A Parallel Multi-objective *Ab initio* Approach for Protein Structure Prediction

David Becerra, Angelica Sandoval, Daniel Restrepo-Montoya, Luis F. Nino

Abstract— **Protein structure prediction is one of the most important problems in bioinformatics and structural biology. This work proposes a novel and suitable methodology to model protein structure prediction with atomic-level detail by using a parallel multi-objective *ab initio* approach. In the proposed model, i) A trigonometric representation is used to compute backbone and side-chain torsion angles of protein atoms ; ii) The Chemistry at HARvard Macromolecular Mechanics (CHARMm) function optimizes and evaluates the structures of the protein conformations; iii) The evolution of protein conformations is directed by optimization of protein energy contributions using the multi-objective genetic algorithm NSGA-II; and iv) The computation process is sped up and its effectiveness improved through the implementation of an island model of the evolutionary algorithm. The proposed model was validated on a set of benchmark proteins obtaining very promising results.**

Keywords: Bioinformatics, Protein Structure Prediction, Ab-initio methods, Parallel evolutionary computation, Multi-objective optimization.

I. Introduction

The Protein Structure Prediction (PSP) problem consists in determining the tertiary structure of a protein from its amino acid sequence alone [15]. PSP methods are based on two groups of principles: i) The theory of evolution; and ii) The laws of physics, biology and chemistry [14].

Although *ab initio* methods are very demanding in terms of computational resources, they are of particular importance because they overcome the inherent problems of comparative methods. Specifically, *ab initio* methods can be applied when there are no known homologous or similar structures related to the target protein [2]. In general, these methods can be used on any amino acid sequence because they only use information about the physical properties of amino acid atoms. Additionally, *ab initio methods* bring a deeper understanding of protein folding mechanisms [13].

Four main aspects have been commonly considered in the development of *ab initio* models [6]: the representation of protein conformations, the implementation of a cost function, modeling of methods to search for protein conformations, and deciding on which metrics to evaluate the error between predicted conformations.

To represent protein conformations and to overcome the high complexity of sampling protein conformations, most methods need a significant reduction in complexity [15]. Methods for reducing protein structure to discrete low-complexity models can be divided into two major classes: lattice [12] and off-lattice models [2].

Once having chosen a protein representation model that reduces the complexity of the conformational search sufficiently, a scoring function that works in the chosen low-complexity space must be defined. Potential energy functions allow evaluating the structure of protein conformations based on an energy value [11].

The next task to determine a protein's native state is to search for the lowest energy conformation in a vast conformational space. Molecular dynamics, Monte Carlo, Statistical Model and Probabilistic Road Maps methods are among the most commonly used search techniques [23].

The final task in the development of *ab initio* methods is defining some metrics to evaluate the similarity between predicted and native conformations. One of the most widely used metrics is the root mean square deviation (RMSD), which measures the similarity between atomic positions with respect to two superposed conformations.

This work proposes a multi-objective approach as a suitable methodology to model the protein structure prediction problem. In this exploration the PSP problem is considered involving multiple objectives where different 3D conformations may produce a trade-off in the funnel landscape [2]. In addition, the folded state of a protein is considered as a small ensemble of conformational structures, which can be modeled using the Pareto optimality [22]. Accordingly, the proposed approach involves the use of a multi-objective genetic algorithm to evolve protein conformations directed by an atom-interaction scoring function. Also, a parallel implementation is used to speed up the computation process and to improve the effectiveness of the algorithm to find the set of solutions arising from multiple executions.

II. The Proposed Approach

In *ab initio* methods, the task of determining a protein's native state can be understood as a search for the lowest energy structure in a vast conformational space [4]. Recently, evolutionary multi-objective formulations to the PSP problem have been introduced and its applicability studied, leading to the conjecture that the PSP problem can be suitably modeled as a multi-objective optimization problem [6].

Advances in the use of multi-objective evolutionary approaches have led to the exploitation of the inherent parallelism in such algorithms. Since the PSP problem is a complex NP-hard, CPU time-consuming process, these features make it a perfect candidate to be solved using a parallel model. By this means, the efficiency (i.e., how well it

All the authors are with the Intelligent Systems Research Laboratory (LISI) and the Algorithms and Combinatorics Research Group (ALGOS), National University of Colombia email:(dcbecerrar, gasondovalp, drestrepom, lfninov)@unal.edu.co

978-1-4244-8305-1/10/$26.00 ©2010 IEEE

performs computationally) and effectiveness (i.e., how good its solutions are) of PSP could be improved by increasing the number of processors allocated to it.

Figure 1 depicts a scheme of the methodology proposed, studied and implemented to model the PSP problem.

A. Representation of Structural Conformations

The tertiary structure of a protein is the spatial structure of its polypeptide chain. In principle, this structure is given by the spatial coordinates of the centers of all atoms in the protein. The tertiary structure of a protein depends on the shape of its backbone and the positions of the R groups relative to the backbone chain. Therefore, the degrees of freedom in a general structural representation are the backbone and side-chain torsion angles (ϕ, ψ, χ, and ω). By convention, both ϕ and ψ are defined as 180^o when the polypeptide is in its fully extended conformation and all peptide groups are in the same plane.

Given that the proposed approach works at an atomic level, off-lattice models were chosen. From the few commonly used conformation representations, the backbone torsion angles (i.e., ϕ and ψ) and side-chain torsion angles (i.e., χ) of a protein were used to represent proteins. Note that ω is not taken into account because it is already fixed at its ideal values.

In this work, backbone torsion angles are restricted by the secondary structure prediction, as shown in Table I [1], [17]. Moreover, side-chain torsion angles are constrained in regions derived from a backbone-independent rotamer library. Specifically, the libraries reported in [19] and [21] were used for this work. Additionally, the number of χ angles used are reported in Table II. Protein conformations are still very flexible under these constraints, and the structure can take on various conformations that are vastly from each other and from the native conformation.

TABLE I
SECONDARY STRUCTURE CONSTRAINTS

Structures	ϕ	ψ
H ($\alpha - helix$)	$[-39^o, -67^o]$	$[-16^o, -57^o]$
E ($\beta - strand$)	$[-130^o, -110^o]$	$[110^o, 130^o]$
C ($coil$)	$[-180^o, 180^o]$	$[-180^o, 180^o]$

TABLE II
NUMBER OF χ ANGLES

Amino acids	χ angles
G, A, P	main chain
S, C, T, V	$\chi_\cdot$
I, L, D, N, H, F, Y, W	$\chi_\cdot, \chi_\cdot$
M, E, Q	$\chi_\cdot, \chi_\cdot, \chi_\cdot$
K, R	$\chi_\cdot, \chi_\cdot, \chi_\cdot, \chi_\cdot$

B. Scoring Functions

In general, an atom energy function can be separated into two main groups, the internal terms, which include bond, angle, and dihedral contributions, and the non-bonded or external terms. Accordingly, in the proposed approach, the different terms of the CHARMm energy function were transformed into several objectives to fit the multi-objective evolutionary optimization formulation of the PSP problem. The first objective takes into account the local interactions, i.e., bonds, angles and torsion interactions between the atoms. In contrast, the second objective considers all the interactions between the atoms that are not connected by a covalent bond, (i.e., non-local interactions).

In this work, the molecular modeling package TINKER (available at http://dasher.wustl.edu/tinker/) is used to calculate the energy value of the CHARMm function. The equation of the CHARMm energy function can be further explored in [20].

C. Search Method

In this work, a multi-objective evolutionary algorithm (MOEA) called Elitist Non-Dominated Sorting Genetic Algorithm (NSGA-II) [9] is used to evolve the protein conformations. NSGA-II uses an elite-preservation strategy and an explicit diversity-preserving mechanism.

In any multi-objective problem, two spaces should be defined: the *decision space*, and the *objective space*. For our problem of concern, the chosen conformational method and the scoring function correspond to the definition of those spaces, respectively so that the proposed search method is the optimizer of the defined objective functions. Additionally, since multi-objective models require a phase of high level of information in order to select one solution from the final dominance set, two different implementations of the decision-making phase based on the Pareto *knees* were computed in this work. In [3], the authors introduced two algorithms to finding the knees in the Pareto front: an angle-based and an utility-based method.

D. Parallel Implementation

Parallelization techniques have been the focus of research in evolutionary algorithms for single objective optimization due to their population-based nature. As a natural extension, parallelization techniques can also be explored for evolutionary multi-objective optimization [8].

In this paper, a parallel framework using the JavaSpaces technology on a compute-server model over an island model as its parallel paradigm is proposed. The goals are to speed up the search process and to improve the precision of the PSP solutions given by the MOEA's. .

III. EXPERIMENTAL FRAMEWORK

This section describes the experimental framework implemented to report the results of the parallel multi-objective approach. Specifically, it provides an analysis of the results obtained over the set of widely studied proteins, which included Met-enkephalin (1PLW), Protein A domain, (1ZDD), Crambin (1CRN), Repressor of Primer (1ROP), Uteroglobin (1UTG). However, due to space restrictions, only the results of the 1ROP protein (63 amino acids) are shown in this paper

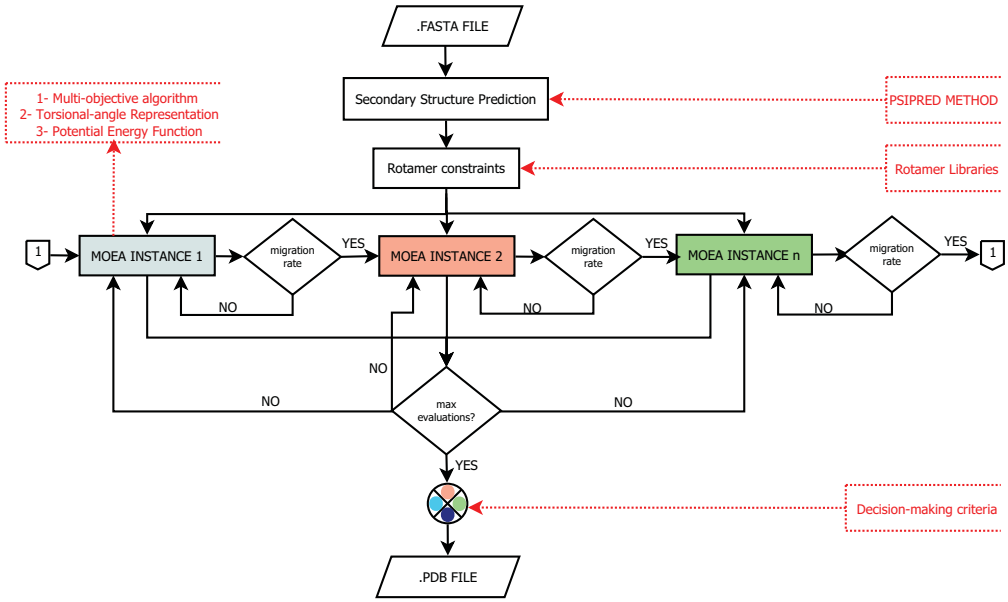


Fig. 1. Proposed parallel framework for multi-objective evolutionary optimization.

but the remaining results are provided as supplementaries files (http://www.biolisi.unal.edu.co/web-servers/psp/).

To thoroughly study the behavior and improvement of the algorithm, the serial counterpart was also implemented. The parameters used in the experimentation are shown in Table III.

TABLE III
ISLAND ARCHITECTURE, MIGRATION POLICIES AND MOEA'S PARAMETERS FOR PSP

Parameter or Policy	Value
Island topology	Ring
Number of islands	4
Migration rates	every 100,000 generations
Population to be migrated	10%
MOEA algorithm	NSGA-II [9]
Number of evaluations	2,000,000
Population size	100
SBX crossover	$\eta_c = 5, p_c = 0.9$
Polynomial mutation	$\eta_m = 10, p_m = 0.01\%$

For each experiment, the following analyses were performed. i) The $rmsdC_\alpha$ and energy values are computed over different protein conformations. These conformations are obtained from different evaluations of the MOEA approach using the implemented decision-making algorithms. Additionally, the results of the $rmsdC_\alpha$ are compared against its serial counterpart by the parallel algorithm. ii) A visual inspection of the predicted protein conformation with regards to its native counterpart (available as supplementary material) was construted. iii) The behavior of each island with respect to the maximum, minimum and average $rmsdC_\alpha$ found in different stages of the algorithm was also studied. iv) Given that the average values could be affected by very high or low $rmsdC_\alpha$ errors, the frequency histograms of these errors in specific intervals are analyzed. v) The dynamics of the Pareto fronts were evaluated in different stages of the algorithm. vi) The analysis of the energy values versus $rmsdC_\alpha$ errors is perform to try to unveil their relationship in the multi-objective algorithm.

A. Results for 1ROP

Based on the decision-making methods, the algorithm matches the crystal structure with an $rmsdC_\alpha$ of 3.7754 $\AA$ and energy -626.75 $kcalmol^{-1}$ (see Table IV). Note that Table IV reports the structures computed in different iterations of the algorithm; specifically, it shows a comparison between the structures in the final archives at different iterations based on different decision-making criteria .

TABLE IV
COMPUTED PROTEIN CONFORMATION FOR 1ROP. (ITER=ITERATIONS, D.M=DECISION MAKER. ERR= ERROR, E= ENERGY, S = SERIAL VERSION, P= PARALLEL VERSION)

Iter$\times 10^{\bullet}$	D.M	Err(S)	E(S)	Err (P)	E (P)
0.5	Utility	5.263	-492.43	3.955	-212.95
	Angles	5.367	-325.35	3.980	-225.44
1	Utility	5.323	-605.21	3.915	-476.07
	Angles	5.331	-598.69	3.925	-477.75
1.5	Utility	5.288	-637.47	3.806	-584.00
	Angles	5.327	-624.97	3.801	-583.88
2	Utility	5.296	-666.17	3.826	-656.40
	Angles	5.296	-665.36	3.775	-626.75

A more detailed inspection of the minimum, maximum and average $rmsdC_\alpha$ values at different iterations of the algorithm is depicted in Figure 2. It is worth to stress that the proposed parallel implementation provides an adequate exploration of the search space; it also improves the precision of the model and the quality of the obtained Pareto fronts.

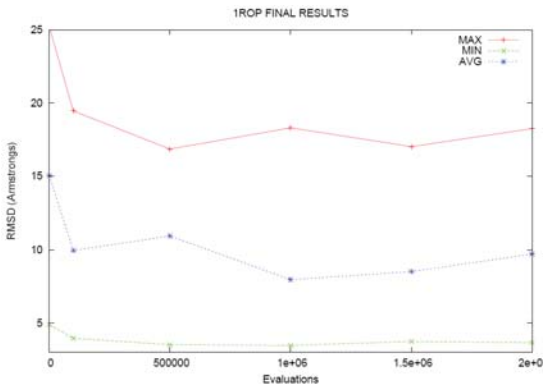


Fig. 2. Minimum, average and maximum RMSD for 1ROP

The conformation with minimum RMSD evolved by the algorithm had an $rmsdC_\alpha$ error of $3.3935\mathring{A}$, which is significantly better than the minimum error reported for this protein.

Note that Figure 3 depicts the amount of individuals belonging to a specific $rmsdC_\alpha$ interval. This measure is highly important because it allows understanding the role of the evolution in minimizing the RMSD errors without showing propensity for outlier values. Specifically, it can be concluded from Figure 3 (a) that the number of protein conformations belonging to the interval [3Å, 5Å] accounts for nearly fifty percent of the population, whereas the number of conformations belonging to this interval at the beginning of the evolution was less than two percent of the population.

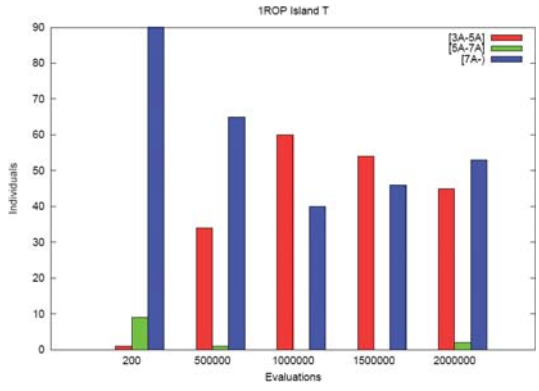
Figure 4 shows the different Pareto fronts found by the algorithm in different iterations. It should be stressed that, although the NSGA-II algorithm finds well distributed fronts and emphasizes the convergence, there are some individuals in the optimal set that present high energy levels. Typically, these individuals show a marked difference in the objectives trade-off.

In the plot shown in Figure 5, the correlation between the energy and $rmsdC_\alpha$ for conformations sampled by the MOEA algorithm is depicted. The plot shows that it is possible to produce structures of lower energy than that of their reported native structure, which is an important point because even if a minimum energy conformation is found by the algorithm, this conformation is not usually close to the native conformation.

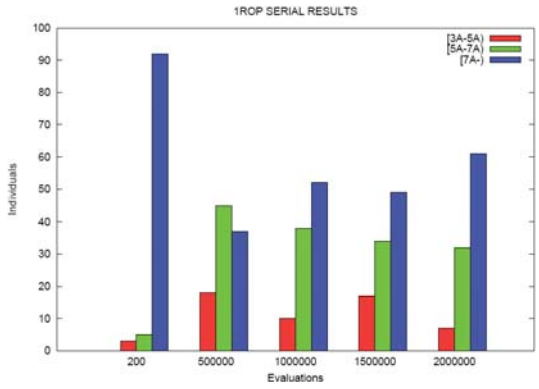
Finally, Table V reports the comparison of the proposed approach versus other approaches for 1ROP.

IV. Conclusions

This work contributes to the solution of the PSP problem as it proposes a novel procedure and algorithm for protein structure prediction based on a parallel multi-objective *ab initio* approach with an atomic level of representation. The proposed approach can certainly benefit from other improvements, such as refinements of the MOEA algorithm and the parallel implementation as well as the addition of new features, such as including amino acid specific information and biological heuristic information.



(a) Final Results Parallel Version



(b) Final Results Serial Version

Fig. 3. Histrogram of RMSD ranges (1ROP), Panel (a) shows the set of dominant solutions obtained by joining the results of the islands. Panel (b) illustrates the result of running a serial instance of the algorithm.

TABLE V
Proposed approach versus other approaches for 1ROP

Algorithm	$rmsdC_\alpha(\mathring{A})$
OUR	3.7754
I-PAES [6]	3.70
Scatter [18]	17.25
HC-GA [5]	5.6
Bhageerath [16]	4.3
CReF [10]	7.1
(1+1)-PAES. [7]	6.31
(1+1)-PAES. [7]	8.665

With respect to the *ab initio* modeling, it can be concluded that the trigonometric representation is a computational feasible spatial representation that facilitates the implementation of the chromosome in a genetic-based approach. Additionally, this spatial representation has biological significance as it allows to model secondary structures and some spatial characteristics of the protein, such as torsion and bond angles. This spatial representation also enables a reconstruction of the protein conformations and their representation in Cartesian coordinates.

An MOEA allowed to find a good set of 3D conformations inside a protein folded state. The use of bonded and non-bonded interactions are appropriate as the objectives to

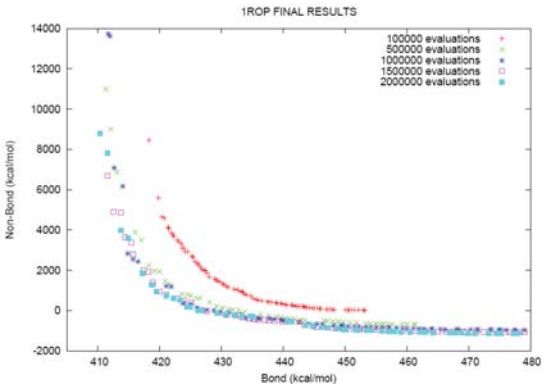


Fig. 4. Pareto Front for the parallel version of the algorithm for 1ROP.

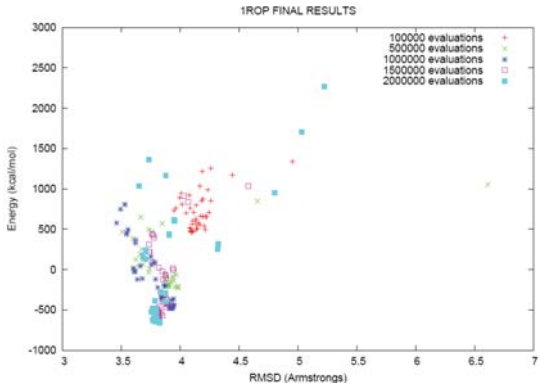


Fig. 5. Relationship between Energy and RMSD for 1ROP.

optimize in the MOEA algorithm as they can be considered as the main forces directing folding toward the native state. In addition, the parallel algorithm provided better PSP results with respect to its serial counterpart.

The use of secondary structure information is fundamental for the accuracy of the predicted structures given the importance of those conformations for the protein folding process in nature. In protein structure prediction methods, the use of a rotamer library allows determining or modeling a structure based on the most likely side-chain conformations, which saves time and produces a more likely structure. The use of these two heuristics causes a significant reduction of the search space but they also have a big impact on the accuracy of the model if wrong predictions are computed.

This work is a contribution to the long distance that is still to be covered to find a final solution to the PSP problem, but it is an important step that could help direct the problem toward new unexplored roads. The results were consistent with outcomes expected for an *ab initio* method. It is clear that *ab initio* methods are still not completely successful in finding the correct protein structure, even for small proteins, but much progress is being made to solve the main difficulties thanks to the use of new algorithms to exploit the available computer power and novel algorithms to tackle the protein structure prediction problem.

REFERENCES

[1] D. Becerra, D. Vanegas, G. Cantor, and L. Niño. An association rule based approach for biological sequence feature classification. In *Proceedings of the Eleventh CEC*, pages 3111–3118. IEEE, 2009.

[2] R. Bonneau and D. Baker. Ab initio protein structure prediction: progress and prospects. *Annual Review of Biophysics and Biomolecular Structure*, 30:173–89, 2001.

[3] J. Branke, K. Deb, H. Dierolf, and M. Osswald. Finding knees in multi-objective optimization. *Lecture Notes in Computer Science*, pages 722–731, 2004.

[4] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes. Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Func. and Genetics*, 21:167, 1995.

[5] L.R. Cooper, D.W. Corne, and M.J. Crabbe. Use of a novel Hill-climbing genetic algorithm in protein folding simulations. *Computational biology and chemistry*, 27(6):575, 2003.

[6] V. Cutello, G. Narzisi, and G. Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of The Royal Society Interface*, 3:139–151, 2006.

[7] 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: from concepts to applications*, page 93, 2008.

[8] D. Dasgupta, D. Becerra, A. Banceanu, F. Nino, and J. Simien3. A Parallel Framework for Multi-objective Evolutionary Optimization. In *Proceedings of the 2010 IEEE World Congress on Computational Intelligence*. IEEE, 2010.

[9] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. *Lecture Notes in Computer Science*, pages 849–858, 2000.

[10] M. Dorn and O.N. de Souza. CReF: a central-residue-fragment-based method for predicting approximate 3D polypeptides structures. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1261–1267. ACM New York, NY, USA, 2008.

[11] Y. Duan and P. A. Kollman. Computational protein folding: From lattice to all-atom. *IBM Systems Journal*, 40:297–309, 2001.

[12] S. Duarte, D. Becerra, L. Nino, and Y. Pinzon. A novel ab-initio genetic-based approach for protein folding prediction. pages 393–400, London, England, 2007. ACM.

[13] F. Allen et al. Blue gene: a vision for protein science using a petaflop supercomputer. *IBM Syst. J.*, 40:310–327, 2001.

[14] A. Fiser, M. Feig, C.L. Brooks, and A. Sali. Evolution and physics in comparative protein structure modeling. *Accounts of Chemical Research*, 35:413–21, 2002.

[15] C.A. Floudas, H.K. Fung, S.R. McAllister, M. Mönnigmann, and R. Rajgaria. Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, 61:966–988, 2006.

[16] B. Jayaram, K. Bhushan, S.R. Shenoy, P. Narang, S. Bose, P. Agrawal, D. Sahu, and V. Pandey. Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. *Nucleic acids research*, 34(21):6195, 2006.

[17] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices1. *Journal of molecular biology*, 292(2):195–202, 1999.

[18] C. Kehyayan, N. Mansour, and H. Khachfe. Evolutionary Algorithm for Protein Structure Prediction. In *Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on*, pages 925–929, 2008.

[19] S.C. Lovell, J.M. Word, J.S. Richardson, and D.C. Richardson. The penultimate rotamer library. *Proteins Structure Function and Genetics*, 40(3):389–408, 2000.

[20] A.D. MacKerel, C.L. Brooks, L. Nilsson, B. Roux, Y. Won, and M. Karplus. CHARMM: the energy function and its parameterization with an overview of the program. *Enzyme*, 1(3dw8):1.

[21] R. More. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, 6(8):1661–1681, 1997.

[22] S.S. Plotkin and J.N. Onuchic. Investigation of routes and funnels in protein folding by free energy functional methods. *Proceedings of the National Academy of Sciences*, 97(12):6509–6514, 2000.

[23] S. Thomas, G. Song, and N.M. Amato. Protein folding by motion planning. *Physical Biology*, 2:S148–S155, 2005.