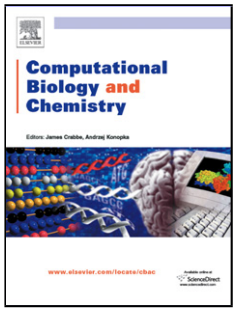


Title: APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction

Author: Bruno Borguesan Mariel Barbachan e Silva Bruno Grisci Mario Inostroza-Ponta Márcio Dorn

PII: S1476-9271(15)30125-0
DOI: http://dx.doi.org/doi:10.1016/j.compbiolchem.2015.08.006
Reference: CBAC 6460

To appear in: *Computational Biology and Chemistry*

Received date: 23-2-2015
Revised date: 5-8-2015
Accepted date: 17-8-2015

Please cite this article as: Bruno Borguesan, Mariel Barbachan e Silva, Bruno Grisci, Mario Inostroza-Ponta, Márcio Dorn, APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction, *Computational Biology and Chemistry* (2015), http://dx.doi.org/10.1016/j.compbiolchem.2015.08.006

PDB

X-ray Diffraction

Resolution <= 2Å

R-Factor <= 0.2

Homology <= 30%

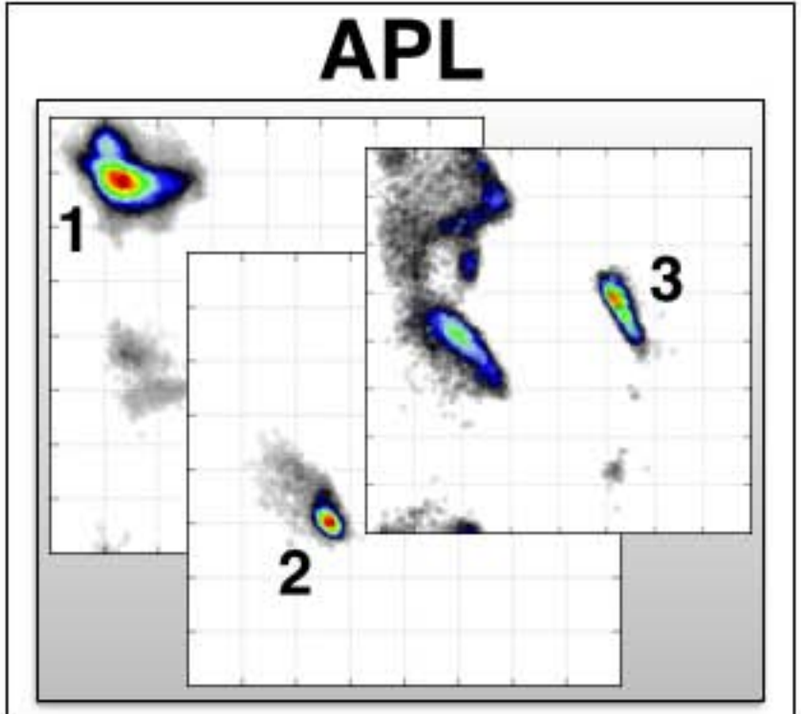
Protein Selection

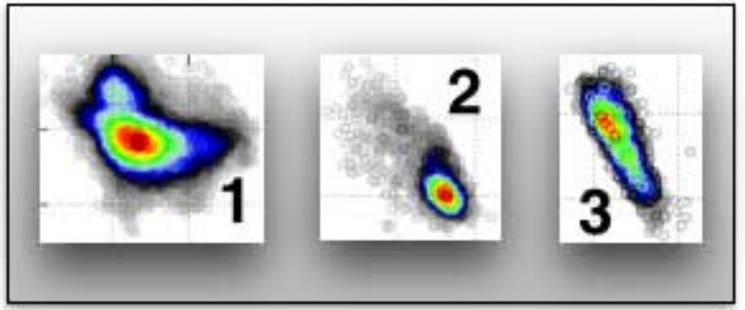
AASS database

B-Factor <= 30Å²

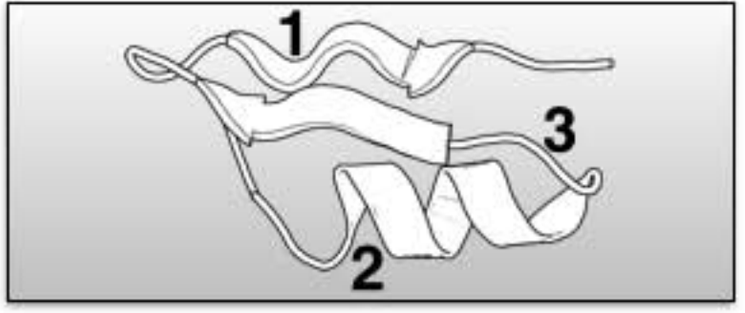
Occupancy = 1

Amino Acid Selection





Metaheuristics



Highlights

- Angle Probability Lists improve knowledge-based protein structure prediction methods;
- Development of knowledge-based methaheuristics for protein tertiary structure prediction;
- First principle methods with database information;
- Amino Acid and Secondary Structure conformational preferences;

APL: an Angle Probability List to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction

Bruno Borguesan[a], Mariel Barbachan e Silva[a], Bruno Grisci[a], Mario Inostroza-Ponta[b], Márcio Dorn*[a]

[a] *Federal University of Rio Grande do Sul, Institute of Informatics, Av. Bento Gonçalves 9500, 91501-970, Porto Alegre, RS, Brazil.*
[b] *Dep. de Ingeniería Informática, Center for Biotechnology and Bioengineering, UdeSantiago, Av. Ecuador 3659, Santiago, Chile.*

Abstract

Tertiary Protein Structure Prediction is one of the most challenging problems in Structural Bioinformatics. Despite the advances in algorithm development and computational strategies, predicting the folded structure of a protein only from its amino acid sequence remains as an unsolved problem. We present a new computational approach to predict the native-like three-dimensional structure of proteins. Conformational preferences of amino acid residues and secondary structure information were obtained from protein templates stored in the *Protein Data Bank* and represented as an *Angle Probability List*. Two knowledge-based prediction methods based on Genetic Algorithms and Particle Swarm Optimization were developed using this information. The proposed method has been tested with twenty-six case studies selected to validate our approach with different classes of proteins and folding patterns. Stereochemical and structural analysis were performed for each predicted three-dimensional structure. Results achieved suggest that the *Angle Probability List* can improve the effectiveness of metaheuristics used to predicted the three-dimensional structure of protein molecules by reducing its conformational search space.

Keywords: three-dimensional protein structure prediction, amino acid conformational preferences, metaheuristics

1. Introduction

Proteins are biological macromolecules responsible for the execution of different and important functions in living systems (Lesk, 2002; Tramontano, 2006). From a structural perspective, protein is an ordered linear chain of building blocks known as amino acid residues. Each protein is defined by its unique sequence of amino acid residues that causes the protein to fold into a particular three-dimensional (3-D) shape. The biochemical function of a protein is close related with its three-dimensional structure (Lehninger et al., 2005). Predicting the folded structure of a protein (PSP problem) only from its amino acid sequence, remains a challenging problem in Computer Science, Mathematics, Physics, Biology and Chemistry (Lander and Waterman, 1999; Wooley and Ye, 2010; Dorn et al., 2014b). The challenge arises due to the combinatorial explosion of plausible shapes that a protein sequence can assume (Levinthal, 1968).

Determining the 3-D structure of a protein is both experimentally expensive (due to the costs associated) and time consuming (Guntert, 2004). The difficulty in determining and finding out the 3-D structure of proteins has generated a significant discrepancy between the volume of sequences of amino acid residues produced by Genome Projects[1] and the number of 3-D structures of proteins determined by experimental methods (X-Ray, NMR, etc). Even though, this small proportion represents a rich source of information to be explored by computational methods for the PSP problem (Greer, 1990; Johnson et al., 1994; Turcotte et al., 2001B; Dorn et al., 2011). Conformational preferences are acquired from protein templates stored in the *Protein Data Bank* (PDB) (Berman et al., 2000) and used in prediction tasks (Hovmoller and Ohlson, 2002; Dorn et al., 2013, 2014a).

The 3-D PSP problem in computational complexity appears to be a NP-complete problem (Guyeux et al., 2014). Several computational strategies and algorithms have been proposed as a solution to the PSP problem. These methods can be classified in four classes (Floudas et al., 2006; Dorn et al., 2014b):

*Corresponding author.
Email address: mdorn@inf.ufrgs.br (Márcio Dorn*)

[1] http://genomics.energy.gov

(*i*) first principle methods without database information (Osguthorpe, 2000); (*ii*) first principle methods with database information (Srinivasan and Rose, 1995; Rohl et al., 2004); (*iii*) threading or fold recognition methods (Bowie et al., 1991; Jones et al., 1992; Bryant and Altschul, 1995; Turcotte et al., 1998) and (*iv*) comparative modeling methods (Sánchez and Sali, 1997; Martí-Renom et al., 2000). Group *ii*, *iii* and *iv* are often referenced as knowledge-based methods. These methods are able to perform fast and effective prediction of protein 3-D structures when template structures and fold libraries are available (Kolinski, 2004). Analysis of the last *Critical Assessment of protein Structure Prediction* (CASP10)[2] experiments reveals that the best results are achieved by methods that use some knowledge from experimental databases (Kryshtafovych et al., 2014a).

A knowledge-based prediction method is fully dependent on the quality of structural models and how they are represented and used (Dorn et al., 2014b). In this article, we propose a strategy to obtain and represent conformational preferences of amino acid residues from experimentally determined protein structures. Conformational preferences of amino acid residues in protein and its secondary structure are obtained from PDB and represented as an *Angle Probability List* (APL) (Dorn et al., 2013). This information represents a rich source of data and can be used in knowledge-based prediction methods. In this paper, we analyze the impact of using APL in metaheuristics.

The paper is organized as follows: Section 2 presents fundamental concepts of protein structure; conformational preferences of amino acid residues of proteins; and metaheuristics. Section 3 shows the developed *Angle Probability List* and the standard implementation of two metaheuristics applied to evaluate our method. In section 4 we discuss the results using our approach in different metaheuristics. Finally, the last section concludes the paper and points out some future works.

2. The Protein Structure Prediction problem

2.1. Proteins and its 3-D Structure

Proteins are composed of ordered linear chains of amino acid residues linked by peptide bonds (Fig. 1) (Liljas et al., 2009; Lehninger et al., 2005; Lesk, 2010). An amino acid residue is a small molecule containing an amino group (H_3N^+), a carboxyl group (COO^-), and a hydrogen atom attached to a central α carbon (C_α). In addition, each amino acid also has a R organic group (also called side-chain) attached to the C_α. The side-chain gives to the amino acid its propriety. A peptide bond is formed when the carboxyl group of one residue reacts with the amino group of other residue, thereby releasing a water molecule (Lehninger et al., 2005).

The linear sequence of amino acid residues is known as the protein's *primary structure*. Frequently, a fragment of amino acid residues adopts the same conformation (Tramontano, 2006). Local segments of the protein main-chain conformation define the *secondary structure*. These structures are defined by the presence of hydrogen bonds between the amino and carboxyl groups of the polypeptide chain. There are preferred conformations like α-helices, β-sheets, β-turns, among others (Lehninger et al., 2005). In different proteins, helices and sheets are combined in many ways to create different spatial arrangements of the polypeptide chain. (Lesk, 2010). This is called the protein *tertiary structure* (3-D) and represents the functional/native state of the protein (Lesk, 2002; Tramontano, 2006).

There are different ways to represent a polypeptide structure (Dorn et al., 2014b): all-atom model (Osguthorpe, 2000), united atom model (Khalili et al., 2005), rotamers (Shapovalov and Dunbrack, 2011), and torsion angles (Cutello et al., 2006; Dorn et al., 2011, 2013). Due to the planarity of the peptide bond, the conformation of a peptide backbone is mainly described by two torsion angles per amino acid residue (Lehninger et al., 2005): ϕ (*phi*) and ψ (*psi*). In this work, we represent a polypeptide chain by its set of main-chain and side-chain torsion angles (ϕ, ψ and χ angles). The main advantage of this representation is the reduced number of variables to control and optimize when predicting the polypeptide structure. The set of consecutive main-chain torsion angles describes the internal rotation of the protein (Lesk, 2002; Scheef and Fink, 2003) and causes the polypeptide to fold into a particular three-dimensional shape.

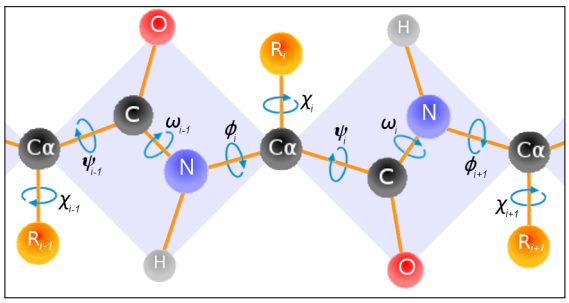


Figure 1: Schematic representation of a peptide bond. C is carbon, O is oxygen, N is nitrogen and H is hydrogen. We represent a polypeptide chain by its set of main-chain and side-chain torsion angles: ϕ (phi), ψ (psi) and χ (chi). The number of χ angles depends on the type of amino acid residue and are represented in the figure by the group R.

[2] http://predictioncenter.org

The peptide bond (C-N bond - Fig. 1 ω angle) is not involved in the molecular rotation, due to its double-bond character. The rotation is only allowed around the N-C_α and C_α-C bonds. The angles of these bonds are known as *phi* (ϕ) and *psi* (ψ) angles respectively (Lodish et al., 1990; Lesk, 2002). The possible conformation of a given polypeptide depends on the amino acid chemical properties. The side-chains of amino acid residues also plays important roles in the conformation of a protein molecule. The number of χ angles in these side-chain depends on the amino acid residue (Liljas et al., 2009) (more details can be found in Section 2.3).

2.2. Metaheuristics and the PSP Problem

Metaheuristics are often used to deal with hard optimization problems (Glover and Kochenberg, 2003; Resende et al., 2010), because of their ability to find satisfactory solutions with less computational effort than exact methods. Metaheuristics designates a class of approximate computational methods that optimizes a problem by an iterative generation process (Goldberg, 1989; Blum and Roli, 2003; Glover and Kochenberg, 2003; Battiti et al., 2008; Talbi, 2009; Mucherino and Seref, 2009; Luke, 2009). This process guides a subordinate heuristic by intelligently combining different concepts for exploring and exploiting the search space (Osman and Kelly, 1996; Osman and Laporte, 1996). Metaheuristics make few or no assumptions about the problem being optimized and can search vast spaces of candidate solutions and apply two strategies of intensification and diversification in the effective exploration of the search space. The first one (intensification) eliminates the search space by examining neighbors of elite solutions, while the last one (diversification) is a stochastic component that explores unvisited regions. Metaheuristics do not guarantee an optimal solution, and they are used to deal with combinatorial optimization problems in which an optimal solution is sought over a discrete search-space (Luke, 2009).

Examples of metaheuristics are: Simulated Annealing (SA) (Kirkpatrick et al., 1983; Granville et al., 1994), Particle Swarm Optimization (PSO) (Kennedy, 2003; Trelea, 2003), Genetic Algorithms (GAs) (Goldberg, 1989), etc. Two of the most popular metaheuristics applied in the field of 3-D protein structure prediction are the Genetic Algorithms (Dandekar and Argos, 1992; Le Grand and Merz Jr., 1993; Sun, 1995; Pedersen and Moult, 1997; Hoque et al., 2006) and Particle Swarm Optimization (PSO)(Meissner and Schneider, 2007; Kondov, 2013). GAs are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics (Luke, 2009). GAs are modeled through the use of a population of individuals representing solutions, which undergo selection in the presence of variation-inducing operators such as *mutation* and *recombination*. For every individual is calculated a fitness value that indicates how good is the solution. For each iteration of the algorithm, called a *generation*, different individuals are combined by chance and the new solution formed by this operation is used in the new population. It is also common to use some mechanism to maintain the variability of the individuals, decreasing the chances of being trapped in a local minimum (Kirkpatrick et al., 1983). Dorn et al. (2013), for example, combines a Genetic Algorithm, structural information from PDB and a Local Search operator for the 3-D protein structure prediction problem. In Dorn et al. (2011) a Genetic Algorithm is combined with a structured population, and it is hybridized with a *path-relinking* procedure that helps the algorithm to escape from the local minimum. Cutello et al. (2006) use a Genetic Algorithm to solve a multi-objective representation of protein structure. Park (2005) uses a Genetic Algorithm for fragment assembly to find low-energy conformations. Hoque et al. (2009) present a comprehensive review of the application of GA in the protein folding problem.

In a Particle Swarm Optimization algorithm, the potential solutions, called particles, fly through the problem space by following the current optimum solutions. Meissner and Schneider (2007) built a PSO algorithm to optimize backbone geometries of proteins considering secondary structure information in the optimization process. In Kondov (2013), a distributed parallel particle swarm optimization algorithm was developed for protein structure prediction problem. Lin and Hsieh (2009) presents a hybrid PSO/GA algorithm to search for the native structure of a protein molecule in a hydrophobic-hydrophilic lattice model representation.

Despite the advances, metaheuristics still have to deal with the challenge of vast conformational search spaces caused by the different combination of amino acid residues. To address this challenge, we developed an approach to obtain conformational preferences of amino acid residues in protein templates based on its secondary structure. This information represented as an *Angle Probability List* was used to reduce the protein conformational search space. Section 3 describes the developed knowledge-based metaheuristics for 3-D protein structure prediction.

2.3. Conformational preferences of amino acid residues in proteins

The ϕ and ψ torsion angles of a protein molecule (Fig. 1) can assume, theoretically, 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 (Hovmoller and Ohlson, 2002). The allowed and prohibited values for the torsion angles ϕ (x-axis) and ψ (y-axis) are graphically demonstrated by the Ramachandran plot (Ramachandran and Sasisekharan, 1968). Despite these prohibited combinations of the *phi* and *psi* torsion angles, proteins can still assume several conformations. The stable arrangement of segments of amino acid residues of the polypeptide shape structural patterns (Lehninger et al., 2005) and represents the secondary structure of a polypeptide (Branden and Tooze, 1998; Scheef and Fink, 2003; Andersen and Rost, 2003). Regularity in the spatial conformation is maintained through these intermolecular interactions. Identical conformations have similar torsion angles values. The two most common secondary structures are α-helices (Pauling et al., 1951) and β-sheets (Pauling and Corey, 1951). There are other periodic conformations (coils and turns), but the α-helix and β-sheets are the most stable and can be considered as the main elements in 3-D structures (Liljas et al., 2009; Andersen and Rost, 2003; Tramontano, 2006).

Amino acid residues in a secondary structure usually adopt a particular set of backbone torsion angles (ϕ and ψ) (Hovmoller and Ohlson, 2002). In this article, we analyze the conformational preferences of amino acid in proteins according to its secondary structure. Figure 2 shows the Ramachandran plot of six most abundant secondary structure conformational states present in 6,650 proteins obtained from PDB (for more details, see Section 3.1).

Knowledge-based protein structure prediction methods are reasoned on the observation that when a new fold is discovered, it is composed of common structural motifs or fragments from super-secondary structures of proteins with known templates (Lesk, 2002; Tramontano, 2006). Conformational preferences (*phi* and *psi*) and secondary structure propensities obtained from experimentally determined proteins can be taken into account by knowledge-based metaheuristics (Dorn et al., 2013, 2014b) developed for the PSP problem. In following sections we described how we combine the conformational preferences of amino acid residues with its secondary structure information to build the *Angle Probability List* (APL), and to apply this into a knowledge-based metaheuristics for the PSP problem.

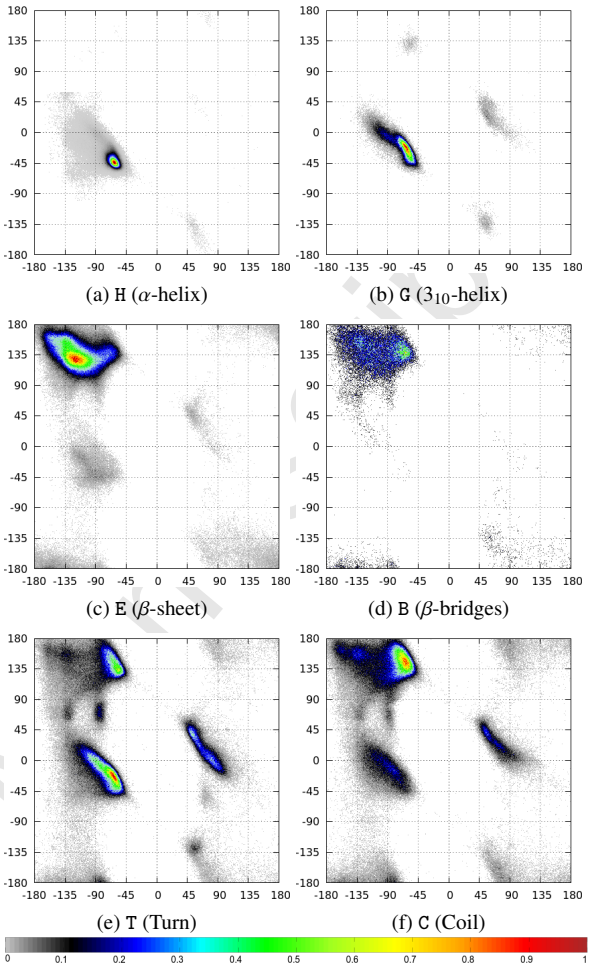


Figure 2: Individual Ramachandran plots for six secondary structures. Torsion angles values were computed from a set of 6,650 protein structures obtained from the PDB. The dark red color marks the densest regions of the Ramachandran plot.

3. Material and Methods

3.1. AASS: A Database of Amino Acid and Secondary Structure conformational preferences

The success of a knowledge-based three-dimensional protein structure prediction approach depends heavily on the quality of the structural protein templates obtained from PDB and how this information is represented and used (Dorn et al., 2014b). We combine the conformational preferences of amino acid residues (*AA*, torsion angles) in proteins with their secondary structure information (*SS*). To that end, we selected a set of 6,650 protein structures from PDB. All 3-D protein structures were experimentally determined by X-ray diffraction with resolution $\leq$ 2.0Å and stored in PDB until December 2014. We remove all structures with *R-factor* greater than 0.2. If homologous protein chains with sequence identity at most 30% were found, only one of

them was retained. From these protein 3-D structures there was a set of 2,670,182 amino acid residues. Another filter was applied for atoms from the backbone, selecting residues with *b-factor* $\leq 30\text{Å}^2$ and occupancy equal to 1, which leaves us with 2,225,475 amino acids to further analysis. Similar parameters to filter PDB data were used before by Hovmoller and Ohlson (2002).

We used STRIDE (Frishman and Argos, 1995; Heinig and Frishman, 2004) to assign the secondary structure of each amino acid residue. STRIDE implements an eight secondary structure model: B, E, H, G, I, b, C and T. STRIDE assigns the shortest α-helix (H) if it contains at least two consecutive i $\rightarrow$ i + 4 hydrogen bonds. The hydrogen bond patterns may be ignored if the ϕ and ψ angles are unfavorable. This definition is also used for 3_{10}-helices (state G with i $\rightarrow$ i + 3 hydrogen bonds) and for π-helices (state I with i $\rightarrow$ i + 5 hydrogen bonds), with the empirical hydrogen bond criterion (Andersen and Rost, 2003). The sheet category does not distinguish between parallel and anti-parallel sheets. The β-sheet (E) is composed by a minimum of two residues in each one of five possible hydrogen bond conformations. Single residue sheets, that is, β-bridges are labeled as B for the three hydrogen bond conformations and as b for the remaining two (Andersen and Rost, 2003). Turns T are assigned according to the ϕ and ψ angles of residue i + 1 and i + 2. The C symbol is used whenever none of the above structure requirements is satisfied. We observe a small number of experimental data, less than 1%, with amino acid residues in I (π-helix) and b (isolated bridge) conformational states. Thus, we only consider six conformational states for further analysis: H (α-helix), G (3_{10}-helix), E (β-sheet), B (β-bridge), T (Turn) and C (Coil).

Table 1 summarizes the conformational pattern of the amino acid residues obtained from the 6,650 selected protein structures. As can be observed, the most common secondary structures in protein are α-helix ($\approx$34%) and β-sheet ($\approx$25%). Turn and coils represent $\approx$35% of the secondary structures. Leucine (LEU - 9.3%), Alanine (ALA - 8.7%), Valine (VAL - 7.5%) and Glycine (GLY - 7.5%) are the most common amino acid residues presented in protein templates and together represent more than 33.0% of the total amino acid residues. Cystine (CYS - 1.2%), Triptophane (TRP - 1.6%), Methionine (MET - 1.7%) and Histidine (HIS - 2.4%) are the amino acid residues with lowest occurrence.

Table 1: Secondary Structure preferences of amino acid residues in proteins. Column 2-7 show the number of amino acid residues belonging to each secondary structure. Line 1-20 show the number of amino acid residues in each secondary structure state. This table summarizes the number of amino acid residues and its secondary structure obtained from 6,650 unique protein chains.

Amino Acid Residue	H (α-helix) Fig. 2a	G (π-helix) Fig. 2b	E (β-sheet) Fig. 2c	B (β-bridge) Fig. 2d	T (Turn) Fig. 2e	C (Coil) Fig. 2f	Total Percentage (%)
ALA	96,475	8,556	37,194	1,621	29,045	22,216	195,107 (8.7%)
ARG	46,801	4,596	25,400	1,620	17,722	16,400	112,539 (5.1%)
ASN	25,032	3,899	15,120	1,328	29,700	18,991	94,070 (4.2%)
ASP	36,695	6,916	18,941	1,407	39,738	25,190	128,887 (5.8%)
CYS	8,074	981	8,640	504	5,332	4,365	27,896 (1.2%)
GLN	35,386	3,513	15,553	890	12,850	10,900	79,092 (3.5%)
GLU	65,101	7,931	25,119	983	22,999	15,764	137,897 (6.2%)
GLY	26,653	4,868	28,336	1,777	58,009	46,626	166,269 (7.5%)
HIS	16,427	2,321	13,715	881	11,541	9,256	54,141 (2.4%)
ILE	45,549	2,860	55,277	2,013	11,696	15,104	132,499 (5.9%)
LEU	93,070	8,430	58,127	2,470	23,127	24,593	209,817 (9.3%)
LYS	46,162	5,003	23,493	1,321	20,505	17,711	114,195 (5.3%)
MET	16,524	1,421	9,609	553	4,611	4,721	37,439 (1.7%)
PHE	30,923	3,961	33,458	1,571	13,476	12,003	95,392 (4.3%)
PRO	14,750	5,804	10,955	1,084	34,846	33,211	100,650 (4.5%)
SER	35,159	6,291	28,804	2,036	29,398	26,257	127,945 (5.7%)
THR	32,805	3,552	38,232	2,190	23,146	24,607	124,532 (5.6%)
TRP	11,825	1,858	10,827	546	5,512	4,534	35,102 (1.6%)
TYR	27,957	3,682	28,312	1,343	13,146	10,660	85,100 (3.8%)
VAL	48,802	3,066	76,472	2,442	16,693	19,431	166,906 (7.5%)
Total	760,170	89,509	561,584	28,580	423,092	362,540	2,225,475
Percentage (%)	(34.1%)	(4.0%)	(25.2%)	(1.3%)	(19.1%)	(16.3%)	(100%)

For each amino acid residue we compute the dihedral angles ϕ and ψ. We developed a database schema to store all achieved structural information from PDB: torsion angles and secondary structure. We analyzed the conformational preference (dihedral angles) according with the amino acid residue and its secondary structure. From the AASS dataset, we compute Ramachandran plots for each set of amino acid residues belonging to a secondary structure and analyze its conformational preferences. For a given secondary structure we can have different conformational preferences (ϕ and ψ) depending on the amino acid residue. This can be clearly observed when we inspect the conformational preference of amino acid residues in *coil* and *turn*. Figure 3 shows the conformational preferences of amino acid residues in *turn* secondary structure.

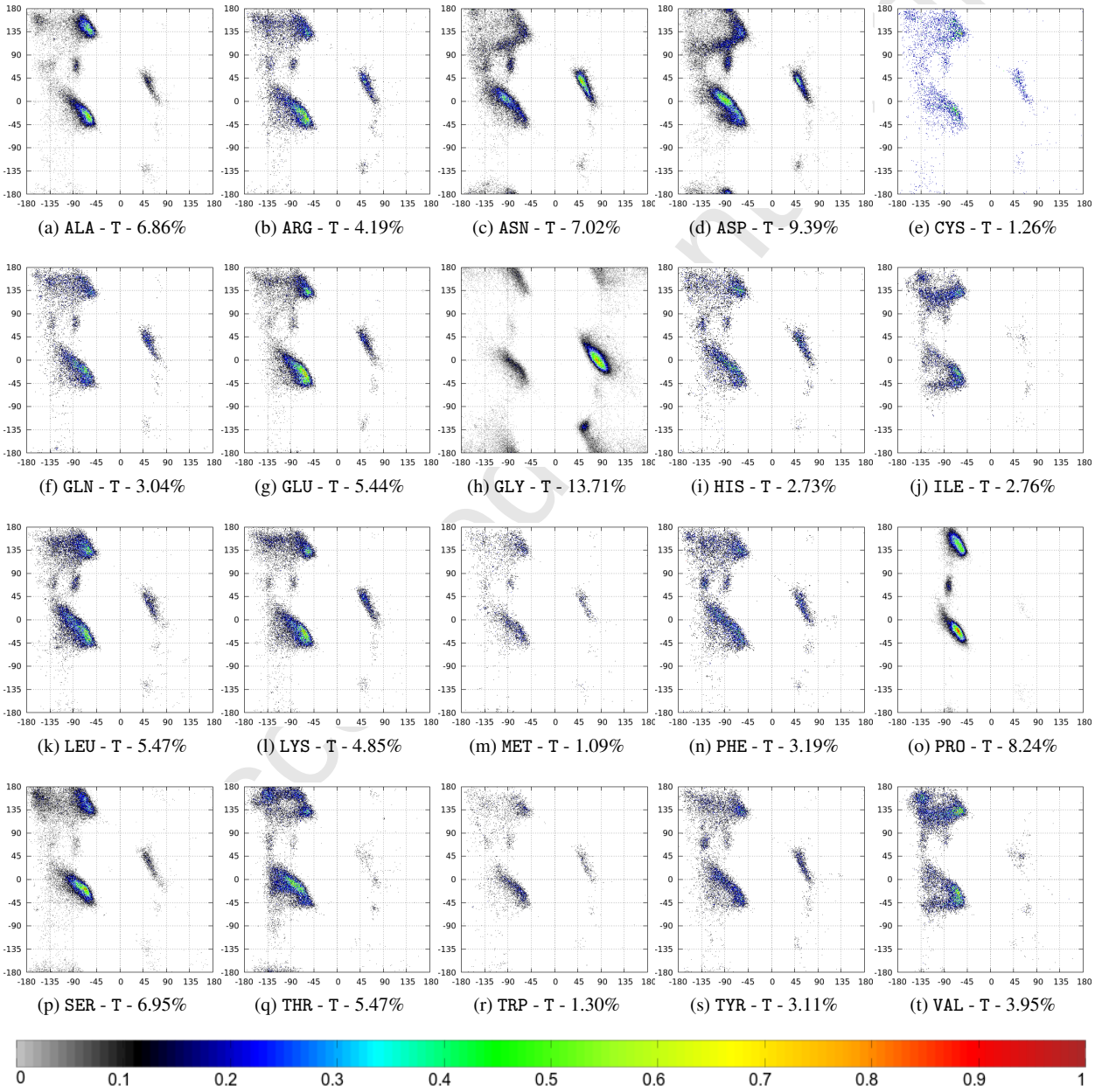


Figure 3: Ramachandran plots for the 20 amino acid residues conformational preference of turn secondary structure (T). The dark red color marks the most densely occupied regions of the Ramachandran plot. The label contains the percentage of each amino acid in the turn secondary structure.

As can be observed, amino acid residues in the same secondary structure have their particular conformational preferences (ϕ and ψ). Figure 4 shows the conformational preferences of amino acid residues in coil secondary structure. We observed that in the same secondary structure exists different preferences for ϕ and ψ torsion angles when we analyzed its occurrence in each one of the 20 amino acid residues. Conformational preferences are crucial to the development of knowledge-based protein structure prediction methods. There are many studies in the literature indicating the presence of conformational patterns in protein sequences and its secondary structure (Xia and Xie, 2002; Moelbert et al., 2004; Ting et al., 2010). This structural information can be used to reduce the protein conformational search space. The conformational preferences of amino acid residues in other secondary structures (α-helix, 3_{10}-helix, β-sheet and β-bridge), can be found in supplementary materials.

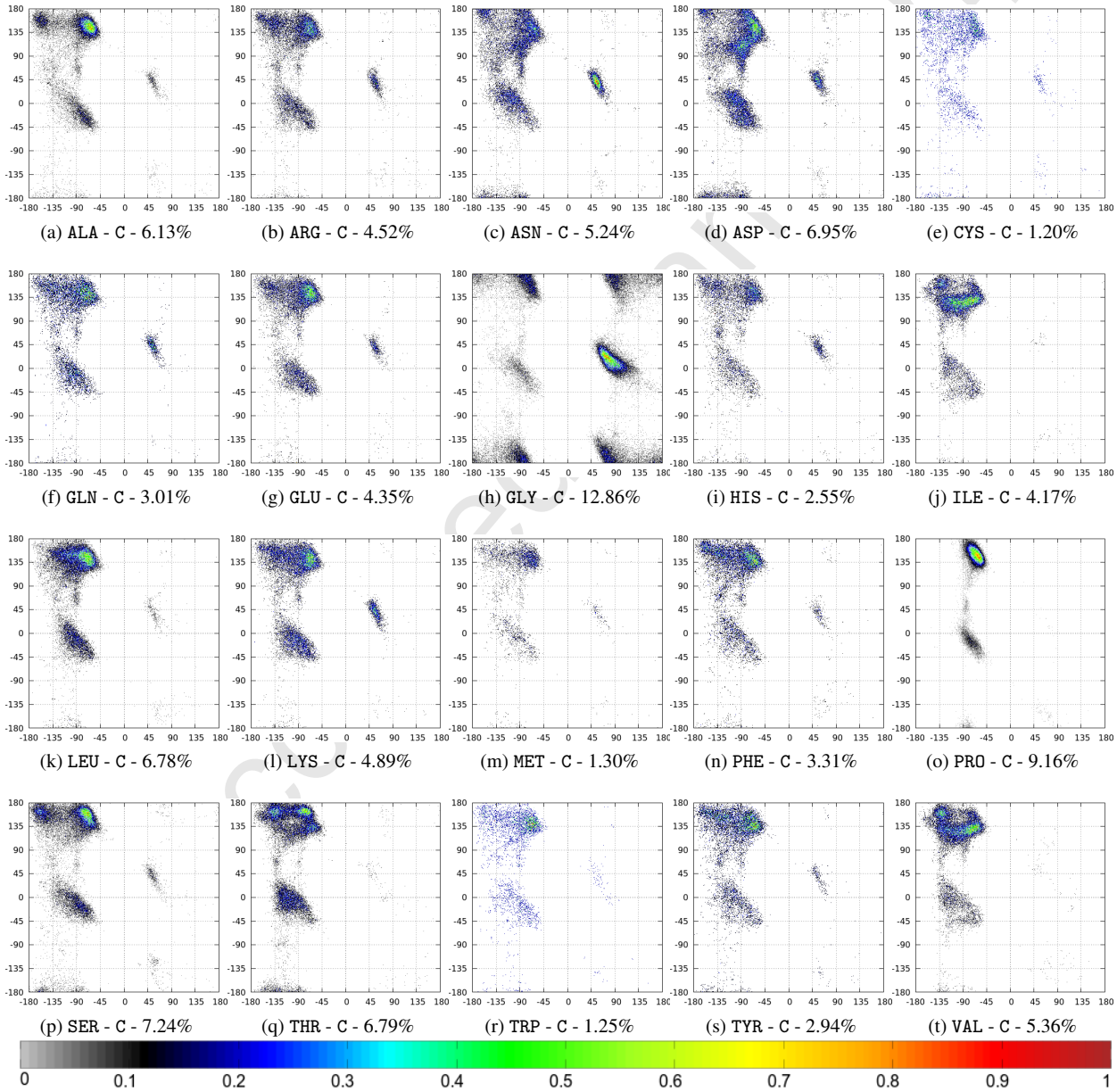


Figure 4: Ramachandran plots for the 20 amino acid residues for coil secondary structure (C). The dark red color marks the most densely occupied regions of the Ramachandran plot. The label contains the percentage of each amino acid in the coil secondary structure.

To use this information in a knowledge-based approach we built an *Angle Probability List* (APL) $H_{aa,s}$ of $[-180, 180] \times [-180, 180]$ cells for each amino acid residue (aa) and secondary structure (s). Each cell (i,j) has the number of times that a given amino acid residue aa in secondary structure s has a pair of torsion angles ($i \leq \phi < i+1$, $j \leq \psi < j+1$). Then, for each amino acid residue and secondary structure we compute the $APL_{aa,s}$ (Eq. 1) that represents the normalized frequency of each pair. Each list was sorted from the highest to the lowest frequency. A higher frequency associated with a pair ϕ and ψ indicates that this combination is more common in nature and should have a higher chance of be selected by a metaheuristic.

$$APL_{aa,s}(i,j) = \frac{H_{aa,s}(i,j)}{\sum(H_{aa,s})}, \quad (1)$$

When we associate these information (type of amino acid residue, secondary structure state and frequency) a valuable information that can be used to predicted new 3-D protein structures is obtained. We computed 120 APLs (20 amino acid residues × 6 secondary structures) and used this information in two different metaheuristics, Genetic Algorithm (Section 3.2) and Particle Swarm Optimization (Section 3.3), to generate more accurate solutions. Figure 5 schematizes the construction of the *Angle Probability List* (APL) from experimental-determined three-dimensional protein structures. APL are used in metaheuristics with the purpose of reducing the conformational search space of proteins. Sections 3.2 and 3.3 describe the implementation of two standard metaheuristics used in our experiments to test the effectiveness of APL.

3.2. Genetic Algorithm

We developed a knowledge-based Genetic Algorithm to search the protein conformational space. APLs are used to speed up the search and prediction of the three-dimensional structure of proteins. A standard implementation of a Genetic Algorithm was combined with a structured population (Ericsson et al., 2002), the APL of each amino acid and the secondary structure obtained from PDB were incorporated and used to generate candidate solutions. Algorithm 1 shows the general structure of the proposed method. We represent an individual as a vector of size n belonging to the domain of real numbers (using their floating point representation). Each position of this vector represents a residue's main-chain and side-chain set of torsion angles (Fig. 1).

Data: The APL, the sequence of AA and the respective sequence of SS.
Result: The best individual

$pop_0 \leftarrow$ *Generate* the first population using the angles values returned by Algorithm 2;
for *i = 1 to NumberofGenerations* **do**
  Sort individuals and define classes A, B and C;
  $pop_i(classA) \leftarrow pop_{i-1}(classA)$;
  for j *= 1 to* $|B|$ **do**
    $parent_1 \leftarrow GetIndividual(classA)$;
    $parent_2 \leftarrow GetIndividual(classB + classC)$;
    $offspring \leftarrow Crossover(parent_1, parent_2)$;
    $offspring \leftarrow DiversityControl(offspring)$;
    $pop_i(classB) \leftarrow add(pop_i(classB), offspring)$;
  end
  for j *= 1 to* $|classC|$ **do**
    $pop_i(classC) \leftarrow$ Generate individuals using the angles values returned by Algorithm 2;
  end
end
$Sort(pop_i)$;
$best \leftarrow top(pop_i)$;
return best;

Algorithm 1: GA for the 3-D PSP Problem.

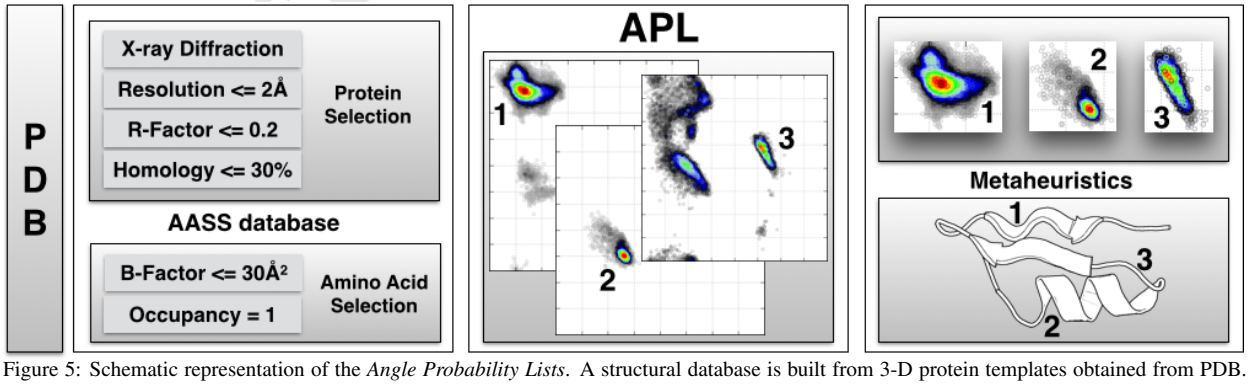


Figure 5: Schematic representation of the *Angle Probability Lists*. A structural database is built from 3-D protein templates obtained from PDB. APLs are computed and used by metaheuristics to improve in the prediction of the three-dimensional structure of proteins. The numbers 1, 2 and 3 represent the β-sheet, α-helix and coil secondary structures, respectively.

For each amino acid residue `aa` of the target protein sequence, torsion angles (ϕ and ψ) are obtained from the `APL` using the steps described in Algorithm 2. Each pair of values for (ϕ, ψ) has an associated probability of being selected.

Data: The APL of a specific AA and the respective SS.
Data: A radius value for the angles range.
Result: The ϕ and ψ angle values for a new individual from the histogram data.

1 *luck* ← *Random* real number in the range 0.0 and 1.0;
2 *edge* ← 0.0;
3 **for** *i = 0 to Number of probabilities* **do**
4 **if** *luck* <= $probability_i$ + *edge* **then**
5 $minimal\_\phi \leftarrow \phi_i - radius$;
6 $maximal\_\phi \leftarrow \phi_i + radius$;
7 $minimal\_\psi \leftarrow \psi_i - radius$;
8 $maximal\_\psi \leftarrow \psi_i + radius$;
9 $aminoacid\_\phi$ ← *Random* number between $minimal\_\phi$ and $maximal\_\phi$;
10 $aminoacid\_\psi$ ← *Random* number between $minimal\_\psi$ and $maximal\_\psi$;
11 break;
12 **else**
13 *edge* ← *edge* + $probability_i$;
14 **end**
15 **end**
16 **return** aminoacid_ϕ, aminoacid_ψ;

Algorithm 2: Getting ϕ and ψ for new individual.

Torsion angle pairs with greater probability are more likely to be selected. Once we have a pair, we add a random real value between − `radius` and + `radius` to each torsion angle ϕ and ψ to reach surrounding regions in the Ramachandran plot (Alg. 2, lines 5-11). We consider a `radius = 1`° to create the initial population. Side-chain torsion angles (χ angles) are generated using our `AASS` dataset, and the angles are selected with amino acid residue dependency, secondary structure dependency and the backbone angles (ϕ and ψ angles) dependency. Population size was fixed on `100` individuals and was structured in classes (Ericsson et al., 2002). The fittest `10%` of the individuals are in class `A`, the `50%` next fit ones are in class `B`, and the remaining `40%` are in class `C`.

The crossover operation, showed in Algorithm 1, creates a new individual called *offspring* using information from two selected parents, Alg. 1 - line 8). $Parent_1$ and $parent_2$ are chosen at random from classes `A` and `B+C`, respectively (Alg. 1 - lines 6 and 7). The *offspring* is composed by 50-70% amino acids residues from $parent_1$. The remaining amino acid residues are obtained from $parent_2$. Then, the *offspring* is added to the population of the next generation.

If the population of a `GA` simulation has lost most of its diversity, the fitness values may not grow further or may take a lot of generations to display some improvement. Taking this in mind, each *offspring* is submitted to a *Diversity Control* (`DC`) operator (Alg. 1 - line 9). The `DC` operator compares the set of dihedral angles of the current *offspring* against the individual solutions from the current population. The *offspring* is maintained if there is no repeated individual; otherwise a new *offspring* is generated.

For the next generation, all individuals from class `A` are automatically promoted. The class `B` is composed of individuals from the crossover function. A new class `C` is entirely created in the same way as the initial population, but with a 2.0° radius. When the new population is complete, the individuals are sorted by their energy value (Section 3.4) in a way that, at the end of each generation, the current best solution is always in the top of the population.

3.3. Particle Swarm Optimization

We implemented a standard Particle Swarm Optimization algorithm that incorporates `APL`. The `PSO` algorithm (Alg. 3) (Eberhart and Kennedy, 1995; Ebenhart, 1995) was inspired by the social behavior of groups of animals such as flocking of birds and a swarm of insects. The algorithm works with a population (swarm) of candidate solutions (particles) generated by the *Angle Probability List*. These particles move in the search space in accordance with a score function (Section 3.4). The movement of the particles is guided by their best position in the search universe, and the best position of the entire swarm. When improved positions are discovered, they will then guide the movements of the swarm. The process is repeated and it is expected that a satisfactory solution will eventually be found. To carry out global optimization of the energy function of proteins the `PSO` was performed as described by Kondov and Berlich (2011). When calculating the velocity, the algorithm takes into account an inertia parameter (constant and equal to 0.4) that along with the previous velocity form the inertial term, which is deterministic. There are two other terms in the equation of particle velocity, both stochastic, which relate to cognitive and social terms. Cognitive term takes into consideration the best position of the particle and a cognitive coefficient (`c1`), while social term takes into consideration the best position of the

swarm and a social coefficient (c2), both coefficients are constant and equal to 2.0. The particle new position is updated by adding the current position and the velocity.

Data: Particles created using the APL.
Result: The best particle

```
particle ← Generate the particles using Algorithm 2;
setOfinitialFitness ← Fitness(Particles);
globalBest ← min(setOfinitialFitness);
c1 = 2.0; c2 = 2.0; w = 0.4; maxTime = 12 hours;
for i in particles do
    i.actualFitness ← Fitness(i);
    if i.actualFitness < i.fitness then
        i.fitness ← actualFitness;
        i.fitnessPosition ← i.actualPosition;
    end
    if i.actualFitness < globalBest.fitness then
        globalBest ← i;
        cognitive ← i.fitnessPosition − i.actualPosition;
        social ← globalBest.position − i.actualPosition;
        term1 ← (c1 ∗ rand ∗ cognitive);
        term2 ← (c2 ∗ rand ∗ social);
        vi ← w ∗ vi−1 + term1 + term2;
        for angle in i.actualPosition do
            positioni = angle + vi;
            i.actualPosition[angle] ← positioni;
        end
    end
    return globalBest
end
```

Algorithm 3: PSO for the 3-D PSP Problem

3.4. Score function

We use a potential energy function to evaluate the candidate solutions of both metaheuristics. This function describes, for the structure of each solution, the internal energy value related to the native/functional state of the protein. The goal for the three-dimensional protein structure prediction problem is to find a conformation with the minimum of potential energy. The energy function incorporates two types of terms: bonded and non-bonded (MacKerrel, 2010). The Rosetta scoring function implemented by PyRosetta (Chaudhury et al., 2010) was used. PyRosetta is a Python-based implementation of the Rosetta (Rohl et al., 2004) molecular modeling suite[3].

Rosetta energy function considers over than 20 energy terms, most of them are derived from knowledge-based potentials (Combs et al., 2013). The function count with Newtonian physics-based terms, including a 6-12 Lennard-Jones potential and a solvation potential. The 6-12 Lennard-Jones potential is split into two terms, an attractive term and a repulsive term, for all *van der* Walls interactions. The Lazaridis-Karplus model (Lazaridis and Karplus, 1999) for implicit solvation is used and penalizes the burial of polar atoms. Interatomic electrostatic interactions are captured through a pair potential and an orientation-dependent hydrogen bond potential for long-range and short-range hydrogen bonding. In addition to the electrostatic terms, Rosetta all-atom scoring function contains terms that dictate side chain conformations according to the Dunbrack rotamer library (Shapovalov and Dunbrack, 2011) for a particular amino acid given a pair of *phi*/*psi* angles, and preference for the *phi*/*psi* angles in a Ramachandran plot (Combs et al., 2013). The final energy value is the sum of all weighted energy terms. These weights were assigned by *talaris2013* (Rohl et al., 2004; Song et al., 2011; Leaver-Fay et al., 2013; O'Meara et al., 2015), which is currently the default weights for scoring full-atom structure in Rosetta.

4. Experiment and Results

We performed two different experiments to test the effectiveness of the *Angle Probability List*s. The first experiment (Section 4.1) aims to test the APL with the two different metaheuristics described in Section 3.2 and 3.3. For the second experiment (Section 4.2) we select a larger set of protein sequences to test APL with the metaheuristic that achieves better results in the first experiment. All algorithms were implemented in Python Language, and tests were executed in a Linux x86_64 environment of a SuperWorkstation with Intel Xeon CPU E5-2650 2.00GHz with 32GB of RAM. The results were analysed in terms of structural and stereochemical quality.

4.1. Experiment I

For the first experiment, a set of 6 target protein sequences were selected from PDB to test the proposed computational strategy. Table 2 presents details of the target protein sequences. Column 2 shows the reference of protein structure, column 3 shows its size and column 4 represents the Secondary Structure Content of each protein. We select these study cases to test our method with different classes of polypeptides with different folding patterns. Due to the complexity of the problem, increasing the size of a selected protein also

[3] https://www.rosettacommons.org

implies increasing the minimum time to achieve a suitable structure, in light of this limitation, larger proteins were not selected.

Table 2: Target protein sequences. The size of the amino acid sequences varies from 33-85 amino acid residues.

PDB ID	Reference	Size	SS Component
1ROP	Banner et al. (1987)	56	2 helices
1ZDD	Starovasnik et al. (1997)	34	2 helices
2KDL	Alexander et al. (2009)	85	3 helices
1UTG	Morize et al. (1987)	70	5 helices
2M7T	Kryshtafovych et al. (2014b)	33	1 helix 1 sheet
1CRN	Teeter (1984)	46	3 helices 1 sheet

Protein sequences of Table 2 were submitted to the metaheuristics described in Sections 3.2 and 3.3, with the objective of validate the APL approach. We also tested both algorithms with angles randomly chosen from an [+180.0°, -180.0°] interval. For each target sequence, the algorithm ran for 15 cycles of 12 hours.

4.1.1. Stereochemical and structural analysis

For stereochemical and structural analysis, we selected the solutions that at the last simulation presents the result with the lowest potential energy. The quality of the predicted structures were evaluated by similarity comparisons with the structures of the experimental proteins obtained from PDB (Eq. 2).

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 amino acid residues in the extremity of structure (N-terminal and C-terminal) were not considered due its high flexibility. The RMSD measure was calculated using PROFIT available in (www.bioinf.org.uk/software/profit).

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

where 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.

Figure 6, shows the best results for the GA and PSO algorithms with and without using APL. RMSD values of the predicted structure with the lowest energy observed among the 15 runs are shown below of each Cartoon representation in Figure 6. We also report the lowest RMSD value (between parenthesis) found at the final generation of each metaheuristic. Through visual inspection, it is possible to notice that predicted structures using the APL are well-formed and has similar fold to the experimental-determined structure. Analysing the RMSD values, is possible to observe that the use of APL contributes to finding better solutions. Predicted structures without using APL present, in some cases, RMSD values close to the predicted structures using APL, nevertheless, when we analyze the structural quality of these structures we note that they are not well-formed.

Table 3 (columns 2-3), presents the energy value of the structures in the Figure 6. In these columns, between parenthesis, it is the energy of the structure with lowest RMSD. Columns 4-5 shows the average energy of all 15 runs. The highlighted results (boldface) represent the best energy found between the two metaheuristics. Observing Table 3, it is possible to affirm that the use of APL in both metaheuristics converged to lower energy values, however, this energy does not always correspond to the lowest RMSD value presented in Figure 6. In terms of energy (Tab. 3), the Genetic

Table 3: Energy values of predicted structures from GA and PSO algorithms with and without the APL. Energy values are expressed in Kcal/mol. Values between parenthesis represent the energy of predicted structure with lowest RMSD (columns 2-3).

PDB ID	Lowest Energy		Average (STD) Energy	
	With APL	Without APL	With APL	Without APL
1ROP-GA	**-73.5** - (-69.8)	332.6 - (355.1)	**-70.1** (±1.9)	353.7 (±13.3)
1ROP-PSO	102.9 - (386.2)	152.8 - (500.4)	139.8 (±17.1)	485.0 (±64.3)
1ZDD-GA	**-40.4** - (-34.4)	230.6 - (230.6)	**-36.0** (±2.6)	274.5 (±33.6)
1ZDD-PSO	52.3 - (160.7)	95.8 - (273.4)	98.4 (±23.2)	239.3 (±40.0)
2KDL-GA	**-55.1** - (-43.2)	300.9 - (343.4)	**-49.6** (±4.0)	321.9 (±16.7)
2KDL-PSO	130.1 - (332.8)	164.8 - (498.8)	163.6 (±23.2)	444.1 (±61.7)
1UTG-GA	**-52.4** - (-50.3)	1139.4 - (1260.4)	**-49.0** (±2.3)	1314.4 (±140.9)
1UTG-PSO	668.3 - (914.1)	1112.9 - (1474.3)	1014.1 (±223.1)	1317.0 (±314.0)
2M7T-GA	**-21.3** - (-19.7)	125.7 - (222.9)	**-17.8** (±2.2)	225.5 (±50.6)
2M7T-PSO	26.5 - (60.5)	26.5 - (191.3)	36.7 (±6.5)	173.1 (±46.3)
1CRN-GA	**-22.0** - (-22.0)	467.7 - (741.7)	**-18.2** (±2.9)	676.7 (±121.3)
1CRN-PSO	77.4 - (339.0)	230.9 - (481.8)	144.4 (±51.9)	478.3 (±102.4)

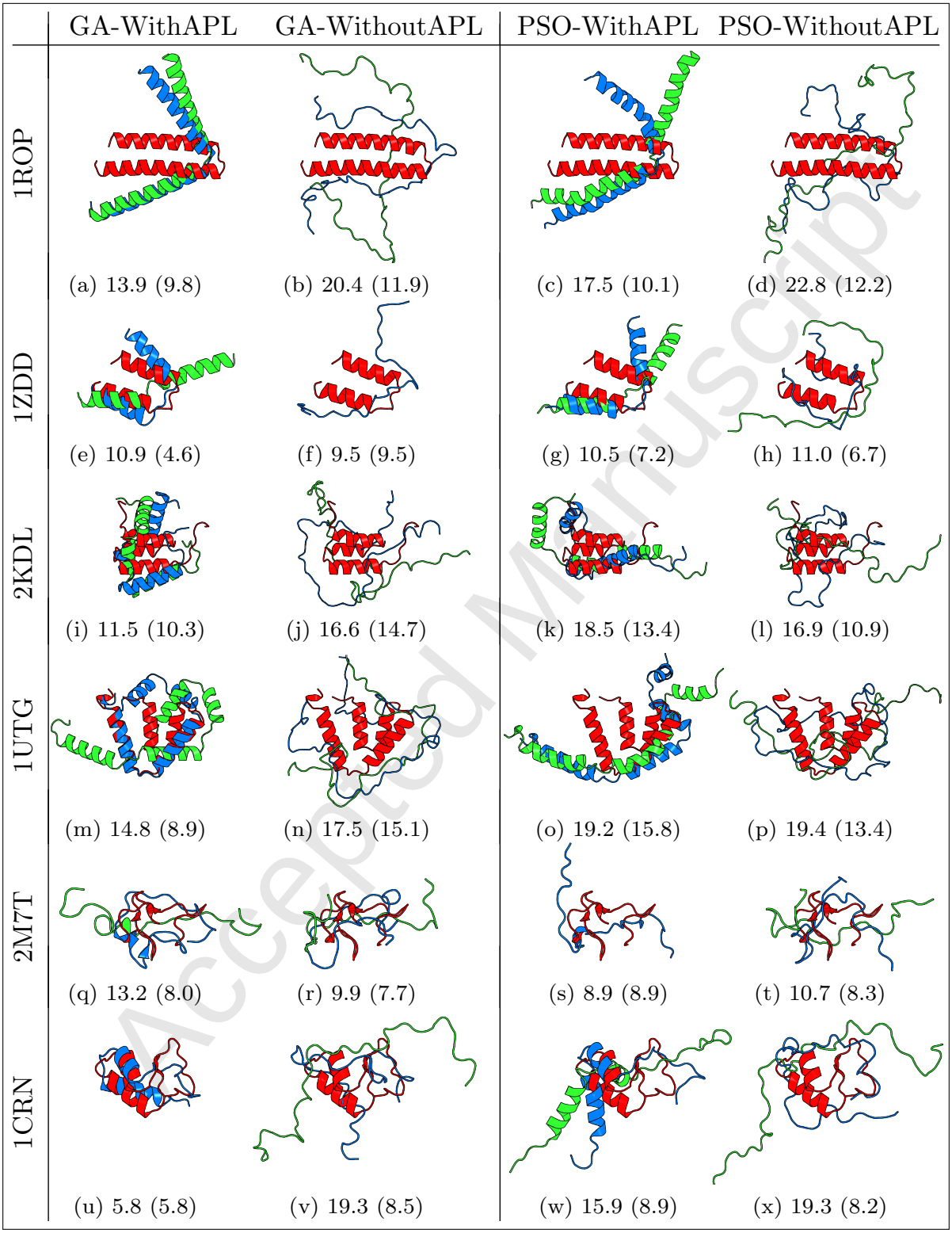


Figure 6: Cartoon representation of the experimental (red), lowest energy (green) and lowest RMSD (blue) 3-D structures for GA and PSO with and without using APL. The C_α of the experimental and the predicted 3-D structure are superimposed. The labels represent the RMSD (Å) value of the lowest energy, and between parenthesis the lowest RMSD (Å) value.

Algorithm using the APL achieved the lowest values in each one of the 6 target protein sequences. In terms of lowest RMSD (Fig. 6), only the structure with PDB ID 2M7T achieved better results without using the APL, all the other cases the GA solution using the APL reached lowest RMSD values. Table 3 demonstrates that the standard deviation (STD) of the average energy in the both metaheuristics presents lower values when APL was used.

Secondary structure analysis of the predicted structure with lowest energy using APL were performed with STRIDE (Heinig and Frishman, 2004). In this analysis we compare the secondary structure contents of the predicted structures against the secondary structure of the experimental structures. We calculate the percentage of correctly classified secondary structure of residues using Q-index (Eq. 3). The secondary structure 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. Table 4 shows the Q-index values for each final state computed by Equation 4.

$$Q_i(\%) = \frac{\#\ of\ AA\ correctly\ predicted_i}{\#\ of\ AA\ in\ class_i} \times 100 \qquad (3)$$

$$Q4(\%) = \frac{\sum \#\ of\ AA\ correctly\ predicted_i}{\sum \#\ of\ AA\ in\ class_i} \times 100 \qquad (4)$$

where $i \epsilon \{H, E, T, C\}$.

As can be observed in Table 4, helices secondary structures was well predicted by both metaheuristics using APL (98.96%), however, sheets secondary structure proved to be difficult to perform in our method. These sheets structures are not well formed due to the coils and turns regions that have a huge search space, even with knowledge from APL. Structures without the APL information, both metaheuristics failed to predict any regular secondary structure (helix or sheet) (Fig. 6).

Distribution of the amino acid residues in the Ramachandran plot and the stereochemical quality of the 3-D structures predicted by our method were analyzed by MolProbity (Davis et al., 2007; Chen et al., 2009). For this analysis, we consider the structure with the lowest energy obtained with and without the APL in both metaheuristics. Table 5 summarizes the achieved results for the stereochemical analysis. Column 2 represents the percentage of poor χ angles, column 3 and 4 show, respectively, the percentage of backbone angles considered outliers and the percentage of angles in the Favorable region of Ramachandran Plot. Column 5 shows the number of steric clashes (clash scores) and column 6 represents the MolProbity general score, which combines the clash score, evaluation of poor χ angles, and Ramachandran plot. The first value of each column represents the structure with the lowest energy using APL and the second value (between parenthesis) represents the structure with the lowest energy computed without the APL. Analysis of the MolProbity general score reveals that better solutions were found when the APL was used. It is also noticeable that the metaheuristic PSO, implemented using the APL, achieved a little higher accuracy in the scope of Ramachandran plots; otherwise the *GA* using APL produced 0% of Poor Rotamers and Ramachandran Outliers, and a better MolProbity score in all tests.

Table 4: Structural analysis for the lowest energy solution of each metaheuristic, GA and PSO, using the Q-Index measure. (P/E) represents the secondary structure proportion of the predicted (P) and the experimental (E) protein structures.

PDB ID (Size)	% Q_H (P/E)	% Q_E (P/E)	% Q_T (P/E)	% Q_C (P/E)	% Q4
1ROP-GA (56)	100.0% (51/51)	–	–	60.0% (3/5)	96.43%
1ROP-PSO (56)	98.04% (50/51)	–	–	100.0% (5/5)	98.21%
1ZDD-GA (34)	100.0% (26/26)	–	60.0% (3/5)	66.67% (2/3)	91.18%
1ZDD-PSO (34)	100.0% (26/26)	–	0.0% (0/5)	100.0% (3/3)	85.29%
1UTG-GA (70)	94.64% (53/56)	–	–	92.86% (13/14)	94.29%
1UTG-PSO (70)	100.0% (56/56)	–	–	92.86% (13/14)	98.57%
1CRN-GA (46)	100.0% (20/20)	0.0% (0/4)	80.0% (4/5)	64.71% (11/17)	76.09%
1CRN-PSO (46)	100.0% (20/20)	0.0% (0/4)	80.0% (4/5)	47.06% (8/17)	69.57%
2KDL-GA (56)	100.0% (36/36)	–	80.0% (4/5)	66.67% (10/15)	89.29%
2KDL-PSO (56)	100.0% (36/36)	–	0.0% (0/5)	100.0% (15/15)	91.07%
2M7T-GA (33)	100.0% (3/3)	0.0% (0/8)	92.31% (12/13)	77.78% (7/9)	66.67%
2M7T-PSO (33)	100.0% (3/3)	0.0% (0/8)	30.77% (4/13)	55.56% (5/9)	36.36%
Average	98.96% (380/384)	0.0% (0/24)	55.36% (31/56)	75.40% (95/126)	85.76%

Table 5: Stereochemical analysis for the lowest energy solution of each algorithm, GA and PSO, performed by MolProbity. Results between parenthesis represents the solution without using the APL and results outside of parenthesis are the structure predicted using the APL. **Boldface** results represent the best result for each test.

Protein ID	Poor Rotamer %	Ramachandran		Score	
		Outlier%	Favorable%	Clash	MolProbity
1ROP-GA	**0.00%** - (24.00%)	**0.00%** - (9.26%)	**100.00%** - (59.26%)	**0.00** - (44.30)	**0.50** - (4.09)
1ROP-PSO	**18.00%** - (34.00%)	**0.00%** - (25.93%)	**100.00%** - (51.85%)	**14.40** - (73.09)	**2.62** - (4.46)
1ZDD-GA	**0.00%** - (15.62%)	**0.00%** - (6.25%)	**100.00%** - (59.38%)	**0.00** - (70.30)	**0.50** - (4.14)
1ZDD-PSO	**6.25%** - (21.88%)	**0.00%** - (28.12%)	**100.00%** - (40.62%)	**15.82** - (80.84)	**2.31** - (4.41)
1UTG-GA	**0.00%** - (32.31%)	**0.00%** - (27.94%)	**98.53%** - (50.00%)	**2.69** - (83.33)	**1.06** - (4.51)
1UTG-PSO	**23.08%** - (43.08%)	**0.00%** - (44.12%)	**100.00%** - (29.41%)	**72.58** - (168.5)	**3.37** - (4.99)
1CRN-GA	**0.00%** - (13.51%)	**0.00%** - (13.64%)	**97.73%** - (68.18%)	**3.09** - (49.38)	**1.16** - (3.88)
1CRN-PSO	**2.70%** - (27.03%)	**0.00%** - (34.09%)	**100.00%** - (40.91%)	**35.49** - (78.70)	**2.36** - (4.47)
2KDL-GA	**0.00%** - (31.25%)	**0.00%** - (9.26%)	**94.44%** - (64.81%)	**0.00** - (41.30)	**0.88** - (4.11)
2KDL-PSO	**10.42%** - (27.08%)	**0.00%** - (22.22%)	**100.00%** - (53.70%)	**25.00** - (60.87)	**2.66** - (4.30)
2M7T-GA	**0.00%** - (8.33%)	**0.00%** - (9.68%)	**96.77%** - (74.19%)	**0.00** - (48.31)	**0.70** - (3.66)
2M7T-PSO	**0.00%** - (8.33%)	**0.00%** - (9.68%)	**100.00%** - (70.97%)	**4.83** - (36.23)	**1.25** - (3.57)

With the first experiment, we observed that among the two standard implementations of metaheuristics, the Genetic Algorithm obtained better results. Thus, we selected the GA to validate our knowledge-based approach with a larger set of proteins sequences. The results obtained with the second experiment is described in Section 4.2.

4.2. Experiment II

To evaluate the effectiveness of the use of the APL we select a larger set of proteins to test our method with different classes of polypeptides with distinct folding patterns and sizes. These structures were tested for the Knowledge-based Genetic Algorithm (GA) and executed for 15 cycles of 24 hours. Table 6 shows the set of 20 target protein sequences used to test our Knowledge-based approach with the GA. In order to validate the APL approach, we also tested the GA with torsion angles randomly chosen from an [+180.0°, -180.0°] interval. Predicted structures were analyzed in terms of stereochemical and structural quality.

4.2.1. Stereochemical and structural analysis

We calculate the root mean square deviation between the position of experimental and predicted three-dimensional structures. Table 7, column 3, 4 and 7 shows respectively the RMSD value of the predicted structure among the 15 runs with the lowest energy, the lowest RMSD and the average RMSD with its standard deviation. The results were organized into two groups P1 to represent solutions using the APL and P2 to represent predicted solutions without the APL.

Table 6: Target protein sequences. The size of the amino acid sequences varies from 14-76 amino acid residues. The bibliographic references for proteins with PDB ID 3P7K and 2PMR were not found.

PDB ID	Reference	Size	SS Component
3P7K	n/a	45	1 helix
2MTW	Cifuentes et al. (2005)	20	1 helix
1WQC	Chagot et al. (2005)	26	2 helices
2P81	Religa et al. (2007)	44	2 helices
1L2Y	Neidigh et al. (2002)	20	2 helices
3V1A	Der et al. (2012)	48	2 helices
2P6J	Shah et al. (2007)	52	3 helices
2F4K	Kubelka et al. (2006)	33	3 helices
1ENH	Clarke et al. (1994)	54	3 helices
2MR9	Nowicka et al. (2015)	44	3 helices
1AIL	Liu et al. (1997)	70	3 helices
2PMR	n/a	76	3 helices
2JUC	Bonet et al. (2008)	59	4 helices
1K43	Pastor et al. (2002)	14	1 sheet
1DFN	Hill et al. (1991)	30	2 sheets
1D5Q	Vita et al. (1999)	27	1 helix 1 sheet
1ACW	Blanc et al. (1996)	29	1 helix 1 sheet
1Q2K	Cai et al. (2004)	31	1 helix 1 sheet
1AB1	Yamano et al. (1997)	46	2 helices 1 sheet
2P5K	Garnett et al. (2007)	63	3 helices 1 sheet

Columns 2, 5 and 6 show, respectively, the lowest energy found by our method among the 15 runs, the energy of the solution with the lowest RMSD and the average energy with its standard deviation. As can be observed, the use of APL contributes to obtaining the best RMSD and energy values. In average, the APL approach achieved better results in all cases. Structures with

Table 7: Algorithm simulation results. P1 represents the results using the APL and P2 without the APL. The **boldface** numbers are the best results in terms of Lowest Energy, Lowest RMSD, Average of Energy and Average of RMSD for each protein. [a](Jayaram et al., 2012), [b](Maupetit et al., 2010), [c](Kapoor and Travesset, 2013), [d](de Sancho and Rey, 2008), [e](DasGupta et al., 2015), [f](Piana et al., 2011), [g](Moktan et al., 2014), [h](Lu et al., 2009). *structures without comparative results.

PDB ID	Low. Energy Kcal/mol	RMSD (Å)	Lowest RMSD (Å)	Energy Kcal/mol	Avg. Energy Kcal/mol	Avg. RMSD (Å)	Ref.
3P7K-P1	**-69.78**	2.09	**1.54**	-68.20	**-68.37** (±0.82)	**1.94** (±0.22)	4.6[e]
3P7K-P2	181.07	10.14	8.59	218.35	217.97 (±16.43)	10.90 (±1.76)	
2MTW-P1	**-22.72**	2.48	**1.48**	-18.64	**-20.77** (±1.70)	**2.11** (±0.43)	*
2MTW-P2	39.46	5.33	4.61	44.25	43.84 (± 2.24)	5.44 (±0.51)	
1WQC-P1	**-20.08**	5.24	**3.49**	-10.99	**-13.95** (±2.20)	**5.38** (±0.62)	2.5[a]
1WQC-P2	77.56	7.30	5.73	94.90	111.37 (±27.54)	8.40 (±1.44)	
2P81-P1	**-45.95**	8.53	**3.90**	-45.03	**-42.00** (±3.05)	**7.31** (±1.49)	3.4[b]
2P81-P2	255.43	12.22	11.58	284.89	281.63 (±12.76)	15.17 (±2.88)	
1L2Y-P1	**-16.96**	5.28	**1.06**	-14.94	**-15.55** (±0.83)	**4.38** (±1.71)	3.1[e]
1L2Y-P2	109.44	4.86	4.20	128.03	178.99 (±54.74)	5.28 (±0.97)	
3V1A-P1	**-61.02**	10.70	9.79	-55.89	**-57.63** (±1.74)	**11.91** (±1.14)	*
3V1A-P2	219.01	16.41	**9.17**	284.90	276.63 (±29.98)	16.55 (±4.99)	
2P6J-P1	**-52.31**	15.18	**11.16**	-44.78	**-46.73** (±3.87)	**14.17** (±1.34)	2.7[c]
2P6J-P2	303.27	16.89	12.19	360.82	340.57 (±17.02)	18.15 (±3.17)	
2F4K-P1	**-30.73**	6.60	**4.90**	-22.26	**-26.83** (±2.23)	**8.12** (±1.49)	0.6[f]
2F4K-P2	112.35	17.06	7.23	148.93	166.44 (±38.22)	9.91 (±2.63)	
1ENH-P1	**-56.08**	14.99	**10.92**	-51.58	**-51.52** (±1.94)	**14.06** (±2.19)	4.6[a]
1ENH-P2	433.12	20.23	15.72	442.54	521.05 (±46.68)	21.29 (±2.98)	
2MR9-P1	**-55.96**	9.22	**7.74**	-50.09	**-51.35** (±2.74)	**9.30** (±0.99)	*
2MR9-P2	217.98	14.88	10.13	373.47	315.07 (±53.92)	14.63 (±2.30)	
1AIL-P1	**-75.07**	19.57	**12.34**	-74.62	**-71.08** (±3.35)	**18.89** (±3.60)	4.4[a]
1AIL-P2	460.27	24.65	16.82	529.00	553.72 (±65.68)	25.91 (±5.44)	
2PMR-P1	**-81.07**	21.54	**19.22**	-78.17	**-77.31** (±2.39)	**23.44** (±2.74)	6.8[a]
2PMR-P2	421.36	24.83	20.78	472.44	465.88 (±23.47)	28.04 (±4.51)	
2JUC-P1	**-46.17**	18.50	**8.14**	-27.08	**-37.72** (±4.99)	**15.18** (±3.31)	3.0[h]
2JUC-P2	353.05	19.37	13.84	415.18	424.90 (±38.42)	19.96 (±3.97)	
1K43-P1	**-11.10**	3.55	**1.39**	-7.77	**-6.96** (±2.07)	**3.08** (±0.65)	1.1[b]
1K43-P2	32.21	3.97	2.97	39.72	38.22 (± 3.76)	4.30 (±0.85)	
1DFN-P1	**-12.40**	10.21	**6.20**	-6.13	**-6.55** (±2.94)	**7.96** (±1.14)	5.0[a]
1DFN-P2	126.04	9.98	8.07	131.18	157.60 (±37.51)	10.79 (±1.53)	
1D5Q-P1	**-28.95**	6.51	**4.08**	-16.93	**-21.65** (±3.45)	**5.51** (±1.22)	0.6[d]
1D5Q-P2	35.13	8.76	6.76	43.52	43.06 (± 3.71)	9.02 (±1.43)	
1ACW-P1	**-12.97**	10.66	**7.99**	-8.95	**-8.77** (±4.25)	**10.83** (±1.65)	5.3[a]
1ACW-P2	109.24	12.00	10.27	158.09	158.62 (±31.88)	11.87 (±0.98)	
1Q2K-P1	**-24.40**	7.59	**5.64**	-18.29	**-19.12** (±2.64)	**7.57** (±1.21)	4.8[a]
1Q2K-P2	78.08	11.53	6.83	85.90	86.51 (± 5.14)	11.01 (±2.39)	
1AB1-P1	**-28.82**	10.10	**5.53**	-23.51	**-25.96** (±2.14)	**10.29** (±1.75)	4.2[a]
1AB1-P2	365.53	20.84	9.88	425.16	458.31 (±68.28)	15.54 (±3.36)	
2P5K-P1	**-54.78**	13.97	**8.77**	-50.19	**-49.75** (±3.51)	**15.50** (±3.13)	2.3[g]
2P5K-P2	387.89	19.16	17.46	468.39	485.03 (±54.04)	23.50 (±4.31)	

simpler fold, like 3P7K and 2MTW, and structures with small size, such as 1K43 and 1L2Y, presents the best results in terms of RMSD. Moreover, structures with a complex folding, such as 2P5K and 1AB1, and structures with larger size, like 2PMR and 1AIL, resulted in higher RMSD, however always presents better folding when using the APL. Highlighted results in Table 7, show the better score between P1 and P2.

Our primary goal in this work was to validate the use of the APL approach in metaheuristics developed for the PSP problem. We implemented a standard version of GA to test the effect of using APL in these methods. Even though the difficult to compare our method with other predictors, column Ref. of Table 7 shows the RMSD values achieved by other prediction methods. All these methods implement more sophisticated metaheuristics. Although, our method achieved comparable results in terms of secondary structure content and RMSD.

Predicted three-dimensional structures were analysed in terms of secondary structure content. We ran STRIDE (Heinig and Frishman, 2004) for the lowest energy predicted structure. We calculated the percentage of the correctly classified secondary structure of residues using Q-index described in Equation 3. Table 8 summarizes the achieved results which reveals similar conclusion as Table 4 (Section 4.1.1). As can be observed APL provides good results for helix prediction (97.01%), and regular results in the prediction of irregular secondary structure such as coil (72.57%).

Visual inspection of predicted structures presented in Figure 7 reveals that helix structures are well formed. When comparing the topology of the protein backbone of the predicted structures against the experimental ones it is possible to observe that the topologies are similar. We observed that the APL approach combined with the standard implementation of GA did not present good results for turn secondary structure, which affect directly the formation of β-sheet secondary structures. Figure 7 shows the predicted structures with the lowest energy (green), the predicted structure with the lowest RMSD (blue) and the experimental structural (red). We observed that for tests without APL there was no formation of any regular secondary structures. Thus, we can affirm that the APL approach corroborates for the prediction of native-like structures.

We also analyzed the stereochemical quality of predicted structures using MolProbity (Davis et al., 2007). Table 9 summarizes the stereochemical results for Clash and *MolProbity* score. Predicted structures with APL have a small number of bad contacts. As can be noticed, structures using APL (highlighted in boldface) achieved better results when compared against predicted structures without

Table 8: Structural analysis using the Q-Index measure. Analysis of the secondary structure contents of the predicted (P) and the experimental (E) protein structures.

PDB ID (Size)	% Q_H (P/E)	% Q_E (P/E)	% Q_T (P/E)	% Q_C (P/E)	% Q4
3P7K (45)	100.0% (44/44)	–	–	100.0% (1/1)	100.0%
2MTW (20)	100.0% (12/12)	–	–	50.0% (4/8)	80.0%
1WQC (26)	94.44% (17/18)	–	–	12.5% (1/8)	69.23%
2P81 (44)	100.0% (27/27)	–	0.0% (0/5)	91.67% (11/12)	86.36%
1L2Y (20)	91.67% (11/12)	–	–	62.5% (5/8)	80.0%
3V1A (48)	100.0% (38/38)	–	66.67% (4/6)	50.0% (2/4)	91.67%
2P6J (52)	100.0% (33/33)	–	28.57% (2/7)	75.0% (9/12)	84.62%
2F4K (33)	66.67% (14/21)	–	–	83.33% (10/12)	72.73%
1ENH (54)	100.0% (38/38)	–	–	87.5% (14/16)	96.3%
2MR9 (44)	100.0% (30/30)	–	55.56% (5/9)	100.0% (5/5)	90.91%
1AIL (70)	98.33% (59/60)	–	–	90.0% (9/10)	97.14%
2PMR (76)	100.0% (63/63)	–	0.0% (0/4)	77.78% (7/9)	92.11%
2JUC (55)	85.71% (30/35)	–	23.08% (3/13)	85.71% (6/7)	70.91%
1K43 (14)	–	0.0% (0/6)	100.0% (5/5)	100.0% (3/3)	57.14%
1DFN (30)	–	0.0% (0/16)	77.78% (7/9)	80.0% (4/5)	36.67%
1D5Q (27)	100.0% (11/11)	50.0% (4/8)	100.0% (2/2)	83.33% (5/6)	81.48%
1ACW (29)	77.78% (7/9)	0.0% (0/10)	40.0% (2/5)	80.0% (4/5)	44.83%
1Q2K (31)	100.0% (11/11)	0.0% (0/8)	75.0% (3/4)	50.0% (4/8)	58.06%
1AB1 (46)	100.0% (20/20)	0.0% (0/4)	60.0% (3/5)	41.18% (7/17)	65.22%
2P5K (63)	100.0% (36/36)	0.0% (0/10)	66.67% (2/3)	92.86% (13/14)	80.95%
Average	97.01% (552/569)	6.45% (4/62)	49.35% (38/77)	72.57% (127/175)	81.65%

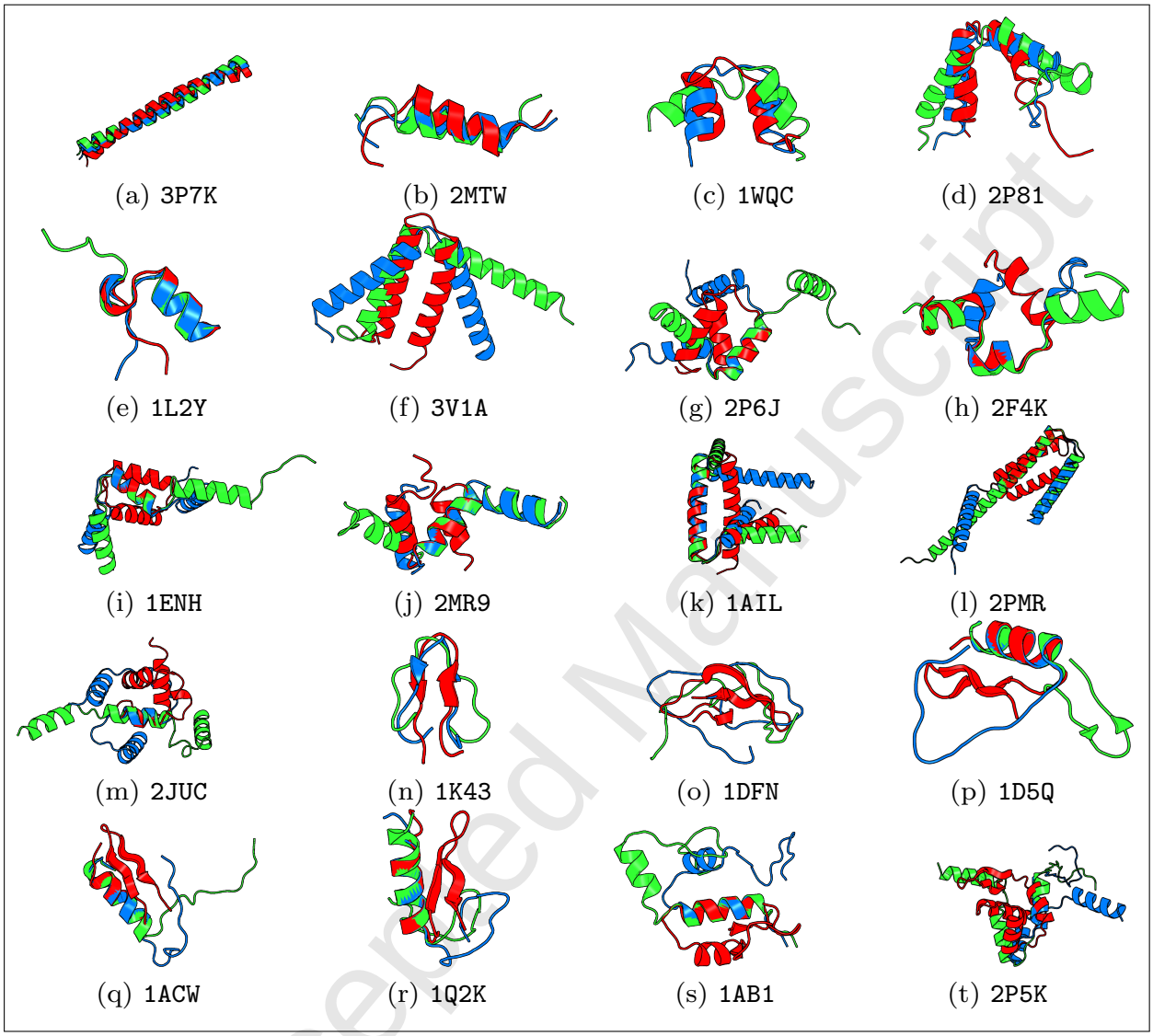


Figure 7: Graphic representation of the experimental (red), lowest Energy (green) and lowest RMSD (blue) structures. The C_α of the experimental and the predicted 3-D structure are fitted. Amino acid side-chain are not shown for clarity. Graphic representation was prepared with PyMOL.

the APL (between parenthesis). It is also noticeable that our method produced 0% of Poor Rotamers and Ramachandran Outliers for all tests using the APL, and great values of Ramachandran Favorable region. The MolProbity score corroborate to conclude that using the APL generate better structures if compared to results without the *Angle Probability List*.

5. Conclusion and further work

Predicting the correct 3-D structure of a protein molecule only from its sequence of amino acid residues is an arduous task. There is an increasing need for new strategies to identify, extract, represent and use structural data from experimentally determined 3-D protein structures. In this paper, we introduce the concept of *Angle Probability Lists* (APL) to represent structural information from the *Protein Data Bank*. We observed that a given secondary structure can have different conformational patterns depending on the type of amino acid residue. APL contains the conformational preference of each amino acid residue according to the amino acid type and its secondary structure. These patterns are crucial to the development of better knowledge-based protein structure prediction methods.

Table 9: Stereochemical analysis for the lowest energy solution of each protein performed by MolProbity. Results between parenthesis represents the solution without using APL and results outside of parenthesis are the structure predicted using the APL. **Boldface** results are the best score for each test.

Protein ID	Poor Rotamer %	Ramachandran		Score	
		Outlier%	Favorable%	Clash	MolProbity
3P7K	**0.00%** - (20.51%)	**0.00%** - (2.33%)	**100.00%** - (72.09%)	**0.00** - (50.06)	**0.50** - (3.99)
2MTW	**0.00%** - (10.53%)	**0.00%** - (**0.00%**)	**94.44%** - (66.67%)	**0.00** - (15.02)	**0.88** - (3.33)
1WQC	**0.00%** - (**0.00%**)	**0.00%** - (**0.00%**)	**100.00%** - (87.50%)	**2.54** - (12.72)	**1.04** - (2.23)
2P81	**0.00%** - (30.00%)	**0.00%** - (4.76%)	**100.00%** - (76.19%)	**1.28** - (78.21)	**0.85** - (4.26)
1L2Y	**0.00%** - (11.76%)	**0.00%** - (11.11%)	**100.00%** - (88.89%)	**0.00** - (39.47)	**0.50** - (3.47)
3V1A	**0.00%** - (17.07%)	**0.00%** - (13.04%)	**100.00%** - (65.22%)	**0.00** - (35.71)	**0.50** - (3.85)
2P6J	**0.00%** - (30.61%)	**0.00%** - (8.00%)	**98.00%** - (72.00%)	**0.00** - (55.50)	**0.50** - (4.17)
2F4K	**0.00%** - (14.29%)	**0.00%** - (**0.00%**)	**100.00%** - (70.97%)	**0.00** - (24.16)	**0.50** - (3.58)
1ENH	**0.00%** - (34.69%)	**0.00%** - (7.69%)	**98.08%** - (75.00%)	**1.06** - (118.27)	**0.81** - (4.50)
2MR9	**0.00%** - (14.29%)	**0.00%** - (9.52%)	**100.00%** - (73.81%)	**1.49** - (44.78)	**0.89** - (3.81)
1AIL	**0.00%** - (20.97%)	**0.00%** - (5.88%)	**100.00%** - (69.12%)	**3.54** - (69.03)	**1.14** - (4.16)
2PMR	**0.00%** - (30.00%)	**0.00%** - (17.57%)	**100.00%** - (54.05%)	**0.80** - (71.09)	**0.75** - (4.39)
2JUC	**0.00%** - (24.07%)	**0.00%** - (5.66%)	**100.00%** - (62.26%)	**0.00** - (57.63)	**0.50** - (4.18)
1K43	**0.00%** - (27.27%)	**0.00%** - (8.33%)	**91.67%** - (**91.67%**)	**0.00** - (76.27)	**1.00** - (3.94)
1DFN	**0.00%** - (16.67%)	**0.00%** - (7.14%)	**92.86%** - (75.00%)	**0.00** - (57.32)	**0.95** - (3.95)
1D5Q	**0.00%** - (9.52%)	**0.00%** - (**0.00%**)	**100.00%** - (84.00%)	**0.00** - (10.72)	**0.50** - (2.97)
1ACW	**0.00%** - (7.41%)	**0.00%** - (7.41%)	**100.00%** - (77.78%)	**9.62** - (21.63)	**1.51** - (3.25)
1Q2K	**0.00%** - (23.08%)	**0.00%** - (**0.00%**)	**100.00%** - (68.97%)	**0.00** - (37.12)	**0.50** - (3.94)
1AB1	**0.00%** - (13.51%)	**0.00%** - (11.36%)	**100.00%** - (68.18%)	**7.75** - (38.76)	**1.42** - (3.78)
2P5K	**0.00%** - (20.34%)	**0.00%** - (11.48%)	**98.36%** - (63.93%)	**0.96** - (48.17)	**0.79** - (4.04)

We validate the proposed method incorporating the APL into two different metaheuristics. As corroborated by experiments, the use of the *Angle Probability List* produced good structures, in terms of structural and stereochemical analysis, when compared with results without using the APL approach. All tests without the use of APL showed bad non-bonded contacts, which suggests that their absence do not produce conditions for correct folding. Structural quality measurements have been made in terms of RMSD for predicted structures achieved with/without the use of APL. As can be observed, in average, the APL approach produced better results in all cases. In the same way, RMSD values of predicted structures achieved without APL reveals ineffectiveness of both metaheuristics when the APL was not used. Despite the fact that not all secondary structures are well formed, we can observe that the topologies of predicted structures are similar to the experimental structure.

The overall contributions of our work are the following: (a) the use of computational techniques and concepts to develop a new algorithm for a relevant biological problem. Determining experimentally the three-dimensional structure of a protein molecule is both expensive and time consuming, this difficulty has generated a significant discrepancy between the number of sequences (Genome Projects) and known 3-D protein structures. Proteins do most of the work in cells and the knowledge of its structure can give us valuable information about its functions (structure→function); (b) the development of a computational strategy to extract and represent structural information from experimentally determined protein structures. As can be observed, the APL reduces the conformational protein search space enabling metaheuristics to find better solutions. The *Protein Data Bank* represents a rich source of information to be explored by computational methods for the PSP problem; (c) the analysis of conformational preferences of amino acid residues in proteins and its use to 3-D protein structure prediction methods. We observed that when we associate the type of amino acid residue and secondary structure we obtain valuable information about the preferences of this amino acid residues; and finally (d) the development and use of two metaheuristic based on Genetic Algorithms and Particle Swarm Optimization to search the three-dimensional protein conformational space using APL. Metaheuristics presents the ability to find satisfactory solutions with less computational effort than exact methods. The search process guides a subordinate heuristic by intelligently combining different concepts for exploring and exploiting the search space. The APL improves the process of explor-

ing the protein search space allowing to find good solutions in shorter time.

There are several research opportunities to be explored in this field, with relevant multidisciplinary applications in Computer Science, Bioinformatics, Biochemistry, and the Medical Sciences. This work also raises interesting research topics, with a range of applications in Computational Biology and Bioinformatics. For instance, one could apply machine learning techniques to learn about different structural information from the structural database and use it to refine mainly coil and turn regions of the predicted structures. The APL model could be improved by using other structural information from the *Protein Data Bank* considering the accessibility of solvent in the protein templates, using the concept of segments of amino acid residues to construct the probability list, improve the selection of protein templates based on the homology of the target sequence with template candidates, etc. Another one could be development and application of parallel metaheuristics to search the three-dimensional protein search space aims to explore in finer details the roughness of the protein energy landscape.

Availability

APL is freely available on the web at `http://sbcb.inf.ufrgs.br/apl`

Acknowledgments

This work was supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9ch), Brazil. MIP was partially funded by Fondecyt Iniciación 11121288 from Conicyt-Chile.

References

Alexander, P.A., He, Y., Chen, Y., Orban, J., Bryan, P.N., 2009. A minimal sequence code for switching protein structure and function. Proc. Natl. Acad. Sci. U.S.A. 106, 21149–21154.

Andersen, C.A.F., Rost, B., 2003. Secondary Structure Assignment: Structural Bioinformatics. chapter 17. p. 341.

Banner, D., Kokkinidis, M., Tsernoglou, D., 1987. Structure of the cole1 rop protein at 1.7 a resolution. J. Mol. Biol. 196, 657–675.

Battiti, R., Brunato, M., Mascia, F., 2008. 1 ed., Springer Verlag, New York.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bath, T., Weissig, H., Shindyalov, I., Bourne, P., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Blanc, E., Fremont, V., Sizun, P., Meunier, S., Van Rietschoten, J., Thevand, A., Bernassau, J., Darbon, H., 1996. Solution structure of P01, a natural scorpion peptide structurally analogous to scorpion toxins specific for apamin-sensitive potassium channel. Proteins: Struct., Funct., Bioinf. 24, 359–369.

Blum, C., Roli, A., 2003. Metaheuristics in combinatorial optimization: overview and conceptual comparison. Comp. Surv. 35, 268.

Bonet, R., Ramirez-Espain, X., Macias, M., 2008. Solution structure of the yeast URN1 splicing factor FF domain: comparative analysis of charge distributions in FF domain structures-FFs and SURPs, two domains with a similar fold. Proteins: Struct., Funct., Bioinf. 73, 1001–1009.

Bowie, J.U., Luthy, R., Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.

Branden, C., Tooze, J., 1998. Introduction to protein structure. 2 ed., Garlang Publishing Inc., New York, USA.

Bryant, S.H., Altschul, S., 1995. Statistics of sequence-structure threading. Curr. Opin. Struct. Biol. 5, 236–244.

Cai, Z., Xu, C., Xu, Y., Lu, W., Chi, C., Shi, Y., Wu, J., 2004. Solution structure of BmBKTx1, a new BKCa1 channel blocker from the Chinese scorpion Buthus martensi Karsch. Biochemistry 43, 3764–3771.

Chagot, B., Pimentel, C., Dai, L., Pil, J., Tytgat, J., Nakajima, T., Corzo, G., Darbon, H., Ferrat, G., 2005. An unusual fold for potassium channel blockers: Nmr structure of three toxins from the scorpion opisthacanthus madagascariensis. Biochem. J. 388, 263–271.

Chaudhury, S., Lyskov, S., Gray, J., 2010. Pyrosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. Bioinformatics 26, 689–691.

Chen, V., Arendall, W., Headd, J., Keedy, D., Immormino, R., Kapral, G., Murray, L., Richardson, J., Richardson, D., 2009. Molprobity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr., Sect. D: Biol. Crystallogr. 66, 12–21.

Cifuentes, G., Salazar, L., Vargas, L., Parra, C., Vanegas, M., Cortes, J., Patarroyo, M., 2005. Evidence supporting the hypothesis that specifically modifying a malaria peptide to fit into HLA-DRbeta1*03 molecules induces antibody production and protection. Vaccine 23, 1579–1587.

Clarke, N., Kissinger, C., Desjarlais, J., Gilliland, G., Pabo, C., 1994. Structural studies of the engrailed homeodomain. Protein Sci. 3, 1779–1787.

Combs, S., DeLuca, S., DeLuca, S., Lemmon, G., Nannemann, D., Nguyen, E., Willis, J., Sheehan, J., Meiler, J., 2013. Small-molecule ligand docking into comparative models with rosetta. Nat. Protoc. 8, 1277–1298.

Cutello, V., Narzisi, G., Nicosia, G., 2006. A multi-objective evolutionary approach to the protein structure prediction problem. J. R. Soc., Interface 3, 139–151.

Dandekar, T., Argos, P., 1992. Potential of genetic algorithms in protein folding and protein eng. simulations. Protein Eng. 5, 637–645.

DasGupta, D., Kaushik, R., Jayaram, B., 2015. From Ramachandran Maps to Tertiary Structures of Proteins. J Phys Chem B Accepted to Publications.

Davis, I., Leaver-Fay, A., Chen, V., Block, J., Kapral, G., Wang, X., Murray, L., Arendall, W., Snoeyink, J., Richardson, J., Richardson, D., 2007. Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 35, W375–W383.

Der, B., Machius, M., Miley, M., Mills, J., Szyperski, T., Kuhlman, B., 2012. Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J. Am. Chem. Soc. 134, 375–385.

Dorn, M., Buriol, L., Lamb, L., 2011. A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy, in: IEEE Congress on Evolutionary Computation, pp. 2709–2716.

Dorn, M., Buriol, L., Lamb, L., 2014a. Moirae: A computational strategy to extract and represent structural information from exper-

imental protein templates. Soft Comput. 18, 773–795.
Dorn, M., Inostroza-Ponta, M., Buriol, L.S., Verli, H., 2013. A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides, in: IEEE Congress on Evolutionary Computation, IEEE, Cancun, MX. pp. 1233–1240.
Dorn, M., Barbachan e Silva, M., Buriol, L.S., Lamb, L.C., 2014b. Three-dimensional protein structure prediction: Methods and computational strategies. Comput. Biol. Chem. 53, Part B, 251 – 276.
Ebenhart, R., 1995. Kennedy. particle swarm optimization, in: Proceeding IEEE Inter Conference on Neural Networks, Perth, Australia, Piscat-away, pp. 1942–1948.
Eberhart, R.C., Kennedy, J., 1995. A new optimizer using particle swarm theory, in: Proceedings of the sixth international symposium on micro machine and human science, New York, NY. pp. 39–43.
Ericsson, M., Resende, M., Pardalos, P., 2002. A genetic algorithm for the weight setting problem in ospf routing. J. Comb. Optim. 6, 299–333.
Floudas, C., Fung, H., McAllister, S., Moennigmann, M., Rajgaria, R., 2006. Advances in protein structure prediction and de novo protein design: A review. Chem. Eng. Sci. 61, 966–988.
Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins 23, 566–579.
Garnett, J., Baumberg, S., Stockley, P., Phillips, S., 2007. A high-resolution structure of the DNA-binding domain of AhrC, the arginine repressor/activator protein from Bacillus subtilis. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 63, 914–917.
Glover, F., Kochenberg, G., 2003. Handbook of meta-heuristics, in: International Series in Operations Research and Management Science. volume 57, p. 570.
Goldberg, D., 1989. 1 ed., Kluwer Academic Publishers, Boston.
Granville, V., Krivanek, M., Rasson, J.P., 1994. Simulated annealing: A proof of convergence. IEEE T. Pattern Anal. 16, 652.
Greer, J., 1990. Comparative modeling methods: application to the family of the mammalian serine protease. Proteins 7, 317–334.
Guntert, P., 2004. Automated nmr structure calculation with cyana. Methods Mol. Biol. 278, 353.
Guyeux, C., Côte, N.M.L., Bahi, J.M., Bienia, W., 2014. Is protein folding problem really a np-complete one? first investigations. J. Bioinf. Comput. Biol. 12, 1350017.
Heinig, M., Frishman, D., 2004. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. 32, W500–2.
Hill, C., Yee, J., Selsted, M., Eisenberg, D., 1991. Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization. Science 251, 1481–1485.
Hoque, M., Chetty, M., Dooley, L., 2006. A guided genetic algorithm for protein folding prediction using 3d hydrophobic-hydrophilic model, in: IEEE Congress on Evolutionary Computation, Vancouver, Canada. pp. 2339–2346.
Hoque, M., Chetty, M., Sattar, A., 2009. Genetic algorithm in *ab initio* protein structure prediction using low resolution model: A review, in: Sidhu, A.S., Dillon, T. (Eds.), Biomedical Data and Applications. volume 224, pp. 317–342.
Hovmoller, T., Ohlson, T., 2002. Conformation of amino acids in protein. Acta Crystallogr. 58, 768–776.
Jayaram, B., Dhingra, P., Lakhani, B., Shekhar, S., 2012. Bhageerath - targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction. J. Chem. Sci. 124, 83–91.
Johnson, M., Srinivasan, N., Sowdhamini, R., Blundell, T., 1994. Knowledge-based protein modeling. Crit. Rev. Biochem. 29, 1–68.
Jones, D., Taylor, W., Thornton, J., 1992. A new approach to protein fold recognition. Nature 358, 86–89.
Kapoor, A., Travesset, A., 2013. Folding and stability of helical bundle proteins from coarse-grained models. Proteins: Struct., Funct., Bioinf. 81, 1200–1211.
Kennedy, J., 2003. The particle swarm: social adaptation of knowledge. IEEE Press, New York.
Khalili, M., Liwo, A., Jagielska, A., Scheraga, H., 2005. Molecular dynamics with the united-residue model of polypeptide chains. ii. langevin and berendsen-bath dynamics and tests on model alpha-helical systems. J. Phys. Chem. B 109, 13798–13810.
Kirkpatrick, S., Gelatt, C., Vecchi, M., 1983. Optimization by simulated annealing. Science 220, 671.
Kolinski, A., 2004. Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. 51, 349–371.
Kondov, I., 2013. Protein structure prediction using distributed parallel particle swarm optimization. Nat. Comput. 12, 29–41.
Kondov, I., Berlich, R., 2011. Protein structure prediction using particle swarm optimization and a distributed parallel approach, in: Proceedings of the 3rd workshop on Biologically inspired algorithms for distributed systems, ACM. pp. 35–42.
Kryshtafovych, A., Fidelis, K., Moult, J., 2014a. CASP10 results compared to those of previous CASP experiments. Proteins: Struct., Funct., Bioinf. 82, 164–174.
Kryshtafovych, A., Moult, J., Bales, P., Bazan, J.F., Biasini, M., Burgin, A., Chen, C., Cochran, F.V., Craig, T.K., Das, R., Fass, D., Garcia-Doval, C., Herzberg, O., Lorimer, D., Luecke, H., Ma, X., Nelson, D.C., van Raaij, M.J., Rohwer, F., Segall, A., Seguritan, V., Zeth, K., Schwede, T., 2014b. Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10. Proteins 82 Suppl 2, 26–42.
Kubelka, J., Chiu, T.K., Davies, D.R., Eaton, W.A., Hofrichter, J., 2006. Sub-microsecond protein folding. J. Mol. Biol. 359, 546–553.
Lander, E., Waterman, M., 1999. The secrets of life: a mathematician's introduction to Molecular Biology. National Academy Press, Washington D. C., USA.
Lazaridis, T., Karplus, M., 1999. Effective energy functions for protein structure prediction. Curr. Opin. Struct. Biol. 10, 139–145.
Le Grand, S., Merz Jr., K., 1993. The application of the genetic algorithm to the minimization of potential energy functions. J. Global Optim. 3, 49–66.
Leaver-Fay, A., O'Meara, M., Tyka, M., Jacak, R., Song, Y., Kellogg, E., Thompson, J., Davis, I., Pache, R., Lyskov, S., Gray, J., Kortemme, T., Richardson, J., Havranek, J., Snoeyink, J., Baker, D., Kuhlman, B., 2013. Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol. 523, 109.
Lehninger, A., Nelson, D., Cox, M., 2005. Principles of Biochemistry. 4 ed., W.H. Freeman, New York, USA.
Lesk, A.M., 2002. Introduction to Bioinformatics. 1 ed., Oxford University Press Inc., New York, USA.
Lesk, A.M., 2010. Introduction to Protein Sci. 2 ed., Oxford University Press, New York.
Levinthal, C., 1968. Are there pathways for protein folding? J. Chim. Phys. Phys.-Chim. Biol. 65, 44–45.
Liljas, A., Liljas, L., Pskur, J., Lindblom, G. amd Nissen, P., Kjeldgaard, M., 2009. Textbook of structural biology. World Scientific Printers, Singapore.
Lin, C., Hsieh, M., 2009. An efficient hybrid taguchi-genetic algorithm for protein folding simulation. Expert Syst. with Applic. 36, 12446–12453.
Liu, J., Lynch, P., Chien, C., Montelione, G., Krug, R., Berman, H., 1997. Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. Nat. Struct. Biol. 4, 896–899.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C.A., Krieger, M., Scott, M., 1990. Molecular Cell Biology. 5 ed., Scientific American Books, W.H. Freeman, New York, USA.

Lu, M., Yang, J., Ren, Z., Sabui, S., Espejo, A., Bedford, M.T., Jacobson, R.H., Jeruzalmi, D., McMurray, J.S., Chen, X., 2009. Crystal structure of the three tandem {FF} domains of the transcription elongation regulator {CA150}. J. Mol. Biol. 393, 397 – 408.

Luke, S., 2009. Essentials of metaheuristics. 1 ed., Lulu.

MacKerrel, A., 2010. Empirical force fields. Springer. chapter 2. pp. 45–69.

Martí-Renom, M., Stuart, A., Fiser, A., Sanchez, A., Mello, F., Sali, A., 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291–325.

Maupetit, J., Derreumaux, P., Tuffery, P., 2010. A fast method for large-scale de novo peptide and miniprotein structure prediction. J. Comput. Chem. 31, 726–738.

Meissner, M., Schneider, G., 2007. Protein folding simulation by particle swarm optimization. Open Struct. Biol. J 1, 1–6.

Moelbert, S., Emberly, E., Tang, C., 2004. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. Protein Sci. 13, 752–762.

Moktan, H., Guiraldelli, M.F., Eyster, C.A., Zhao, W., Lee, C.Y., Mather, T., Camerini-Otero, R.D., Sung, P., Zhou, D.H., Pezza, R.J., 2014. Solution structure and DNA-binding properties of the winged helix domain of the meiotic recombination HOP2 protein. J. Biol. Chem. 289, 14682–14691.

Morize, I., Surcouf, E., Vaney, M.C., Epelboin, Y., Buehner, M., Fridlansky, F., Milgrom, E., Mornon, J.P., 1987. Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 A resolution. J. Mol. Biol. 194, 725–739.

Mucherino, A., Seref, O., 2009. Modeling and solving real life global optmization problems with meta-heuristic methods. Adv. Mod. Agr. Syst. 25, 1.

Neidigh, J., Fesinmeyer, R., Andersen, N., 2002. Designing a 20-residue protein. Nat.Struct.Biol. 9, 425–430.

Nowicka, U., Zhang, D., Walker, O., Krutauz, D., Castaneda, C., Chaturvedi, A., Chen, T., Reis, N., Glickman, M., Fushman, D., 2015. DNA-damage-inducible 1 protein (Ddi1) contains an uncharacteristic ubiquitin-like domain that binds ubiquitin. Structure 23, 542–557.

O'Meara, M., Leaver-Fay, A., Tyka, M., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., Kuhlman, B., 2015. A combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. J. Chem. Theory Comput. .

Osguthorpe, D., 2000. Ab initio protein folding. Curr. Opin. Struct. Biol. 10, 146–152.

Osman, I., Kelly, J., 1996. Metaheuristics. An overview. 1 ed., Kluwer, Boston.

Osman, I., Laporte, G., 1996. Metaheuristics: A bibliography. Ann. Oper. Res. 63, 511.

Park, S., 2005. A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. Genome Inf. 16, 104–113.

Pastor, M., Lpez de la Paz, M., Lacroix, E., Serrano, L., Prez-Pay, E., 2002. Combinatorial approaches: A new tool to search for highly structured beta-hairpin peptides. Proc. Natl. Acad. Sci. 99, 614–619.

Pauling, L., Corey, R., 1951. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. U. S. A. 37, 251–256.

Pauling, L., Corey, R., Branson, H., 1951. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U. S. A. 37, 205–211.

Pedersen, J., Moult, J., 1997. Protein folding simulations with genetic algorithms and a detailed molecular description. J. Mol. Biol. 269, 240–259.

Piana, S., Lindorff-Larsen, K., Shaw, D., 2011. How robust are protein folding simulations with respect to force field parameterization? Biophys. J. 100, L47 – L49.

Ramachandran, G., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. 23, 238–438.

Religa, T., Johnson, C., Vu, D., Brewer, S., Dyer, R., Fersht, A., 2007. The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain. Proc. Natl. Acad. Sci. U.S.A. 104, 9272–9277.

Resende, M., Ribeiro, C., Glover, F., Martí, R., 2010. Scatter search and path-relinking: Fundamentals, advances, and applications, in: Gendreau, M., Potvin, J.Y. (Eds.), Handbook of Metaheuristics. Springer, New York, pp. 87–107.

Rohl, C., Strauss, C., Misura, K., Baker, D., 2004. Protein structure prediction using rosetta. Methods Enzymol. 383, 66–93.

Sánchez, R., Sali, A., 1997. Advances in comparative protein-structure modeling. Curr. Opin. Struct. Biol. 7, 206–214.

de Sancho, D., Rey, A., 2008. Energy minimizations with a combination of two knowledge-based potentials for protein folding. J. Comput. Chem. 29, 1684–1692.

Scheef, E., Fink, J., 2003. Fundamentals of protein structure: Structural Bioinformatics. chapter 2. p. 15.

Shah, P., Hom, G., Ross, S., Lassila, J., Crowhurst, K., Mayo, S., 2007. Full-sequence computational design and solution structure of a thermostable protein variant. J. Mol. Biol. 372, 1–6.

Shapovalov, M., Dunbrack, R., 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19, 844–858.

Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J., Baker, D., 2011. Structure-guided forcefield optimization. Proteins 79, 1898–1909.

Srinivasan, R., Rose, G., 1995. Linus - a hierarchic procedure to predict the fold of a protein. Proteins 22, 81–99.

Starovasnik, M.A., Braisted, A.C., Wells, J.A., 1997. Structural mimicry of a native protein by a minimized binding domain. Proc. Natl. Acad. Sci. U.S.A. 94, 10080–10085.

Sun, S., 1995. A genetic algorithm that seeks native states of peptides and proteins. Biophys. J. 69, 340–355.

Talbi, E.G., 2009. 1 ed., Wiley, Hoboken.

Teeter, M.M., 1984. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. Proc. Natl. Acad. Sci. U.S.A. 81, 6014–6018.

Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M.I., Dunbrack Jr, R.L., 2010. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. PLoS Comput. Biol. 6, e1000763.

Tramontano, A., 2006. Protein structure prediction: concepts and applications. 1 ed., John Wiley and Sons, Inc., Weinheim, Germany.

Trelea, I., 2003. The particle swarm optimization algorithm: convergence analysis and parameter selection. Inform. Process. Lett. 85, 317.

Turcotte, M., Muggleton, S., Sternberg, M., 1998. Application of inductive logic programming to discover rules governing the three-dimensional topology of protein structure, in: Proceedings of the International Workshop on Inductive Logic Programming, pp. 53–64.

Turcotte, M., Muggleton, S., Sternberg, M., 2001B. The effect of relational background knowledge on learning of protein three-dimensional fold signatures. Machine Learning 43, 81–96.

Vita, C., Drakopoulou, E., Vizzavona, J., Rochette, S., Martin, L., Menez, A., Roumestand, C., Yang, Y., Ylisastigui, L., Benjouad, A., Gluckman, J., 1999. Rational engineering of a miniprotein that reproduces the core of the cd4 site interacting with hiv-1 envelope glycoprotein. Proc.Natl.Acad.Sci.USA 96, 13091–13096.

Wooley, J., Ye, Y., 2010. A historical perspective and overview of protein structure prediction. Springer. chapter 1. pp. 1–43.
Xia, X., Xie, Z., 2002. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. Mol. Biol. Evol. 19, 58–67.
Yamano, A., Heo, N., Teeter, M., 1997. Crystal structure of Ser-22/Ile-25 form crambin confirms solvent, side chain substate correlations. J. Biol. Chem. 272, 9597–9600.