Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 28, Num. 02 (2021) 11-24

REVIEW ARTICLE

Ab Initio Protein Structure Prediction Using Evolutionary Approach: A Survey

Predição *Ab Initio* da Estrutura de Proteínas Usando Abordagens Evolutivas: Uma Revisão da Literatura

Lucas Siqueira[1]*, Sandra Venske[1]

Abstract: The *ab initio* Protein Structure Prediction (PSP) problem is to determine the three-dimensional structure of a protein only from its primary structure. Misfolding of a protein causes human diseases. Thus, the knowledge of the structure and functionality of proteins, combined with the prediction of their structure is a complex problem and a challenge for the area of computational biology. The metaheuristic optimization algorithms are naturally applicable to support in solving NP-hard problems. These algorithms are bio-inspired, since they were designed based on procedures found in nature, such as the successful evolutionary behavior of natural systems. In this paper, we present a survey on methods to approach the *ab initio* protein structure prediction based on evolutionary computing algorithms, considering both single and multi-objective optimization. An overview of the works is presented, with some details about which characteristics of the problem are considered, as well as specific points of the algorithms used. A comparison between the approaches is presented and some directions of the research field are pointed out.

Keywords: Protein Structure Prediction — Evolutionary Algorithms — *Ab Initio* Modeling — Bioinformatics

Resumo: O problema da predição da estrutura de proteínas na modelagem *ab initio* consiste em determinar a estrutura tridimensional de uma proteína utilizando apenas sua estrutura primária. O dobramento incorreto de uma proteína é a causa de algumas doenças humanas. Assim, o conhecimento da estrutura e funcionalidade das proteínas, aliado à previsão de sua estrutura, é um problema complexo e um desafio para a área da biologia computacional. Os algoritmos de otimização metaheurística são naturalmente aplicáveis para auxiliar na resolução de problemas NP-difíceis. Esses algoritmos são bioinspirados, pois foram projetados com base em procedimentos encontrados na natureza, como o comportamento evolutivo bem-sucedido de sistemas naturais. Neste artigo, é apresentada uma pesquisa sobre métodos de abordagem de predição *ab initio* de estrutura de proteínas com base em algoritmos de computação evolutiva, considerando a otimização mono e multi-objetivo. Uma visão geral dos trabalhos é apresentada, com alguns detalhes sobre quais características do problema foram consideradas, bem como pontos específicos dos algoritmos utilizados. Adicionalmente, é apresentada uma comparação entre as abordagens e são apontadas algumas tendências do campo de pesquisa.

Palavras-Chave: Predição da Estrutura de Proteínas — Algoritmos Evolutivos — Modelagem *Ab Initio* — Bioinformática

[1] *Midwestern Paraná State University (UNICENTRO), Guarapuava - Paraná, Brasil*
***Corresponding author**: lucascorsique@gmail.com

DOI: http://dx.doi.org/10.22456/2175-2745.111993 • **Received:** 14/05/2021 • **Accepted:** 22/06/2021

1. Introduction

Proteins have a major significance on the complex biomolecular structures and they act in several mechanisms that surrounds life. Since proteins perform various functions in living cells, understanding these functions and associated structures can assist in the development of drugs, crops and even synthetic biofuels [1, 2]. The Protein Structure Prediction (PSP) problem is to determine the three-dimensional structure (the native conformation) of a protein given its primary structure, i.e., a sequence of amino acids [3].

Some human disorders caused by protein misfolding include Alzheimer's disease, neurodegenerative diseases and certain types of cancer [4]. Thus, the prediction of the protein structure is a challenge for the area of computational biology [4].

Due to advances in large-scale sequencing technologies, there is naturally a growth in protein sequence information.

X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR) are commonly applied methods to define protein structures. Such methods are, however, time consuming and are not available for all proteins [1]. An alternative to be explored is to consider computational approaches to determine protein structure and function.

Computational methods for PSP can be separated into three category: *homology*, *threading* (or *fold recognition*), and *ab initio* (also known as *de novo* prediction) [4]. In homology modeling there is a comparison of sequence similarity. The threading approach is a procedure to mount the sequences on a set of template structures (mostly based on secondary structure). The *ab initio* modeling performs prediction of the tertiary structure of the protein only considering some characteristics and properties of its amino acids. In this way, this is a template-free approach.

Lattice and off-lattice are two models that are normally used to represent a protein computationally [5]. The lattice models represent the amino acids in a grid of points. Exact methods can be applied to this model due to its simplification [6]. The off-lattice models are designed to represent proteins more realistically (although they are still limited) as they have more degrees of freedom in space [7].

The metaheuristic optimization algorithms are naturally applicable in solving NP-hard problems. These algorithms are bio-inspired, that is, inspired by nature, since they were designed based on procedures found in the ecosystem [8] Metaheuristic algorithms comprise genetic algorithms (GA) [9], simulated annealing (SA) [10], particle swarm optimization (PSO) [11], ant colony optimization algorithm (ACO) [12], artificial bee colony (ABC) algorithm [13], harmony search (HS) [14], cuckoo search (CS) algorithm [15], differential evolution (DE) [16], artificial immune systems (AIS) [17], and many others.

The objective of this work is to present a survey on approaches of *ab initio* protein structure prediction based on evolutionary computing algorithms, considering both single and multi-objective optimization. An overview of the works is presented, with some details about which characteristics of the problem are considered, as well as specific points of the algorithms used.

Evolutionary algorithms (EAs) are robust and are global optimization techniques. Evolutionary computation algorithms are metaheuristic algorithms widely used and these methods have been applied efficiently in different types of problems, for both single and multi-objective approaches [18]. These characteristics justify the choice of EAs to compose this survey in addressing a problem that still has an open solution - the protein structure prediction.

Just the approaches mentioned are already quite broad and each of them could generate an independent paper. Thus, the main ideas and results are reported here, with references to the original articles, which contain complete detailed information on the techniques and experiments considered and mentioned in this research.

The rest of this paper is organized as follows: in Section 2 some important concepts about the protein structure prediction problem are presented. Sections 3 to 6 present some works that solve the PSP problem with evolutionary approaches. In Section 7 are pointed directions observed in this study. Finally, Section 8 presents the conclusions.

2. Protein Structure Prediction Problem

In this section we present some basic concepts and definitions related to the protein structure prediction problem and how the study of this problem can be conducted using computational approaches. First, an overview of the problem is presented. In second place, potential energy functions, also called force fields, are defined, highlighting some of the most commonly used functions in the literature. Next, computational methods used to treat PSP are presented: homology, threading and *ab initio*. A final section is devoted to some common quality metrics used in the PSP.

2.1 Problem overview

The three-dimensional structure of a protein defines its biochemical function [2].

[5] showed that this structure is determined by its amino acid sequence. He noted that, once denatured, the protein unfolds its “native” conformation *in vitro* when denaturing agents are removed from the process.

The three-dimensional structure of a protein can be obtained experimentally (X-ray crystallography and nuclear magnetic resonance techniques). Although these techniques seem like good alternatives, they cost a lot of time and resources, and may not be available for all proteins [1].

Under appropriate conditions, a protein folds into the same stable structure. Considering all the conformations that the protein can explore, the stable structure - also called native, has the global free-energy minimum [5].

An accurate potential energy function is required to analyze a conformation. The potential energy function computes the energy for a given protein conformation using a searching algorithm. Some potential energy functions (sometimes called empirical force fields) are presented in Section 2.2.

The space of conformations of a protein is large. The way to calculate the free energy for each of the possible conformations is still deficient, coupled with the problem of understanding how the process occurs in nature itself [5].

The application of computational methods (with computer resources) aims to overcome some of the difficulties that are related to experimental approaches. The computational methods for PSP are commonly classified into three categories [4]: homology or comparative modeling, threading or fold recognition and *ab initio* or *de novo* approach. The section 2.3 presents an overview of these methods.

2.2 Potential energy functions

The mathematical models used to construct the potential energy function of a protein are relatively simple. This simplicity combines computational speed and molecular mechanics. The energy potential function considers the smallest particles in the model (usually atoms) and treat them as point masses centered on the nucleus of each atom in the molecules present in the considered system. Therefore, interactions between atoms in the protein system are represented by the potential energy function [19].

In general, Newton's equations of motion are used by molecular force fields to describe the physical interactions between atoms. Normally, the force field considers covalent bonds and noncovalent interactions, such as electrostatic interactions, the van der Waals interactions, and sometimes, hydrogen bonds and hydrophobic interactions. The force fields use some parameters which were obtained through experimental studies using small organic molecules. There are several software packages that predict protein conformations using computer simulations [19].

The potential energy functions vary in their degree of representation and in the precision of the approximation and, therefore, also vary in their complexity and accuracy. To improve the accuracy, it is necessary to add more parameters to the model, consequently increasing its complexity [20].

The force fields have performed useful results in the field of biological molecules, but still have limitations. One example is related to the fixed atoms center point charge method, where the force fields leave out the intermolecular and intramolecular charge transfer and electrostatic polarization. This restriction of force fields means that they are not sufficient to calculate the electrostatic polarization [21, 22].

Examples of potential energy functions are CHARMM [23, 24], OPLS/AA [25, 26], AMBER [27], and GROMOS [28]. The formula for CHARMM is presented as an example in Equation 1.

$$\begin{aligned} E = {} & \sum_{bounds} K_b(b-b_0)^2 + \sum_{UB} K_{UB}(S-S_0)^2 \\ & + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} K_\Phi[1+cos(n\Phi-\gamma)] \\ & + \sum_{impropers} K_{imp}(\varphi-\varphi_0)^2 \\ & + \sum_{nb} \varepsilon_{ij}\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{e r_{ij}} \qquad (1) \end{aligned}$$

The terms *bonds, UB, angles, dihedrals and impropers* in Equation 1 describe the molecule. They are depicted by [3]:

- the bond length, b;
- valence angle, θ;
- distance between atoms separated by two covalent bonds, S;
- dihedral or torsion angle, Φ;
- improper angle, φ;
- the bond force constant and equilibrium distance, K_b and b_0;
- the valence angle force constant and equilibrium angle, K_θ, and θ_0;
- the Urey-Bradley force constant and equilibrium distance, K_{UB} and S_0;
- the dihedral angle force constant, multiplicity, and phase angle, K_Φ, n, and γ; and
- the improper force constant and equilibrium improper angle, K_{imp} and φ_0.

The last terms of Equation 1 are the parameters that describe the interactions between atoms i and j and are represented by [3]:

- the partial atomic charges, q_i;
- the Lennard Jones well-depth, ε_{ij};
- the minimum interaction radius, $R_{min_{ij}}$ (used to estimate the van der Waals interactions); and
- distance between atoms i and j, r_{ij}.

In most works that use the multi-objective approach to address the PSP problem is used a combination of bonded and non-bonded energy functions as objectives to be minimized. In [29] an experiment was conducted in which it is demonstrated that these two energy functions are conflicting, which justifies its approach as multi-objective optimization. In some recent approaches the authors consider the effect of the solvent as a third objective. [30] experimentally inspected the degree of conflict between the three objectives and showed that the effect of the solvent is strongly in conflict with the energy functions.

2.3 Computational methods

Three categories represent the computational methods for dealing with PSP: homology or comparative, threading or fold recognition, and *ab initio* or *de novo* [31].

Homology benefits when two proteins share a common ancestor. The fundamental approach to structure prediction for an unknown structure using homology is to perform a pairwise sequence alignment against each sequence in protein sequence databases. There is homology when more than 30 % identity is observed between the sequences [1].

Threading is used mainly when a sequence under study has no matches but may have folds in common with proteins whose structure is known. In this method, an input sequence is analyzed considering subfragments and "threaded" onto a library of known folds. Threading uses scoring functions that allow one to assess the compatibility of the analyzed sequence with known structures [32].

The works listed in this research are arranged under the *ab initio* methodology.

The *ab initio* modeling allows to use some information from the amino acid sequence. The secondary structure (SS) of a protein refers to the location of α-helices, β-sheets and turns in the amino acid sequence. Algorithms can use these amino acid properties as input in order to obtain the tertiary structure of the protein.

Torsion angle model is a type of data model for representation of tertiary structure of proteins. It uses a measure of the rotation about a bond, generally considered to be between -180 and +180 degrees.

Torsion angles are sometimes called dihedral angles. *Phi* (ϕ) is the angle of rotation of a peptide backbone about the bond between the nitrogen and the $_{\alpha}$-carbon atoms, whereas *psi* (ψ) is the angle of rotation about the bond between the $_{\alpha}$-carbon and the carbonyl carbon atoms. Not all combinations of *phi* and *psi* are possible; many combinations are prohibited due to steric collisions between atoms. The Ramachandran plot shows the allowed values for the *phi* and *psi* angles [33].

The predicted secondary structure and torsion angles of a residue provide *local* structural information along the amino acid sequence. The *global* structural properties of a residue should provide information on, for example, its position and orientation relative to covalently bonded sequence neighbors. In general, the parameters that measure the exposure of a residue to the solvent are the most common general structural properties considered.

A common representation of the protein structure is called *full-atom*, which specifies the positions of all non-hydrogen atoms, considering ϕ, ψ, ω[1] and χ angles[2]. The representation leaves out the hydrogen atoms because their positions can be inferred from the molecular structure. A general formula for all-atom force field consists of terms bonded and non-bonded terms (Equation 1) [34].

Another representation is called *backbone-only* model, for which it is calculated only three atomic positions per residue using ϕ, ψ and ω. In this representation the amino acid side chains are not used [34].

Another representation is called *coarse-grained* and is described by a similar formula as an full-atom representation. In this representation additional expressions are considered and are used to describe the energy of coarse-grained models. During the coarse-graining process some atoms are removed and the degrees of freedom related to them are averaged out. The coarse-grained methods remove atoms and do not use some details of some interactions in their process. The goal here is to focus on general features. These modifications used in the method reduces the number of degrees of freedom of biomolecules, and smoothed the energy landscape of the system. This causes a reduction in the computational cost [34, 35].

Some works in state-of-art mention the CASP (Critical Assessment of protein Structure Prediction) competition. The objective of CASP is to assess the current state of the art of structure and function prediction methods, identifying limitations and pointing out opportunities for new developments. CASP experiments examine the prediction of structures considering template-based and free-modeling categories [5].

2.4 Quality metrics

To complement the prediction of the structure of a protein (by computational methods, for example) it is necessary to carry out the comparison with the original native structure. Some metrics are taken to measure the similarity between the expected conformation and the native structure: RMSD (Root Mean Square Deviation), GDT (Global Distance Test) and TM-score (Template Modeling score) [19] are the most commonly used.

RMSD is a metric that assesses the degree of similarity between two structures, and is computed using the Equation 2 [19].

$$RMSD(a,b) = \sqrt{\frac{\sum_{i=1}^{n} \mid r_{ai} - r_{bi} \mid^2}{n}} \qquad (2)$$

where r_{ai} and r_{bi} are the positions of the atom i in the structures a and b, respectively, and n is the number of atoms.

The RMSD value can be measured in *Angstroms*, symbolized by Å (most commonly used) or Nanometers (nm). Identical structures have a value of RMSD = 0 Å while their value increases as structures become more divergent [36]. RMSD can be measured by considering all the atoms of the structures (RMSD$_{all-atoms}$) or only carbon-α (RMSD$_{c\alpha}$).

The GDT [37, 38] measures the similarity between two proteins x and y with equal primary structures (amino acid sequences) but different tertiary structures. GDT is calculated as the largest set of amino acid residues' α carbon atoms in x falling within a defined cutoff distance d_0 of their position y. In order to define all intermolecular stabilization interactions, d_0 = 0.5 nm is usually defined. High values of the GDT metric indicate a better fit between two conformations. GDT of 100 means the folds are the same.

TM-score [39, 40] is another measure of protein similarity that is more accurate than RMSD and GDT. TM-score gives pairs of residues at shorter distances higher weights than those at greater distances and normalized by the length of the target proteins. The value of the TM-score varies between 0 and 1, with 1 indicating the best fit between two conformations. Values below 0.2 correspond to unrelated conformations. The TM score of structures with the same fold is greater than 0.5.

3. Genetic Algorithm Approaches

Genetic algorithms [9] are evolutionary algorithms that act on a fixed-length data structure and use operators (mutation and crossover) to perform variations in solutions [41]. In this section are presented some works for solving the PSP problem using genetic algorithms. The works are first ordered in single-objectives and multi-objectives and then in chronological order.

[44] uses a genetic algorithm to deal with PSP taking the primary structure as input. A hybrid approach is used, which consists of the combination of the GA with a refinement step. This refinement is added to the evolutionary process to assist

[1] The torsion angle measured the chemical bond that connects two amino acids.

[2] The side chain dihedral angles.

Method	Optimization	Adaptive	Local search	Energy function	Number of proteins
PSAGC [42]	Single-objective	No	Yes	ECEPP/2	3
NOMAD-PSP [43]	Single-objective	No	Yes	CHARMM	3
GA hybrid [44]	Single-objective	No	No	ECEPP	1
CSSGA [45]	Single-objective	Yes	No	GROMOS	6
SOGA-PSP [46]	Single-objective	Yes	No	CHARMM	2
NSGA-II [47]	Multi-objective (2)	No	No	CHARMM	1

Table 1. Genetic Algorithms of *ab initio* off-lattice protein structure prediction.

the balance and stability of a structure. The representation used was the full-atom torsion angle. The force field is calculated using ECEPPAK package and the protein tested is PDB (Protein Data Bank) id 1Q2K. The predicted structures were assessed using two measures of similarity, TM-Score and RMSD.

In [42] is applied a hybrid algorithm using Simulated Annealing with Genetic Algorithm, called PSAGC (Parallel Simulated Annealing with Genetic Crossover). The procedures combined with local search (SA) are performed in parallel. The algorithm considers the solutions represented considering the dihedral angles in the range of [-180°, 180°]. PDB id 1PLW, C-peptide and PTH(1-34) are the proteins tested. For the energy function was used ECEPP/2. The results were presented in terms of energy.

[43] proposes a method called NOMAD-PSP (Nonlinear Optimization for Mixed Variables and Derivatives algorithm for PSP). NOMAD-PSP is based on two algorithms: Generalized Pattern Search (GPS) and Mesh Adaptive Direct Search (MADS). The full-atom torsion angle was used for protein representation. The proposed approach uses CHARMM and it is tested considering three proteins (PDB ids 1PLW, 2MLT and 1ZDD). It compares the results for 1PLW, in terms of energy and RMSD, with other algorithms in the literature.

[45] proposes CSSGA (Crowding-based Steady-State Genetic Algorithm) which applies a k-nearest neighbors surrogate modeling strategy. CSSGA improves the quality of proteins structures predicted using two similarity criteria. A crowding-based steady-state GA is applied without increasing the number of exact fitness evaluations. The first similarity criterion is based on the phenotypes, using the metrics of the alpha carbons of hydrophobic residues. The second criterion is genotypic and is measured using the Euclidean distance between the chromosomes. The adaptive scheme, according to the authors, is detailed in [48]. The proposed method uses the full-atom representation. The molecular force field used is GROMOS96 and the quality of generated structures was considered using RMSD metric. The tested proteins used are 23ALA, and PDB ids 1E0N, 1AMB, 1VII, 1L2Y, 1E01 (with sizes between 23 and 37 amino acids).

In [46] is proposed an algorithm that combines techniques of Self-Organization with Genetic Algorithm for PSP problem (SOGA-PSP). To automate the selection of parameter values (crossover and mutation rates), the influence of auto-adaptation is used to design genetic operators in order to optimize the protein prediction process. The protein representation used was the full-atom torsion angle. This approach uses CHARMM and presents an analysis using two proteins (PDB ids 1PLW and 3DGJ). The results were analyzed in terms of minimal energy value and RMSD metric.

NSGA-II (Elitist Non-Dominated Sorting Genetic Algorithm) [47] uses a parallel multi-objective *ab initio* approach. Two objectives were considered: the first objective takes into account the local interactions, while the second objective considers all the interactions between the atoms that are not connected by a covalent bond. The algorithm evolves the protein conformations applying an elite-preservation strategy and an explicit diversity-preserving mechanism. Besides that, the island model is used with the evolutionary algorithm. For the representation of the proteins the full-atom model was used. For the function optimization and evaluating the structures of the protein conformations, was utilized the equations of CHARMM. The protein tested was the PDB id 1ROP. The results were compared with the literature using the RMSD metric.

Genetic Algorithms approaches for PSP problem are summarized in Table 1. First, works with a single and then multi-objective approach are listed (Optimization). The only work with multi-objective optimization in this section uses 2 objectives. Table 1 also shows information about the use of adaptive method, local search, name of the energy function used and the number of proteins tested.

4. Immune Algorithm Approaches

Artificial Immune Systems (AISs) are biological-inspired algorithms that attempt to explore elements of immunology to plan scientific applications based on the immune system. AIS is based on the concept of intelligent bottom-up methodology, in which reasoning operates at the local level of cells and molecules and adaptation appears at the global level [43]. Most AISs that inspired optimization algorithms are based on the applications of clonal selection and hypermutation, and known as clonal selection algorithms [49]. This section shows some works that use AISs to handle the PSP problem. The works are first ordered in single-objectives and multi-objectives and then in chronological order.

In [50] is proposed an approach using an hybrid immune

Method	Optimization	Adaptive	Local search	Energy function	Number of proteins
IMMALG-Direct [50]	Single-objective	No	No	CHARMM	4
I-PAES [51]	Multi-objective (2)	No	No	CHARMM	5
I-PAES [29]	Multi-objective (2)	No	No	CHARMM	5

Table 2. Immune Algorithms of *ab initio* off-lattice protein structure prediction.

algorithm and a quasi-Newton method. The authors chose to start the evolutionary search from a population of "promising protein conformations" produced by the global optimizer (named Direct). The proposed approach, named IMMALG-Direct, uses CHARMM, is tested in PDB ids 1PLW, 1POLY, 1ROP and 1BDC and presents other results of literature. The authors use the primary and secondary structures. IMMALG-DIRECT represents the protein using torsion angles, where the backbone angles are selected based on the Ramachandran plot; angle ω is set to the standard value of 180°. The results were compared with the literature in terms of RMSD and energy values.

I-PAES [51, 29, 52] is a modified version of PAES (Pareto Archived Evolutionary Strategy) [53] applied to the PSP and uses immune inspired operators. The authors use CHARMM force field and show experimentally that the interactions energies (bond and non-bond atoms) are in conflict, adopting a multi-objective approach to the problem. Backbone torsion angles are bounded in regions derived from secondary and supersecondary structure prediction. The authors use torsion angle representation where bond lengths and angles are fixed at their ideal values and the angle ω is set to the standard value of 180°. The degrees of freedom are the backbone and side-chain torsion angles. A set of small [51] and medium size [29] protein sequences (5-70 residues) is tested, comparing results in terms of energy and Pareto front with the literature.

The approaches cited in this section for tertiary structure prediction using AIS are summarized in Table 2. Works are listed by optimization approach: first single and then multi-objective (Optimization). In the case of those with multi-objective optimization, it is also mentioned how many objectives are considered (2 or 3). In addition, Table 2 also informs if the work involves any adaptive method, local search, which energy function is used and the number of proteins tested.

5. Differential Evolution Approaches

The Differential Evolution (DE) was developed by Storn and Price in 1995 aiming for better results with a different approach from the one utilized in genetic algorithms and evolution strategies. It is a stochastic direct search method that emerged from attempts to solve Chebychev's polynomial adjustment problem. Kenneth Price introduces the idea of vector differences to disturb the vector population (individuals) resulting in a method that requires few control variables, is fast converging, easy to use and robust [16]. Some works with DE applied to the PSP problem are listed in this section. The works are first ordered in single-objectives and multi-objectives and then in chronological order.

In [54] the PSP problem is studied based on two algorithms: SaDE (Self-Adaptive Differential Evolution) and RGA (Real-coded Genetic Algorithm). Different crossovers and mutations are tested. Three parameters of Differential Evolution (crossover rate, mutation factor and mutation strategies) are the focus of the adaptive process in the evolutionary process using SaDE approach. The approach is tested using ECEPP/2 and ECEPP/3 force fields. In their other work [55], the PSP problem is solved using a proposed approach named DCSaDE-LS (Diversity Controlled Self-Adaptive Differential Evolution with Local Search). DCSaDE-LS, a modified version of SaDE, adopts a fuzzy system to regulate individual diversity and local search, therefore preserving the balance between exploration and exploitation. The full-atom torsion angle was used for protein representation. For energy functions ECEPP/2, ECEPP/3 and CHARMM were used. In both [54] and [55], 1PLW is the tested protein and the results are presented using RMSD metric.

In [56] is used the DE algorithm applying two diversification strategies (Generation Gap - GG and Gaussian Perturbation - GP) to handle the protein structure prediction problem. The approaches, named DE$_{GG}$ and DE$_{GP}$, employ the backbone and side-chain model with CHARMM force field. To test their approaches, the authors use 1PLW, 1ZDD and 1CRN proteins. The results were presented in terms of RMSD and energy values.

In [57] is proposed LUE (Lipschitz UnderEstimation), an approach for conducting exploration in conformational feature space with Lipschitz underestimation. The method is applied to *ab initio* protein structure prediction based on the Lipschitz estimation theory, using DE and Metropolis Monte Carlo algorithms (Rosetta framework). The representation used was based on the coarse-grained model. LUE is tested on 15 small-to-medium proteins (PDB ids 1VII, 1ENH, 2JUJ, 1GYZ, 2MU2, 1AIL, 4ICB, 2EZK, 3GWL, 2MRF, 1FD4, 1GB1, 1AOY, 2MIT and 1I6C). The results are presented using RMSD and TM-Score metrics.

A modification of DE using the surrogate approach and gene expression programming (GEP) is proposed for PSP in [58]. In the approach, named SGDE, the GEP is used to generate a diversified set of configurations, whereas the surrogate model supports DE to find the best set of configurations. Besides that, covariance matrix adaptation evolution strategy (CMAES) is assumed in order to explore the search space. SGDE uses adapting methods to optimize the F and CR pa-

Method	Optimization	Adaptive	Local search	Energy function	Number of proteins
SaDE/RGA [54]	Single-objective	Yes	No	ECEPP/2, ECEPP/3	1
DCSaDE-LS [55]	Single-objective	Yes	Yes	CHARMM, ECEPP/2, ECEPP/3	1
DE$_{GG-GP}$ [56]	single-objective	No	No	CHARMM	3
LUE [57]	Single-objective	No	No	ROSETTA	15
SGDE [58]	Single-objective	Yes	Yes	AMBER	1
MODE-P [59]	Multi-objective (2)	No	No	CHARMM	3
ADEMO/D [60]	Multi-objective (2)	Yes	No	CHARMM	5

Table 3. Differential Evolutionary Algorithms of *ab initio* off-lattice protein structure prediction.

rameters of DE. In this study the authors used the AMBER force field. The SGDE was tested on 1GK4 protein using an full-atom model and RMSD metric.

MODE-P (Multi-Objective Differential Evolution for PSP problem) [59] uses a DE-based approach to deal with the PSP problem. MODE-P identifies non-dominated solutions, stores them and then includes them in the population. The storage procedure used is based on Pareto Archived Evolution Strategy (PAES). The protein model is based on off-lattice and an internal coordinates representation. CHARMM force field is used. The two objectives were internal (bonded) and external (interaction or non-bonded) energy functions. MODE-P is tested on the Met-Enkephalin peptide (PDB id 1PLW) and two others protein sequences (PDB ids 1CRN and 1ZDD), presenting the RMSD values of predicted conformations.

In [60], PSP was modeled as a multi-objective optimization problem and adopts the Adaptive Differential Evolution algorithm for Decomposition-based Multi-objective Problems (ADEMO/D) in its optimizing platform. ADEMO/D incorporates problem decomposition concepts and mechanisms of adaptation of mutation strategies. Since the approach is multi-objective, the algorithm implements decision making and four different methods of decision maker have been tested. The energy function used was CHARMM. Bond and non-bond atoms interactions were the two objective functions. ADEMO/D uses an off-lattice model based on the torsion angles and the secondary structure constraints to model conformations. The RMSD metric is used to assess the similarity between the predicted conformations and the native structures. The tested proteins are PDB ids 1PLW, 1ZDD, 1CRN, 1ROP and 1CTF.

A summary of DE methods for PSP problem is shown in Table 3. Works are listed in order by optimization approach, with single-objective first and and then multi-objective works (Optimization). For multi-objective approaches, the table also mentions how many objectives are considered (2 or 3). Table 3 also shows information about adopted adaptive methods, local search, energy function used and the number of proteins tested.

6. Other Evolutionary Approaches

There are other evolutionary approaches to the PSP problem besides Genetic Algorithm, Artificial Immune System and Differential Evolution. In this section, the works using Cuckoo Algorithm [61, 15], Evolution Strategy [62], Particle Swarm Optimization [11] and some generic evolutionary algorithms are presented. Some authors do not specify the name of the evolutionary algorithm used. In these *generic* approaches, the algorithms are stochastic and population-based, use selection of the best individuals and apply mutation and recombination operators.

Cuckoo Algorithm is a recent evolutionary optimization algorithm which is inspired by lifestyle of a bird family, called cuckoo. This search algorithm is biologically inspired by the way in which this type of bird looks for nests where they could lay eggs [15].

The Particle Swarm Optimization (PSO) algorithm is a population-based search algorithm that was inspired by the social behavior of birds within a flock. Individuals, or particles, represent a potential solution and move in a search space. Changes in the position of the particles in the search space reflect the social tendency of individuals to imitate the success of individuals neighboring the environment [41].

Evolution Strategies (ES) [62] is a stochastic optimization algorithm inspired by the biological theory of evolution by natural selection. ES generally evolves a Gaussian distribution and repeats the procedures of generating a population of candidate solutions from the search distribution and learning the distribution parameters from the generated samples.

In [63] is proposed a single-objective algorithm implemented in Protpred-GROMACS, a framework that uses a generic evolutionary approach for PSP in structural and energetic context. It is studied the use of a structural fitness (hydrophobic solvent accessible surface area) which is compared with an energy fitness (protein potential energy). ProtPred adopts the full-atom model with internal coordinates for both the backbone and side-chains to represent the conformations. ProtPred uses GROMACS analysis tools and CHARMM force field. Results are presented considering the

RMSD metric for a set of 5 proteins (PDB ids 1VII, 1EON, 1A11, 1PLW, 1UAO).

In [64] is presented a hybrid generic EA that incorporates strategies used in state-of-the-art *ab initio* protocols. The method incorporates the coarse-grained representation used in Rosetta package. This representation uses only ϕ, ψ and ω angles and uses protocols describes by CASP competition. This EA also applies the molecular fragment replacement technique, which according to the authors, helps in the quality of the prediction of the structure. The hybrid part of the proposal incorporates a local search that makes use of the molecular fragment replacement technique. The authors analyze the crossover operators and implement a novel homologous 1-point crossover. They concludes that the use of crossover with mutation is beneficial in navigating the protein energy surface. The proposal is evaluated on 10 proteins (with sizes ranging from 61 to 123) using RMSD metric.

[65] proposes the PITAGORAS-PSP approach (Parallel Implemented procedure with Template information, *Ab initio* Global Optimization, and Rotamer Analysis and Statistics for Protein Structure Prediction). PITAGORAS-PSP provides a reduction of the search space by using the dependent rotamer library and includes new heuristics using a generic EA and the PAES algorithm. Three objectives are considered: the bond energy, the non-bond energy and the difference with the initial conformation 3D structure. The proposed method used the full-atom representation. AMBER force field is used. PITAGORAS-PSP is applied to a benchmark set with four proteins in CASP8 and the results are shown using GDT-TS and RMSD metrics.

In [66] is proposed the MEAMT (Multi-objective evolutionary algorithms with many tables), an algorithm that deal with PSP problem with four objectives. The authors propose a methodology for the evolutionary algorithm to address the many-objectives approach to the PSP problem. MEAMT uses four interaction energy terms as objectives: van der Waals interaction, electrostatic interactions, solvation contribution, and hydrogen bond interaction. These terms comprise weighting functions that combine these energy terms two by two, three by three and four by four. A full-atom description and the CHARMM force field is used in algorithm and the results were presented using the RMSD and GDT-TS metrics. The algorithm were tested through extensive benchmark tests with 32 proteins.

[67] design a Multi-Objective Diversity Controlled Self Adaptive Cuckoo Algorithm (MODCSA-CA) in order to solve the PSP problem. MODCSA-CA uses a modified SaDE (Self-Adaptive Differential Evolution) algorithm to verify population diversity and a local search is applied to preserve a balance between explore and exploit cycles. The two objectives are the bonded and non-bonded terms of potential energy function. CHARMM force field is used with an internal coordinate's representation – the torsion angles - with backbone and side-chain torsion angles to model proteins. Seven proteins are considered in experimental tests (PDB ids 1CRN, 1CTF, 1PLW, 1ROP, 1ZDD, 2L56, 2MLT) and the results are presented in RMSD metric but only in graphics, not in exact values.

[68] uses a three multi-objective evolutionary algorithm with preference, called MO3-P. The tree objectives were bond energy, non-bond energy and solvent accessible surface area (SASA). MO3-P uses ES approach where at every generation, a parent reproduces two offspring by applying a local mutation operator and a global mutation operator, respectively. The preference information is used in the survival criteria of the individuals, with emphasis on the exploration of search process. The authors conclude that the use of the preference information can diversify the solutions. For the representation the full-atom model was used. The proteins tested were PDB ids 1WQC, 2F4K, 2P6J, 3P7K, 3V1A using the CHARMM force field. The results are presented in terms of RMSD and energy.

[69] proposes an archive information assisted MOEA (that was named AIMOES), as a three-objective evolution algorithm. The three energy terms considered are: bond energy, non-bond energy, and solvent accessible surface area. AIMOES uses ES approach to control evolutionary process. AIMOES makes use of a strategy that reuses the experiences obtained previously in the evolutionary process to increase the effectiveness in the search for conformations. The full-atom torsion angle was used for representation. CHARMM force field is used in AIMOES to evaluate 25 proteins, with lengths ranging from 30 to 91 amino acids. RMSD metric is used to evaluate the quality of the predicted structures.

MO3 [30] is a multi-objective algorithm where is considered the effect of solvent, adopting a solvent-accessible surface area as the third objective (in addition to bond and non-bond energy). MO3 uses the framework of ES to design their multi-objective evolutionary algorithm and PAES. For the mutation, it also utilizes an achieve method. The proteins were represented using the full-atom model. For the calculation of force fields, it was used CHARMM, and the protein tested were PDB ids 1ZDD, 1E0M, 1ROP and 1CRN. The performance of MO3 is evaluated using protein targets up to 345 residues taken from the 11th CASP experiment. The results are presented in terms of energy values, besides GDT-TS and RMSD metrics.

In [70], the PSP problem is modeled as a multi-objective optimization problem using Particle Swarm Optimization (MOPSO). The proposal is based on a full-atom torsion angle representation and CHARMM in a three-objective optimization. The three objectives were bond, non-bond and dDFIRE (dipolar distance-scaled, finite ideal-gas reference) energy functions. MOPSO is tested with twelve proteins using the RMSD and GDT metrics. The authors favorably compare their results with six other works in the literature.

Table 4 summarizes the main characteristics of the works cited in this section. Works are listed by optimization approach: first single and then multi-objective (Optimization). In the case of those with multi-objective optimization, it is

Method	**Optimization**	**Adaptive**	**Local search**	**Energy function**	**Number of proteins**
ProtPred [63]	Single-objective	No	No	CHARMM	5
Hybrid EA [64]	Single-objective	No	Yes	ROSETTA	10
PITAGORAS-PSP [65]	Multi-objective (3)	No	Yes	AMBER	4
MEAMT [66]	Many objectives (4)	No	No	CHARMM	32
MODCSA-CA [67]	Multi-objective (2)	Yes	Yes	CHARMM	7
MO3-P [68]	Multi-objective (3)	No	No	CHARMM	5
AIMOES [69]	Multi-objective (3)	No	No	CHARMM	25
MO3 [30]	Multi-objective (3)	No	No	CHARMM	4
MOPSO [70]	Multi-objective (3)	No	No	CHARMM	12

Table 4. Other Evolutionary Algorithms of *ab initio* off-lattice protein structure prediction.

also mentioned how many objectives are considered (2 or 3). Just one of the works considers 4 objectives and fits the many objective optimization classification [71]. Table 4 also presents the adaptive method and local search used, in addition to the energy function applied and the number of proteins tested.

7. Directions

The use of evolutionary approaches has been growing in recent years for the most diverse problems, especially for the NP-hard, as is the case of the PSP problem.

Genetic Algorithms and Differential Evolution with single-objective modeling remain the most used, even recently, in case of PSP problem using *ab initio* off-lattice modeling.

In terms of adaptive parameter techniques, most do not use this feature. Only about 30% of the works use some type of adaptation method to assist the search process.

Likewise, local search is applied to the minority of the listed works (also about 30%).

Only three works implement parallelism. As force field, CHARMM is used in about 66% of the approaches.

It has been observed in several works that certain proteins are tested more frequently than others.

PDB id 1PLW (*Met-enkephalin*) is a peptide for which a substantial amount of experiments has been done. 1PLW is a peptide with only five amino acids, 22 variable backbone and side-chain torsion (or dihedral) angles and 75 atoms. Table 5 shows the results for 1PLW considering energy (in kcal mol^{-1}), RMSD (in Å) and number of Fitness Function Evaluation (FFE) for the works considered. The two approaches with the best results in terms of energy, both for single-objective (DE$_{GG-GP}$ [56] and DCSaDE-LS [55]) and multi-objective optimization (MODE-P [59] and ADEMO/D [60]), use the Differential Evolution algorithm.

PDB id 1ZDD (*Disulphide-stabilized mini protein A domain*) is a peptide with two α-helices structures and 34 amino acids with 179 angles to be optimized. Table 6 presents some results for 1ZDD. NOMAD-PSP [43], MO3 [30] and ADEMO/D [60] obtained the best results in terms of minimizing energy for this protein.

PDB id 1ROP (*Repressor of primer*) is a dimer, and each monomer consists almost completely of two α-helices, is composed of 56 residues and forms an α-turn secondary structure. Table 7 lists good results for this peptide, especially for the single-objective approach for the IMMALG-Direct [50].

Algorithm	**Energy (kcal mol^{-1})**	**RMSD$_{C_\alpha}$ (Å)**	**FFE**
Single-objective optimization			
DE$_{GG-GP}$ [56]	-35.82	1.98	5.0×10^5
DCSaDE-LS [55]	-30.57	0.23	1.5×10^5
NOMAD-PSP [43]	-30.14	1.55	2.5×10^5
IMMALG-Direct [50]	-20.47	-	3.0×10^4
PSAGC [42]	-11.10	-	-
Multi-objective optimization			
MODE-P [59]	-33.11	1.814	-
ADEMO/D [60]	-30.43	1.77	2.0×10^5
I-PAES [29]	-20.56	1.74	2.5×10^5
SaDE/RGA [54]	-12.42	-	2.8×10^4

Table 5. Results for 1PLW in terms of energy, RMSD and number of Fitness Function Evaluation (FFE). Numerical values reported by the original papers. ’-’ indicates that a value is not available.

Algorithm	Energy (kcal mol^{-1})	RMSD$_{C\alpha}$ (Å)	FFE
Single-objective optimization			
NOMAD-PSP [43]	-1460.75	3.87	2.5×10^5
DE$_{GG-GP}$ [56]	-1156.95	5.69	5.0×10^5
Multi-objective optimization			
MO3 [30]	-1347.46	6.13	-
ADEMO/D [60]	-1301.38	2.14	2.0×10^5
MODE-P [59]	-1050.85	3.84	-
I-PAES [29]	-1037.83	2.27	2.5×10^5

Table 6. Results for 1ZDD in terms of energy, RMSD and number of Fitness Function Evaluation (FFE). Numerical values reported by the original papers. '-' indicates that a value is not available.

Algorithm	Energy (kcal mol^{-1})	RMSD$_{C\alpha}$ (Å)	FFE
Single-objective optimization			
IMMALG-Direct [50]	–1117.24	-	3.0×10^4
Multi-objective optimization			
ADEMO/D [60]	-723.40	4.48	2.0×10^5
I-PAES [29]	-661.48	3.70	2.5×10^5
MO3 [30]	-579.49	5.23	-

Table 7. Results for 1ROP in terms of energy, RMSD and number of Fitness Function Evaluation (FFE). Numerical values reported by the original papers. '-' indicates that a value is not available.

PDB id 1CRN (*Crambin*) is a 46-residue protein with two α-helices and a pair of β-strands and 191 angles to be optimized. Table 8 presents the results based on the works listed. Here the best result was obtained by the multi-objective algorithm (ADEMO/D [60]), followed by the single-objective approach of DE$_{GG-GP}$ [56]. The best RMSD values were also found by the multi-objective algorithms.

The multi-objective approach considering three objectives is relatively new, with the first work appearing in 2017 (MO3 [30]). The only work found with a many objective approach to the PSP problem is the MEAMT [66], which considers four objectives.

A tendency to solve the PSP problem is to consider solvent effect. Several works have been gradually incorporating solvation term as an objective, with promising results: CSSGA [45], ProtPred [63], MEAMT [66], MO3-P [68], MO3 [30], AIMOES [69]. Table 9 presents numerical comparison between some of these algorithms and I-PAES [51], a classic reference to *ab initio* PSP problem that does not use any solvent information. The RMSD values found by the I-PAES are quite competitive, which can also be seen in the numerical comparison made for the proteins most commonly used in the literature (Tables 5 to 8). Results of Table 9 are extracted from the original papers and are shown considering eight peptides with different sizes. RMSD metric is used. For most of the peptides considered, approaches with solvent information reduce the RMSD value.

Optimized proteins have lowest RMSD values. However, results of literature showed that in some cases lower energy values are not associated with lower RMSD values. Some works attribute this fact to the challenges imposed by the PSP problem and that further investigation is needed in the modeling of the problem, as in the case of force fields, for example [72, 60].

8. Conclusions

Proteins are responsible for many different biological functions. The protein structure prediction is an important problem of Bioinformatics, classified as NP-hard, and in the *ab initio* approach can be formulated as a minimization problem, in which it is intended to find the global minimum of a function that estimates the free energy of a protein conformation. Proteins that are related to diseases are of high value for research mainly in the Health area, as they provide a molecular framework with information on pathological processes. This forms the necessary basis for drug development.

In this paper, we present a survey on *ab initio* protein structure prediction approaches based on evolutionary algorithms. Genetic Algorithm, Immune Algorithm, Differential Evolution and other evolutionary methods were considered with single and multi-objective optimization. We present an overview of some works, including specific features and points of the problem modeling and also of the algorithms used. Some numerical results were included for the most tested proteins in the literature.

Despite the evolution of the problem modeling and the

Algorithm	**Energy (kcal mol^{-1})**	**RMSD$_{C\alpha}$ (Å)**	**FFE**
Single-objective optimization			
DE$_{GG-GP}$ [56]	260.12	8.60	5.0×10^5
Multi-objective optimization			
ADEMO/D [60]	253.25	6.06	2.0×10^5
MO3 [30]	363.91	6.91	-
MODE-P [59]	408.53	5.55	-
I-PAES [29]	410.03	4.43	2.5×10^5

Table 8. Results for 1CRN in terms of energy, RMSD and number of Fitness Function Evaluation (FFE). Numerical values reported by the original papers. '-' indicates that a value is not available.

Algorithm	**1WQC (26)**	**2P81 (44)**	**1L2Y (20)**	**3V1A (48)**	**2P6J (52)**	**1ENH (54)**	**1AB1 (46)**
I-PAES	5.23	6.81	3.77	4.31	10.26	11.13	9.09
MEAMT	3.67	-	3.64	-	-	6.56	-
MO3-P	3.25	-	-	3.62	6.74	-	-
MO3	4.65	4.30	3.44	2.23	5.96	11.99	7.52
AIMOES	-	3.77	-	2.32	5.43	5.75	6.23

Table 9. Comparison between algorithms that use solvent information and I-PAES. Results are shown in terms of RMSD (Å) considering eight peptides with different sizes. Numerical values reported by the original papers. '-' indicates that a value is not available.

computational methods applied, the PSP problem is still a challenge. Adaptation, local search and parallelism are still under-explored techniques in addressing the *ab initio* PSP problem. Changes in the computational modeling of the problem, such as considering the effect of the solvent, appear to be a promising trend.

Author contributions

- Lucas Siqueira: Conceptualization, Methodology, Validation, Formal analysis, Writing - Original Draft.
- Sandra Venske: Validation, Writing - Review and Editing, Supervision, Project administration.

References

[1] RANGWALA, H.; KARYPIS, G. Introduction to protein structure prediction. In: ____. **Introduction to Protein Structure Prediction**. Nova Jersey: John Wiley & Sons, Ltd, 2010. cap. 1, p. 1–13.

[2] NELSON, D. L.; COX, M. M. **Lehninger Principles of Biochemistry**. 4. ed. New York: W. H. Freeman, 2004.

[3] TRAMONTANO, A. **Protein Structure Prediction**: Concepts and applications. Nova Jersey: John Wiley and Sons, 2006.

[4] JANA, N. D.; DAS, S.; SIL, J. **A Metaheuristic Approach to Protein Structure Prediction**. [S.l.]: Springer, 2018. v. 31. (Emergence, Complexity and Computation, v. 31).

[5] TRAMONTANO, A. **The Ten Most Wanted Solutions in Protein Bioinformatics**. Boca Raton: CRC Press, 2005. (Chapman & Hall/CRC Mathematical and Computational Biology).

[6] NUNES, L. F. et al. An integer programming model for protein structure prediction using the 3d-hp side chain model. **Discrete Applied Mathematics**, Elsevier, Amsterdã, v. 198, p. 206 – 214, 2016.

[7] SILVA, R. S. P. R. S. A multistage simulated annealing for protein structure prediction using rosetta. In: **Computer on the Beach**. [S.l.]: Universidade do Vale do Itajaí - UNIVALI, 2018. p. 850–859.

[8] GENDREAU, M.; POTVIN, J.-Y. **Handbook of Metaheuristics**. 2. ed. Nova Iorque: Springer, 2010.

[9] HOLLAND, J. H. **Adaptation in Natural and Artificial Systems**. 2. ed. Ann Arbor: University of Michigan Press, 1975.

[10] KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **Science**, American Association for the Advancement of Science, Washington, v. 220, n. 4598, p. 671–680, 1983.

[11] KENNEDY, J.; EBERHART, R. C. Particle swarm optimization. In: INTERNATIONAL CONFERENCE ON NEU-

RAL NETWORKS, 1995, Perth. **Proceedings of the [...]**. Piscataway: IEEE, 1995. p. 1942–1948.

[12] DORIGO, M.; STÜTZLE, T. **Ant Colony Optimization**. Holland: Bradford Company, 2004.

[13] KARABOGA, D.; BASTURK, B. Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. In: MELIN, P. et al. (Ed.). **Foundations of Fuzzy Logic and Soft Computing**. Berlin, Heidelberg: Springer-Verlag, 2007. p. 789–798.

[14] GEEM, Z.; KIM, J.; LOGANATHAN, G. A new heuristic optimization algorithm: Harmony search. **Simulation**, SAGE Publications Ltd, Thousand Oaks, v. 76, n. 2, p. 60–68, jun. de 2001.

[15] YANG, X.-S.; DEB, S. Engineering optimisation by cuckoo search. **International Journal of Mathematical Modelling and Numerical Optimisation**, Inderscience Publishers, Genève, v. 1, n. 4, p. 330–343, 2010.

[16] STORN, R.; PRICE, K. Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. **Journal of Global Optimization**, Kluwer Academic Publishers, Hingham, MA, USA, v. 11, n. 4, p. 341–359, 1997.

[17] HUNT, J. E.; COOKE, D. E. Learning using an artificial immune system. **Journal of Network and Computer Applications**, Elsevier, v. 19, n. 2, p. 189 – 212, 1996.

[18] GUJRATHI, A.; BABU, B. **Evolutionary computation**: Techniques and applications. Nova Jersey: Apple Academic Press, 2016.

[19] XU, Y.; XU, D.; LIANG, J. **Computational Methods for Protein Structure Prediction and Modeling**: Volume 1: Basic characterization. Nova Iorque: Springer, 2006.

[20] HARRISON, J. et al. Review of force fields and intermolecular potentials used in atomistic computational materials research. **Applied Physics Reviews**, College Park, v. 5, p. 031104, set. de 2018.

[21] XU, P. et al. Advancement of polarizable force field and its use for molecular modeling and design. In: ____. **Advance in Structural Bioinformatics**. Dordrecht: Springer Netherlands, 2015. p. 19–32.

[22] SNEHA, P.; George Priya Doss, C. Chapter seven - molecular dynamics: New frontier in personalized medicine. In: DONEV, R. (Ed.). **Personalized Medicine**. Cambridge: Academic Press, 2016, (Advances in Protein Chemistry and Structural Biology, v. 102). p. 181–224.

[23] BROOKS, B. R. et al. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. **Journal of Computational Chemistry**, Cambridge, v. 4, n. 2, p. 187–217, 1983.

[24] MACKEREL, A. D. et al. CHARMM: The energy function and its parameterization with an overview of the program. In: ____. Chichester: John Wiley & Sons, 1998. v. 1, p. 271–277.

[25] JORGENSEN, W. L.; TIRADO-RIVES, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. **Journal of the American Chemical Society**, Washington, v. 110 6, p. 1657–66, 1988.

[26] JORGENSEN, W. L.; MAXWELL, D. S.; TIRADO-RIVES, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. **Journal of the American Chemical Society**, Washington, v. 118, p. 11225–11236, 1996.

[27] CORNELL, W. D. et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. **Journal of The American Chemical Society**, v. 117, p. 5179–5197, 1995.

[28] GUNSTEREN, W. F. van; BERENDSEN. **Groningen Molecular Simulation (GROMOS) Library Manual**. Groningen: Biomos, 1987.

[29] CUTELLO, V.; NARZISI, G.; NICOSIA, G. A multi-objective evolutionary approach to the protein structure prediction problem. **Journal of the Royal Society Interface**, v. 3, n. 6, p. 139–151, 2006.

[30] GAO, S. et al. Incorporation of solvent effect into multi-objective evolutionary algorithm for improved protein structure prediction. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, Piscataway, v. 15, n. 4, p. 1365–1378, jul. de 2018.

[31] SHAIK, N. et al. **Essentials of Bioinformatics**: Volume i: Understanding bioinformatics: Genes to proteins. Nova Iorque: Springer International Publishing, 2019.

[32] PEVSNER, J. **Bioinformatics Genomics and Functional**. 3. ed. Hoboken: Wiley Blackwell, 2015.

[33] BERG, J.; TYMOCZKO, J.; STRYER, L. **Biochemistry**. New York: W. H. Freeman, 2007. (Biochemistry (Berg)).

[34] LI, D.; HWANG, K.; HUANG, Y. Coarse grained modeling of biopolymers and proteins: Methods and applications. **International Journal of Applied Mechanics**, New Jersey, v. 1, n. 1, p. 113–136, mar. de 2009.

[35] KMIECIK, S. et al. Coarse-grained protein models and their applications. **Chemical Reviews**, Washington, v. 116, n. 14, p. 7898–7936, 2016.

[36] MOSS, D.; JELASKA, S.; PONGOR, S. **Essays in Bioinformatics**. Amsterdã: IOS Press, 2005. v. 368. (NATO Science Series: Life and Behavioural Sciences, v. 368).

[37] ZEMLA, A. Lga: A method for finding 3d similarities in protein structures. **Nucleic Acids Research**, Oxford, v. 31, n. 13, p. 3370–3374, jul. de 2003.

[38] DEZA, M.; DEZA, E. **Encyclopedia of Distances**. Berlin: Springer-Verlag, 2014.

[39] ZHANG, Y.; SKOLNICK, J. Scoring function for automated assessment of protein structure template quality. **Proteins: Structure, Function, and Bioinformatics**, Hoboken, v. 57, n. 4, p. 702–710, 2004.

[40] HAN, X.; LI, L.; LU, Y. Selecting near-native protein structures from predicted decoy sets using ordered graphlet degree similarity. **Genes**, Basel, v. 10, n. 2, 2019.

[41] ENGELBRECHT, A. P. **Computational Intelligence**: An introduction. 2. ed. Hoboken: Wiley Publishing, 2007.

[42] LIU, Y.-L.; TAO, L. An improved parallel simulated annealing algorithm used for protein structure prediction. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS, 2006, Dalian. **Proceedings of the [...]**. Piscataway: IEEE, 2006. p. 2335–2338.

[43] NICOSIA, G.; STRACQUADANIO, G. Generalized pattern search and mesh adaptive direct search algorithms for protein structure prediction. In: INTERNATIONAL CONFERENCE ON ALGORITHMS IN BIOINFORMATICS, 7., Philadelphia, PA. **Proceedings of the [...]**. Berlin, Heidelberg: Springer-Verlag, 2007. (WABI'07), p. 183–193.

[44] YOUSEF, M.; ABDELKADER, T.; ELBAHNASY, K. A hybrid model to predict proteins tertiary structure. In: INTERNATIONAL CONFERENCE ON COMPUTER ENGINEERING AND SYSTEMS (ICCES), 12., Cairo. **Proceedings of the [...]**. Piscataway: IEEE, 2017. p. 85–91.

[45] CUSTÓDIO, F. L.; BARBOSA, H. J. C.; DARDENNE, L. E. Full-atom ab initio protein structure prediction with a genetic algorithm using a similarity-based surrogate model. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION, 2010, Barcelona. **Proceedings of the [...]**. Piscataway: IEEE, 2010. p. 1–8.

[46] VENKATESAN, A. et al. Computational approach for protein structure prediction. **Healthcare Informatics Research**, Seoul, v. 19, n. 2, p. 137–147, 2013.

[47] BECERRA, D. et al. A parallel multi-objective ab initio approach for protein structure prediction. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2010, Hong Kong. **Proceedings of the [...]**. Piscataway: IEEE, 2011. p. 137–141.

[48] DAVIS, L. (Ed.). **Handbook of Genetic Algorithms**. New York: Van Nostrand Reinhold, 1991.

[49] GONG, M.; JIAO, L.; ZHANG, L. Baldwinian learning in clonal selection algorithm for optimization. **Information Sciences**, Elsevier, Amsterdã, v. 180, n. 8, p. 1218–1236, 2010.

[50] ANILE, A. M. et al. Determination of protein structure and dynamics combining immune algorithms and pattern search methods. **Natural Computing**, Basingstoke, v. 6, n. 1, p. 55–72, 2007.

[51] CUTELLO, V.; NARZISI, G.; NICOSIA, G. A class of pareto archived evolution strategy algorithms using immune inspired operators for ab-initio protein structure prediction. In: WORKSHOPS ON APPLICATIONS OF EVOLUTIONARY COMPUTATION (EVOWORKSHOPS), 3., 2005, Lausanne, Switzerland. **Proceedings of the [...]**. Berlin, Heidelberg: Springer-Verlag, 2005. (Lecture Notes in Computer Science (LNCS), v. 3449), p. 54–63.

[52] CUTELLO, V.; NARZISI, G.; NICOSIA, G. Computational studies of peptide and protein structure prediction problems via multiobjective evolutionary algorithms. In: KNOWLES, J. et al. (Ed.). **Multiobjective Problem Solving from Nature**. [S.l.]: Springer-Verlag, 2008, (Natural Computing Series). p. 93–114.

[53] ANGELINE, P. J. et al. (Ed.), 1999, Washington. **The Pareto Archived Evolution Strategy**: A new baseline algorithm for pareto multiobjective optimisation, v. 1. Congress on Evolutionary Computation, Piscataway: IEEE Press, 2002. 98–105 p.

[54] SUDHA, S.; BASKAR, S.; KRISHNASWAMY, S. Protein tertiary structure prediction using evolutionary algorithms. **International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)**, IASIR, Georgia, USA, v. 3, n. 3, p. 338–348–595, 2013.

[55] SUDHA, S. et al. Protein structure prediction using diversity controlled self-adaptive differential evolution with local search. **Soft Computing**, Nova Iorque, v. 19, n. 6, p. 1635–1646, jun. de 2015.

[56] NARLOCH, P. H.; PARPINELLI, R. S. Diversification strategies in differential evolution algorithm to solve the protein structure prediction problem. In: MADUREIRA, A. M. et al. (Ed.). **Intelligent Systems Design and Applications**. Cham: Springer International Publishing, 2017. p. 125–134.

[57] HAO, X.; ZHANG, G.; ZHOU, X. Guiding exploration in conformational feature space with lipschitz underestimation for ab-initio protein structure prediction. **Computational Biology and Chemistry**, v. 73, p. 105 – 119, abr. de 2018.

[58] RAKHSHANI, H. et al. Speed up differential evolution for computationally expensive protein structure prediction problems. **Swarm and Evolutionary Computation**, v. 50, p. 100493, 2019.

[59] VENSKE, S. M.; GONÇALVES, R. A.; DELGADO, M. R. Differential evolution to multi-objective protein structure prediction. In: BIOINFORMATICS 2012 - INTERNATIONAL CONFERENCE ON BIOINFORMATICS MODELS, METHODS AND ALGORITHMS, 2012, Vilamoura. **Proceedings of the [...]**. Setúbal: SciTePress - Science and Technology Publications, 2012. p. 295–298.

[60] VENSKE, S. M. et al. ADEMO/D: An adaptive differential evolution for protein structure prediction problem. **Expert Systems with Applications**, Amsterdã, v. 56, p. 209 – 226, 2016.

[61] YANG, X.; DEB, S. Cuckoo search via lévy flights. In: WORLD CONGRESS ON NATURE BIOLOGICALLY INSPIRED COMPUTING (NABIC), 2009, Coimbatore. **Proceedings of the [...]**. Piscataway: IEEE, 2010. p. 210–214.

[62] BEYER, H.-G.; SCHWEFEL, H.-P. Evolution strategies: A comprehensive introduction. **Natural Computing**, Nova Iorque, v. 1, p. 3–52, mar. de 2002.

[63] FACCIOLI, R. et al. A mono-objective evolutionary algorithm for protein structure prediction in structural and energetic contexts. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION, 2012, Brisbane. **Proceedings of the [...]**. Piscataway: IEEE, 2012. p. 1–7.

[64] OLSON, B.; JONG, K. D.; SHEHU, A. Off-lattice protein structure prediction with homologous crossover. In: ANNUAL CONFERENCE ON GENETIC AND EVOLUTIONARY COMPUTATION, 15., 2013, Amsterdam, The Netherlands. **Proceedings of the [...]**. New York, NY, USA: ACM, 2013. (GECCO ’13), p. 287–294.

[65] CALVO, J.; ORTEGA, J.; ANGUITA, M. PITAGORAS-PSP: Including domain knowledge in a multi-objective approach for protein structure prediction. **Neurocomputing**, Amsterda, v. 74, n. 16, p. 2675 – 2682, 2011.

[66] BRASIL, C. R. S.; DELBEM, A. C. B.; SILVA, F. L. B. da. Multiobjective evolutionary algorithm with many tables for purely ab initio protein structure prediction. **Journal of Computational Chemistry**, v. 34, n. 20, p. 1719–1734, 2013.

[67] RAMYACHITRA, D.; AJEETH, A. Modcsa-ca: A multi objective diversity controlled self adaptive cuckoo algorithm for protein structure prediction. **Gene Reports**, Amsterdã, v. 8, p. 100 – 106, 2017.

[68] SONG, Z. et al. A preference-based multi-objective evolutionary strategy for ab initio prediction of proteins. In: INTERNATIONAL CONFERENCE ON PROGRESS IN INFORMATICS AND COMPUTING (PIC), 2017, Nanjing. **Proceedings of the [...]**. Piscataway: IEEE, 2017. p. 7–12.

[69] SONG, S. et al. AIMOES: Archive information assisted multi-objective evolutionary strategy for ab initio protein structure prediction. **Knowledge-Based Systems**, v. 146, p. 58 – 72, 2018.

[70] SONG, S. et al. Adoption of an improved PSO to explore a compound multi-objective energy function in protein structure prediction. **Applied Soft Computing**, v. 72, p. 539 – 551, nov. 2018.

[71] ISHIBUCHI, H.; TSUKAMOTO, N.; NOJIMA, Y. Evolutionary many-objective optimization: A short review. In: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION (IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE), 2008, Hong Kong. **Proceedings of the [...]**. Piscataway: IEEE, 2008. p. 2419–2426.

[72] SHATABDA, S. et al. How good are simplified models for protein structure prediction? **Advances in Bioinformatics**, Londres, v. 2014, apr 2014.