Diversification Strategies in Differential Evolution Algorithm to Solve the Protein Structure Prediction Problem

Pedro Henrique Narloch and Rafael Stubs Parpinelli(✉)

Graduate Program in Applied Computing, Departament of Computer Science, State University of Santa Catarina, Joinville, Brazil
pedro.narloch@gmail.com, rafael.parpinelli@udesc.br

Abstract. The protein structure prediction is considered as one of the most important open problems in biology and bioinformatics due the huge amount of plausible shapes that a protein can assume. The objective of this paper is to apply the Differential Evolution (DE) algorithm employing two simple diversification strategies known as generation gap and Gaussian perturbation to solve the protein structure prediction problem in the backbone and side-chain model. To test our approaches the 1PLW, 1ZDD and 1CRN proteins were used and the standard DE algorithm was compared with DE using the diversification approaches and with some state-of-art algorithms. Also, the genotypic diversity was analyzed during the algorithm run, showing the impacts generated by the diversification mechanisms. Despite its simplicity, the proposed approaches achieved competitive results.

Keywords: Protein structure prediction · Bioinformatics · Differential evolution

1 Introduction

Proteins are macromolecules which have important biological functions in every living organism when the tree-dimensional conformation is reached. These macromolecules are composed by an unique amino acid sequence, known as the primary structure, which influences the protein to fold into a three-dimensional shape [2]. Nowadays, the methods to determine an existing protein's tertiary structure are the nuclear magnetic resonance and the crystallography X-ray [9]. Although these methods can determine the native conformation of a determined protein, they are too expensive [1].

Different representations were created to solve the protein structure prediction (PSP) problem. The prediction only by the amino acid sequence is called *Ab Initio* prediction and it's one of the most challenging problems in bioinformatics because of its complexity even for small proteins [10]. The high complexity associated with this problem is due the huge amount of plausible shapes that a

A.M. Madureira et al. (eds.), *Intelligent Systems Design and Applications*, Advances in Intelligent Systems and Computing 557,
DOI 10.1007/978-3-319-53480-0_13

protein can assume. Hence, the protein structure prediction (PSP) is labeled as a NP-complete problem [11].

Due limitations of exact algorithms to solve this class of problems, metaheuristics became a viable way to explore the search space and find possible conformations in a plausible time. In recent literature, different approaches of Evolutionary Computing (EC) algorithms have been used to solve the PSP problem in atomic representations as in [2,6,9].

The standard Differential Evolution (DE) algorithm is a population-based EC algorithm that has been chosen to solve the PSP problem in the present work. The DE algorithm is considered a good algorithm to solve problems from continuous optimization [14]. However, it is known that it loses its diversity very quickly, increasing the chance of getting stuck in a local optimum when employed in a high multimodal problem like the PSP. Hence, with diversity control strategies it is possible to slow down the convergence, aiming to escape from local optima [4]. In this work, two simple diversification strategies are used with the DE algorithm: the Generation Gap and the Gaussian Perturbation.

Four different approaches were used to solve the PSP problem. One of them is the standard Best/1/Bin DE algorithm, while others are called as Generation Gap (GG), Gaussian Perturbation (GP) and the combination of them (GG-GP). To verify the efficiency of the proposed approaches, the genotypic diversity is analysed. With this analysis it is possible to verify the behaviour of algorithms and the impact of such strategies in the results obtained. Furthermore, results are compared with recent literature which used the same representation model and the same energy function.

The next sections are organized as follows. In Sect. 2 the PSP problem is described and related works are discussed. Section 3 explains the DE approaches to solve the PSP problem. Section 4 exhibit the results obtained in our test cases. Finally, Sect. 5 contains the conclusion of this work and some future directions.

2 Protein Structure Prediction

Proteins are made from amino acids chains where each amino acid is composed by an amino group (H_3N^+), a carboxyl group (COO^-) and a hydrogen atom attached to a central carbon (C_α) [2]. Each amino acid has a side chain attached to the C_α, distinguishing each one of the 20 different amino acids known in nature.

A protein can be depicted into four different well defined structures. The primary structure is formed by a linear sequence of amino acids, the secondary structure represents the local structure found in the backbone conformation, the tertiary structure considers the protein's final conformation (including the amino acids side chain) and determine its biological function. The quaternary structure represents interactions among proteins to accomplish specific functions.

To evaluate if a protein is near its native state, the Anfinsen's thermodynamic hypothesis declares that a native three-dimensional protein shape has the lowest

free energy. In the present work, the energy is obtained by the CHARMM force field [3] which is one of the most popular energy functions [10] shown in Eq. 1.

$$E_{total} = \sum_{bonds} K_b(b - b_0)^2 + \sum_{UB} K_{UB}(S - S_0)^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi(1 + \cos(\eta - \delta)) + \sum_{impropers} K_{imp}(\varphi - \varphi_0)^2 + \sum_{nonbond} \epsilon[(\frac{R_{minij}}{R_{ij}})^{12} - (\frac{R_{minij}}{R_{ij}})^6] + \frac{q_i q_j}{\epsilon_1 r_{ij}} \quad (1)$$

From Eq. 1, $\mathbf{E_{total}}$: is the total energy value; **bonds** measures the energy according to the bond stretching between two atoms; Urey-Bradley (**UB**) represents the interactions between pairs of atoms; **angle** is the sum among all angles in the structure; **dihedrals** is the energy associated with the torsion angles; **impropers** values are associated to deformations of improper torsion angles; **nonbond** values are related to Van der Waals and Charge-Charge energy. Van der Waals is the energy from interactions between nonbonded angles from attraction and repulsion. Charge-Charge varies according to the distance among atoms.

Different types of protein's atomic representations emerged with different levels of abstraction. Some commonly used are: **(a)** all-atom three-dimensional coordinates; **(b)** all-heavy-atom coordinates; **(c)** backbone atom coordinates + side-chain centroids; **(d)** C_α coordinates; **(e)** backbone and side-chain torsion angles.

As this work employs the backbone and side-chain torsion angles model, it is known that each residue has a defined number of torsion angles that is needed to be optimized. Each amino acid has three backbone angles (ϕ, ψ, and ω) and a particular number of side chain angles (χ_i) as shown in Table 1. A backbone classification was employed to identify the secondary structure and

Table 1. χ angles for each amino acid

Aminoacid	χ angles
GLY, ALA, PRO	Backbone
SER, CYS, THR, VAL	χ_1
ILE, LEU, ASP, ASN, PHE, TYR, TRP	χ_1, χ_2
MET, GLU, GLN	χ_1, χ_2, χ_3
LYS, ARG	χ_1, χ_2, χ_3, χ_4

Table 2. DSSP 8-class classification.

Secondary structure	ϕ bounds	ψ bounds
H (α-helix)	$[-67°, -47°]$	$[-57°, -37°]$
B (β-bridge)	$[-130°, -110°]$	$[110°, 130°]$
E (β-strand)	$[-130°, -110°]$	$[110°, 130°]$
G (3-10-helix)	$[-59°, -39°]$	$[-36°, 16°]$
I (pi-helix)	$[-67°, -47°]$	$[-80°, -60°]$
T (turn)	$[-180°, 180°]$	$[-180°, 180°]$
S (bend)	$[-180°, 180°]$	$[-180°, 180°]$
U (undefined)	$[-180°, 180°]$	$[-180°, 180°]$

the recommended bounds for each type of structure. These angles are shown in Table 2 and are based on the full DSSP 8-class classification [12].

2.1 Related Works

Some related works apply bio-inspired algorithms, *Ab Initio* prediction and CHARMM energy function. In [9] was applied a GA with two different approaches for diversity control: the random immigrants technique, which replaces a percentage of individuals from the population for new randomly generated individuals, and an extension called simplified self-organizing random immigrants with dynamic replacement rate. In [13] A bacterial foraging optimization algorithm (BFOA) was applied for the 1PLW protein using CHARMM energy function. Although most of related works tried to improve the diversity, in [17] a GA was combined with Hill Climbing in a parallel grid environment, improving the exploitation capacity.

In [15] the NSGA-II was employed to solve the PSP problem using island models. Also, a modified PAES algorithm is proposed in [7,8] with immune inspired operators (cloning and hyper mutation). Another multi-objective approach was proposed in [18] using the DE algorithm. However, the DE was modified to be adaptive varying the mutation mechanism. This technique is known as probability matching, associating a percentage at each DE version to execute the mutation process. If the mutation process is successful, than its chance to be selected again increases. This is the approach that found the lowest energy value using CHARMM for the 1PLW, 1ZDD and 1CRN proteins. The multi-objective formulation in these works is slicing the CHARMM energy function in two terms: bonded and non-bonded.

3 Methods

In this work each individual is formed by a set of angles representing amino acids. An individual is structured as a vector and its size changes according to the number of amino acids in each protein. Figure 1 illustrates the structure of an individual for the 1PLW protein which have 5 amino acids. Note that the ω angle is not in our representation because its value is set always to 180°.

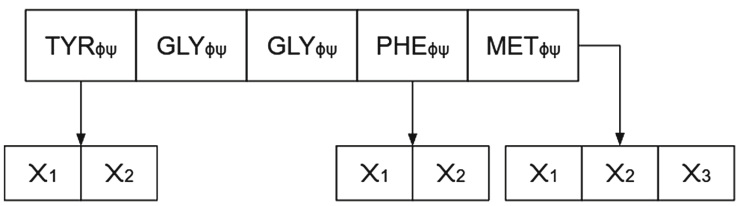

Fig. 1. Graphical representation of 1PLW individual.

The standard DE algorithm is population-based and at each new iteration an offspring is generated by a mutation operator to replace the current population if it achieves a better solution. In order to modify this routine and improve the diversity in the population, we have used the generation gap mechanism [16]. This mechanism is commonly used in EC and only a fraction of the population is replaced by the offspring according to a parameter G which varies between 0 and 1. The maintained individuals are selected at random from the current population. This model is called generation gap DE (DE_{GG}).

To create new individuals, the DE algorithm uses a mutation mechanism, combining values from different individuals. There are different approaches for mutation in DE and, in this work, the $\text{DE}_{Best/1/Bin}$ version was selected. This standard version of DE always uses the best individual in the population to combine with two other random individuals. The $\text{DE}_{Best/1/Bin}$ approach creates a new individual using $\boldsymbol{w} = \boldsymbol{x_{best}} + F \cdot (\boldsymbol{x_{rand1}} - \boldsymbol{x_{rand2}})$, where $\boldsymbol{w}$ represents the new generated individual and F is a threshold that need to be set.

Hence, another modification was done in the mutation operator of the standard $\text{DE}_{Best/1/Bin}$ algorithm. For the two randomly selected individuals, a gaussian perturbation technique is applied. With the gaussian perturbation, $\boldsymbol{x_{rand1}}$ and $\boldsymbol{x_{rand2}}$ are considered the mean and the standard deviation is defined between 0 and 1. This model is called gaussian perturbation DE (DE_{GP})

To verify how these approaches impact the diversity of solutions during the optimization process, this work uses a genotypic diversity measure for continuous domains [5]. The Eq. 2 shows how to calculate the genotypic diversity.

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

where D is the size of the solution vector, N is the population's size and x the individual (or solution vector). The NMDF is a normalization factor which corresponds to the maximum diversity value so far. The genotypic diversity starts with 1 which is the maximum value and when it reaches 0 it corresponds to the full convergence of the population. With this measurement it is possible to verify the diversity level during each iteration. This is an important measure to verify if the algorithm is getting trapped in a local optima and, consequently, getting a premature convergence.

Besides the function evaluation given by CHARMM and the genotypic diversity measure given by Eq. 2, there is another important metric which is considered in this work: the root mean square deviation (RMSD). The RMSD is a measure given in Å(angstrom) which compares the atomic distance between proteins and verifies if the final predicted conformation reached the native conformation. When the RMSD is near 0 means that the predicted protein is very similar to the native protein.

4 Experiments, Results and Analysis

In the current work three different proteins were used as problem instances: 1PLW, 1ZDD and 1CRN. The smallest protein used is known as *Met-Enkephalin* (1PLW) with only 5 amino acids and 22 angles to be optimized, without any well defined secondary structure. The 1ZDD is a protein which have two well defined α-helices structures and it contains 34 amino acids with 179 angles to be optimized. The biggest protein used in this work has 46 amino acids and 191 angles to be optimized, known as 1CRN. The 1CRN protein has two well defined α-helices and two β-sheets as secondary structures.

The experiments were conducted with 4 different algorithm configurations: the standard $\text{DE}_{Best/1/Bin}$, the standard DE with generation gap mechanism (DE_{GG}), the standard DE with gaussian perturbation (DE_{GP}) and DE_{GG-GP} which combines the standard DE with generation gap and gaussian perturbation. This work also compares the results obtained with another works found in the literature. All approaches use the atomic representation and the CHARMM energy function calculated with Tinker Molecular Dynamics Package.

The DE parameters used in this work are recommended by [18], with a population size of 100 individuals, the mutation factor (F) is set to 0.5, the crossover factor is 1 and the number of function evaluations is 500.000. The parameters G, which controls the generation gap was empirically set to 0.8, and the GSD which is responsible for the standard deviation was empirically set to 0.1.

For each protein and each approach 10 runs were done. Table 3 contains the results obtained for 1PLW, 1ZDD and 1CRN proteins.

The first column indicates each protein and the second column identifies each algorithm. Column 3 represents the minimum energy found in all runs and column 4 the RMSD_α from each minimum energy. Finally, column 5 represents the average minimum energy with the standard deviation. All DE approaches were developed using C++ language in an Intel core i7 with 8GB RAM.

For *Met-Enkephalin* (1PLW) all four developed DE approaches got similar results. However, the DE_{GG-GP} reached -35.82 kcal mol^{-1} with RMSD_α of 1.98Å. These values are competitive with the state-of-art ADEMO/D algorithm proposed in [18]. The lowest RMSD_α was reached by NSGA-II [15] with 1.26Å.

For 1ZDD protein the results obtained showed significant differences among all four DE approaches. It is possible to notice that the DE_{GP} and DE_{GG-GP} reached better results for minimum energy values when compared with standard $\text{DE}_{Best/1/Bin}$ and DE_{GG}. The DE_{GP} reached $-1,216.40$ kcal mol^{-1} with a RMSD of 2.36 Å, becoming competitive with the state-of-art algorithm found in literature known as ADEMO/D [18] and NSGA-II [15].

Analysing the results for 1CRN protein, the best approach was the DE_{GP} with energy value of 166.83 kcal mol^{-1}. Comparing with the four DE approaches developed in this work, the DE_{GP} was better than $\text{DE}_{Best/1/Bin}$ when considering the energy value and the standard deviation showing that the gaussian perturbation improved the results obtained. Comparing the DE_{GP} with ADEMO/D [18], our approach achieved lower energy besides the RMSD_α was bigger than the results found in the literature.

Table 3. Results obtained.

Protein	Version	Min. energy	$RMSD_\alpha$	Avg. energy
1PLW	$DE_{Best/1/Bin}$	−34.69	1.90Å	-28.58 ± 3.00
	DE_{GG}	−33.95	1.99Å	-26.35 ± 2.77
	DE_{GP}	−32.10	1.63Å	-27.96 ± 1.91
	DE_{GG-GP}	**-35.82**	1.98Å	-30.47 ± 4.44
	ADEMO/D [18]	−30.43	1.77Å	–
	BFOA [13]	−19.10	3.60Å	–
	I-PAES [7]	−20.56	2.83Å	–
	NSGA-II [15]	−22.73	**1.26Å**	–
	SSORIGA [9]	42.82	–	46.23 ± 1.64
1ZDD	$DE_{Best/1/Bin}$	−955.68	2.65Å	-508.52 ± 262.99
	DE_{GG}	−796.68	4.25Å	$95.03 \pm 1,503.92$
	DE_{GP}	$-1,216.40$	2.36Å	$-1,086.99 \pm 105.74$
	DE_{GG-GP}	$-1,156.95$	5.69Å	-983.97 ± 119.80
	ADEMO/D [18]	**-1,301.38**	**2.14Å**	–
	I-PAES [8]	$-1,052.09$	2.27Å	–
	NSGA-II [15]	$-1,218.57$	3.81Å	–
1CRN	$DE_{Best/1/Bin}$	818.04	7.89Å	$2,483.87 \pm 3,424.94$
	DE_{GG}	594.07	13.12Å	$1,608.72 \pm 1,152.38$
	DE_{GP}	**166.83**	10.71Å	288.52 ± 86.67
	DE_{GG-GP}	260.12	8.60Å	464.60 ± 133.59
	ADEMO/D [18]	253.25	6.06Å	–
	I-PAES [8]	509.09	**4.43 Å**	–
	NSGA-II [15]	262.68	7.32Å	–
	SSORIGA [9]	503.56	–	535.09 ± 20.98

Figure 2 shows both the energy and the genotypic diversity convergence over generations for all developed DE approaches. Note that all approaches converged very quickly for 1PLW protein. At generation 1,000 the energy function stabilized. However, when the genotypic diversity is plotted, there is diversity in the population after the generation 1,000 for DE_{GG-GP} while for $DE_{Best/1/Bin}$ the diversity ended earlier, given no possibility to create different individuals. Because of the small size of this protein, even using a diversity control mechanism the $DE_{Best/1/Bin}$ and DE_{GG-GP} got very similar values.

Analysing the convergence for 1ZDD protein, the energies of $DE_{Best/1/Bin}$ and DE_{GG} stabilizes around generations 2,500 and 3,000, respectively, while the energies of DE_{GP} and DE_{GG-GP} are still decreasing at generation 5,000, when all algorithms end. This behaviour is related with the diversity in the population. Note that for the approaches which have converged earlier, worst energy values were obtained and the diversity was lost prematurely. However, for DE_{GP} and

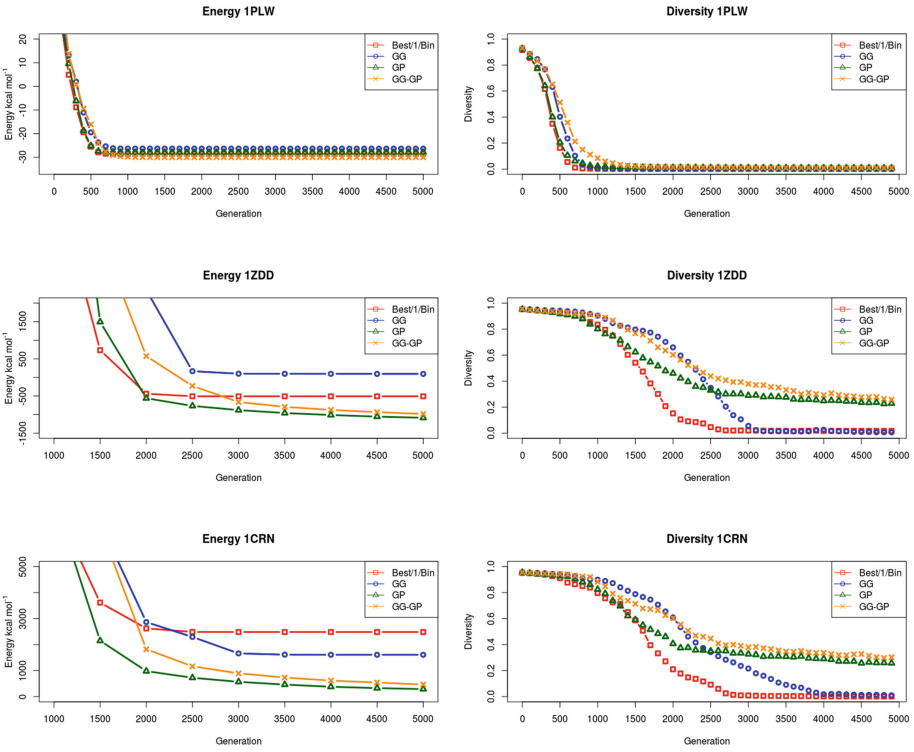

Fig. 2. Energy (left) and genotypic diversity (right) for each sequence.

DE_{GG-GP}, that have reached better energy results, the diversity is not over in the last generation.

The convergence analysis made for 1ZDD protein can also be made for 1CRN protein, where the approaches with diversity maintenance routines achieved better results avoiding premature convergence.

Overall the diversification techniques helped the standard DE to obtain lower average energy values mainly for the two biggest instances (1ZDD and 1CRN). It was verified through the GDM index that the genotypic maintenance is an important factor that need to be considered in PSP problem.

5 Conclusions and Future Research

This work applied four different DE approaches to solve some PSP problem instances using the torsion angles model and the CHARMM energy function. Two diversification strategies were used in order to avoid premature convergence: the generation gap and the gaussian perturbation.

Despite there are many works in the literature solving some PSP problem, none of them made an analysis of the diversity of solutions during the optimization process. As proposed in this work, the genotypic diversity was analyzed using

the GDM index. With this index was possible to relate the genotypic diversity with the energy convergence, verifying that the versions in which maintained the diversity got better results in all three proteins.

Although the genotypic diversification strategies increased the population's diversity, the gaussian perturbation always got the best energy values in comparison with the standard DE and the generation gap version. All four algorithms were also compared to state-of-art algorithms found in literature that used CHARMM as energy function. The DE_{GP} version showed to be competitive in all three proteins, 1PLW, 1ZDD and 1CRN, even being a much simpler approach than the works found in the literature.

Also, it was verified that when the algorithm ends, the diversity for bigger proteins is about 40%. This indicates that exploitation routines, like local search algorithms could be used to explore this diversity aiming to reach better energy values. Another future research could be the use of GPUs for energy minimization, possible granting higher speed ups when comparing with CPU approaches, providing bigger amounts of function evaluations and longer convergences.

As our developed approaches showed to be much simpler than the literature ones and reached competitive results, it is possible to use a famous Occam's razor statement: when you have two competing theories that make exactly the same predictions, the simpler one is the better.

References

1. Benítez, C.M.V., Parpinelli, R.S., Lopes, H.S.: An ecologically-inspired parallel approach applied to the protein structure reconstruction from contact maps. In: Genetic and Evolutionary Computation Conference, GECCO 2016, Denver, CO, USA, pp. 1299–1306 July, 2016
2. 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)
3. Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M., Karplus, M.: CHARMM: the biomolecular simulation program. J. Comput. Chem. **30**(10), 1545–1614 (2009)
4. Corriveau, G., Guilbault, R., Tahan, A., Sabourin, R.: Review and study of genotypic diversity measures for real-coded representations. IEEE Trans. Evol. Comput. **16**(5), 695–710 (2012)
5. Corriveau, G., Guilbault, R., Tahan, A., Sabourin, R.: Review of phenotypic diversity formulations for diagnostic tool. Appl. Soft Comput. **13**(1), 9–26 (2013)
6. Custodio, F.L., Barbosa, H.J., Dardenne, L.E.: A multiple minima genetic algorithm for protein structure prediction. Appl. Soft Comput. **15**, 88–99 (2014)
7. Cutello, V., Narzisi, G., Nicosia, G.: A multi-objective evolutionary approach to the protein structure prediction problem. J. Roy. Soc. Inter. **3**(6), 139–151 (2006)

8. Cutello, V., Narzisi, G., Nicosia, G.: Computational studies of peptide and protein structure prediction problems via multiobjective evolutionary algorithms. In: Knowles, J., Corne, D., Deb, K., Chair, D.R. (eds.) Multiobjective Problem Solving from Nature. Natural Computing Series, pp. 93–114. Springer, Heidelberg (2008)
9. Do, O., Tragante, V., Tinos, R.: A self-organizing genetic algorithm for protein structure prediction. Learn. Nonlinear Models **8**(3), 135–147 (2010)
10. 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)
11. Guyeux, C., CoTe, N.M.L., Bahi, J.M., Bienia, W.: Is protein folding problem really a NP-Complete one? first investigations. J. Bioinform. Comput. Biol. **12**(01), 1350017 (2014)
12. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12), 2577–2637 (1983)
13. Pal, A.: Ab-initio protein structure prediction using bacterial foraging optimization algorithm. Ph.D. thesis, Jadavpur University KOLKATA (2014)
14. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution: A Practical Approach to Global Optimization. Natural Computing Series. Springer, Berlin, New York (2005)
15. Romero, D.C.B.: A multi-objective Ab-initio model for protein folding prediction at an atomic conformation level. Ph.D. thesis, Universidad Nacional de Colombia. Facultad de Ingeniera. Departamento de Ingeniera de Sistemas y Computacin (2010)
16. Sarma, K.: Generation gaps revisited. Found. Genet. Algorithms (FOGA 2) **2**, 19 (1993)
17. Tantar, A.A., Melab, N., Talbi, E.G., Parent, B., Horvath, D.: A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. Future Gener. Comput. Syst. **23**(3), 398–409 (2007)
18. Venske, S.M., Gonalves, 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)