed scenario, testing different communication protocols and their impacts on the system, such as time performance, robustness and reliability. The hierarchical structure of the Optimization step could also be tested about the number of Optimization Agents employed and the applied metaheuristic. It is still possible to test the proposed method with longer protein sequences with more complex folding patterns.

ACKNOWLEDGMENT

This work was partially supported by grants from FAPERGS (002021-25.51/13, PRONUPEQ-16/2551-0000520-6), MCT/CNPq (311022/2015-4), Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES/SticAmSud - 88881.117607/2016-01), Brazil.

This Research is supported by Microsoft under a Microsoft Azure for Research Award. MIP thanks CEBIB and DICYT project 061619IP, VRIDEI-USACH.

REFERENCES

[1] M. Dorn, M. Barbachan e Silva, L.S. Buriol, L.C. Lamb, Three-dimensional protein structure prediction: Methods and computational strategies. Comp. Biol. and Chemistry, vol. 53, 2014, pp. 251-276.

[2] A. Lesk. Introduction to bioinformatics. Oxford University Press, 2013.

[3] C.B. Anfinsen, Principles that govern the folding of protein chains. Science, vol. 182, n. 96, 1973, pp. 223-230.

[4] A. Tramontano, Protein structure prediction. Belmont, Weinheim: John Wiley and Sons, 2006, pp. 208.

[5] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis, On the complexity of protein folding. J. Comput. Biol., v. 5, n. 3, p. 423-465, 1998.

[6] D.J. Osguthorpe, Ab initio protein folding. Curr. Opin. Struct. Biol., vol. 10, n. 2, 2000, pp. 146-152.

[7] C.A. Rohl, C.E. Strauss K.M.S. Misura, D. Baker, Protein structure prediction using Rosetta. Methods Enzymol., vol. 383, n. 2, 2004, pp. 66-93.

[8] J.U. Bowie, R. Luthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure. Science, vol. 253, n. 5016, 1991, pp. 164-170.

[9] M.A. Martí-Renom, A. Stuart, A. Fiser, A. Sanchez, F. Mello, A. Sali, Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct., vol. 29, n. 16, 2000, pp. 291-325.

[10] A. Tramontano, A.M. Lesk, Protein structure prediction. John Wiley and Sons, Inc, Weinheim, 2006.

[11] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bath, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank. Nucleic Acids Res., vol. 28, n. 1, 2000, pp. 235-242.

[12] J.R. Desjarlais; N.D. Clarke, Computer search algorithms in protein modification and design. Curr. Opin. Struct. Biol., v. 8, n. 4, p. 471-475, 1998.

[13] S. Chaudhury, S. Lyskov, J.J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics, v. 26, n. 5, p. 689-691, 2010.

[14] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, v. 22, n. 12, p. 2577-2637, 1983.

[15] B. Borguesan, M. Barbachan e Silva, B. Grisci, M. Inostroza-Ponta, M. Dorn, APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. Comput. Biol. Chem., 2015.

[16] M. Wooldridge, An introduction to multiagent systems. John Wiley & Sons, 2009.

[17] N.R. Jennings, K. Sycara, M. Wooldridge, A roadmap of agent research and development.Auton. Agent. Multi Agent Syst., v. 1, n. 1, p. 7-38, 1998.

[18] M. Wooldridge, N.R. Jennings, Intelligent agents: Theory and practice. Knowl. Eng. Rev., v. 10, n. 02, p. 115-152, 1995.

[19] E. Merelli, et al., Agents in bioinformatics, computational and systems biology. Brief. Bioinform., v. 8, n. 1, p. 45-59, 2007.

[20] F. Campeotto, A. Dovier, E. Pontelli, Protein structure prediction on GPU: a declarative approach in a multi-agent framework. Parallel Processing (ICPP), 2013 42nd International Conference on. IEEE, 2013. p. 474-479.

[21] E.A. Kabat, T. Wu, The influence of nearest-neighboring amino acid residues on aspects of secondary structure of proteins. Attempts to locate α-helices and β-sheets. Biopolymers, v. 12, n. 4, p. 751-774, 1973.

[22] T. Lipinski-Paes, O. Norberto De Souza, MASTERS: A general sequence-based MultiAgent System for protein TERtiary Structure prediction. Electron. Notes Theor. Comput. Sci., v. 306, p. 45-59, 2014.

[23] F.H. Stillinger, T. Head-Gordon, C.L. Hirshfeld, Toy model for protein folding. Phys. Rev. E, v. 48, n. 2, p. 1469, 1993.

[24] P.P.G. Prez, H.I. Beltrn, A. Rojo-Domnguez, M. Eduardo, S. Gutirrez, Multi-agent systems applied in the modeling and simulation of biological problems: A case study in protein folding. World Acad. Sci. Eng. Technol., v. 58, p. 128, 2009.

[25] K.A. Dill, S. Bromberg, K. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, H.S. Chan, Principles of protein folding–a perspective from simple exact models. Protein Sci., v. 4, n. 4, p. 561, 1995.

[26] I. Muscalagiu, A. Iordan, M. Osaci, M. Pnoiu, Modeling and simulation of the protein folding problem in DisCSP-Netlogo. Global Journal on Technology, v. 2, 2012.

[27] I. Muscalagiu, H.E. Popa, M. Panoiu, V. Negru, Multi-agent systems applied in the modelling and simulation of the protein folding problem using distributed constraints. Multiagent System Technologies. Springer Berlin Heidelberg, 2013. p. 346-360.

[28] M. Yokoo, E.H. Durfee, T. Ishida, K. Kuwabara, The distributed constraint satisfaction problem: Formalization and algorithms. IEEE Trans. Knowl. Data Eng., v. 10, n. 5, p. 673-685, 1998.

[29] P.A. Bates, L.A. Kelley, R.M. MacCallum, M.J. Sternberg, Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins: Struct., Funct., Bioinf., v. 45, n. S5, p. 39-46, 2001.

[30] A. Garro, G. Terracina, D. Ursino, A multi-agent system for supporting the prediction of protein structures. Integr. Comput-Aid. E. , v. 11, n. 3, p. 259-280, 2004.

[31] H. Jin, I.C. Kim, Plan-Based coordination of a multi-agent system for protein structure prediction. In: Artificial Intelligence and Simulation. Springer Berlin Heidelberg, 2005. p. 224-232.

[32] J. Dro, A. Petrowski, P. Siarry, E. Taillard, Metaheuristics for hard optimization: methods and case studies. Springer Science & Business Media, 2006.

[33] S. Luke, Essentials of metaheuristics. Lulu, ed. 1, 2009, pp. 227.

[34] S. Kirkpatrick, Optimization by simulated annealing: Quantitative studies.J. Stat. Phys., v. 34, n. 5-6, p. 975-986, 1984.

[35] M.E. Gregori, J.P. Camara, G.A. Bada, A jabber-based multi-agent system platform. Proceed. of the fifth international joint conference on Autonomous agents and multiagent systems. ACM, 2006. p. 1282-1284.

[36] FIPA. Abstract architecture specification. TR SC00001L, 2002.

[37] Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. Proteins: Struct., Funct., Bioinf., v. 57, n. 4, p. 702-710, 2004.

[38] M. Heinig, D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res., v. 32, n. suppl 2, p. W500-W502, 2004.

[39] L.L.C Schrödinger , The PyMOL Molecular Graphics System, Version 1.3r1. August, 2010.

[40] B. Borguesan, M. Inostroza-Ponta, M. Dorn, NIAS-Server: Neighbors Influence of Amino acids and Secondary Structures in Proteins. Journal of Computational Biology. August 2016, ahead of print.

[41] L. L. Correa, B. Borguesan, C. Farfan, M. Inostroza-Ponta, M. Dorn. A Memetic Algorithm for 3-D Protein Structure Prediction Problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics, v. PP, n 99, p. 1-17, 2016.

[42] L. L. Correa, M. Dorn. Multi-Agent Systems in Three-Dimensional Protein Structure Prediction. In: D. F. Adamatti. Multi-Agent-Based Simulations Applied to Biological and Environmental Systems. 1ed. Hershey: IGI Global, 2016, v.1 p.241-278

[43] I. Muscalagiu, H.E. Popa, M. Panoiu, V. Negru, (2013). Multi-agent systems applied in the modelling and simulation of the protein folding problem using distributed constraints. In Multiagent system technologies (pp. 346-360). Springer.

[44] A. Garro, G. Terracina, D. Ursino, D. A multi-agent system for supporting the prediction of protein structures. Integr. Comput. Aid. E., 11 (3), 259-280, 2004.