Multi-Objective Differential Evolution Algorithms for the Protein Structure Prediction Problem

Pedro Henrique Narloch
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
phnarloch@inf.ufrgs.br

Mathias J. Krause
Department of Mathematics
Karlsruhe Institute of Technology
Kalrsruhe, Germany
mathias.krause@kit.edu

Márcio Dorn
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
mdorn@inf.ufrgs.br

***Abstract*—The structural analysis of proteins is an essential step for understanding their biological function. However, the process of the structural determination of these molecules is expensive and time-consuming. In order to reduce these factors, computational methods might be a provocative approach, despite the complexity associated with it. Over the decades, different computational approaches were proposed as well as different energy force fields. As the force fields consider conflicting terms in its composition, multi-objective optimization approaches showed to be suitable to the Protein Structure Prediction problem. In this way, the objective of the current work is to evaluate and compare three multi-objective algorithms, the Non-Dominated Sorting Genetic Algorithm in its second version, the Generalized Differential Evolution in its third version, and the Differential Evolution Multi-Objective. We split the *score3* energy function provided by Rosetta into a bi-objective problem. The first objective considers only the non-bonded *van der Waals*, while the second one is composed of bonded-terms and a secondary structure reinforcement score. Moreover, structural information provided by the Angle Probability List is considered, since this kind of information proved to be reliable in single-objective approaches. Results obtained are analyzed using GDT and RMSD metrics, showing the better capability of Differential Evolution based methods for the problem.**

***Index Terms*—multi-objective optimization, protein structure prediction, structural bioinformatics, evolutionary algorithms**

I. Introduction

Proteins are vital molecules for every living organism since these macromolecules act in different biological functions. Their three-dimensional shape dictates which kind of biological function it will assume, including harmful behaviors if the misfolding process occurs [1]. In order to determine the three-dimensional shape of a protein, and its function, there two experimental techniques: X-ray crystallography and Nuclear Magnetic Resonance (NMR). Despite the efficiency of these techniques, they are expensive and time-consuming. In this sense, different researchers have been working together to predict the structure of the protein by computational means, creating one of the most challenging problems in the Structural Bioinformatics known as Protein Structure Prediction (PSP).

The prediction of proteins is mainly based on the Anfinsen thermodynamics hypothesis [2], which states that the native state (functional state) of a protein is reached only by the sequence of its amino acids (primary structure) and environmental conditions. In this way, the minimum possible free energy represents the native structure (functional form). In order to create a possible solver for the PSP problem, there are needed three primary definitions: **(i)** the computational representation of the molecule; **(ii)** a way to measure the energy of these molecules; and **(iii)** an algorithmic way to explore the conformational search space [3]. Although there are different types of molecular representation and various ways to measure the energy of a protein, there is no efficient algorithm that can thoroughly explore the conformational search space finding the global minimum possible energy. This restriction is related to the increasing amount of possible conformations accordingly to the protein's size. From the Computer Science perspective, the PSP problem is considered an NP-Hard problem [4], turning it computationally expensive.

In light of this fact, metaheuristics became interesting approaches to the problem due to their capability of finding good solutions in a huge search space [5], even though these methods do not guarantee the best solution at the end of the optimization process. For the PSP problem, different metaheuristics where explored [6] [7] [8] [9] [10], but it is still not known whether algorithm can solve the problem entirely. Moreover, recent studies have pointed that the PSP problem might be a multi-objective problem, since there are conflicts among the energy terms in the force fields [11], requiring a more extensive exploration of the capabilities of different multi-objective metaheuristics for the problem.

In order to provide a larger body of analysis regarding multi-objective algorithms for the PSP problem, this work evaluates three different multi-objective approaches (NSGA-II [12], DEMO [13], and GDE3 [14]) using the centroid (*score3*) energy function provided by the PyRosetta [1] [15], something not yet covered by the literature. Besides, the Angle Probability List (APL) [7] is used as a source of information to enhance the generated solutions. Different studies pointed [7] [9] [6] [16] the efficiency of APL for the problem. Results obtained by our study showed that the Differential Evolution (DE) variations achieved better results in comparison with the Genetic Algorithm (GA) algorithm.

The overall structure of the current work takes the shape of 4 main sections. The Section II explores the main concepts

[1] http://www.pyrosetta.org

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

regarding the problem and the algorithm used. Also, related works are exposed and discussed. The methodology adopted is then described in Section III, while in Section IV results are explored and discussed. Finally, in Section V we conclude our work, highlighting the main points of our research, the future works, and the limitations of the current study.

II. Preliminaries

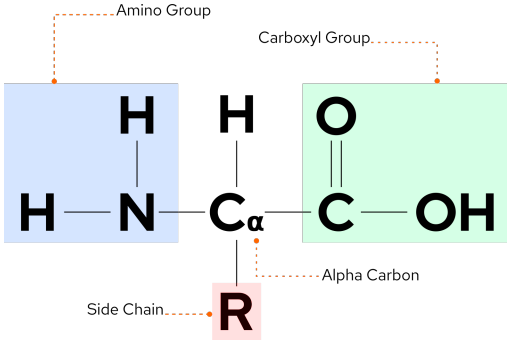
A. Protein Structure Prediction

Proteins are macromolecules composed of a chain of amino acids. This linear sequence of amino acids is known as the protein's primary structure, and together with the environmental conditions, it dictates the final functional structure of each sequence [2]. During the folding process, the amino acids are arranged into space, creating local geometrical shapes called protein's secondary structure. There are different secondary structures, depending on the type of classification, but the most common ones are the α-helices and β-sheets. At the end of the folding process, the global adjustment of the secondary structures forms the functional three-dimensional shape, known as protein's tertiary structure and native conformation. In order to exert specific biological functions, proteins can cooperate among themselves while forming what is called the quarternary structure of proteins [1].

The main objective of the PSP problem is to find the tertiary structure of a protein, using a determined molecular representation and a way of measuring the quality of the predicted molecule. As explained in [3], it is possible to qualify the types of predictions in four classifications accordingly to the prediction method. The current work can be classified as a *de novo* method since we follow the thermodynamic hypothesis in combination with the APL as a source of information, improving the quality of predicted structures.

Molecular Representation: For template-free approaches, different degrees of molecular representation and energy measures can be used. Most of the works use the torsion angles representations [10] [17] [18] [6] [11] [19] [8], allowing reduction of the computational cost while keeping the biological plausibility. The torsion angles are found among the amino acids. All amino acids have the same backbone structure with three torsion angles: ϕ ($\mathtt{N{-}C_\alpha}$), ψ ($\mathtt{C_\alpha{-}C}$), and ω ($\mathtt{C{-}N}$) as shown in Fig. 1.

Fig. 1: Molecular Structure of an amino acid.



Additionally, there are the χ angles representing the side-chain, which distinguishes each amino acid. The number of χ angles differ from each amino acid, ranging from 0 to 4 angles. Excepting the ω angle, usually set to $+180.0°$ value due to its bonding planarity, all other angles can assume values between $-180.0°$ and $+180.0°$. As in this work we use a centroid energy function, χ angles are not taken in consideration.

Energy Function: In order to quantify the structure of a protein, it is needed to use some energy function that describes the different forces regarding atomic interactions. For this purpose, it would be interesting to use quantum mechanics measurements, but they are impracticable due to the high computational cost of it. Instead of quantum mechanics, classical physics measurements are used in order to determine the energy of a protein [11]. Different works have proposed different force fields, but in general, they consider bonded and non-bonded terms. Among all force fields, the Rosetta force field [20] showed promising results and one of the most used for high-performance predictors. The Rosetta force field contains more than 18 terms, considering classical physics measurements and knowledge-based terms.

Angle Probability List: The APL was proposed in [7] based on the study of conformational preferences of amino acids, further analyzed in [21]. Thus far, a growing number of studies have shown that the APL improves the capability of finding more accurate structures than pure *ab initio* (without any source of information) approaches [6] [7] [22]. The APL is composed by several high-quality structures found in the Protein Data Bank (PDB[2]) [23], with resolution $\leq$ 2.5Å. With this information, the APL creates a histogram matrix of $[-180, 180] \times [-180, 180]$ of the angle preferences for each amino acid and the secondary structure. Authors of APL also provided a web interface for APL generation called Neighbors Influence of Amino acids and Secondary structures (NIAS[3]) [24], which is open and free to use.

B. Multi-Objective Optimization

Besides the importance of single-objective studies and their benchmarks, most of them do not accurately describe real-world problems. Real-world problems are better depicted by a multi-objective formulation, representing different aspects of the problem that might conflict with each other [25]. The multi-objective formulation can be defined a minimization (or maximization) of m objective functions: $\min(f(x)_1, f(x)_2, f(x)_3, \ldots f(x)_m)$, with $x \in X$, where X defines the set of feasible solutions (x) in the decision space.

As the m possible objectives are related to different aspects of the problem, they might conflict, leading to a non-trivial decision of which solution is the best one in a set of solutions (something that does not occur in single-objective optimization). In this sense, the concept of Pareto dominance is essential to multi-objective optimization, allowing the direct comparison of two possible solutions for the problem [26]. In

[2]www.rcsb.org
[3]http://sbcb.inf.ufrgs.br/npas/

this sense, a solution is non-dominated if no other solution in the data set offers a better value for one objective without incurring a worse value for other objectives.

With the application of the Pareto dominance concept, it is possible to select a set of feasible candidates that better describe a solution for the problem. Thus, leaving the decision-maker agent in a better position to choose which solution should be selected [27].

C. Related Works

Over the years, different approaches have been proposed to solve the PSP problem. Although substantial progress was obtained, the PSP problem is still an open problem in structural bioinformatics. Initial evidences regarding the multi-objectiveness of the PSP problem was revealed in [11] [28], where the CHARMM energy function was split into two different objectives, one considering the bonded force while the second considered the non-bonded. The optimization method was the IPAES algorithm, an evolutionary strategy for multi-objective problems enhanced with imuno-inspired operators. Following the same approach for the energy function, an Adaptive Differential Evolution Multi-Objective based on Decomposition (ADEMO/D) was proposed in [19] and compared with the IPAES algorithm. Results obtained by both works showed that the multi-objective formulation of the problem could reach similar structures using different evolutionary algorithms.

As the multi-objective formulation of the problem showed to be interesting, other approaches appeared, considering more than two objectives in their formulation. In [29] the NSGA-II algorithm was used with a three-objective formulation of the problem, considering bonded terms as the first objective, the *van der Waals* as the second objective, and other non-bonded terms as the third objective. Authors compared their work with the IPAES approach used in [11], qualifying the NSGA-II, one of the most well known multi-objective algorithms, as a competitive algorithm with state-of-art approaches. Another three-objective approach was proposed in [30], considering bonded, non-bonded, and structural differences as three objectives, and by [31] which considered the solvent information as one of the objectives.

More recently, in [32] the DEMO algorithm was tested and compared with other research works, including the IPAES [11] and ADEMO/D [19] approaches. This work had the objective of evaluating for the first time a multi-objective approach using the APL information. Also, the DEMO algorithm was never tested in the PSP problem. The results obtained by the authors showed that DEMO could reach reasonable solutions regarding RMSD and GDT metrics.

Although significant contributions were reported in different works, only in [30] some structural information that did not come from APL was explored. The structural information has been employed in different single-objective works and one multi-objective approach [32], showing their relevance to the problem. Among different sources of information, the APL showed to be a crucial source of information for better final conformations. Evidences could be found in [7] with the canonical versions of GA and Particle Swarm Optimization, in [16] with different versions of the DE algorithm, and in [32] where a multi-objective version of DE (DEMO) was used.

In general, it is possible to identify a gap between the single and multi-objective approaches, where relevant sources of information were barely used in multi-objective approaches. Also, we did not found in the body of research some comparative work that used the same energy approach with the three algorithms used in this work (NSGA-II, DEMO, and GDE3), showing another contribution of this paper.

III. Methodology

A. Energy Function

In order to evaluate the structures and guide the search mechanisms, we have selected the Rosetta *score3* as energy function. Energy functions provided by the Rosetta modeling software are the most used ones in high-resolution predictors. As proposed in [32], and following the research line of other works, we are going to split the energy function into two objectives. One objective containing the *van der Waals* term, while the second one considers the bonded and knowledge-based terms. This definition follows the intuitive argument presented in [11] and in [19], where bonded (local interaction) and non-bonded (non-local) interactions are in conflict. In addition to the second objective, we include the secondary structure reinforcement score, promoting solutions with well-formed secondary structures as determined by the DSSP [33] included in the PyRosetta package.

$$\begin{aligned} f(1) = \quad & E_{vdW}, \\ f(2) = \quad & E_{cenpack} + E_{pair} + E_{env} + E_{cbeta} + \\ & E_{rg} + E_{hs\_pair} + E_{ss\_pair} + E_{rsigma} + \\ & E_{sheet} + SS_{reinforcement} \end{aligned} \tag{1}$$

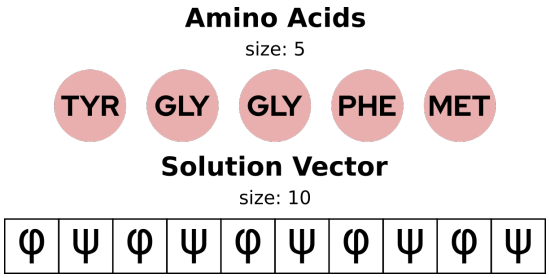
where E_{vdW} stands for *van der Waals* forces and composes the first objective function. Other bonded and knowledge-based terms ($\mathrm{E}_{cenpack}$, E_{pair}, E_{env}, E_{cbeta}, E_{rg}, E_{hs_pair}, E_{ss_pair}, E_{rsigma}, and E_{sheet}) compose the second objective function. In order to benefit solutions with matching secondary structures the $\mathrm{SS}_{reinforcement}$ is add to the second objective function. The same objective formulation was used in [32] with the DEMO algorithm.

B. Molecular Representation

As the *score3* is a centroid based energy function, our computational representation of proteins is composed by the dihedral angles related to the backbone of amino acids, in other words, the ϕ and ψ angles must be optimized (Fig. 2), while the ω fixed to $+180^{\circ}$ due to its planar characteristic, and χ angles are not optimized.

In this sense, it is possible to determine the dimensionality of each solution as $2N$, where N is the size of the primary structure (quantity of amino acids).

Fig. 2: Computational Representation.



C. Search Mechanisms

As the objective, and contribution, of the current work, is to compare different multi-objective metaheuristics using APL as a source of information, we have selected two versions of the DE algorithms (GDE3 [14] and DEMO [13]), and the well-known NSGA-II [12] as search mechanisms. It is essential to state that only DEMO has been tested with the *score3* energy function in a bi-objective formulation using APL [32]. At the same time, the NSGA-II and GDE3 were not yet evaluated with it, neither with the *score3* energy function or with APL information. We used the jMetalPy framework [34] to prototype this work since NSGA-II and GDE3 were available. We have added the DEMO algorithm and modeled the PSP problem inside of the framework. The APL information is used only in the population initialization procedure.

- **NSGA-II**:The NSGA-II algorithm [12] is one of the most well-known algorithms applied to multi-objective optimization. As the NSGA-II be a GA, it keeps all the mechanisms such as selection, crossover, and mutation. The main difference from the canonical GA and the NSGA-II is related to the process of creating the new population. At the first moment, all solutions are evaluated and organized with the non-dominated sorting algorithm. Each solution receives a ranking position according to the non-dominated sorting algorithm without replacing any solution during the recombination process. In this way, the offspring is formed by the union of the current population individuals and the newly generated ones. After the recombination, mutation, and all solutions have their ranking assigned, and the best solutions have to be selected to create the next generation. For this purpose, solutions from the best rankings are copied to the next generation until the number of solutions is equal to the predefined population size. As a tie-breaker for solutions of the same ranking, the crowding distance metric defines which solution will be discarded or not. With this procedure, the algorithm keeps the elitism characteristic by selecting the well-ranked solutions as well as the diversity since the crowding distance metric is used as a tie-breaker, maintaining solutions with different characteristics.
- **GDE3**: To create an algorithm that works for constrained, unconstrained, and many-objectives problem, the generalized differential evolution (GDE) was proposed, modifying only the selection criterion in comparison with the canonical DE [35]. In 2005 the third version of the GDE [14] was launched, considering the NSGA-II strategies for non-dominated sorting and crowding distance. As in NSGA-II, the offspring grows two times the population size, requiring a selection of the individuals that compose the next generation. It follows the same step as the NSGA-II, where the individuals with the best non-domination rank go to the next generation, using the crowding distance as a tie-breaker. However, the GDE3 algorithm modified the diversity preservation technique, where each time an individual is dropped from the population, the crowding distance is re-calculated, something that does not happen in the NSGA-II or DEMO algorithm. The GDE algorithm also uses the feasibility of the generated solution, deciding if the solution dominates or not the parent. However, in our case, all solutions are feasible. In this sense, the definition of dominance is only related to the objective space, and not the constraint space.
- **DEMO**: The Differential Evolution Multi-Objective was firstly proposed in [13] as a new DE version for multi-objective problems. The DEMO is based in the NSGA-II algorithm, but using the differential evolution mutation and crossover operators. As the canonical version of DE, a new individual is generated by the mutation and crossover mechanism. What differentiates the DEMO and the canonical DE proposed by Storn and Price [36] is the replacement operator and a truncation mechanism. For the replacement step, the algorithm takes into consideration the dominance factor. Three different possibilities could happen, **(i)** the new generated individual dominates the parent, replacing the parent in the population; **(ii)** the parent dominates the new individual, discarding the new individual, and **(iii)** no dominance state. If there is no dominance, the new individual is appended into the population, enabling the mutation mechanism to choose the new individuals in its process right away. As the population can grow during the mutation process (being possible of achieving the size of 2 times the population size), the truncation step kicks in, ranking the solutions with the non-dominated sorting algorithm and crowding distance (as in NSGA-II). In this way, the best-ranked solutions compose the population of the next generation. The size of this set obeys the population size parameter. This process goes through until a stop-criterion is met.

As exposed, the three algorithms are quite similar to each other. As the NSGA-II be one of the most famous multi-objective algorithms due to its efficiency, both GDE3 and DEMO are based on its non-dominated sorting and crowding distance strategies to determine the new generation of possible solutions. Although GDE3 and DEMO being very similar, it is essential to state that GDE was designed to work with more

than two objectives, and considering the constraint violation as information to decide which solution will be discarded or not. That is not the case of DEMO. The idea of using newly generated individuals in the mutation mechanism, and the immediate replacement of dominated parents, are the core of the algorithm, as stated by their authors [13].

IV. Results and Analysis

In order to compare different algorithms, we have selected the same test set with eight proteins as in [32]. The selection of the proteins took into consideration different sizes (ranging from 29 to 73 amino acids) and secondary structures (Tab. I). To keep it a fair comparison, the parameters of GDE3 and DEMO were kept the same as in [32], with $F = 0.5$ and $CR = 0.9$. The NSGA-II algorithm follows the number of individuals of 100 and fitness evaluations of 1 million, such as the GDE3 and DEMO. For DE approaches, the classical rand/1/bin mutation ($\vec{v}_i^{g+1} = \vec{x}_{r1}^{g} + F \cdot (\vec{x}_{r2}^{g} - \vec{x}_{r3}^{g})$) version is used. For NSGA-II, the mutation and crossover mechanisms are the ones described in its original publication [12]. For each protein, 30 runs were done because of the stochastic behavior of the algorithms.

TABLE I: Target protein sequences [32].

PDB ID	Size	Secondary Structure
1ACW	29	$\alpha + \beta$
1ZDD	34	α
2MR9	44	α
2P81	44	α
1CRN	46	$\alpha + \beta$
1ENH	54	α
1ROP	63	α
1AIL	73	α

In order to compare the predicted structures with the experimental model, we have chosen two metrics, the Root Mean Square Deviation (RMSD) and the Global Distance Test (GDT). The RMSD is widely adopted in different works, and it compares the C_α (central carbon atom that bond the amino acid side-chain) positions between two structures. The Equation 2 displays how the RMSD metric is calculated.

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

where r_{ai} and r_{bi} are the ith atom in a set of n atoms from the compared structures (a and b). This RMSD equation returns a distance measure in Å. The closer the value of 0, the more similar the structures are. As the RMSD could be very sensitive to coils (very flexible structures), we also used the GDT index to qualify our results. The GDT compares two structures by superimposing the C_α, as is done in the RMSD. However, the GDT calculates different positions accordingly to different distance cutoffs, reducing the sensitivity of loops and coils. The GDT measurement returns a percentage value describing how similar the structures are. In this way, the higher the percentage obtained, the more similar the structures are.

As the Pareto front could return multiple solutions, we selected the best solution regarding the GDT index of each run, leading us to evaluate 30 solutions for each protein, for each algorithm. In addition, we run the *FastRelax* protocol provided by the Rosetta package. This protocol makes small backbone and sidechain movements, reducing possible steric clashes within atoms, thus reducing energy.

The results obtained to compare the algorithms are found in Tab. II with four columns. The first one indicates the PDB identification of each predicted structure, followed by the tested algorithms in the second column. Columns 3 and 4 bring the GDT and RMSD, respectively. Highlighted cells express the best values accordingly to the measurement index (GDT and RMSD).

TABLE II: Results obtained from the three algorithms: NSGA-II, GDE3, and DEMO

PDB ID	Algorithm	GDT (%)	RMSD (Å)
1ACW	NSGA-II	$57.93(43.70 \pm 7.94)$	$3.81(6.47 \pm 1.58)$
	GDE3	$65.51(50.75 \pm 6.03)$	$3.63(6.56 \pm 1.72)$
	DEMO	$62.75(48.89 \pm 6.07)$	$3.82(7.17 \pm 1.81)$Å
1AIL	NSGA-II	$31.42(22.91 \pm 3.54)$	$7.07(10.30 \pm 1.38)$
	GDE3	$67.14(48.10 \pm 8.36)$	$3.25(6.77 \pm 2.81)$
	DEMO	$61.71(48.80 \pm 6.97)$	$3.14(7.40 \pm 2.55)$Å
1CRN	NSGA-II	$42.60(31.04 \pm 4.58)$	$6.23(9.24 \pm 1.50)$
	GDE3	$56.52(40.55 \pm 7.08)$	$5.13(9.62 \pm 2.69)$
	DEMO	$50.00(38.81 \pm 4.71)$	$6.32(9.31 \pm 2.79)$Å
1ENH	NSGA-II	$39.62(26.86 \pm 3.79)$	$6.84(10.14 \pm 1.31)$
	GDE3	$73.70(46.22 \pm 8.73)$	$3.10(8.09 \pm 2.79)$
	DEMO	$72.96(49.07 \pm 7.82)$	$4.29(7.47 \pm 1.67)$Å
1ROP	NSGA-II	$42.14(26.65 \pm 5.33)$	$5.96(10.86 \pm 1.85)$
	GDE3	$68.92(45.22 \pm 8.21)$	$2.74(7.05 \pm 2.07)$
	DEMO	$66.42(46.01 \pm 6.96)$	$3.04(6.80 \pm 1.61)$Å
1ZDD	NSGA-II	$61.17(41.50 \pm 9.27)$	$3.47(6.04 \pm 1.29)$
	GDE3	$87.05(60.76 \pm 9.58)$	$1.79(4.05 \pm 1.16)$
	DEMO	$93.52(62.72 \pm 12.92)$	$1.19(4.01 \pm 1.98)$Å
2MR9	NSGA-II	$43.18(32.34 \pm 4.94)$	$6.16(8.29 \pm 1.29)$
	GDE3	$71.36(48.10 \pm 8.41)$	$3.00(7.09 \pm 1.87)$
	DEMO	$70.45(48.57 \pm 8.82)$	$2.62(7.21 \pm 1.87)$Å
2P81	NSGA-II	$37.72(30.50 \pm 4.12)$	$5.85(8.94 \pm 1.30)$
	GDE3	$69.09(53.51 \pm 6.06)$	$4.78(7.10 \pm 1.34)$
	DEMO	$69.09(49.81 \pm 6.61)$	$5.06(7.72 \pm 1.44)$Å

A. GDT and RMSD Analysis

When comparing the results among the three algorithms, two aspects are relevant to the problem. In the first step, it is possible to identify the superiority of DE approaches in comparison with the NSGA-II. Although the NSGA-II be one of the most used algorithms for different multi-objective problems, GDE3 achieved better GDT and RMSD values for all eight proteins. In comparison with NSGA-II, DEMO was better in 6 of 8 proteins regarding RMSD and better in all

proteins regarding GDT. The only two proteins that NSGA-II achieved competitive results are 1ACW and 1CRN, where both structures have β-sheets in their compositions. For α-helices structures, both GDE3 and DEMO achieved superior performance.

The second comparison is between the two DE-based algorithms. Although similar results were obtained when comparing both GDT and RMSD, the GDE3 algorithm found more accurate structures in the majority part of the tested proteins. This behavior can be related to the similarity of the two algorithms, where the significant difference is in the selection of individuals, with DEMO including newly generated individuals in the process, and GDE3 does not.

In order to statistically validate the performance of the GDE3 algorithm, the *Wilcoxon Signed Rank Test* is applied considering the GDT value since it better describes the similarity between the predicted and already determined structures. The first column of Tab. III displays the PDB identification of each predicted protein followed by the algorithm and $p-value$ in columns 2 and 3 respectively. With $p - value$ lower than 0.05 indicates that there is a significant difference between GDE3 and the compared method. Otherwise, the methods can be considered as statistically equivalent.

TABLE III: Wilcoxon Rank Test comparing NSGA-II and DEMO with GDE3 algorithm.

PDB ID	Algorithm	$p - value$ (GDT)
1ACW	NSGA-II	0.0001
	DEMO	0.1504
1AIL	NSGA-II	0.0000
	DEMO	0.6288
1CRN	NSGA-II	0.0000
	DEMO	0.3234
1ENH	NSGA-II	0.0000
	DEMO	0.0460
1ROP	NSGA-II	0.0000
	DEMO	0.4556
1ZDD	NSGA-II	0.0000
	DEMO	0.3493
2MR9	NSGA-II	0.0000
	DEMO	0.5769
2P81	NSGA-II	0.0000
	DEMO	0.0256

With the results summarized in Tab. III, we have a clear indication of the better performance achieved by GDE3 in comparison with the NSGA-II algorithm in all eight proteins. When comparing it with the DEMO algorithm, only two cases (1ENH and 2P81) have *p-value* lower than 0.05, thus qualifying them as equivalents in other 6 cases.

B. Visual Comparison

In addition to the analysis made considering RMSD and GDT, the visual comparison among the predicted structures found by each algorithm, and the structural one can be found in Fig. 3 and Fig 4. In the figures there are four structures aligned, the best structures in terms of GDT (red for NSGA-II, blue for GDE3, and orange for DEMO) found by each algorithm and the experimental one (green). It is also possible to see that, in some cases, the NSGA-II did not aligned the α-helixes (i.e. 1ROP, 2MR9, 2P81) as DE did, justifying the results demonstrated in Tab. II. Another interesting fact is that all approaches had problems identifying the β-sheet (depicted in arrows in Fig.3) structures for 1ACW and 1CRN. This is related to the difficulty of aligning the atoms properly in order to form the required hydrogen bonding that composes the β-sheet structure. This issue is observed in different works in the literature [6] [11] [19] [32].

With the results obtained and discussed in this section, it is possible to identify the capacity of the three different algorithms (NSGA-II, GDE3, and DEMO) for the problem regarding GDT, RMSD, and visual comparison. The Wilcoxon rank test confirmed the superiority of GDE3 in comparison with the NSGA-II algorithm with the proposed test scenario, while statistically equivalent with DEMO.

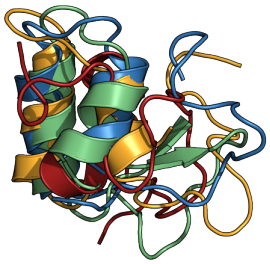
V. Conclusions and Future Works

Proteins are essential molecules for every living being, exerting different biological functions when in its tertiary structure. Due to the importance of the structural understanding of these molecules, and the high cost and complexity associated with the experimental determination, computational approaches became a possible solution for the problem. Although different works made advances in the literature, the prediction of protein structures is still an open problem in structural bioinformatics. Furthermore, some studies exposed the need for using the multi-objective formulation of the problem due to the conflict between the different terms that a force field can have.

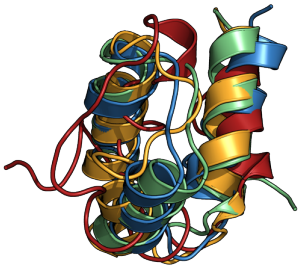
In light of these facts, this work has the objective of comparing three different multi-objective metaheuristics (NSGA-II, GDE3, and DEMO) using the *score3* energy function provided by the Rosetta modeling software. Moreover, we used the Angle Probability List as a source of information for the population initialization procedure, providing a piece of better comparison information for future works. Results obtained by our simulations showed the superiority of GDE3 and DEMO in comparison with the NSGA-II algorithm, something very similar to what occurs in their canonical single-objective approaches to the problem. The GDE3 and DEMO showed to be very competitive between themselves in terms of GDT and RMSD, as confirmed by the Wilcoxon rank test. When comparing them visually with the crystallized structure, it is possible to see that the methods could find very similar structures. However, none of them could correctly identify the β-sheets in 1ACW and 1CRN.

(a) 1ACW - $\beta-sheets$ are shown as arrows

(b) 1AIL

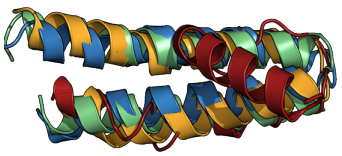


(c) 1CRN - $\beta-sheets$ are shown as arrows

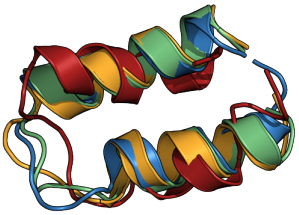


(d) 1ENH

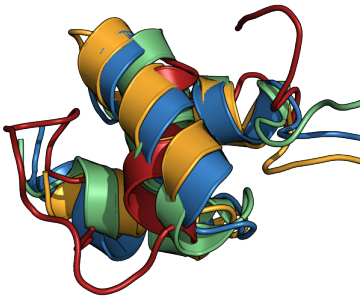
Fig. 3: Cartoon representation (part. 1)



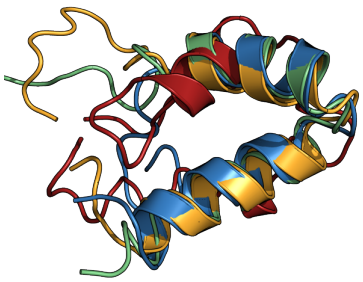
(a) 1ROP



(b) 1ZDD



(c) 2MR9



(d) 2P81

Fig. 4: Cartoon representation (part. 2).

Notwithstanding the relatively limited number of algorithms and test-set, this work offers valuable data about the capacity of multi-objective algorithms enhanced with problem domain knowledge. However, several improvements can be made in order to find better solutions. For further works, it would be interesting to test different energy function compositions, including solvent surface information. Also, as these algorithms being sensitive to parameter control, the development of self-adaptive versions of them could be a way to avoid the parameter generalization. Finally, the investigation of how to better determine β-sheets, something not yet explored, could be interesting for the research area.

ACKNOWLEDGMENT

This work was supported by grants from MCT/CNPq [*311611/2018-4*], FAPERGS [*19/2551000190 6-8*], Alexander von Humboldt-Stiftung (AvH) [*BRA 1190826 HFST* CAPES-P] - Germany, and was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. CAPES PROBRAL [88881.198766/2018- 01] - Brazil.

REFERENCES

[1] G. Walsh, *Proteins: Biochemistry and Biotechnology*. Wiley, 2014.

[2] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, pp. 223–230, 1973.

[3] M. Dorn, M. B. E Silva, L. S. Buriol, and L. C. Lamb, "Three-dimensional protein structure prediction: Methods and computational strategies," *Comput. Biol. Chem.*, vol. 53, pp. 251–276, 2014.

[4] C. Guyeux, N. M.-L. Côté, J. M. Bahi, and W. Bienie, "Is Protein Folding Problem Really a NP-Complete One ? First Investigations," *J. Bioinf. Comput. Biol.*, vol. 12, 2014.

[5] J. Dréo, A. Pétrowski, P. Siarry, and E. Taillard, *Metaheuristics for Hard Optimization*, 1st ed. Berlin/Heidelberg: Springer-Verlag, 2006.

[6] P. H. Narloch and M. Dorn, "A knowledge based differential evolution algorithm for protein structure prediction," in *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*. Springer, 2019, pp. 343–359.

[7] B. Borguesan, M. B. E Silva, B. Grisci, M. Inostroza-Ponta, and M. Dorn, "APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction," *Comput. Biol. Chem.*, vol. 59, pp. 142–157, 2015.

[8] R. S. Silva and R. S. Parpinelli, "A self-adaptive differential evolution with fragment insertion for the protein structure prediction problem," in *International Workshop on Hybrid Metaheuristics*. Springer, 2019, pp. 136–149.

[9] L. d. L. Corrêa, B. Borguesan, M. J. Krause, and M. Dorn, "Three-dimensional protein structure prediction based on memetic algorithms," *Computers and Operations Research*, vol. 91, pp. 160–177, 2018.

[10] M. Dorn, M. Inostroza-Ponta, L. S. Buriol, and H. Verli, "A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides," *IEEE Congress on Evolutionary Computation*, pp. 1233–1240, 2013.

[11] V. Cutello, G. Narzisi, and G. Nicosia, "Computational Studies of Peptide and Protein Structure Prediction Problems via Multiobjective Evolutionary Algorithms," *Multiobjective Problem Solving from Nature*, pp. 93–114, 2008.

[12] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[13] T. Robič and B. Filipič, "Differential evolution for multiobjective optimization," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 520–533.

[14] S. Kukkonen and J. Lampinen, "GDE3: The third Evolution Step of Generalized Differential Evolution," *IEEE Congress on Evolutionary Computation*, vol. 1, pp. 443–450, 2005.

[15] S. Chaudhury, S. Lyskov, and J. J. Gray, "PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta," *Bioinformatics*, vol. 26, pp. 689–691, 2010.

[16] P. H. Narloch and M. Dorn, "A knowledge based self-adaptive differential evolution algorithm for protein structure prediction," in *International Conference on Computational Science*. Springer, 2019, pp. 87–100.

[17] P. H. Narloch and R. S. Parpinelli, *Diversification strategies in differential evolution algorithm to solve the protein structure prediction problem*, 2017, vol. 557.

[18] ——, "The protein structure prediction problem approached by a cascade differential evolution algorithm using rosetta," *Brazilian Conference on Intelligent Systems*, 2017.

[19] S. M. Venske, R. A. Gonçalves, E. M. Benelli, and M. R. Delgado, "ADEMO/D: An adaptive differential evolution for protein structure prediction problem," *Expert Systems with Applications*, vol. 56, pp. 209–226, 2016.

[20] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein Structure Prediction Using Rosetta," 2004, pp. 66–93.

[21] R. Ligabue-Braun, B. Borguesan, H. Verli, M. J. Krause, and M. Dorn, "Everyone Is a Protagonist: Residue Conformational Preferences in High-Resolution Protein Structures," *J. Comput. Biol.*, 2017.

[22] M. Oliveira, B. Borguesan, and M. Dorn, "SADE-SPL: A Self-Adapting Differential Evolution algorithm with a loop Structure Pattern Library for the PSP problem," in *IEEE Congress on Evolutionary Computation*, 2017, pp. 1095–1102.

[23] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res*, vol. 28, pp. 235–242, 2000.

[24] B. Borguesan, M. Inostroza-Ponta, and M. Dorn, "NIAS-Server: Neighbors Influence of Amino acids and Secondary Structures in Proteins," *J. Comput. Biol.*, vol. 24, pp. 255–265, 2017.

[25] J. Handl, D. B. Kell, and J. Knowles, "Multiobjective optimization in bioinformatics and computational biology," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 2, pp. 279–291, 2007.

[26] M. T. Emmerich and A. H. Deutz, "A tutorial on multiobjective optimization: fundamentals and evolutionary methods," *Natural Computing*, vol. 17, no. 3, pp. 585–609, 2018.

[27] K. Deb, "Multi-objective optimization," in *Search methodologies*. Springer, 2014, pp. 403–449.

[28] V. Cutello, G. Narzisi, and G. Nicosia, "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*. Springer, 2005, pp. 54–63.

[29] J. C. Calvo, J. Ortega, and M. Anguita, "Comparison of parallel multi-objective approaches to protein structure prediction," *Journal of Supercomputing*, vol. 58, no. 2, pp. 253–260, 2011.

[30] ——, "PITAGORAS-PSP: Including domain knowledge in a multi-objective approach for protein structure prediction," *Neurocomputing*, vol. 74, no. 16, pp. 2675–2682, 2011.

[31] S. Gao, S. Song, J. Cheng, Y. Todo, and M. C. Zhou, "Incorporation of Solvent Effect into Multi-Objective Evolutionary Algorithm for Improved Protein Structure Prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1365–1378, 2018.

[32] P. H. Narloch and M. Dorn, "Differential Evolution Multi-Objective for Tertiary Protein Structure Prediction," in *International Conference on the Applications of Evolutionary Computation (Part of EvoStar), in press*. Springer, 2020.

[33] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577–2637, 1983.

[34] A. Benitez-Hidalgo, A. J. Nebro, J. Garcia-Nieto, I. Oregi, and J. Del Ser, "jMetalPy: a Python Framework for Multi-Objective Optimization with Metaheuristics," *Swarm and Evolutionary Computation*, vol. 51, mar 2019.

[35] M. Vasile and L. Ricciardi, "Generalized Differential Evolution for Numerical and Evolutionary Optimization," vol. 663, no. 2508, pp. 223–252, 2015.

[36] R. Storn and K. Price, "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," *Journal of Global Optimization*, vol. 11, pp. 341–359, 1997.