RESEARCH

Using conformational preferences of amino acid residues and meta-heuristics to predict 3-D protein structures

Bruno Grisci[1], Bruno Borguesan[1]
, Márcio Dorn[1*]
and Mario Inostroza-Ponta[2]

Abstract

Tertiary protein structure prediction *in silico* is currently the most challenging problem in Structural Bioinformatics. In this article we describe a new computational strategy to predict the three-dimensional structures of proteins. The proposed method is an hybrid genetic algorithm with a greedy local search strategy to explore the main chain torsion angles frequency of each amino acid residue of target sequence. The novelty of the proposal comes from the use of the angles frequency information together with a greedy local search mainly focused on the most variable secondary structure regions of a target protein sequence. The proposal was tested with five protein sequences whose sizes vary from 14 to 40 amino acids residues. Stereochemical and structural analysis were performed for each predicted three-dimensional structure and the results were compared to their corresponding experimental ones suggesting that the developed method produces accurate predictions.

Keywords: protein structure prediction; knowledge-base methods; metaheuristics; structural bioinformatics

1 Introduction

Proteins are biological macromolecules that are responsible for the execution of different and important functions in the cell [1, 2]. From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acid residues. Each protein is defined by its unique sequence residues that causes the protein to fold into a particular three-dimensional (3-D) shape. 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 [3, 4]. The challenge arises due to the combinatorial explosion of plausible shapes that a protein sequence can assume [5].

The linear sequence of amino acid residues is known as the protein primary structure. The polypeptide chain is very flexible and can assume a large number of spatial conformations. 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 some preferred conformations like alpha-helices, beta-sheets, beta-turns, among others [7]. In different proteins, helices and sheets are combined in many ways, to create different spatial arrangements of the chain, and also distinct patterns of interactions between helices and sheets [6]. This is called the protein tertiary structure and represents the functional/native state of the protein. Finally, many proteins contain more than one subunit, these may be multiples copies of the same polypeptide chain. These assemble of subunits of a macromolecule is called the quaternary structure.

The difficulty in determining and finding out the 3-D structure of proteins has generated a large discrepancy between the volume of data (sequences of amino acid residues) generated by the Genome Projects[1] and the number of 3-D structures of proteins which are known. Currently, there are about 102,000 of 3-D protein structures stored in the PDB and about 174 millions of "protein sequences". Even tough this proportion (≈0.06%) is really small, it represents a rich

*Correspondence: mdorn@inf.ufrgs.br
[1]Institute of Informatics, Federal University of Rio grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre, RS, Brazil
Full list of author information is available at the end of the article

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

source of information to be explored to predict the tertiary structure of protein sequences. Conformational preferences of amino acid residues and common secondary structure information can be acquired from protein templates stored in PDB. These common secondary structures have the particularity that some of them are more conserved, in terms of its conformation, than others (for example helices and turns) [7]. Having together known 3-D protein structures and more conserved secondary structure, build a rich source of data that combined in a proper way can be used to create an algorithm to predict the 3-D shape of a protein.

Determining the 3-D structure of a protein is both experimentally expensive (due to the costs associated to crystallography, electron microscopy or NMR), and time consuming [8]. The 3-D PSP problem was also classified as a NP-complete problem [9]. Several computational strategies and algorithms have been proposed as a solution to the PSP problem [10, 11, 12, 2]. These methods can be classified in one of four classes [13]: (*i*) first principle methods without database information [12]; (*ii*) first principle methods with database information [14, 15]; (*iii*) threading or fold recognition methods [16, 17, 18, 19] and (*iv*) comparative modelling methods [20, 21]. Group *iv* and *iii* are often referenced as knowledge-based methods. These methods are able of performing fast and effective prediction of protein 3-D structures when known template structures and fold libraries are available [22]. Analysis of the last Critical Assessment of protein Structure Prediction (CASP)[2] experiments reveals that the best results are achieved by methods which combine principles of the four groups. In this paper we propose a novel algorithm to acquire and uses structural information from PDB. Both conformational preferences of amino acid residues and secondary structure information from protein templates are used. The proposed method is an hybrid genetic algorithm with a greedy local search strategy to explore the main chain torsion angles frequency of each amino acid residue of target sequence. The novelty of the proposal comes from the use of the angles frequency information together with a greedy local search mainly focused on the most variable secondary structure regions of a target protein sequence.

The remainder of the article is structured as follows. Section 2 present basic concepts of protein structure; conformational preferences of amino acid residues in proteins; and meta-heuristics. Section 3 shows the proposed hybrid genetic algorithm with a greedy local search strategy and the developed structural database. Section 4 reports several results illustrating the effectiveness of our method. Section 5 concludes and points out directions for further research.

[2] http://predictioncenter.org

2 Preliminaries

Proteins, structure and representation

A protein can be described by an ordered linear chain of amino acid residues linked by a peptide bond. This 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). Due the planarity of the peptide bond, the conformation of a peptide backbone is mainly described by two torsion angles per amino acid residue [7]: ϕ (phi) and ψ (psi). The set of consecutive main-chain torsion angles describe the internal rotation of the protein [1, 23] and causes the polypeptide to fold into a particular three-dimensional shape.

The Bond C-N has a double bond, getting with the N a partial positive charge and with atom O a partial negative charge not allowing 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[1, 24]. 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 peptide backbone[25, 1, 23]. As a consequence, the possible conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties. 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) [26]. 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 [27].

There are many ways to represent a polypeptide structure. In this article we represent a polypeptide chain by its set of main-chain and side-chain torsion angles. Torsion angles are among the most important local structural parameters that control protein folding, providing the flexibility required for the polypeptide backbone to adopt a certain fold. The main advantage of this representation is the reduced number of variables to control and optimize in order to predict the polypeptide structure.

Conformational preferences of amino acid residues

Secondary structures in proteins are defined by the presence of hydrogen bonds between the amino and carboxyl groups of the polypeptide chain. There are two most common secondary structures: α-helices [28] and β-sheets [29]. 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 present in 3-D structures.

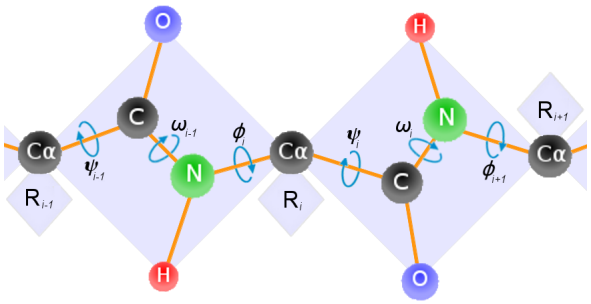


Figure 1: Schematic representation of a peptide. `C` is carbon, `O` is oxygen, `N` is nitrogen and `H` is hydrogen. We represent a polypeptide chain by its set of main-chain and side-chain torsion angles: ϕ (phi), ψ (psi) and χ. The number of χ angles depends on the type of amino acid residue.

α-`helices` are stabilized by one hydrogen bond between the Nitrogen (`N`) atom of a peptide bond and the Oxygen (`O`) atom of the carboxyl group in the fourth amino acid residue of the `N`-terminal region [7, 29]. Each successive turn of helix is held with the adjacent turns by three or four hydrogen bonds. These hydrogen bonds when combined, ensure the stability of the helical structure. The number of amino acid residues in an α-helix is highly variable and may be in the range of 5 to 40 amino acid residues - commonly, α-helices present 10 amino acid residues [29]. The amino acid residues present in a α-helix have their torsion angles (ϕ and ψ) ranging `-73.0` $\leq \phi \leq$ `-54.0` and `-52.0` $\leq \psi \leq$ `-30.0` (Fig. 3a).

β-`sheets` occurs when the polypeptides structures are arranged side by side and they form a regular structure similar to a series of sheets (β-sheets) [28]. The β-sheets consist of extended polypeptide chains with neighboring chains extending parallel/anti-parallel to each other. The amine and carboxyl groups of peptide bonds point towards each other in the same plane, so hydrogen bonding can occur between adjacent polypeptide chains. Amino acid residues in a β-sheet state have their torsion angles in ranging `-172.0` $\leq \phi \leq$ `-46.0` and `91.0` $\leq \psi \leq$ `176.0` (Fig. 3c).

In this work we use `STRIDE` [30, 31] to compute the secondary structure of a protein molecule. `STRIDE` implements a eight secondary structure model: `B`, `E`, `H`, `G`, `I`, `b`, `C` and `T`. This program 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 [32]. The sheet category does not distinguish between parallel and anti-parallel sheets. The minimal sheet (`E`) is composed of two residues each in 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 [32]. 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 met.

Metaheuristics and the PSP problem

When dealing with hard optimization problems, metaheuristics are often used [33, 34] because of their ability to find satisfactory solutions with less computational effort than exact methods. However, metaheuristics do not guarantee an optimal solution. They are used to deal with combinatorial optimization problems in which an optimal solution is sought over a discrete search-space. Among the most important metaheuristics we can highlight Tabu Search (`TS`) [34], Simulated Annealing (`SA`) [35, 36], Genetic Algorithms (`GA`) [37, 38], Particle Swarm Optimization (`PSO`) [39, 40] and Memetic Algorithms (`MA`) [41, 42].

`GA`s are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics [35]. `GA`s are modelled through the use of a population of individuals representing solutions, that undergo selection in the presence of variation-inducing operators such as *mutation* and *recombination*. A fitness function is used to evaluate individuals, and reproductive success varies with fitness. For every individual is calculated a fitness value that indicates how good the solution is [35, 36]. 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 [35].

Genetic algorithms are not new in the field of 3-D protein structure prediction problem, and have already been used in many works on the subject of `GA` [43, 44, 45, 46, 47, 48]. Dorn et.al. [49] combines a genetic algorithm, structural information from `PDB` and a Local Search operator for the 3-D protein structure prediction problem. In Dorn el.al. [50] a genetic algorithm is combined with a structured population, and it is hybridized with a path-relinking procedure that helps the algorithm to scape from local minima. Cutello et al. [51] use a genetic algorithm for solving a multi-objective representation of a protein structure. Park [52] uses a genetic algorithm for fragment assembly to find low-energy conformations. Hoque et.al. [53]

present a comprehensive review of the application of GA in the protein folding problem. Despite the advances, methods that use GAs still have to deal with the challenge of very large conformational search spaces caused by different combination of amino acid residues. In order to address these two challenges we developed a knowledge-base GA which incorporates structural information from the PDB and use it to reduce the protein conformational search space. Section 3 describes the developed meta-heuristic for 3-D protein structure prediction.

3 Material and Methods

In this section we describe the developed methods and algorithms for the PSP problem. Figure 2 shows a schematic representation of the proposed method used to predict the 3-Dstructure of a given protein sequence. It can be divided in two parts: build the conformation database and the hybrid genetic algorithm.

3.1 Structural Database

We selected a set of 6,650 protein structures from the PDB. All structures were experimentally determined by X-ray diffraction with resolution $\leq$ 2.0Å in the PDB until 21 Dec 2013. We remove all structures with R-factor greater than 0.2 and with at most 30% of sequence homology. At this point we have a set of 2,670,182 amino acid residues. We only consider residues (all atoms from the backbone) with b-factor $\leq$ 30Å^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 Hovmöller and Ohlson [26]. We use STRIDE [31] to compute the secondary structure assignments of the remaining amino acid residues. For each amino acid residue we compute the dihedral angles phi and psi. We developed a database schema using PostgreSQL[3] to store all achieved structural information from the PDB.

We analyse the structural information (dihedral angles and secondary structure) stored in own structural data bank. We observe a small number of experimental data with amino acid residues in conformational state of I (π-helix) and b (isolated bridge). Thus, we only consider six conformational states for further analysis: H (α-helix), G (π-helix), E (β-sheet), B (β-bridge), T (Turn) and C (Coil). Table summarizes the number of amino acid residues present in each secondary structure. We compute Ramachandran plots for each set of amino acid residues belonging to a secondary structure and analyse its conformational preferences. Figure 3 shows the Ramachandran plot generated for each secondary structure.

[3] http://www.postgresql.org

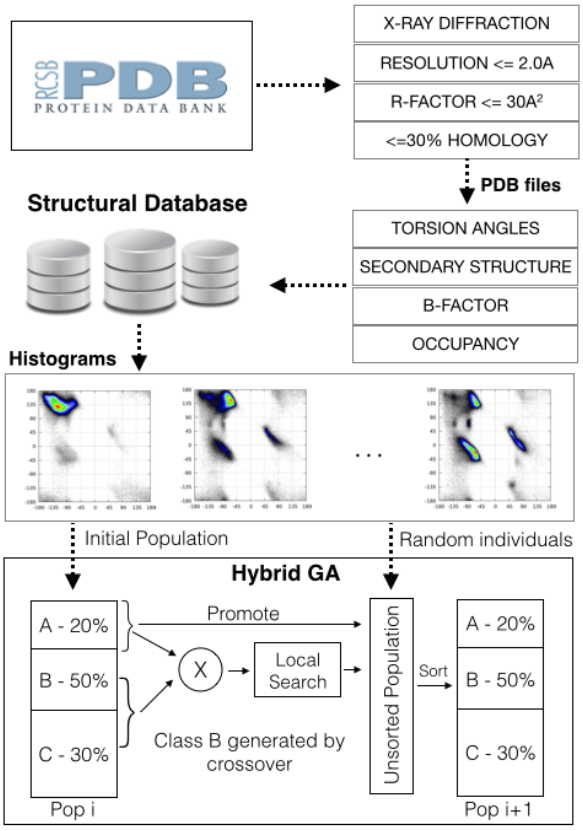


Figure 2: Schematic representation of the proposed method. A structural database is built from 3-D protein templates obtained from the PDB. Histograms are computed and used by a GA alg.

Torsion angles ϕ and ψ can have any value between -180° and +180° [7]. However, some combinations are prohibited because *inter-atomic* clashes that can occur [26, 7]. The most favorable regions can be observed in Figure 3. We also perform a more detailed analysis on the pairs (ϕ,ψ) of torsion angles in each secondary structure. We build a histogram matrix $H_{a,s}$ of 361 $\times$ 361 squares for each amino acid residue in each secondary structure. Each square $H_{a,s}(i,j)$ has the number of times that a given amino acid residue a 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 torsion angles probability matrix $AP_{a,s}$ (function 1) that represents the normalized frequency $f(i,j)$ of each square. Table 1 shows a general vision of the number of templates obtained from PDB.

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

Table 1: Template data base. Column 1-6 show the number of amino acid residues belonging to each secondary structure. Line 1-20 shows the number of amino acid residues in each secondary structure state. Ramachandran plots were computed for each set of amino acid residues belonging to a secondary structure (Fig. 3).

Amino Acid Residues	H (α-helix) Fig. 3a	G (π-helix) Fig. 3b	E (β-sheet) Fig. 3c	B (β-bridge) Fig. 3d	T (Turn) Fig. 3e	C (Coil) Fig. 3f	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%)

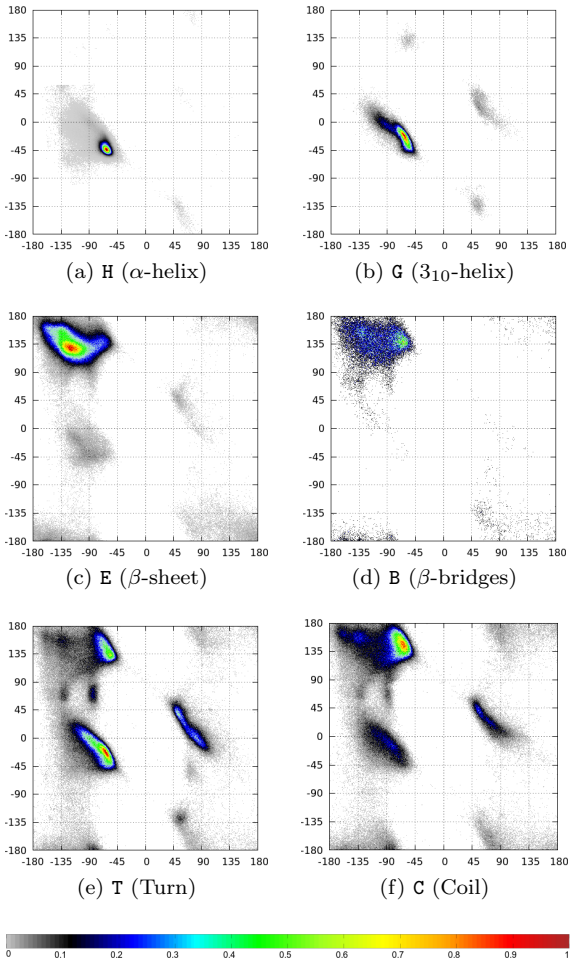
In order to analyse the histogram matrices computed, for each matrix the squares were sorted from the highest to the lowest frequency. Each square was color-coded using color schema (Fig. 3). The area in each color contains 10% of all the amino acids in the respective plot. We observe that for the same secondary structure there are different preferences for ϕ and ψ torsion angles when we analyse its occurrence in each one of the 20 amino acid residues. Figure 4 shows the torsion angles preferences for ALA, GLY, ASP, VAL, PHE and TRP in secondary structure as T (turn). As can be observed in the figure, for the same secondary structure state there are different preferred regions in the Ramachandran Plot. It clearly depends on the amino acid residue that is being considered. When we associate these two information (type of amino acid residue and secondary structure state) we obtain a valuable information that can be used to predicted new 3-D protein structures. In our GA, this information is used to generate the individuals (candidate solution) of the population. Section 3.2 shows the proposed genetic algorithm that incorporates the structural information obtained from the PDB.

3.2 Genetic Algorithm

This section describes the proposed GA. We developed a local greedy strategy to speed up the search by improving candidate solutions [49]. The genetic algorithm is combined with a structured population [54], and it is hybridized with the greedy local search procedure. Structural information obtained from the PDB was incorporated and used by the GA to generate candidate solutions. Algorithm 2 shows the general structure of the GA. 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 represent main-chain and side-chain torsion angles of the polypeptide.

Score function

In order to distinguish between good and bad solutions the GA needs a fitness/score function. This function describes, for the structure of each solution, the internal energy value that can be related to the native or 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. This energy function incorporates two types of terms: bonded and non-bonded. The bonded terms (bonds, angles and torsions) are covalently linked. The bonded terms constrain bond lengths and angles near their equilibrium values. The bonded terms also include a torsional potential (torsion) that models the periodic energy barriers encountered during bond rotation. The non-bonded potential term include: ionic

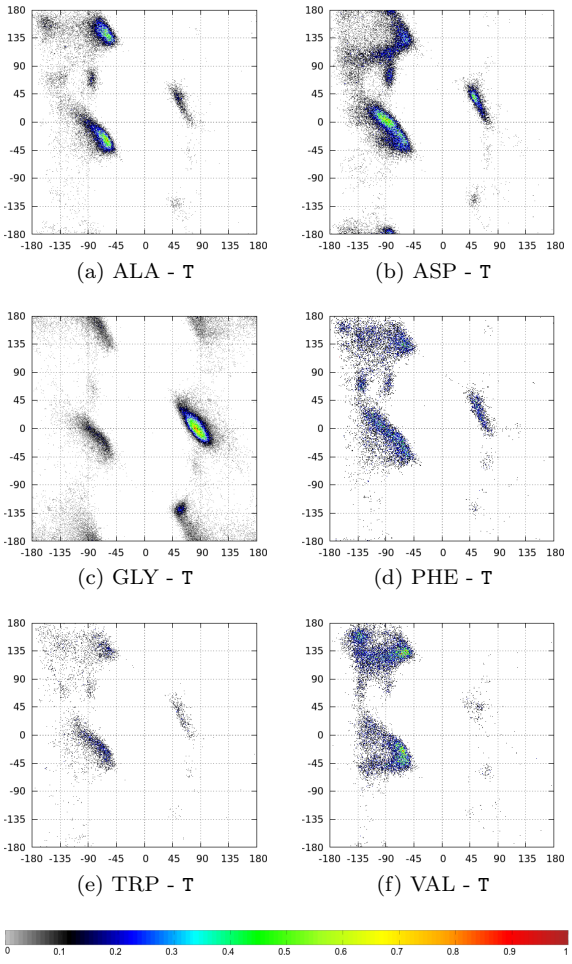


(a) `H` (α-helix) (b) `G` (3_{10}-helix)

(c) `E` (β-sheet) (d) `B` (β-bridges)

(e) `T` (Turn) (f) `C` (Coil)

Figure 3: Individual Ramachandran plots for each secondary structure. The dark red colour marks the most dense regions of the Ramachandran plot.



(a) ALA - `T` (b) ASP - `T`

(c) GLY - `T` (d) PHE - `T`

(e) TRP - `T` (f) VAL - `T`

Figure 4: Individual Ramachandran plots for the amino acid residues `ALA`, `ASP`, `GLY`, `PHE`, `TRP` and `VAL` for secondary structure state of turn (`T`). The dark red colour marks the most densely occupied regions of the Ramachandran plot. As can be observed even for the same secondary structure state we have different preferred regions in Ramachandran Plot depending on the type of amino acid residue we are considering.

bonds, hydrophobic interactions, hydrogen bonds, *van der* Waals forces, and dipole-dipole bonds. We use the `Amber99` force field [55] implemented by library `MMTK` [56] to compute the energy of a conformation.

Initial population

In order to generate the `GA` initial population we proceed as follows: for each amino acid `a` of the target protein sequence, torsion angles (ϕ, ψ) (Fig. 1) are obtained from the histogram data using the steps described in Algorithm 1. Each pair of values for (ϕ, ψ) has an associated probability of being selected, ranging from `0.0` to `1.0`. Torsion angles 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. We consider a `radius = 1` to create the initial population. Side-chain torsion angles (χ angles) are generated by `PeptideBuilder` library [57] using `Dunbrack rotamers` [58]. Population size was fixed on `100` individuals and was structured in classes [54]. The fittest `20%` of the individuals are in class `A`, the `50%` less fit ones are in class `B`, and the remaining `30%` are in class `C` [49, 50]. Figure 2 - `Hybrid GA` shows the schema of the proposed metaheuristic.

Crossover operator
The crossover operation, showed in Algorithm 2, creates a new individual called *Offspring* using information from two selected parents. $Parent_1$ and $Parent_2$ are chosen at random from classes A and B+C, respectively. With a 60% of probability, Offspring receives an amino acid from $Parent_1$, and with a 40% of probability, from $Parent_2$. Then the *Offspring* is added to the population of the next generation.

Local search operator
On each offspring obtained the algorithm apply a Greedy Local Search strategy. It takes as input a solution and changes the main-chain torsion angles of each amino acid with a probability of $p = 0.05$. If the angle is selected to be modified, its neighborhood is visited within a range of one degree (ϕ - 1 $\leq \phi \leq \phi$ + 1 or ψ - 1 $\leq \psi \leq \psi$ + 1) with the utilization of a greedy strategy that adds or subtracts a random value between the range 0.1 to 1.0 to the angle and then calculates the energy of this new structure. The local search stops when there is no improving in the fitness or it reach a region outside the one degree neighborhood.

Data: A list with a probability of a tuple of angles (ϕ, ψ) being chosen for a specific amino acid and secondary structure.
Data: A radius value that defines the size of the angles range.
Result: The ϕ and ψ angle values for a new individual from the histogram data.
1 $luck \leftarrow Random$ real number in the range 0.0 and 1.0;
2 $edge \leftarrow 0.0$;
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 $aminoacid\_\phi \leftarrow Random$ real number in the range minimal_ϕ and maximal_ϕ;
10 $aminoacid\_\psi \leftarrow Random$ real number in the range minimal_ψ and maximal_ψ;
11 break;
12 **else**
13 $edge \leftarrow edge + probability_i$;
14 **end**
15 **end**
16 return aminoacid_ϕ, aminoacid_ψ;

Algorithm 1: Getting ϕ and ψ values for new individual

Computing the next population
In order to build the next generation, all individuals from class A are automatically promoted (Fig. 2). Individuals from the local search procedure are inserted as well in the next population (class B). A new class C is entirely created in the same way as the initial population with a radius of 10Å°. This is important to preserve the population diversity. When the new population is complete, the individuals are sorted by their fitness value in a way that, at the end of each generation, the current best solution is always in the top of the population. An illustration of this process is showed in Figure 2. Class C is generated using PDB structural information represented as histograms.

Data: A protein given as a sequence of amino acids and the respectives secondary structures.
Result: The best individual
1 $Pop_0 \leftarrow Generate$ the first population using the angles values returned by Algorithm 1;
2 Sort individuals and define classes A, B and C;
3 **for** $i = 1$ *to NumberofGenerations* **do**
4 $Pop_i(A) \leftarrow Pop_{i-1}(A)$;
5 **for** $j = 1$ *to* $|B|$ **do**
6 $Parent_1 \leftarrow getIndividual(A)$;
7 $Parent_2 \leftarrow getIndividual(B + C)$;
8 $Offspring \leftarrow Crossover(Parent_1, Parent_2)$;
9 $Pop_i(B) \leftarrow add(Pop_i(B), Offspring)$;
10 **end**
11 $Pop_i(B) \leftarrow LocalSearch(Pop_i(B))$;
12 **for** $j = 1$ *to* $|C|$ **do**
13 $Pop_i(C) \leftarrow Generate$ individuals using the angles values returned by Algorithm 1;
14 **end**
15 **end**
16 $sort(Pop_i)$;
17 $best \leftarrow top(Pop_i)$;
18 return best;

Algorithm 2: GA for the 3-D PSP Problem

4 Experiments and Results

Model and target proteins

The amino acid sequences of five proteins obtained from the PDB were used to test the proposed algorithm: 1K43 (Fig. 5a), 1L2Y (Fig. 5b), 2BF9 (Fig. 5c), 3E7R (Fig. 5d) and 1FME (Fig. 5e). The algorithm was implemented in Python and tested with protein sequences whose sizes vary from 14-40 amino acid residues. It was run six times for each protein on an Intel Xeon CPU E5-2407 2.20GHz x8, 32 GB and 2TB computer. Each prediction took 72 hours of CPU time. Table 2 presents details of the target protein sequences. Column 2 shows the target amino acid sequences and column 4 shows its SCOP classification[4]. Stereo-chemical and structural analysis were performed for each predicted three-dimensional structure.

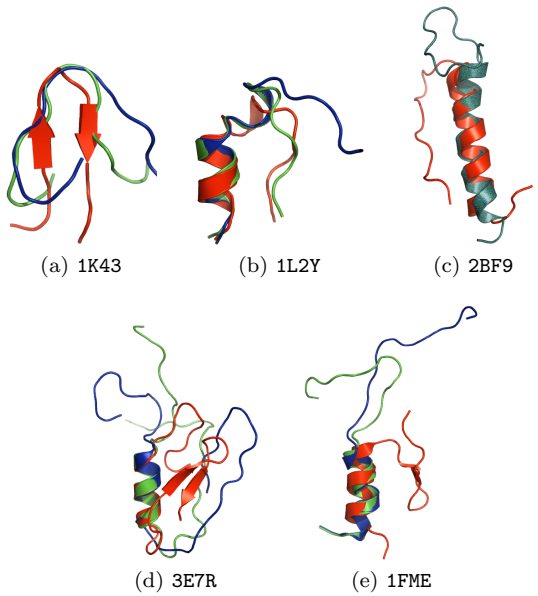
Stereo-chemical and structural analysis

For Stereo-chemical and structural analysis we selected the class of solutions that at the last GA simulation

[4] http://scop.mrc-lmb.cam.ac.uk/scop

Table 2: Target protein sequences. The size of the amino acid sequences vary from 14-40 amino acid residues.

PDB ID	Target Sequence	Res. Length	SCOP Class
1K43	RGKWTYNGITYEGR	14	Designed
1L2Y	NLYIQWLKDGGPSSGRPPPS	20	Designed
2BF9	GPSQPTYPGDDAPVEDLIRFYNDLQQYLNVVTRHR	35	Peptides
3E7R	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY	40	-
1FME	EQYTAKYKGRTFRNEKELRDFIEKFKGR	28	Designed



(a) 1K43 (b) 1L2Y (c) 2BF9

(d) 3E7R (e) 1FME

Figure 5: Ribbon representation of the experimental (red), lowest RMSD (green) and predicted (blue) 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.

presents the solution 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 the PDB (Eq.2).

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

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.

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[5]. Table 3, column 4, 5 and 6 shows respectively the RMSD of predicted structure predicted in run that achieved the lowest energy, the lowest RMSD achieved between all five runs of the proposed algorithm, and finally the average RMSD of five runs. The predicted 3-D protein structure of proteins 1K43 and 1L2Y present RMSD values around 3.5Å. The predicted 3-D protein structure of 2BF9 and 3E7R present RMSD values around 6.0Å. Study cases 3E7R and 1FME present higher RMSD. This result is somewhat expected given that these case studies show a more complex folding pattern when compared with the other test cases. By visual inspection (Fig. 5), it is noticeable that the individual helices and other secondary structures are well formed in most of the case studies.

Secondary structure assessment were analysed using Promotif [59]. We analyse pattern of hydrogen bonds that define the secondary structure of the predicted structures using the schema described in Section 2. We compare the secondary structure contents of the predicted 3-D protein structures against the secondary structure of the native structures. Table 4 summarizes the achieved results, "E" and "P" denote respectively the experimental and the predicted structure. We had observed that the secondary structure of the structures predicted by our method are comparable to their experimental structures.

The largest difference between the secondary structure elements of the predicted and experimental structures was observed in case studies 1K43, 3E7R, 1FME. Through visual inspection of these case studies (Fig. 5) we can observe that β-sheets regions are not well formed, this in turn occurs because of the presence of distortions in the coil regions. If we compare the topology of the protein backbone of the predicted structures against the experimental 3-D ones we can observe that the topologies are comparable (Fig. 5). The distribution of the amino acid residues in the Ramachandran plot and the stereo-chemical quality of the 3-D structures predicted by our method were analysed by Procheck [60]. Ramachandran plots are used to visualize backbone dihedral angles ϕ against ψ of amino acid residues in protein structure. Table 5 summarizes the obtained Ramachandran plot values for the experimental and predicted structures.

[5] www.bioinf.org.uk/software/profit

Table 3: Target protein sequences and its SCOP classification. Case studies were selected in order to test the developed strategy with different protein folding patterns.

PDB ID	Lowest Energy	Average Energy (Kcal/mol^{-1})	RMSD (Å) (Lowest Energy)	Lowest RMSD (Å) (5 runs)	Average RMSD (Å) (5 runs)	Average Number of Generations
1K43	-1,243.50	-1,209.37 (±25.56)	3.74	2.63	3.49 (±0.56)	1,591.33 (±26.86)
1L2Y	-342.30	-291.60 (±40.98)	3.42	1.73	4.05 (±1.551)	657.00 (±17.84)
2BF9	-2,563.90	-2,336.38 (±398.13)	5.85	5.85	9.31 (±2.71)	246.00 (±2.09)
3E7R	-1,386.70	1,487.83 (±83.96)	7.31	6.96	7.83 (±0.91)	156.00 (±2.28)
1FME	-1,960.50	-1,870.37 (±65.63)	11.25	7.57	9.88 (±1.65)	366.50 (±4.50)

Table 4: Analysis of the secondary structure contents of the predicted and the native 3-D protein structures.

PDB ID	β-sheet	α-helix	3^{10}-helix	Others
1K43-E	6 (42.90%)	0 (0.00%)	0 (0.00%)	8 (57.10%)
1K43-P	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
1L2Y-E	0 (0.00%)	7 (35.00%)	4 (20.00%)	9 (45.00%)
1L2Y-P	0 (0.00%)	6 (30.00%)	3 (15.00%)	11 (55.00%)
2BF9-E	0 (0.00%)	18 (51.40%)	0 (0.00%)	17 (48.60%)
2BF9-P	0 (0.00%)	19 (54.30%)	0 (0.00%)	16 (45.70%)
3E7R-E	10 (25.00%)	10 (25.00%)	0 (0.00%)	20 (50.00%)
2E7R-P	0 (0.00%)	8 (20.00%)	0 (0.00%)	32 (80.00%)
1FME-E	4 (14.30%)	10 (35.70%)	0 (0.00%)	14 (50.00%)
1FME-P	0 (0.00%)	8 (28.60%)	2 (7.10%)	18 (64.30%)

We observe that in all of 3-D predicted structures, the amino acid residues are located in the most favourable regions (columns 2 and 3) of the map (favourable or additional allowed region) (Tab. 5). 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.

5 Conclusion and further work

Predicting the correct 3-D structure of a protein molecule is an arduous task. There is an increasing need for new strategies to extract, represent and use structural data from experimentally determined 3-D protein structures. In this paper, we present a new algorithm for the 3-D PSP problem. The developed method acquires structural information from protein template obtained from the PDB. These information is then used as input in a search strategy based on a Genetic Algorithm. The search strategy combines a genetic algorithm with a structured population and it is hybridized by a greedy local search procedure. As corroborated by experiments, the developed method can produce accurate predictions, where the 3-D protein structures are comparable to their corresponding experimental ones. When compared with other first principle prediction methods (*ab initio*) that use database information, our method presents advantages in terms of demanded time to produced native-like 3-D structures of proteins.

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 (Fig. 2); (b) the development of a computational strategy to extract and represent structural information from experimentally determined protein structures (Section 3.1); (c) the analysis of conformational preferences of amino acid residues in proteins and its use to 3-D protein structure prediction methods (Section 3.1); and finally (d) the development and use of a meta-heuristic based on genetic algorithms with local search operator to search the three-dimensional protein search space.

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 different machine learning techniques to extract structural information from structural database and use it to refine coil regions of the predicted structures. Another one could be development and application of other metaheuristics to search the three-dimensional protein search space. Finally, we could test the method with other classes of proteins and with longer protein sequences with more complex folding patterns.]

Acknowledgements

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.

Author details

[1]Institute of Informatics, Federal University of Rio grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre, RS, Brazil. [2]Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Av. Ecuador 3659, Santiago, Chile.

References

1. Lesk, A.M.: Introduction to Bioinformatics, 1st edn., p. 308. Oxford University Press Inc., New York, USA (2002)
2. Tramontano, A.: Protein Structure Prediction: Concepts and Applications, 1st edn., p. 208. John Wiley and Sons, Inc., Weinheim, Germany (2006)

Table 5: Numerical Ramachandran plot values for the experimental and predicted structures.

PDB ID	Most Favorable	Most Allowed	Generously Allowed	Disallowed	Number of amino acid Residues
1K43-E	6 (66.7%)	3 (33.3%)	0 (0.0%)	1 (0.0%)	10
1K43-P	9 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	9
1L2Y-E	10 (90.0%)	1 (9.1%)	0 (0.0%)	0 (0.0%)	11
1L2Y-P	11 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	11
2BF9-E	27 (93.1%)	2 (6.9%)	0 (0.0%)	0 (0.0%)	29
2BF9-P	25 (89.3%)	3 (10.7%)	0 (0.0%)	0 (0.0%)	28
3E7R-E	25 (83.3%)	5 (16.7%)	0 (0.0%)	0 (0.0%)	30
3E7R-P	25 (83.3%)	5 (16.7%)	0 (0.0%)	0 (0.0%)	30
1FME-E	15 (62.5%)	9 (37.5%)	0 (0.0%)	0 (0.0%)	24
1FME-P	22 (91.7%)	2 (8.3%)	0 (0.0%)	0 (0.0%)	24

3. Lander, E.S., Waterman, M.S.: The Secrets of Life: a Mathematician's Introduction to Molecular Biology, p. 300. National Academy Press, Washington D. C., USA (1999). Chap. 1
4. Wooley, J.C., Ye, Y.: 1. In: Xu, Y., Xu, D., Liang, J. (eds.) A historical perspective and overview of protein structure prediction, pp. 1–43. Springer, ??? (2010)
5. Levinthal, C.: Are there pathways for protein folding? J. Chim. Phys. Phys.-Chim. Biol. **65**(1), 44–45 (1968)
6. Lesk, A.M.: Introduction to Protein Science, 2nd edn., p. 455. Oxford University Press, New York (2010)
7. Lehninger, A.L., Nelson, D.L., Cox, M.M.: Principles of Biochemistry, 4th edn., p. 1100. W.H. Freeman, New York, USA (2005)
8. Guntert, P.: Automated nmr structure calculation with cyana. Methods Mol. Biol. **278**, 353 (2004)
9. Crescenzi, P., Goldman, D., Papadimitriou, C.H., Piccolboni, A., Yannakakis, M.: On the complexity of protein folding. J. Comput. Biol. **5**(3), 423–466 (1998)
10. Bujnicki, J.M.: Protein structure prediction by recombination of fragments. ChemBioChem **7**(1), 19–27 (2006)
11. Moult, J.A.: Decade of casp: progress, bottlenecks an prognosis in protein structure prediction. Curr. Opin. Struct. Biol. **15**(3), 285–289 (2005)
12. Osguthorpe, D.J.: Ab initio protein folding. Curr. Opin. Struct. Biol. **10**(2), 146–152 (2000)
13. Floudas, C.A., Fung, H.K., McAllister, S.R., Moennigmann, M., Rajgaria, R.: Advances in protein structure prediction and de novo protein design: A review. Chem. Eng. Sci. **61**(3), 966–988 (2006)
14. Rohl, C.A., Strauss, C.E., Misura, K.M.S., Baker, D.: Protein structure prediction using rosetta. Methods Enzymol. **383**(2), 66–93 (2004)
15. Srinivasan, R., Rose, G.D.: Linus - a hierarchic procedure to predict the fold of a protein. Proteins **22**(2), 81–99 (1995)
16. Bowie, J.U., Luthy, R., Eisenberg, D.: A method to identify protein sequences that fold into a known three-dimensional structure. Science **253**(5016), 164–170 (1991)
17. Jones, D.T., Taylor, W.R., Thornton, J.M.: A new approach to protein fold recognition. Nature **358**(6381), 86–89 (1992)
18. Bryant, S.H., Altschul, S.: Statistics of sequence-structure threading. Curr. Opin. Struct. Biol. **5**(2), 236–244 (1995)
19. Turcotte, M., Muggleton, S.H., Sternberg, M.J.E.: 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 (1998)
20. Martí-Renom, M.A., Stuart, A., Fiser, A., Sanchez, A., Mello, F., Sali, A.: Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. **29**(16), 291–325 (2000)
21. Sánchez, R., Sali, A.: Advances in comparative protein-structure modeling. Curr. Opin. Struct. Biol. **7**(2), 206–214 (1997)
22. Kolinski, A.: Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. **51**, 349–371 (2004)
23. Scheef, E.D., Fink, J.L.: 2. In: Bourne, P.E., Weissig, H. (eds.) Fundamentals of protein structure: Structural Bioinformatics, p. 15 (2003)
24. Lodish, H., Berk, A., Matsudaira, P., Kaiser, C.A., Krieger, M., Scott, M.P.: Molecular Cell Biology, 5th edn., p. 970. Scientific American Books, W.H. Freeman, New York, USA (1990)
25. Branden, C., Tooze, J.: Introduction to Protein Structure, 2nd edn., p. 410. Garland Publishing Inc., New York, USA (1998)
26. Hovmoller, T.Z., Ohlson, T.: Conformation of amino acids in protein. Acta Crystallogr. **58**(5), 768–776 (2002)
27. Ramachandran, G.N., Sasisekharan, V.: Conformation of polypeptides and proteins. Adv. Protein Chem. **23**, 238–438 (1968)
28. Pauling, L., Corey, R.B.: The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. U. S. A. **37**(5), 251–256 (1951)
29. Pauling, L., Corey, R.B., Branson, H.R.: The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U. S. A. **37**(4), 205–211 (1951)
30. Frishman, D., Argos, P.: Knowledge-based protein secondary structure assignment. Proteins **23**(4), 566–579 (1995)
31. Heinig, M., Frishman, D.: Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res. **32**(Web Server issue), 500–2 (2004)
32. Andersen, C.A.F., Rost, B.: In: Bourne, P.E., Weissig, H. (eds.) Secondary Structure Assignment: Structural Bioinformatics, p. 341 (2003). Chap. 17
33. Resende, M.G.C., Ribeiro, C.C., Glover, F., Martí, R.: Scatter search and path-relinking: Fundamentals, advances, and applications. In: Gendreau, M., Potvin, J.-Y. (eds.) Handbook of Metaheuristics, pp. 87–107. Springer, ??? (2010)
34. Glover, F.W., Kochenberg, G.A.: Handbook of meta-heuristics. In: International Series in Operations Research and Management Science vol. 57, p. 570 (2003)
35. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671 (1983)
36. Granville, V., Krivanek, M., Rasson, J.-P.: Simulated annealing: A proof of convergence. IEEE T. Pattern Anal. **16**(6), 652 (1994)
37. Goldberg, D.E., 1st edn. Kluwer Academic Publishers, Boston (1989)
38. Haupt, R.L., Haupt, S.E., 1st edn. John Wiley and Sons, Inc, New York (1998)
39. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization, p. 1942. IEEE Press, New York (1995)
40. Kennedy, J., Eberhart, R.C., Shi, Y., 1st edn. Morgan Kaufmann, San Diego (2001)
41. Chi-Keong, G., Ong, Y., Tan, K.C., 1st edn. Springer, Heidelberg (2009)
42. Schwefel, H.P., 1st edn. Wiley, New York (1995)
43. Dandekar, T., Argos, P.: Potential of genetic algorithms in protein folding and protein eng. simulations. Protein Eng. **5**(7), 637–645 (1992)
44. Hoque, M.T., Chetty, M., Dooley, L.S.: A guided genetic algorithm for protein folding prediction using 3d hydrophobic-hydrophilic model. In: IEEE Congress on Evolutionary Computation, Vancouver, Canada, pp. 2339–2346 (2006)
45. Le Grand, S.M., Merz Jr., K.M.: The application of the genetic algorithm to the minimization of potential energy functions. J. Global Optim. **3**(1), 49–66 (1993)
46. Pedersen, J.T., Moult, J.: Protein folding simulations with genetic

algorithms and a detailed molecular description. J. Mol. Biol. **269**(2), 240–259 (1997)
47. Sun, S.: A genetic algorithm that seeks native states of peptides and proteins. Biophys. J. **69**(2), 340–355 (1995)
48. Unger, R., Moult, J.: On the applicability of genetic algorithms to protein folding. In: The Twenty-Sixth Hawaii International Conference on System Sciences, pp. 715–725 (1993)
49. Dorn, M., Inostroza-Ponta, M., Buriol, L.S., Verli, H.: A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: IEEE Congress on Evolutionary Computation, pp. 1233–1240. IEEE, Cancun, MX (2013)
50. Dorn, M., Buriol, L.S., Lamb, L.C.: A hybrid genetic algorithm for the 3-d protein structure prediction problem using a path-relinking strategy. In: IEEE Congress on Evolutionary Computation (CEC), pp. 2709–2716 (2011)
51. Cutello, V., Narzisi, G., Nicosia, G.: A multi-objective evolutionary approach to the protein structure prediction problem. J. R. Soc., Interface **3**(6), 139–151 (2006)
52. Park, S.: A study of fragment-based protein structure prediction: biased fragment replacement for searching low-energy conformation. Genome Inf. **16**(2), 104–113 (2005)
53. Hoque, M.T., Chetty, M., Sattar, A.: 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 vol. 224, pp. 317–342 (2009)
54. Ericsson, M., Resende, M.G.C., Pardalos, P.M.: A genetic algorithm for the weight setting problem in ospf routing. J. of Combinatorial Optimization **6**, 299–333 (2002)
55. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C.: Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins **65**, 712–725 (2006)
56. K Hinsen, K.: The molecular modeling toolkit: A new approach to molecular simulations. J. Comp. Chem. **21**(2), 79–85 (2000)
57. Tien, M.Z., Sydykova, D.K., Meyer, A.G., Wilke, C.O.: Peptidebuilder: A simple python library to generate model peptides. PeerJ. **1**(80e) (2013)
58. Shapovalov, M.S., Dunbrack, R.L.: A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure **19**(6), 844–858 (2011)
59. Hutchinson, E.G., Thornton, J.M.: Promotif: A program to identify and analyze structural motifs in proteins. Protein Sci. **5**(2), 212–220 (1996)
60. Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M.: Procheck: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. **26**(2), 283–291 (1993)