2013 IEEE Congress on Evolutionary Computation
June 20-23, Cancún, México

A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides

Márcio Dorn
Institute of Informatics
UFRGS
Porto Alegre, Brazil
mdorn@inf.ufrgs.br

Mario Inostroza-Ponta
Departamento de Ingeniería Informática
Universidad de Santiago
Santiago, Chile
mario.inostroza@usach.cl

Luciana S. Buriol
Institute of Informatics
UFRGS
Porto Alegre, Brazil
buriol@inf.ufrgs.br

Hugo Verli
Center for Biotechnology
UFRGS
Porto Alegre, Brazil
hverli@cbiot.ufrgs.br

***Abstract*—Three-dimensional (3-D) protein structure determination has become an important area of research in structural bioinformatics. Proteins are responsible for the execution of different functions in the cell. Understanding the 3-D structure provides important information about the protein function. Many computational methodologies for the protein structure prediction were developed along the last 20 years, but the problem still challenges researchers because the complexity and high dimensionality of its large search space. In this article we present a strategy for reducing the search space explored by heuristic methods for solving the problem taken into consideration previous occurrences of amino acid residues in a well known protein database (PDB). We propose a genetic algorithm that takes advantages of this kind of information, reducing considerable the search space, allowing the algorithm to save time with less promising solutions. A simple Local Search operator helps the GA to intensify the search of the 3-D protein conformational space. We demonstrate the effectiveness of the strategy with a set of experimental results.**

I. Introduction

Proteins or polypeptides are long sequences of 20 different amino acid residues that in physiological conditions adopt a unique 3-D structure [31]. This structure gives us important information about the function of the protein in the cell. Predicting the correct 3-D structure of a protein molecule is an intricate and arduous task. The Protein Structure Prediction (PSP) and Protein Folding (PF) problems[1] are classified in computational complexity theory as NP-complete problems [11], [20], [21], [33], [38], i.e, they are among the hardest problems in terms of computational requirements. This complexity is due to the folding process of a protein being highly selective. A long amino acid chain ends up in one out of a huge number of 3-D conformations. In contrast, the conformational preferences of a single amino acid residues is weak. Thus, the high selectivity of protein folding is only possible through the interaction of many residues. Therefore, non-local interactions play an important role in protein three-dimensional structure, as local sequence-structure relationships are not absolute [41].

The prediction of the 3-D structure of polypeptides based only on the amino acid sequence (primary structure) is a problem that has, over the last 40 years, challenged computer scientists, biochemists, mathematicians and biologists [2]. The 3-D PSP problem [10] is one of the main research problems in Structural Bioinformatics. The main challenge is to understand how the information encoded in the linear sequence of amino acid residues is translated into the 3-D structure, and from this acquired knowledge, to develop computational methodologies that can correctly predict the native structure of a protein molecule.

Many methods and algorithms have been proposed, tested and analyzed over the years as a solution to this complex problem, see e.g [8], [12], [22], [25], [26], [29], [36], [39], [43], [46]–[49], [51]–[55]. In the literature, one can find several classifications of the 3-D protein structure prediction methods. Floudas et al. [19] classifies the computational methods for protein structure prediction into four groups: (1) first principle (*ab initio*) methods without database information, (2) first principle methods with database information, (3) comparative homology, and (4) fold recognition. *Ab initio* methods (first principle methods without database information) can obtain novel and unknown protein folds. Nevertheless, the complexity and the high dimensionality of the search space [38] even for a small protein molecule makes the problem intractable [33]. The direct simulation of protein folding in atomic details, as used in Molecular Dynamics (MD)[2], is not tractable [50], for large proteins of medical and scientific interest, due to high computational costs, despite the efforts towards the development of distributed computing platforms. On the other hand, comparative homology modelling does not present such problems; however, it can only predict structures of protein sequences which are similar or nearly identical to other sequences of known structures. Fold recognition *via* threading, in turn, is limited to the fold library derived from the Protein Data Bank (PDB) structures [3].

Considering the computational complexity of the PSP problem, current 3-D protein structure prediction methods make use of a wide range of optimization algorithms [28]. Metaheuristics are used in order to provide near optimal solutions. In addition, considering the limitations of the four classes of protein structure prediction methods, researchers have recently

[1]Protein folding is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil.

[2]MD is a simulation procedure in which the protein system is placed into a random conformation and then the system reacts to force atoms to exert on each other. The model assumes that, as a result of these forces, atoms move in a Newtonian manner. The trajectory of the system should lead to the native conformation.

978-1-4799-0454-9/13/$31.00 ©2013 IEEE

developed hybrid methods which combine principles of these four classes. For example, the accuracy presented by homology modelling methods is combined with the capacity of *ab initio* methods in predicting novel folds [13], [18]. In order to reduce the complexity and the high dimensionality of the conformational search space inherent to *ab initio* methods, information about structural motifs found in known protein structures can be used to construct approximate conformations. These approximate conformations are expected to be sufficient to allow latter refinements by means of Molecular Mechanics (MM) such as MD simulation [50]. In a refinement step, global interactions between all atoms in the molecule (including e.g. non-bond interactions) are evaluated and deviations in the polypeptide main-chain and side-chain torsion angles can be corrected [18]. These in turn reduce the total time spent by *ab initio* methods - which usually start from a fully extended conformation of a polypeptide - to fold a sequence of unknown structure [7]. The first principle methods that make use of database information cover this class of methods. Such methods use previous protein structural information from existing databases in order to construct starting point protein structures.

In this work we propose a GA that considers the conformation of amino acid residues [23] obtained from experimental-determined 3-D protein structures. This information was gathered from the PDB and used to reduce the conformational search space of the target protein sequences. A simple Local Search operator was developed in order to help the GA to intensify the exploration of certain regions of the 3-D protein conformational space. The article is organized as follows. In Section II some basic concepts related to the PSP problem are introduced. Next, in Section III the strategy proposed in this paper is presented and applied in a genetic algorithm. Experimental results are reported an analysed in Section IV. Finally, the conclusions of this work are reported.

II. Preliminaries

A. Protein and Structure representation

In nature there are 20 distinct amino acid residues, each one with its own chemical properties (including size, charge, polarity, hydrophobicity, or the tendency to avoid water packing) [35]. Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic character. The importance of the physical properties of the side chains comes from the influence they have on the amino acid residues interactions in the structure. The distribution of the hydrophilic and hydrophobic amino acids are important to determine the tertiary structure of the polypeptide. A peptide is a molecule composed of two or more amino acid residues chained by a chemical bond called the peptide bond. This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule (H_2O). Two or more linked amino acid residues are referred to as a peptide, and larger peptides are generally referred to as polypeptides or proteins [10], [32]. In a peptide or polypeptide all atoms from the group R are referred to as side-chain and the remaining atoms are referred to as the protein backbone. The specific characteristics of the peptide bond have important implications for the 3-D fold that can be adopted by proteins.

The peptide bond (C-N) has a double bond and is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds N-C_α and C_α-C. These bonds are known as Phi (ϕ) and Psi (ψ) angles and are free to rotate [32], [35]. This freedom is mostly responsible for the conformation adopted by the protein backbone. However, the rotational freedom around the ϕ (N-C_α) and ψ (C_α-C) angles is limited by steric hindrance between the side-chain of the amino acid residue and the protein backbone [6], [32], [45]. As a consequence, the possible conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties.The peptide bond itself tends to be planar, with two allowed states: *trans*, $\omega \simeq 180°$(usually) and *cis*, $\omega \simeq 0°$(rarely) [6], [32]. The sequence of ϕ, ψ and ω angles of all residues in a protein defines the backbone conformation or fold [23]. The angles ϕ and ψ can have any value between $-180°$ and $+180°$. However, some combinations are prohibited by steric interferences between atoms from the main-chain and atoms from the side-chain (two atoms cannot occupy the same space) [23]. The allowed and prohibited values for the torsion angles ϕ and ψ are graphically demonstrated by the map of Sasisekharan-Ramakrishnan-Ramachandran, or simply Ramachandran map [42]. In this work we represent the 3-D structure of a protein using only the main-chain and side-chain torsion angles.

B. Energy function

An energy function describes the internal energy of the protein and its interactions with the environment in which it is inserted. In Protein Structure Prediction the goal is to find a 3-D structure with the global minimum of free energy that corresponds to the native or functional state of the protein [39], [49]. The energy function used for the PSP problem contain elements accounting for van der Waals force, electrostatics, solvation, and hydrogen bonding. A potential energy function incorporates two main 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 terms also include a torsional potential (torsion) that models the periodic energy barriers encountered during bond rotation. The non-bonded potential includes: ionic bonds, hydrophobic interactions, hydrogen bonds, van der Waals forces, and dipole-dipole bonds. van der Waals force is usually described by the equation for Lennard-Jones 6-12 potential [5]. There is a variety of potential energy functions used in protein structure prediction. In this work we use the AMBER potential energy function [9] (Eq. 1).

$$
\begin{aligned}
E_{\text{total}} &= \sum_{\text{bonds}} \frac{1}{2} K_b (b - b_0)^2 \\
&+ \sum_{\text{angles}} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \sum_{\text{torsions}} \frac{1}{2} K_\eta (1 + \cos(\eta_\omega - \gamma)) \\
&+ \sum_{\text{j=1}}^{N-1} \sum_{\text{i=j+1}}^{N-1} \left\{ \epsilon_{i,j} \left[\left(\frac{R_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}
\end{aligned}
\tag{1}
$$

where:
Bonds: represent the energy between covalently bonded atoms.

Angles: represent the energy due to the geometry of electron orbitals involved in covalent bonding.
Torsions: represent the energy for twisting a bond due to bond order (e.g. double bonds) and neighboring bonds or lone pairs of electrons.
The last term represents the non-bonded energy between all atom pairs, which can be decomposed into van der Waals and electrostatic energies.

III. Search strategy

The algorithm proposed in this work takes advantage of the a-priori knowledge that is stored in the PDB database [3] for each of the amino acids. In theory the pair of angles (ϕ, ψ) can take any real value between -180°and 180°, which builds a very large search space. However, if we plot the information stored in the PDB database we can see that different amino acids have different distribution of angles. In Figure 1 we show the distribution of angles for amino acids Proline (Fig. 1a) and Glycine (Fig. 1b). It is possible to observe that the torsion angles of amino acid residues are concentrated in some regions. This information can be taken into account when exploring the search space, avoiding evaluating solutions with angles within improbably ranges.

In order to take advantage of this knowledge we follow the next procedure: first, we build an histogram matrix H_a of $361x361$ cells for each amino acid residue. Each cell (i, j) has the number of times that a given amino acid residue a has a pair of main-chain torsion angles ($i \leq \phi < i+1$, $i \leq \psi < i+1$). In order to increase the representation of more dense regions of the torsion angles pairs (ϕ and ψ), the values of all eight adjacent neighbours cells of each pair (i, j) is also taken into consideration (2). Then, for each amino acid residue we compute the torsion angles probability matrix AP_a using Function 3.

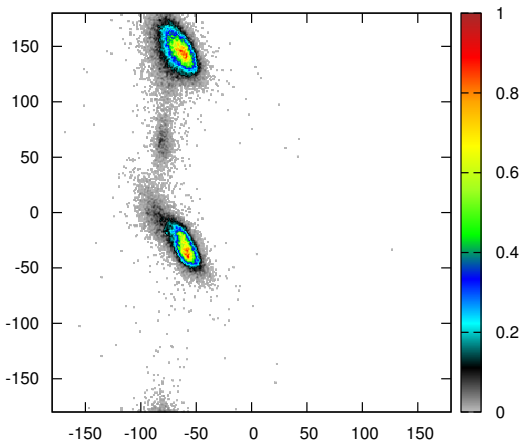
$$H'_a(i,j) = \sum_{r=i-1}^{i+1} \sum_{s=j-1}^{j+1} H_a(r,s) \tag{2}$$

$$AP_a(i,j) = \frac{H'_a(i,j)}{\sum_{\forall x,y} H'_a(x,y)} \tag{3}$$

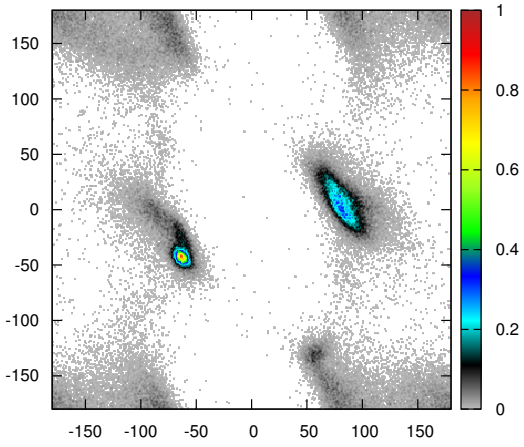
We proposed the use of the torsion angles probability matrix as a knowledge base local search strategy. The goal is to find the proper pair of angles (ϕ,ψ) that produces the best three-dimensional protein structure. We build an angle probability list (APL_a) of existing pair of angles of a given amino acid residue from which an angle cell can be selected with a probability of $AP_a(i, j)$.

A. Genetic Algorithm with Local Search

We developed a genetic algorithm (GA) for the 3-D PSP problem. The GA uses a structured population (previously used in other works [14], [17]) and incorporates the information of the AP matrix in the search process. In the GA, each individual represents a solution for the problem and it is composed of a set of torsion angles pairs (ϕ,ψ) and a set of side-chain torsion angles χ which depends on the amino acid residue type. The population is composed of n individuals and



(a) Proline



(b) Glycine

Fig. 1: Distribution of torsion angles for amino acids residues Proline (a) and Glycine (b) considering all their occurrences in PDB. It is possible to see that in the case of Proline angles are concentrated in two small regions. For the case of Glycine there are two regions that concentrate a high number of angles, but the rest are spread in the space.

they are sorted according to their fitness value (energy function shown in II-B). Then, the population is divided in three groups or "castes": **(A)** 30% of the best performing individuals, **(B)** 40% of the second best performing individuals and **(C)** 30% of the worst performing individuals. Figure 2 shows the schema of the GA detailed in Algorithm 1. The information provided by the AP_a matrices is used in three places of the GA: (1) in the generation of the initial population (line 2 of Algorithm 1), (2) in the local search (line 10 of Algorithm 1) and (3) in the generation of individuals of group C (line 13 of Alg. 1).

Initial Population: during the generation of the initial population each individual is built as follows: for each amino acid a of the target protein sequence, torsion angles (ϕ, ψ) are randomly chosen from the APL_a. Thus, pairs that appear more often have a higher chance of being selected. Once a pair has been selected from APL_a, we add a random real value between -1 and 1 to each angle ϕ and psi to reach surrounding regions of the pair. The χ torsion angles were randomly selected from

Algorithm 1 GA with Local Search for the 3-D PSP Problem

1: **Input:** A protein given as a sequence of amino acids;
2: $Pop^0 \leftarrow$ Generate initial population using APL;
3: Sort individuals and define groups A, B and C;
4: **for** $i = 1$ to $NGen$ **do**
5: $\quad Pop^i(A) \leftarrow Pop^{i-1}(A)$
6: $\quad$ **for** $j = 1$ to $|B|$ **do**
7: $\quad\quad P_1 \leftarrow getIndividual(A)$;
8: $\quad\quad P_2 \leftarrow getIndividual(B + C)$;
9: $\quad\quad Offspring \leftarrow$ **Crossover**(P_1, P_2)
10: $\quad\quad Offspring \leftarrow$ **LocalSearch**$(Offspring)$
11: $\quad\quad add(OffSpring, Pop^i(B))$
12: $\quad$ **end for**
13: $\quad$ **for** $j = 1$ to $|C|$ **do**
14: $\quad\quad Pop^i(C) \leftarrow$ Generate Individual using APL
15: $\quad$ **end for**
16: $\quad$ **sort**(Pop^i), $best \leftarrow top(pop^i)$
17: **end for**
18: **return** $best$.

intervals computed from the Dunbrack rotamers library [15], [16].

Crossover Operator: the crossover operation (line 9) produces a new offspring using the information provided by the two selected parents. First, parents P_1 and P_2 are chosen at random from groups A and $B + C$, respectively. For defining each amino acid of the offspring the algorithm uses the information either from P_1 or P_2 with a probability of 0.7 and 0.3, respectively. The offspring is then inserted in the population of the next generation.

Local Search Operator: a Local Search procedure is applied to every offspring generated by the crossover. This procedure takes as input a solution and for each amino acid residue it perturbates the torsion angles (main-chain and side-chain) with a probability of $p = 0.10$. In the case that an amino acid was chosen to be modified, we visit the neighbourhood of the angles with a size of one degree ($\phi - 1 \leq \phi \leq \phi + 1$, $\psi - 1 \leq \psi \leq \psi + 1$) using a greedy strategy in the surrounding neighbourhood, increasing by a small random amount the angle ϕ and, afterwards, decreasing it. The same procedure is applied on angle ψ and side-chain torsion angles. At any time if a better solution is found, we continue the process until no improvement is found. Because the evaluation function is computationally expensive, this local search procedure consumes a considerable long time, since each new solution must be fully evaluated to get the new fitness value.

Conformation of the Next Population: individuals from class A are automatically promoted to the next generation. All solutions resulted from the local search procedure are inserted in the next population. Finally, the individuals of class C are deleted, and new ones are randomly generated in the same way as the initial population (line 14). Once the population is complete, solutions are sorted by their fitness values (line 16). At the end of each generation, the best performing solution is always at the top of the population.

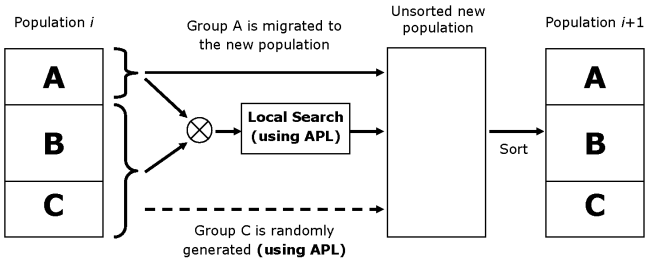


Fig. 2: Schema of one iteration of GA used for the PSP problem. The APL is also used in the generation of the initial population.

IV. Experiments and Results

A. Model and target proteins

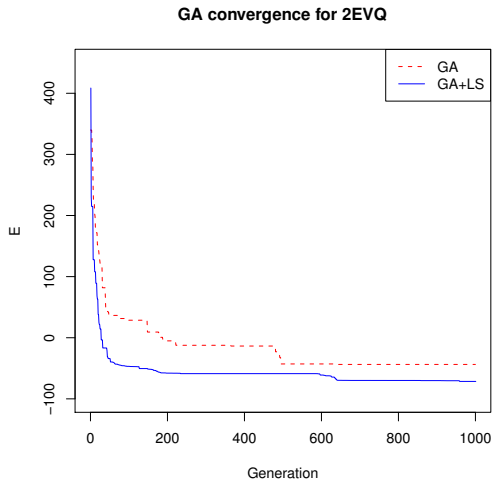
The amino acid sequences of six proteins were obtained from the PDB [3] and used as case studies in our experiments: 2EVQ [1] (12 residues, Fig. 4a); 1K43 [40] (14 residues, Fig. 4b); 1RPV [44] (17 residues, Fig. 4c); 1L2Y [37] (20 residues, Fig. 4d); 1DEP [27] (15 residues, Fig. 4e) and 1ACW [4] (29 residues, Fig. 4f). The 3-D structure of all of these proteins were experimentally determined and are stored in the PDB. Figure 4 - magenta shows the 3-D structure of each target protein. These case studies were selected in order to test our method with different classes of polypeptides with different folding patterns [34]. The six target protein sequences were submitted to the proposed method in order to predict their 3-D structures. In Section IV-B we show and analyse the time costs of the developed search strategy. Structural analysis of the predicted 3-D structures are presented in Section IV-C. We analyse the root mean square deviation (RSMD) of the predicted 3-D structures when compared with its corresponding native structure. We also analyse the stereo-chemical quality of secondary structure of the predicted conformations.

B. Computational Results

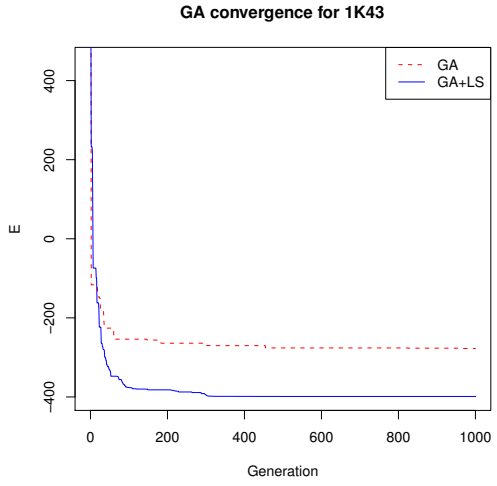
The computational tests have two main goals: to measure the effectiveness of the use of the APL (quality of the solutions) and to measure the contribution of the Local Search in the GA. We ran the GA without and with the local search six times for each protein and then collected the results shown in Table I. The stopping criteria of the GA was two hours (7200 seconds) or 1000 generations, whichever it is reached first. Almost all the runs reached 1000 generations, with the exception of the runs with Local Search for proteins 1RPV ($\sim$930 generations) and 1ACW ($\sim$750 generations). The proposed algorithm was implemented using the NAB language[3] and tested were ran on a Linux PC Intel I3-2100 CPU 3.10GHz x 4 cores and 4GB of memory. The energy calculation was performed using OpenMP to take advantage of the multiple core configuration.

The Local Search allows the GA to speeds up the convergence to much better solutions with the same restrictive times (see Fig. 3). The exploration of a smaller search space allows the algorithm to spend more time visiting useful solutions. This is specially important in problems like 3-D PSP, since

[3] *http://ambermd.org*



(a) 2EVQ



(b) 1K43

Fig. 3: Convergence curves of the GA without LS (red dash line) and with LS (blue line) for proteins 2EVQ and 1K43.

the evaluation function takes a high proportion of time. From table I it is possible to see that the percentage of time used by the local search corresponds to an average of 80% of the total running time.

C. Structural analysis

For biochemical and structural analysis we selected the class of solutions that at the last GA simulation presents the solution with the lowest potential energy. The quality of the predicted 3-D structures were evaluated by similarity comparisons with the structures of the experimental proteins obtained from the PDB (Eq.4). Quality measurements have been made in terms of the root mean square deviation (RMSD) between the position of the C_α atoms of the predicted and the experimental structures. The RMSD measure was calculated using PROFIT[4].

[4] *http://www.bioinf.org.uk/software/profit*

TABLE I: Results of the six runs of the GA without and with Local Search. For each run we show the $RMSD$, Energy (E), time in seconds (t). We also show the percentage of time used by the Local Search procedure.

	No Local Search			Local Search		
Protein	RMSD	E	$t(sec)$	RMSD	E	$t(sec)$
2EVQ-1	3.6	-15.59	473	2.59	-71.47	1813(78.3%)
2EVQ-2	3.12	-38.28	473	3.58	12.77	1587(75.7%)
2EVQ-3	2.63	-10.39	436	3.58	-29.09	1733(76.6%)
2EVQ-4	3.2	-43.66	483	3.32	-63.92	1614(76.7%)
2EVQ-5	2.71	-26.02	527	3.64	12.29	1548(76.1%)
2EVQ-6	3.03	-36.94	470	3.04	-64.06	1736(77.7%)
AVE	3.05	-28.48	477	3.29	-33.91	1672(76.9%)
1K43-1	2.19	-368.96	592	4.6	-398.93	2924(80.1%)
1K43-2	3.5	-371.09	605	4.7	-392.97	3354(81.1%)
1K43-3	4.84	-277.49	584	2.69	-446.21	3129(81.1%)
1K43-4	4.9	-355.71	581	5.94	-357.51	2430(77.8%)
1K43-5	5.81	-255.29	562	3.08	-413.27	2581(78.7%)
1K43-6	3.64	-337.41	604	3.88	-406.17	2521(78.1%)
AVE	4.15	-327.66	588	4.15	-402.51	2823(79.7%)
1RPV-1	1.71	-597.27	1488	0.86	-650.89	7204(88.1%)
1RPV-2	1.73	-488.24	1050	0.54	-726.55	7209(88.0%)
1RPV-3	1.23	-587.75	1051	0.7	-784.03	7202(86.9%)
1RPV-4	1.49	-561.8	1055	0.67	-717.45	7204(88.3%)
1RPV-5	1.36	-577.48	1027	0.82	-710.27	7205(86.8%)
1RPV-6	1.56	-621.71	1030	0.88	-726.36	7206(86.6%)
AVE	1.51	-572.37	1117	0.75	-719.26	7205(87.5%)
1L2Y-1	5.09	-89.22	974	4.12	-144.2	6043(84.8%)
1L2Y-2	5.52	-104.13	948	5.76	-186.53	5729(83.8%)
1L2Y-3	3.54	-46.91	960	4.96	-238.48	5418(83.7%)
1L2Y-4	3.33	-84.35	1000	4.33	-227.75	5057(83.4%)
1L2Y-5	4.52	-124.19	989	3.97	-258.7	4937(83.2%)
1L2Y-6	3.66	-47.36	941	4.09	-211.55	4788(82.6%)
AVE	4.28	-82.69	969	4.54	-211.2	5329(83.6%)
1DEP-1	0.85	-114.84	850	1.04	-204.71	4508(84.1%)
1DEP-2	0.76	-93.46	801	0.76	-191.88	4494(85.0%)
1DEP-3	1.07	-72.91	808	0.38	-196.98	4363(84.1%)
1DEP-4	0.62	-21.36	772	0.61	-182	4331(84.6%)
1DEP-5	0.54	-101.23	841	0.99	-217.26	4244(83.8%)
1DEP-6	0.72	-142.74	758	1.29	-185.1	4496(84.7%)
AVE	0.76	-91.09	805	0.85	-196.32	4406(84.4%)
1ACW-1	9.33	-82.97	1480	11.59	-142.69	7229(89.9%)
1ACW-2	11.7	-22.72	1452	11.8	-115.8	7208(92.7%)
1ACW-3	9.15	67.89	1484	10.14	-137.82	7202(90.5%)
1ACW-4	9.63	96.16	1453	9.9	-110.13	7206(85.0%)
1ACW-5	9.41	-4.75	1443	10.14	-190.55	7201(84.9%)
1ACW-6	10.98	26.67	1456	12.96	-133.47	7204(84.9%)
AVE	10.03	13.38	1461	11.09	-138.41	7208(87.9%)

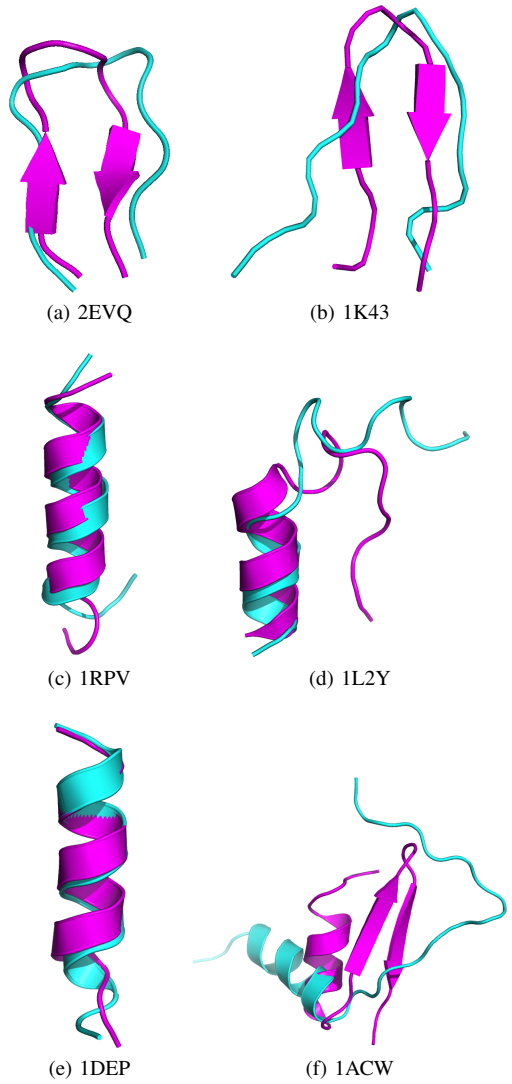
$$\mathbf{RMSD}(a,b) = \sqrt{\left(\sum_{i=1}^{n} \|r_{ai} - r_{bi}\|^2\right)/n}, \qquad (4)$$

were r_{ai} and r_{bi} are vectors representing the positions of the same atom i in each of two structures, a and b respectively, and where the structures a and b are optimally superimposed. Table I (Column 1, 5) shows the RMSD value achieved with each GA run with and without the local search operator, respectively.

The predicted 3-D protein structure with the lowest RMSD was the protein with PDB ID = 1DEP (≈0.38Å- Fig. 4e-Cyan), followed by, 1RPV (≈0.54Å- Fig. 4a-Cyan), 2EVQ (≈2.59Å- Fig. 4e-Cyan), 1K43 (≈2.69Å- Fig. 4b-Cyan), 1L2Y (≈3.97Å- Fig. 4d-Cyan) and 1ACW (≈9.9Å- Fig. 4f-Cyan). By visual inspection (Fig. 4), it is noticeable that the individual helices and other secondary structures are well formed in most of the case studies.

D. Secondary structure analysis

Secondary structure analysis were performed with Promotif [24]. We run Promotif in order to analyse the patterns of hydrogen bonds that define the secondary structure of the predicted 3-D structures. In this analysis we compare the secondary structure contents of the predicted 3-D protein structures against the secondary structure of the native structures.



(a) 2EVQ (b) 1K43 (c) 1RPV (d) 1L2Y (e) 1DEP (f) 1ACW

Fig. 4: Ribbon representation of the experimental (magenta) and predicted (cyan) 3-D structures. The C_α of the experimental and the predicted 3-D structure are fitted. Amino acid side chains are not shown for clarity. Graphic representation was prepared with PYMOL (*http://www.pymol.org*).

Table II, columns 1-5, summarizes the obtained results with Promotif. This analysis reveals that the secondary structure of the structures predicted by our method are comparable to their experimental structures. This can be observed, for example, when we examine the predicted structure of 1DEP-P which presents 80.0% (against 80.0% of the experimental 3-D structure (1DEP-E)) of the amino acid residues in a α-helix state, 0.0% (against 0.0% of the experimental) in a β-sheet state, and 20.0% (against 20.0% of the experimental 3-D structure) representing other irregular structures. The secondary structure similarity between the predicted and experimental 3-D structures can be also observed in case studies 1RPV and 1L2Y. The largest difference between the secondary structure elements of the predicted and experimental structures is observed in case studies 2EVQ, 1K43 and 1ACW. Through visual inspection of Figure 4 we can observe that β-sheets regions are not well formed for this three study cases, this in turn occurs because of the presence of distortions in the coil region.

TABLE II: Structure Analysis of the predicted proteins. For each protein we the values for the Experimental (-E) and the predicted (-P) structure. In the Ramachadran Plot Statistics we show (A) Most favoured regions, (B) Additional allowed regions, (C) Generously allowed regions and (D) Disallowed regions

	Secondary Structure Analysis				Ramachandran Plot Statistics			
Protein	Strand	α-helix	310 Helix	Other	(A)	(B)	(C)	(D)
2EVQ-E	50.0%	0.0%	0.0%	50.0%	87.5%	12.5%	0	0
2EVQ-P	0.0%	0.0%	0.0%	100.0%	100%	0	0	0
1K43-E	42.9%	0.0%	0.0%	57.1%	66.7%	33.3%	0	0
1K43-P	0.0%	0.0%	0.0%	100.0%	88.9%	11.1%	0	0
1RPV-E	0.0%	64.7%	0.0%	35.3%	86.7%	13.3%	0	0
1RPV-P	0.0%	64.7%	0.0%	35.3%	93.3%	6.7%	0	0
1L2Y-E	0.0%	35.0%	20.0%	45.0%	90.9%	9.1%	0	0
1L2Y-P	0.0%	45.0%	30.0%	25.0%	100%	0	0	0
1DEP-E	0.0%	80.0%	0.0%	20.0%	91.7%	8.3%	0	0
1DEP-P	0.0%	80.0%	0.0%	20.0%	91.7%	8.3%	0	0
1ACW-E	34.5%	24.1%	0.0%	41.4%	84%	16%	0	0
1ACW-P	0.0%	14.3%	0.0%	85.7%	96%	4%	0	0

E. Stereo-chemical analysis

The distribution of the amino acid residues in the Ramachandran plot[5], and the stereo-chemical quality of the 3-D structures predicted by our method were analysed by Procheck[6] [30]. Table II, columns 6-9, summarizes the numerical Ramachandran plot values for the experimental and predicted structures. We observe that in all of 3-D predicted structures, the amino acid residues are located in the most favourable regions of the map (favourable or additional allowed region). The percentage of residues in the "core" regions (most favourable regions) is one of the better guides to analyse the stereo-chemical quality of the predicted 3-D protein structures. When we compare the results obtained with the 3-D structure predicted by our method against the experimental structures we observe that these structures are comparable in terms of stereo-chemical quality.

V. Conclusion

The study of proteins and the prediction of their three-dimensional (3-D) structures is one of the key research problems in Structural Bioinformatics. Predicting the three-dimensional structure of a protein that have no templates in the Protein Data Bank is a very hard and sometimes virtually intractable task. In this paper, we introduced a novel search strategy for the 3-D PSP problem. The search strategy combines a genetic algorithm with a structured population and it is hybridized with a Local Search procedure. The developed

[5]We use the Ramachandran plot to visualize backbone dihedral angles ϕ against ψ of amino acid residues in protein structure.

[6]*www.ebi.ac.uk/thornton-srv/software/PROCHECK*

search strategy allows an efficient mechanisms for protein structure prediction. This is achieved by the use of a Local Search operator which allows the GA to scape from local minima. In the case at hand (the PSP problem) this occurs when torsion angles are modified by the GA. As corroborated by experiments, the developed method can produce good approximate 3-D protein structure, where the 3-D protein structures are comparable to their corresponding experimental ones.

The overall contributions of our work are the following: first, the combination of metaheuristics techniques and the consideration of real information to guide the search to develop a new and effective algorithm for a relevant biological problem (the 3-D PSP problem). Second, the use of the Local Search operator allows to speed up the algorithm to converge to better solutions. Although the latter has been shown in several other applications, it is important to highlight its contribution in a very computational expensive problem as PSP. More advances techniques for Local Search will be explored in the future, in order to take more advantage of proposed APL.

Finally, Protein Structure Prediction is a relevant problem and further research remains to be done. The development of new strategies, the adaptation and investigation of new methods and the combination of existing and state-of-the-art computational methods and techniques to the 3-D PSP problem is clearly needed.

ACKNOWLEDGMENT

The authors would like to thank CAPES N°12216127[7], CNPq[8] and CONICYT/FONDECYT N°11121288[9].

REFERENCES

[1] N. Andersen, K. Olsen, R. M. Fesinmeyer, X. Tan, F. M. Hudson, L. Eidenschink, and S. Farazi, "Minimization and optimization of designed beta-hairpin folds," *J. Am. Chem. Soc.*, vol. 128, no. 18, pp. 6101–6110, 2006.

[2] A. Baxevanis and B. Quellette, *Bioinformatics: A practical guide to the analysis of genes and proteins*, 2nd ed. New York, USA: John Wiley and Sons, Inc., 1990.

[3] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bath, H. Weissig, I. Shindyalov, and P. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[4] E. Blanc, V. Fremont, P. Sizun, S. Meunier, J. Van Rietschoten, A. Thevand, J. Bernassau, and H. Darbon, "Solution structure of p01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel," *Proteins*, vol. 24, pp. 359–369, 1996.

[5] F. E. Boas and P. B. Harbury, "Potential energy functions for protein design," *Curr. Opin. Struct. Biol.*, vol. 17, no. 2, pp. 199–204, 2007.

[6] C. Branden and J. Tooze, *Introduction to protein structure*, 2nd ed. New York, USA: Garlang Publishing Inc., 1998.

[7] A. Breda, D. Santos, L. Basso, and O. Norberto de Souza, "Ab initio 3-d structure prediction of an artificially designed three-a-helix bundle via all-atom molecular dynamics simulations," *Genet. Mol. Res.*, vol. 6, no. 2, pp. 901–910, 2007.

[8] J. Bujnicki, "Protein structure prediction by recombination of fragments," *ChemBioChem*, vol. 7, no. 1, pp. 19–27, 2006.

[7] *http://www.capes.gov.br*

[8] *http://www.cnpq.br*

[9] *http://www.conicyt.cl*

[9] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz Jr., D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, 1995.

[10] T. E. Creighton, "Protein folding," *Biochem. J.*, vol. 270, pp. 1–16, 1990.

[11] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the complexity of protein folding," *J. Comput. Biol.*, vol. 5, no. 3, pp. 423–466, 1998.

[12] V. Cutello, G. Narzisi, and G. Nicosia, "A multi-objective evolutionary approach to the protein structure prediction problem," *J. R. Soc., Interface*, vol. 3, no. 6, pp. 139–151, 2006.

[13] M. Dorn, A. Breda, and O. Norberto de Souza, "A hybrid method for the protein structure prediction problem," *Lect. Notes Bioinf.*, vol. 5167, pp. 47–56, 2008.

[14] M. Dorn, L. Buriol, and L. Lamb, "A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy," in *2011 IEEE Congress on Evolutionary Computation (CEC)*, 2010, pp. 2709–2716.

[15] R. Dunbrack Jr. and F. Cohen, "Bayesian statistical analysis of protein side-chain rotamer preferences," *Protein Sci.*, vol. 6, no. 8, pp. 1661–1681, 1997.

[16] R. Dunbrack Jr. and M. Karplus, "Backbone-dependent rotamer library for proteins: application to side-chain prediction," *J. Mol. Biol.*, vol. 230, no. 2, pp. 543–574, 2003.

[17] M. Ericsson, M. G. C. Resende, and P. M. Pardalos, "A genetic algorithm for the weight setting problem in ospf routing," *Journal of Combinatorial Optimization*, vol. 6, pp. 299–333, 2002.

[18] H. Fan and A. Mark, "Refinement of homology-based protein structures by molecular dynamics simulation techniques," *Protein Sci.*, vol. 13, no. 1, pp. 211–220, 2004.

[19] C. Floudas, H. Fung, S. McAllister, M. Moennigmann, and R. Rajgaria, "Advances in protein structure prediction and de novo protein design: A review," *Chem. Eng. Sci.*, vol. 61, no. 3, pp. 966–988, 2006.

[20] A. S. Fraenkel, "Complexity of protein folding," *Bull. Math. Biol.*, vol. 55, no. 6, pp. 1199–1210, 1993.

[21] W. Hart and S. Istrail, "Robust proofs of np-hardness for protein folding: general lattices and energy potentials," *J. Comput. Biol.*, vol. 4, no. 1, pp. 1–22, 1997.

[22] A. Hildebrand, M. Remmert, A. Biegert, and J. Soding, "Fast and accurate automatic structure prediction with hhpred," *Proteins*, vol. 77, no. S9, pp. 128–132, 2009.

[23] T. Hovmoller and T. Ohlson, "Conformation of amino acids in protein," *Acta Crystallogr.*, vol. 58, no. 5, pp. 768–776, 2002.

[24] E. Hutchinson and J. Thornton, "Promotif: A program to identify and analyze structural motifs in proteins," *Protein Sci.*, vol. 5, no. 2, pp. 212–220, 1996.

[25] D. Jones, "Predicting novel protein folds by using fragfold," *Proteins*, vol. 45, no. S5, pp. 127–132, 2001.

[26] D. Jones, W. Taylor, and J. Thornton, "A new approach to protein fold recognition," *Nature*, vol. 358, no. 6381, pp. 86–89, 1992.

[27] H. Jung, R. Windhaber, D. Palm, and K. Schnackerz, "Nmr and circular dichroism studies of synthetic peptides derived from the third intracellular loop of the beta-adrenoceptor," *FEBS Lett.*, vol. 358, no. 2, pp. 133–136, 1995.

[28] J. Klepeis, M. Pieja, and C. Floudas, "Hybrid global optimization algorithms for protein structure prediction: alternating hybrids," *Biophys. J.*, vol. 84, pp. 869–882, 2003.

[29] E. Krieger, K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus, "Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in casp8," *Proteins*, vol. 77, no. S9, pp. 114–122, 2009.

[30] R. Laskowski, M. MacArthur, D. Moss, and J. Thornton, "Procheck: a program to check the stereochemical quality of protein structures," *J. Appl. Crystallogr.*, vol. 26, no. 2, pp. 283–291, 1993.

[31] A. Lehninger, D. Nelson, and M. Cox, *Principles of Biochemistry*, 4th ed. New York, USA: W.H. Freeman, 2005.

[32] A. M. Lesk, *Introduction to Bioinformatics*, 1st ed. New York, USA: Oxford University Press Inc., 2002.

[33] C. Levinthal, "Are there pathways for protein folding?" *J. Chim. Phys. Phys.-Chim. Biol.*, vol. 65, no. 1, pp. 44–45, 1968.

[34] A. Liljas, L. Liljas, J. Pskur, P. Lindblom, G. amd Nissen, and M. Kjeldgaard. Singapore: World Scientific Printers, 2001.

[35] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, and M. Scott, *Molecular Cell Biology*, 5th ed. New York, USA: Scientific American Books, W.H. Freeman, 1990.

[36] J. A. Moult, "Decade of casp: progress, bottlenecks an prognosis in protein structure prediction," *Curr. Opin. Struct. Biol.*, vol. 15, no. 3, pp. 285–289, 2005.

[37] J. Neidigh, R. Fesinmeyer, and N. Andersen, "Designing a 20-residue protein," *Nat. Struct. Biol.*, vol. 9, no. 6, pp. 425–430, 2002.

[38] J. Ngo, J. Marks, and M. Karplus, "The protein folding problem and tertiary structure prediction," in *Computational complexity, protein structure prediction and the Levinthal Paradox*, K. Merz Jr and S. Grand, Eds. Boston, USA: Birkhauser, 1997, pp. 435–508.

[39] D. Osguthorpe, "Ab initio protein folding," *Curr. Opin. Struct. Biol.*, vol. 10, no. 2, pp. 146–152, 2000.

[40] M. Pastor, M. Lopez de la Paz, E. Lacroix, L. Serrano, and E. Perez-Paya, "Combinatorial approaches: a new tool to search for highly structured beta-hairpin peptides," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 614–619, 2002.

[41] S. Rackovsky, "Global characteristics of protein sequences and their implications," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 19, pp. 8623–8626, 2010.

[42] G. Ramachandran and V. Sasisekharan, "Conformation of polypeptides and proteins," *Adv. Protein Chem.*, vol. 23, pp. 238–438, 1968.

[43] C. Rohl, C. Strauss, K. Misura, and D. Baker, "Protein structure prediction using rosetta," *Methods Enzymol.*, vol. 383, no. 2, pp. 66–93, 2004.

[44] M. Scanlon, D. Fairlie, D. Craik, D. Englebretsen, and M. West, "Nmr solution structure of the rna-binding peptide from human immunodeficiency virus (type 1) rev," *Biochemistry*, vol. 34, no. 26, pp. 8242–8249, 1995.

[45] E. Scheef and J. Fink, *Fundamentals of protein structure: Structural Bioinformatics*, P. Bourne and H. Weissig, Eds., 2003.

[46] K. Simons, I. Ruczinki, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker, "Improved recognition of native-like structures using a combination of sequence-dependent and sequence-independent features of proteins," *Proteins: Struct., Funct., Bioinf.*, vol. 34, no. 1, pp. 82–95, 1999.

[47] R. Srinivasan and G. Rose, "Linus - a hierarchic procedure to predict the fold of a protein," *Proteins*, vol. 22, no. 2, pp. 81–99, 1995.

[48] ——, "Ab initio prediction of protein structure using linus," *Proteins*, vol. 47, no. 4, pp. 489–495, 2002.

[49] A. Tramontano, *Protein structure prediction*, 1st ed. Weinheim, Germany: John Wiley and Sons, Inc., 2006.

[50] W. van Gunsteren and H. Berendsen, "Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry," *Angew. Chem., Int. Ed. Engl.*, vol. 29, no. 9, pp. 992–1023, 1990.

[51] S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative tasser simulations," *BMC Biol.*, vol. 5, no. 17, pp. 1–10, 2007.

[52] J. Xu, J. Peng, and F. Zhao, "Template-based and free modeling by raptor11 in casp8," *Proteins*, vol. 77, no. S9, pp. 133–137, 2009.

[53] Y. Zhang, "Template-based modeling and free modeling by i-tasser in casp7," *Proteins*, vol. 69, no. 8, pp. 108–117, 2007.

[54] ——, "I-tasser server for protein 3d structure prediction," *BMC Bioinf.*, vol. 9, no. 40, pp. 1–8, 2008.

[55] H. Zhou and J. Skolnick, "Protein structure prediction by pro-sp3-tasser," *Biophys. J.*, vol. 96, no. 6, pp. 2119–2127, 2009.