



A Knowledge Based Differential Evolution Algorithm for Protein Structure Prediction

Pedro H. Narloch  and Márcio Dorn ^(✉) 

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

Abstract. Three-dimensional protein structure prediction is an open-challenging problem in Structural Bioinformatics and classified as an NP-complete problem in computational complexity theory. As exact algorithms cannot solve this type of problem, metaheuristics became useful strategies to find solutions in viable computational time. In this way, we analyze four standard mutation mechanisms present in Differential Evolution algorithms using the Angle Probability List as a source of information to predict tertiary protein structures, something not explored yet with Differential Evolution. As the balance between diversification and intensification is an essential fact during the optimization process, we also analyzed how the Angle Probability List might influence the algorithm behavior, something not investigated in other algorithms. Our tests reinforce that the use of structural data is a crucial factor to reach better results. Furthermore, combining experimental data in the optimization process can help the algorithm to avoid premature convergence, maintaining population diversity during the whole process and, consequently, reaching better conformational results.

Keywords: Protein structure prediction · Differential Evolution · Structural Bioinformatics

1 Introduction

Proteins are essential molecules for every living organism due to biological functions they provide [1]. Their biological functions are directly related with a stable three-dimensional (3D) structure, called as protein's tertiary structure. As their structures being so important, if a protein folds unexpectedly, it can be harmful to the biological system. In light of these facts, the determination of the three-dimensional protein structure is vital for the understanding of how life goes on. However, the experimental determination of these structures made by Nuclear Magnetic Resonance (NMR) and X-ray crystallography are not cheap, neither straightforward. Hence, computational methods could be interesting to reduce these costs and shorten the difference between known sequences and

already determined structures. In this way, the Protein Structure Prediction (PSP) problem became a critic and challenging problems in Structural Bioinformatics [2].

There are different manners to computationally approach the PSP problem, each one varying the degree of freedom. Due to these large number of possible conformations a protein can fold, the PSP is considered, according the computational complexity theory, an NP-complete problem [3]. Since exact algorithms are inefficient for these class of problems, bio-inspired algorithm became interesting, although they do not guarantee an optimum solution. In this way, some works have aggregated domain-based knowledge to boost metaheuristics and get better and native-like protein structures. Besides these improvements and the significant number of researches done [4], the PSP still an open, challenging problem.

Among different evolutionary algorithms, the Differential Evolution algorithm (DE) has been showing good results not only in important competitions [5], but also in PSP applications [6, 7]. Due to these facts, in this paper, our objective is to analyze four classical DE mutation mechanisms with domain-based knowledge provided by an *Angle Probability List* (APL) [8], using the preference of amino acids in the population initialization procedure. Besides this source of knowledge has shown promising results in other evolutionary algorithms, it has not be used with the DE algorithm yet. Moreover, a populational diversity metric [9] is employed to verify if the APL might influence the balance between DE intensification and diversification capacities, something not observed yet (besides its importance). The next sections in this paper are organized as follows. Section 2 presents a literature review of the problem, the classical DE algorithm, and related works. The method, metrics, and tools are described in Sect. 3. In Sect. 4 the results obtained by the different approaches are discussed. Finally, the conclusions and future works are given in Sect. 5.

2 Preliminaries

2.1 Proteins, Structure and Representation

From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acids. The thermodynamic hypothesis created by Anfinsen [10] states that the protein's conformation is given by the lowest free energy of the entire system. In this way, it is possible to assume that a protein folds into its tertiary structure purely by its amino acid sequence (primary structure) and environment conditions. Over the last years, different computational strategies were applied to the PSP problem in order to achieve the minimum of its free energy. As proteins are complex molecules, the computational representation of them is not a trivial task. Thus, there were proposed different manners to describe a protein and to simulate factors which contribute to the folding process. The most real computational representation includes all atoms in the system, where each atom has its atomic coordinates in a three-dimensional space. However, as this type of representation is very similar to real proteins,

it becomes computationally expensive due to the significant number of atoms it can assume. A less expensive, but equally important representation, is the dihedral angles representation where each amino acid has a determined number of angles to be set. This representation overcomes the all-atom problem, and it maintains the protein characteristics.

A polypeptide (or protein) is composed by a set of amino acids chained by a chemical bond called peptide bond. All amino acids found in proteins have the same backbone structure, composed by an amino-group (N), a central carbon atom called by alpha-carbon (C_α), four hydrogens (H) and a carboxyl-group (C). What differs each amino acid is their side-chain composition which can vary in 20 different types. The peptide bond, responsible for bonding two amino acids, is formed by the C-N interaction, forming the ω angle which tends to be planar. There are two other angles which are free to rotate in the space: the ϕ angle which rotates around N- C_α and the ψ angle which rotates around the C_α -C bond, varying from -180° to $+180^\circ$. The number of side-chain angles (χ) varies according to each amino acid ranging from 0 to 4 angles. The set of consecutive torsion angles represent the internal rotations of a polypeptide main chain. A single polypeptide may contain multiple secondary structures. α -helix and β -sheet are the most stable secondary structures and they can be considered as the principal elements present in 3D structures of proteins. There is another type of regular secondary structure, known as β -turn, that does not occur so frequently as α -helices and β -strands. The β -turn structure is a set of short segments and are often connect two β -strands.

There are different functions which calculate the protein free energy according to its computational representation. The *Rosetta energy function* [11] is one well-known all-atom high-resolution strategy used in different high-performance predictors [12]. Nowadays, there are more than 18 energy terms that compose the *Rosetta energy function*, and most of them are composed by knowledge-based potentials. The final energy is the sum of all these terms organized in five classes as shown in Eq. 1.

$$E_{Rosetta} = \begin{cases} E_{physics-based} + E_{inter-electrostatic} \\ + E_{H-bonds} + E_{knowledge-based} + E_{AA} \end{cases} \quad (1)$$

where $E_{physics-based}$ calculates the 6–12 Lennard-Jones interactions and Solvation potential approximation, $E_{inter-electrostatic}$ stands for inter-atomic electrostatic interactions and $E_{H-bonds}$ hydrogen-bond potentials. In $E_{knowledge-based}$ the terms are combined with knowledge-based potentials while the free energy of amino acids in the unfolded state is in E_{AA} term.

2.2 Angle Probability List - APL

Given the complexity related to the prediction problem, it is reasonable to enhance the search strategy with already known structural data from a determined database. As discussed in [13], residues can assume different torsion angles values accordingly to the secondary structure they might assume, thus being

valuable information which should be used in order to reduce the search space and improve the search capabilities. With this in mind, an *Angle Probability List* (APL) is proposed by [8] based on the conformational preferences of amino acids according to its secondary structure using high-quality information from the Protein Data Bank (PDB) [14]. To compose the knowledge-database, a set of 11,130 structures with resolution $\leq 2.5\text{\AA}$ stored in PDB until December 2014 were selected. An APL is built from a histogram matrix ($H_{aa,ss}$) of $[-180, 180] \times [-180, 180]$ for each amino acid residue (aa) and secondary structure (ss). It is possible to generate different combinations of amino acids considering degrees of the neighborhood from the reference one. As promising results were obtained using the APL, a web tool called NIAS¹ (*Neighbors Influence of Amino acids and Secondary structures*) is available to compute APLs [15]. Figure 1 illustrates the conformational preference of three amino acid residues in coil secondary structure: Glycine (GLY), Asparagine (ASP) and Proline (PRO).

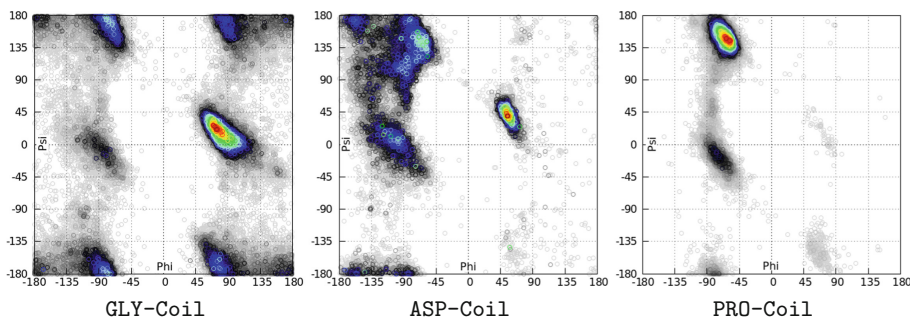


Fig. 1. Example of APL's for an amino acid sequence “GNP” with secondary structure “CCC”. The dark red color marks the densest regions of the Ramachandran plot. The boldface letters represent the reference amino acids and their SS (Color figure online).

2.3 Differential Evolution

With the intention to handle nonlinear and multimodal cost functions, Storn and Price developed in 1997 one of the best optimization algorithms since then: the *Differential Evolution* (DE) [16]. The DE is a population-based evolutionary algorithm composed of four steps: initialization, mutation, crossover, and selection. Initially, a population of NP solution vectors with D dimensions is randomly generated. During the optimization, the algorithm iterates a defined number of generations over the mutation, crossover, and selection process. In the mutation step, for each target vector \mathbf{x}_i^g , a mutant vector \mathbf{v}_i^{g+1} is generated according to a

¹ <http://sbcinf.ufgrs.br/nias>.

mutation strategy. One of the initial formulations proposed by Storn and Price is called $DE/rand/1$ (Eq. 2).

$$DE/rand/1 : \mathbf{v}_i^{g+1} = \mathbf{x}_{r_1}^g + F \cdot (\mathbf{x}_{r_2}^g - \mathbf{x}_{r_3}^g) \quad (2)$$

where g represents the generation, \mathbf{x}_{r_n} a random solution from the current population, and $F > 0$ a parameter for scaling the difference between vectors. All selected vectors are mutually exclusive and different from the target vector.

During the mutation process, a crossover mechanism selects dimensions from the mutant vector which are mixed by the employed mutation strategy, creating a trial vector \mathbf{u}_i^g . Generally, DE applications use a binomial crossover scheme. In this case, the dimension is mutated whenever a randomly generated number is less than the crossover rate (CR) parameter. The binomial crossover scheme is expressed by Eq. 3.

$$\mathbf{u}_{i,d}^g = \begin{cases} \mathbf{v}_{i,d}^g & \text{if } d = d_{rand} \text{ or } \text{rand}[0,1] \leq CR, \\ \mathbf{x}_{i,d}^g & \text{otherwise} \end{cases} \quad (3)$$

where d_{rand} is any random dimension to guarantee at least one modification and rand an uniform random number between 0 and 1.

After mutation and crossover stages, the trial vector is passed by a score function to evaluate the new solution. In this stage, the selection mechanism act to determines whether the target or the trial vector will compose the population in the next generation (iteration). It is possible to describe the selection operator as shown in Eq. 4.

$$\mathbf{x}_i^{g+1} = \begin{cases} \mathbf{u}_i^g & \text{if } f(\mathbf{u}_i^g) \leq f(\mathbf{x}_i^g), \\ \mathbf{x}_i^g & \text{otherwise} \end{cases} \quad (4)$$

where $f(x)$ is the score function to be minimized. In this way, the solution with the best fitness value will be part of the offspring, and consequently, part of the new population in the next generation. Algorithm 1 presents the classical Differential Evolution scheme for a minimization problem.

Besides the straightforward implementation of classical DE, as shown in Algorithm 1, with few parameters to be selected, it has been getting high ranks in different optimization competitions in a wide variety of objective functions when compared with other evolutionary computation techniques [5].

2.4 Related Works

Metaheuristics are broadly used in hard optimization problems due to their capacity to reach feasible results since exact algorithms cannot handle NP-Complete problems [17]. Some of the well-known methods are Genetic Algorithms (GA), Simulated Annealing (SA), Differential Evolution (DE), Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and others [17]. As the protein structure prediction be considered an NP-Complete problem [3], different methods were applied to predict protein tertiary structure.

Algorithm 1. Classical Differential Evolution

Data: NP, F and CR
Result: The best individual in population
 Generate initial population with NP individuals
while $g \leq \text{number of generations}$ **do**
 for each i **individual in population** **do**
 select three random individuals (x_{r1}, x_{r2}, x_{r3})
 $d_{rand} \leftarrow$ select a random dimension to mutate
 for each d **dimension** **do**
 if $d = d_{rand}$ **or** $\text{random} \leq CR$ **then**
 $u_{i,d} \leftarrow x_{r1,d}^g + F \cdot (x_{r2,d}^g - x_{r3,d}^g)$
 else
 $u_{i,d} \leftarrow x_{i,d}$
 end
 end
 if $u_{i,fitness} \leq x_{i,fitness}$ **then**
 add u_i in the offspring
 else
 add x_i in the offspring
 end
 end
 population \leftarrow offspring
 $g \leftarrow g + 1$
end

As bio-inspired methods could be easily modified, many variations of the same base-algorithm compose different approaches to the problem. In [6] the DE algorithm was tested using two mechanisms, known as generation gap and Gaussian mutation, in order to maintain the populational diversity during the optimization process. Another version of DE is proposed in [18] where the optimization process is divided in four slices and each part a different mutation mechanism is used. A multi-objective formulation is combined with the self-adaptive DE in [19] where the energy function is divided by bonded and non-bonded terms. It is important to notice that none of these works have used APL information to enhance their search algorithms and all of them have competitive results despite the number of objectives. In [7] a self-adaptive differential evolution (SADE) is combined with a library of amino acid segments based on different motifs present in proteins. Reported results have shown that DE is a good meta-heuristic to find plausible proteins conformation and it can be enhanced with domain-based knowledge.

Another well known evolutionary algorithm used to predict tertiary structures is the GA and its different versions. In [20] a knowledge-based Genetic Algorithm was proposed with the intention to reduce the search space. The used information uses torsion angles intervals for amino acids based on previous occurrences in experimentally determined proteins, a very similar approach provided by APL. Another knowledge-based GA was proposed in the APL paper [8] and compared with a PSO algorithm using the same source of information. The results obtained by this work showed that algorithms enhanced with APL get better energy values and conformational similarity when compared with the experimentally known structure. Among different algorithms found in literature, only two of them [6, 18] in some way monitored the populational diversity gen-

erated by different mechanisms. Also, is noteworthy that when methods use some problem-domain information, the search algorithm reaches better values of energy and more similar structural conformation. In this way, this paper test four different DE mutation mechanisms with APL and analyze if there is an impact on the algorithm behavior.

3 Materials and Methods

Tertiary protein structure prediction is not an easy task to do, and there are different ways to approach the problem. In light of the conclusions presented by the Anfinsen's thermodynamic hypothesis [10], it is possible to declare that proteins reach stability with the minimum free energy. Considering the challenges presented by the problem, there are three essential components to create a PSP solver: **(a)** a way to computationally represents the protein structure; **(b)** an energy function to evaluate the protein, and **(c)** an algorithm to explore the search space in order to find a solution with the minimum possible energy [4].

Scoring Function:

In order to create a solver for the PSP problem, we used a *Python-based* interface [21] to interact with a state-of-art molecular modeling suite known as *Rosetta* [11]. With this interface, it is possible to calculate the free energy of each possible protein using the *score3 energy function*². The *score3 energy function* uses a centroid-based representation for the side chains of each amino acid, reducing not only the computational cost for energy calculation but also the representation vector used in the search algorithm. Hence, the problem dimensionality is $2N$, where N is the length of the primary structure. Moreover, a new term is added to the fitness function in order to benefit well-formed secondary structures. To identify the formed secondary structures, an implementation of DSSP [22] by *PyRosetta* is used during the optimization process. In this way, every time that search algorithm finds a solution which the secondary structure matches with the one given as an input, a reinforcement score is assigned to the *score3* value. On the other hand, if the perturbation made by the algorithm does not correctly find the secondary structure, a punishment is ascribed to the solution. Equation 5 formulates the fitness function used in this work.

$$E_{total} = E_{score3} + E_{SecondaryStructure} \quad (5)$$

It is important to mention that choosing an energy function to guide a search algorithm in PSP task is not a trivial effort. The rules that govern the biochemical processes and relations are only partially known, which makes it harder to design efficient computational strategies for these situations. There is not a function which correctly describes the potential energy of a real system, which implies that different energy functions could lead to different final structural results. Although the energy function is the fitness function which guides our search

² <https://www.rosettacommons.org>.

algorithm, the *Root Mean Square Deviation* (RMSD) might be considered as well. The RMSD is a measurement which compares the distance (in angstroms) among atoms in two structures. In our case, we use it to compare a final solution with the already known structure. Equation 6 presents the RMSD_α , where only C_α atoms are compared.

$$\text{RMSD}(a, b) = \sqrt{\frac{\sum_{i=1}^n |r_{ai} - r_{bi}|^2}{n}} \quad (6)$$

where r_{ai} and r_{bi} are the i th atoms in a group of n atoms from structures a and b . The closer RMSD is from 0 Å more similar are the structures.

Search Strategy: Over the years, different approaches were proposed to predict tertiary protein structures. However, a different approach was proposed in [8] where conformational preferences of amino acid residues can be used to improve the search algorithm, leading to better values of energy and structural quality. The *Angle Probability List* (Sect. 2.2) has been tested with a *Biased Random-Key Genetic Algorithm*, *Particle Swarm Optimization* [8], and a memetic algorithm [2]. With this in mind, our approach combines the data found with APL and four different mutation strategies in the *Differential Evolution* algorithm using the *PyRosetta score3* energy function. The mutation mechanisms are listed in Table 1. As used in other works, the information will be used only in the initialization of the population, where random individuals are generated based on the APL.

Table 1. Classical mutation strategies in DE.

Approach	Equation
$\text{DE}_{\text{best}/1/\text{bin}}$	$\mathbf{v}_i^{g+1} = \mathbf{x}_{\text{best}}^g + F \cdot (\mathbf{x}_{r2}^g - \mathbf{x}_{r3}^g)$
$\text{DE}_{\text{rand}/1/\text{bin}}$	$\mathbf{v}_i^{g+1} = \mathbf{x}_{r1}^g + F \cdot (\mathbf{x}_{r2}^g - \mathbf{x}_{r3}^g)$
$\text{DE}_{\text{curr-to-rand}}$	$\mathbf{v}_i^{g+1} = \mathbf{x}_i^g + F1 \cdot (\mathbf{x}_{r1}^g - \mathbf{x}_i^g) + F2 \cdot (\mathbf{x}_{r2}^g - \mathbf{x}_{r3}^g)$
$\text{DE}_{\text{curr-to-best}}$	$\mathbf{v}_i^{g+1} = \mathbf{x}_i^g + F1 \cdot (\mathbf{x}_{\text{best}}^g - \mathbf{x}_i^g) + F2 \cdot (\mathbf{x}_{r2}^g - \mathbf{x}_{r3}^g)$

Furthermore, a populational diversity measure (Eq. 7) is applied to verify if the usage of APL information has some impact in the diversity during the optimization process, something not analyzed in any work. This metric was proposed in [9], and it can be used in continuous-domain problems.

$$\text{GDM} = \frac{\sum_{i=1}^{N-1} \ln \left(1 + \min_{j[i+1, N]} \frac{1}{D} \sqrt{\sum_{k=1}^D (x_{i,k} - x_{j,k})^2} \right)}{\text{NMDF}} \quad (7)$$

where D represents the dimensionality of the solution vector, N is related to the population size and x the individual (or solution vector). The NMDF is a normalization factor which corresponds to the maximum diversity value so far.

The population diversity metric starts with the value 1, meaning maximum diversity and as this index tends to 0 means that individuals are getting closer (without considering the fitness function). This type of measurement is helpful in the understanding the algorithm convergence behavior. It also helps to identify if the algorithm is getting trapped in local optima, which leads to premature convergence.

4 Experiments and Analysis

In order to compare the four different mutation strategies and APL usage, we have chosen ten proteins to compose our experiment. In Table 2 all proteins are organized by alphabetical order with their size (amino acid quantity) and secondary structure types. These proteins were selected based on literature works [7,8]. For our simulation we used the same DE parameters (listed in Table 3) based on literature works [6,18,19] with 1 million of fitness evaluations which corresponds to 10 thousand generations.

Table 2. Target protein sequences.

PDB ID	Size	Secondary Structure Content
1AB1	46	One sheet/Two helices
1ACW	29	One sheet/Two helices
1CRN	46	One sheet/Two helices
1ENH	54	Three helices
1ROP	63	Two helices
1UTG	70	Five helices
1ZDD	35	Two helices
2MR9	44	Three helices
2MTW	20	One helix
2P81	44	Two helices

Table 3. DE parameters.

Parameter	Value	Description
NP	100	Population size
CR	1	Crossover factor
F	0.5	Mutation factor

With the intention to keep a fair comparison among different versions, each mutation mechanism started with the same initial population. In this way, it is possible to ensure that none of the mechanism was benefited by randomness when created the initial population. For each protein, in each mutation mechanism, thirty experiments were done in the same environment. The results obtained are listed in Table 4, organized by protein and mutation mechanism.

Adopting the Angle Probability List: Analyzing the general energy results listed in Table 4, one can observe that in all predicted proteins, the lowest energy found is always obtained by a mechanism that used the APL as a source of information. Nonetheless, some solutions found by approaches that not used the APL

Table 4. Results obtained for target proteins using different mutation mechanisms.

PDB	Strategy	Energy	
		With APL	Without APL
1AB1	DE _{rand/1/bin}	-98.00(-75.48 ± 9.54)	-129.64(53.73 ± 118.64)
	DE _{best/1/bin}	-152.24(-95.32 ± 18.48)	-61.21(11.44 ± 40.25)
	DE _{curr-to-rand}	-169.14(-109.07 ± 17.29)	-161.30(-103.47 ± 15.76)
	DE _{curr-to-best}	-158.14(-122.57 ± 15.64)	-112.02(-54.29 ± 29.50)
1ACW	DE _{rand/1/bin}	-148.22(-25.17 ± 41.97)	-56.26(3.36 ± 29.10)
	DE _{best/1/bin}	-133.85(-88.22 ± 39.33)	-31.23(50.65 ± 38.53)
	DE _{curr-to-rand}	-135.75(-63.13 ± 24.92)	-151.10(-95.40 ± 33.96)
	DE _{curr-to-best}	-160.84(-111.85 ± 26.69)	-116.13(-32.59 ± 50.66)
1CRN	DE _{rand/1/bin}	-95.03(-72.76 ± 6.13)	-114.07(103.40 ± 91.24)
	DE _{best/1/bin}	-136.18(-93.92 ± 16.06)	-97.59(18.49 ± 62.02)
	DE _{curr-to-rand}	-188.41(-113.55 ± 23.59)	-145.69(-103.30 ± 15.22)
	DE _{curr-to-best}	-173.95(-129.20 ± 23.17)	-119.92(-58.94 ± 25.79)
1ENH	DE _{rand/1/bin}	-343.13(-334.83 ± 3.08)	-303.81(110.16 ± 191.09)
	DE _{best/1/bin}	-364.38(-348.84 ± 7.92)	-170.09(-65.82 ± 61.22)
	DE _{curr-to-rand}	-376.11(-363.21 ± 10.90)	-330.66(-275.93 ± 32.43)
	DE _{curr-to-best}	-368.94(-359.37 ± 5.06)	-263.78(-185.41 ± 31.73)
1ROP	DE _{rand/1/bin}	-498.18(-485.32 ± 6.59)	-290.06(122.87 ± 237.18)
	DE _{best/1/bin}	-471.52(-458.66 ± 6.13)	-224.12(-46.93 ± 102.87)
	DE _{curr-to-rand}	-484.88(-475.80 ± 3.14)	-415.73(-331.45 ± 36.06)
	DE _{curr-to-best}	-477.11(-468.65 ± 4.64)	-308.32(-195.35 ± 64.94)
1UTG	DE _{rand/1/bin}	-514.55(-487.69 ± 10.24)	-406.34(276.78 ± 248.75)
	DE _{best/1/bin}	-516.13(-497.01 ± 9.29)	-208.48(4.16 ± 81.38)
	DE _{curr-to-rand}	-545.70(-533.13 ± 8.03)	-381.62(-299.04 ± 53.70)
	DE _{curr-to-best}	-536.09(-515.88 ± 9.49)	-313.49(-183.90 ± 73.30)
1ZDD	DE _{rand/1/bin}	-233.00(-225.00 ± 3.78)	-241.13(-80.29 ± 131.59)
	DE _{best/1/bin}	-232.28(-225.54 ± 3.66)	-164.58(-79.73 ± 51.30)
	DE _{curr-to-rand}	-245.71(-236.38 ± 4.22)	-231.66(-216.58 ± 11.51)
	DE _{curr-to-best}	-240.61(-231.89 ± 4.05)	-226.97(-185.55 ± 29.04)
2MR9	DE _{rand/1/bin}	-287.20(-264.20 ± 11.33)	-241.53(-22.16 ± 166.52)
	DE _{best/1/bin}	-282.84(-270.72 ± 6.96)	-153.54(-70.01 ± 36.29)
	DE _{curr-to-rand}	-296.22(-289.38 ± 3.28)	-269.38(-230.85 ± 16.97)
	DE _{curr-to-best}	-290.33(-283.44 ± 4.76)	-218.43(-157.42 ± 25.07)
2MTW	DE _{rand/1/bin}	-109.56(-102.87 ± 3.45)	-107.41(-100.68 ± 5.22)
	DE _{best/1/bin}	-95.02(-90.62 ± 2.12)	-97.16(-65.23 ± 20.75)
	DE _{curr-to-rand}	-104.58(-98.74 ± 2.88)	-103.79(-100.00 ± 1.99)
	DE _{curr-to-best}	-101.91(-94.70 ± 2.53)	-100.69(-92.48 ± 8.22)
2P81	DE _{rand/1/bin}	-249.80(-236.02 ± 5.37)	-260.32(22.46 ± 129.73)
	DE _{best/1/bin}	-252.24(-242.28 ± 4.69)	-164.87(-92.15 ± 45.06)
	DE _{curr-to-rand}	-266.89(-252.41 ± 10.84)	-251.31(-227.03 ± 14.42)
	DE _{curr-to-best}	-257.72(-251.24 ± 3.70)	-237.72(-184.42 ± 31.72)

could be considered outliers, e.g., the energy obtained by $DE_{rand/1/bin}$ for protein 1AB1. This emphasizes that APL information is an important factor that might be present in PSP solvers since the use of this database enhance the algorithm and help the search mechanism to find solutions with lower energy values. As shown in Fig. 2, it is possible to notice that approaches which used a more significant amount of individuals to compose the new solution have maintained higher levels of diversity during the optimization process and, consequently, better energy values ($DE_{curr-to-rand}$ and $DE_{curr-to-best}$). This behavior can be related to the better populational quality generated by APL, leading to a better combination among different individuals thus, avoiding the premature convergence. In $DE_{rand/1/bin}$ and $DE_{best/1/bin}$ this behavior was not observed, showing that using three individuals to compose a new one might be influenced by selective pressure, where local optima influenced the algorithm's evolution contributing to the premature convergence. This pattern was observed in all cases. Therefore, only one convergence comparison is plotted as an example.


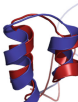

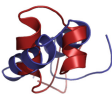
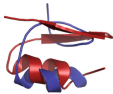
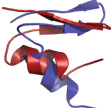
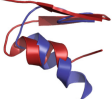
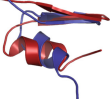
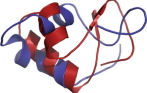
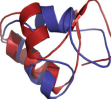
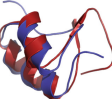
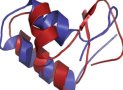
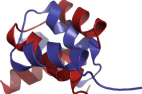
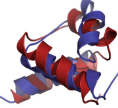
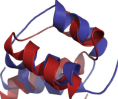
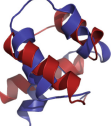

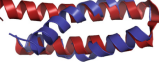

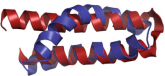
Among the for methods that used APL as a source of information, the methods $DE_{curr-to-rand}$ and $DE_{curr-to-best}$ achieved better energy results. A post-hoc non parametric test (Dunn's Test) was used to verifies the null hypothesis (Table 5), comparing all four methods that used APL. The better energy values obtained can be related to the diversity maintenance during the whole optimization process as shown in Fig. 2. Moreover, presented results keep showing a decreasing tendency, even in the last generation, showing that it is possible to reach better energy values if the optimization had continued. In contrast, the $DE_{best/1/bin}$ had a premature convergence, which made the evolution impossible to happen. Finally, with the results obtained and discussed in this section, it is possible to ensure that APL information does guide the search algorithm to better results in three factors: energy, RMSD and diversity indexes (depending on the employed strategy). In general, the $DE_{curr-to-rand}$ showed to be better than other three mutation strategies, or at least equivalent, not only in energy values but also in diversity maintenance. In Table 6 the minimum energy solutions (blue) are compared with the experimental structures (red) found in PDB. It is possible to notice that similar secondary structures are found in all cases besides the alignment in some cases. Achieved results are comparable in terms of folding organization with state-of-the-art prediction methods, corroborating the effectiveness of our proposal.

Besides the $DE_{curr-to-rand}$ showed to be better, or at least equivalent, than other methods in energy terms, it is not the case when the RMSD value of the minimum energy found is compared with the e.g. $DE_{best/1/bin}$. As energy functions are approximations methods to computationally evaluate the potential energy of a protein, and the search space has multimodal characteristics, it is expected that proteins with lower energy values can have bigger RMSD results, meaning that the global optimum was not found, since the minimum potential energy might describes the native conformation of a protein (0 Angstrom in comparison with the experimental data).

Table 5. Dunn’s Multiple Comparison Test for all versions which used APL as source of information. Values with values down to 0.05 means that there is statistical significance between a pair of methods.

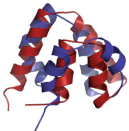
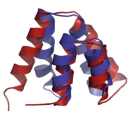
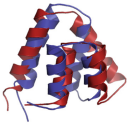
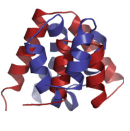

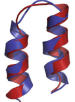
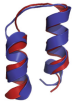
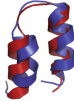
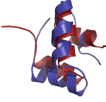
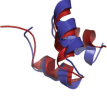
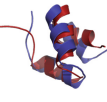
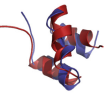




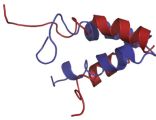
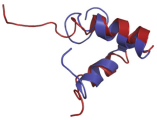
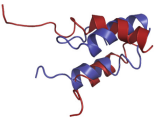
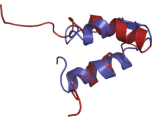
Protein		$DE_{rand/1/bin}$	$DE_{best/1/bin}$	$DE_{curr-to-rand}$
1AB1	$DE_{best/1/bin}$	0.00	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	1.00	0.00
1ACW	$DE_{best/1/bin}$	0.00	—	—
	$DE_{curr-to-rand}$	0.04	0.03	—
	$DE_{curr-to-best}$	0.00	0.00	0.00
1CRN	$DE_{best/1/bin}$	0.00	—	—
	$DE_{curr-to-rand}$	0.00	0.01	—
	$DE_{curr-to-best}$	0.00	0.00	0.14
1ENH	$DE_{best/1/bin}$	0.00	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	0.71
1ROP	$DE_{best/1/bin}$	0.00	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	0.00
1UTG	$DE_{best/1/bin}$	0.12	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	0.00
1ZDD	$DE_{best/1/bin}$	1.00	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	0.02
2MR9	$DE_{best/1/bin}$	0.78	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	0.03
2MTW	$DE_{best/1/bin}$	0.00	—	—
	$DE_{curr-to-rand}$	0.40	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	0.00
2P81	$DE_{best/1/bin}$	0.01	—	—
	$DE_{curr-to-rand}$	0.00	0.00	—
	$DE_{curr-to-best}$	0.00	0.00	1.00

Table 6. Cartoon representation of experimental structures (red) compared with lowest energy solutions (blue) found by each mutation mechanism.

PDB	$DE_{rand/1/bin}$	$DE_{best/1/bin}$	$DE_{curr-to-rand}$	$DE_{curr-to-best}$
1AB1	 7.42Å	 3.59Å	 4.58Å	 9.53Å
1ACW	 4.87Å	 5.53Å	 3.72Å	 3.57Å
1CRN	 7.40Å	 6.15Å	 2.40Å	 6.00Å
1ENH	 6.24Å	 2.94Å	 4.99Å	 5.86Å
1ROP	 8.54Å	 1.90Å	 3.98Å	 3.43Å

(continued)

Table 6. (continued)

PDB	$DE_{rand/1/bin}$	$DE_{best/1/bin}$	$DE_{curr-to-rand}$	$DE_{curr-to-best}$
1UTG	 5.03Å	 3.90Å	 3.94Å	 11.69Å
1ZDD	 2.78Å	 1.27Å	 2.48Å	 2.65Å
2MR9	 4.11Å	 3.14Å	 3.27Å	 3.26Å
2MTW	 5.90Å	 5.43Å	 6.84Å	 5.08Å
2P81	 8.11Å	 1.56Å	 5.58Å	 2.37Å

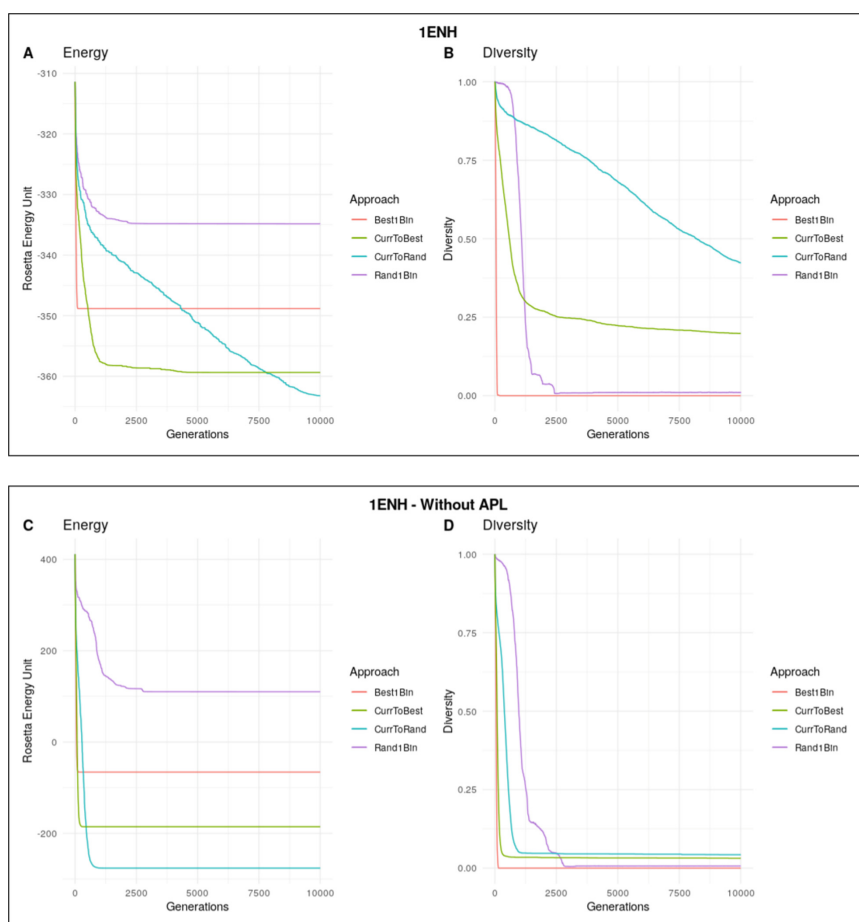


Fig. 2. PDB ID 1ENH convergence of energy and diversity for all mutation mechanisms with and without APL. **A** and **B** plots are the mean results obtained using APL whereas **C** and **D** were obtained without APL information.

5 Conclusion

Meaningful progress in protein structure prediction area has happened in the last decade. However, it is still necessary the development of methods which enhance search algorithms with well structured biological data. In this paper we have tested one valuable source of information called APL with the DE algorithm in four different mutation mechanisms, comparing the DE behavior with and without APL. In order to compare these different versions, ten proteins with different secondary structures were predicted using *score3* energy function provided by *PyRosetta* package. Results showed that the combination of well-structured data with the differential evolution algorithm achieve better conformational results in comparison with the same algorithm without experimental data, even if the

data is used only to create the initial population. The overall contributions of our work are the following: (a) the use of computational techniques and concepts to develop a new algorithm for a relevant biological problem; (b) the analysis of conformational preferences of amino acid residues in proteins and its use to 3D protein structure prediction methods. We observed that when we associate the type of an amino acid residue and secondary structure, it is possible to obtain valuable information about the preferences of this amino acid residues; and finally (c) the development and evaluation of different DE versions to search the three-dimensional protein conformational space using APL.

Aside the better results obtained by all approaches that used APL, the knowledge database used in this work improved the DE exploration capacity in two different versions: $DE_{curr-to-rand}$ and $DE_{curr-to-best}$, helping them to avoid premature convergence. In 8 of 10 cases, the $DE_{curr-to-rand}$ got better energy results than other three versions. This version also has shown higher diversity indexes during the whole optimization process. Furthermore, as this approach demonstrates high diversity index, even in the end of the optimization process, it would be interesting to enhance the mechanism with exploitation capabilities in order to explore this diversity and get even better results. Another interesting application would be the combination of niching methods to find multiple local minima since PSP is considered a multimodal problem.

As in other works that used APL as a source of information, we used the data only to create the initial population. Thus, new methodologies on how to use the information would be a significant advance in the PSP area, since structural data could be obtained in different ways such as in angle probability lists or contact maps. In future works it would be interesting to join a multi modal approach with a multi objective DE algorithm using the knowledge provided by APL.

Acknowledgements. This work was supported by grants from FAPERGS [16/2551-0000520-6], MCT/CNPq [311022/2015-4; 311611/2018-4], CAPES-STIC AMSUD [88887.135130/2017-01] - Brazil, Alexander von Humboldt-Stiftung (AvH) [BRA 1190826 HFST CAPES-P] - Germany. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

References

1. Walsh, G.: Proteins: Biochemistry and Biotechnology. Wiley, Hoboken (2014)
2. Corrêa, L.d.L., Borguesan, B., Krause, M.J., Dorn, M.: Three-dimensional protein structure prediction based on memetic algorithms. *Comput. Oper. Res.* **91**, 160–177 (2018)
3. Guyeux, C., Côté, N.M.L., Bahi, J.M., Bienie, W.: Is protein folding problem really a NP-complete one? First investigations. *J. Bioinf. Comput. Biol.* **12**, 1350017–1–1350017–24 (2014)
4. Dorn, M., E Silva, M.B., Buriol, L.S., Lamb, L.C.: Three-dimensional protein structure prediction: methods and computational strategies. *Comput. Biol. Chem.* **53**, 251–276 (2014)
5. Das, S., Mullick, S.S., Suganthan, P.N.: Recent advances in differential evolution-an updated survey. *Swarm Evol. Comput.* **27**, 1–30 (2016)

6. Narloch, P.H., Parpinelli, R.S.: Diversification strategies in differential evolution algorithm to solve the protein structure prediction problem. In: Madureira, A.M., Abraham, A., Gamboa, D., Novais, P. (eds.) ISDA 2016. AISC, vol. 557, pp. 125–134. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53480-0_13
7. Oliveira, M., Borguesan, B., Dorn, M.: SADE-SPL: a self-adapting differential evolution algorithm with a loop structure pattern library for the PSP problem. In: IEEE Congress on Evolutionary Computation, pp. 1095–1102 (2017)
8. Borguesan, B., E Silva, M.B., Grisci, B., Inostroza-Ponta, M., Dorn, M.: APL: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. *Comput. Biol. Chem.* **59**, 142–157 (2015)
9. Corriveau, G., Guilbault, R., Tahan, A., Sabourin, R.: Review of phenotypic diversity formulations for diagnostic tool. *Appl. Soft Comput. J.* **13**, 9–26 (2013)
10. Anfinsen, C.B.: Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973)
11. Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D.: Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004)
12. Tai, C.H., Bai, H., Taylor, T.J., Lee, B.: Assessment of template-free modeling in CASP10 and ROLL. *Proteins Struct. Funct. Bioinf.* **82**, 57–83 (2014)
13. Ligabue-Braun, R., Borguesan, B., Verli, H., Krause, M.J., Dorn, M.: Everyone is a protagonist: residue conformational preferences in high-resolution protein structures. *J. Comput. Biol.* **25**, 451–465 (2017)
14. Berman, H.M., et al.: The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000)
15. Borguesan, B., Inostroza-Ponta, M., Dorn, M.: NIAS-Server: neighbors influence of amino acids and secondary structures in proteins. *J. Comput. Biol.* **24**, 255–265 (2017)
16. Storn, R., Price, K.: Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11**, 341–359 (1997)
17. Du, K.-L., Swamy, M.N.S.: Search and Optimization by Metaheuristics. *Techniques and Algorithms Inspired by Nature*. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-41192-7>
18. Narloch, P., Parpinelli, R.: The protein structure prediction problem approached by a cascade differential evolution algorithm using ROSETTA. In: Proceedings-2017 Brazilian Conference on Intelligent Systems, BRACIS 2017 (2018)
19. Venske, S.M., Gonçalves, R.A., Benelli, E.M., Delgado, M.R.: ADEMO/D: an adaptive differential evolution for protein structure prediction problem. *Expert Syst. Appl.* **56**, 209–226 (2016)
20. Dorn, M., Inostroza-Ponta, M., Buriol, L.S., Verli, H.: A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. In: IEEE Congress on Evolutionary Computation, pp. 1233–1240 (2013)
21. Chaudhury, S., Lyskov, S., Gray, J.J.: PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010)
22. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983)