Fragment-Based Computational Protein Structure Prediction

Nashat Mansour, Meghrig Terzian
Department of Computer Science and Mathematics
Lebanese American University, Lebanon
e-mail: nmansour@lau.edu.lb, meghrig.terzian@lau.edu

***Abstract*—Proteins consist of sequences of amino acids that fold into 3-dimensional structures. The 3-dimensional configuration determines a protein's function. Hence, it is very important to determine the correct structure in order to identify the wrong folding that indicates a disease situation. Computational protein structure prediction methods have been proposed in order to alleviate the enormous time taken by wet-lab methods. This paper presents a fragment-based protein tertiary structure prediction method which employs the CHARMM36 energy model. The method is based on a two-phase Scatter Search algorithm that minimizes the energy function. Backbone fragments are extracted from the Robetta server and side chains are, extracted from the Dunbrack Library. The results show that the algorithm produces tertiary structures with promising root mean square deviations.**

Keywords-protein structure prediction; scatter search; CHARMM36; protein fragments.

I. INTRODUCTION

Proteins are macromolecules found in all biological organisms. They are composed of a sequence of amino acids and are involved in a wide variety of functions within cells including cell structure, cell motility, cell signaling, enzyme catalysis, and substance transport. For example, enzymatic proteins, such as pepsin, are fundamental for the metabolism and accelerate the rates of biochemical reactions. The various functions are determined by the 3-dimensional folding that is based on the unique sequence of amino acids.

Predicting protein tertiary structure provides information about the functionality, localization and interactions between proteins and consequently contributes in drug design and disease prevention associated with protein misfold. The laboratory experimental methods for protein structure prediction, mainly X-ray crystallography and nuclear magnetic resonance, consume a lot time and are error-prone. Hence, computational methods may offer an alternative. Computational approaches for protein structure prediction lie in two groups. The first group, comparative modeling, predicts structures using proteins of known structures as templates [1]-[4]. The second, *ab initio*, predicts structures using the amino acid sequence of the structure to be predicted [5]-[7].

Ab initio approaches are based on Anfinsen's theory stating that the lowest energy value protein conformation is the most stable one [8]. *Ab initio* methods are divided into two classes. The first is fragment-based and the second biophysics-based. Fragment-based methods employ database information, whereas biophysics-based methods do not [5]. A typical *ab initio* method starts with random conformations, generates substitute conformations using heuristics, calculates their energies, and keeps on generating substitute conformations until the ending criterion is reached, where the solution is the conformation with the lowest energy. The efficiency of *ab initio* methods depends on the utilized energy function accuracy and the search algorithm efficiency.

The protein structure prediction problem is NP-Complete. Hence, there is a need for heuristic methods. The main challenge of structure prediction methods is the search space vastness. To limit the search space, a number of models, such as the Hydrophobic-Polar model [9], UNRES model [10], and dihedral angles model [6] have been developed. But, limiting the search space by simplifying the structure model may limit the quality of the predicted protein structure.

Atom-based *ab initio* methods either use fragment databases or are pure. Pure ab initio methods do not employ any prior information. Examples of such published pure *ab initio* work are based on scatter search algorithm [11] and a genetic algorithm [12].

Fragment-based protein structure prediction methods employ peptide fragments, secondary structures and statistical information from the Protein Data Bank (PDB) structures to predict protein tertiary structures. The basic principle behind this method is the presence of a strong relationship between an amino acid sequence and structure [4]. A typical *ab initio* fragment-based method starts with generating fragments from the PDB. Then, heuristics are used to optimize conformations and generate native like structures by using energy functions and evaluation methods.

Iterative Threading Assembly Refinement (I-TASSER) is a unified meta server for protein structure and function prediction [13]. Fragments are utilized to assemble well-aligned structural regions of the segments with unaligned regions. Starting from the query sequence, I-TASSER uses Basic Local Alignment Search Tool (BLAST) to identify sequence homologs. Then, the homologs are aligned using multiple sequence alignment to form a sequence profile and are utilized for predicting secondary structures which are threaded by gathering top template hits from ten threading programs.

The University College London (UCL) bioinformatics group developed several algorithms to tackle protein structure prediction and function annotation including

Fragment-based protein folding (FRAGFOLD) for prediction of tertiary structure [14]. FRAGFOLD starts the folding simulation with supersecondary fragment selection for each position in the query sequence. The energy function utilized includes terms for short-range, long-range, solvation, steric clashes, and hydrogen bonds with their corresponding weights. The energy minimization phase is conducted using a Simulated Annealing approach [15].

ROSETTA [16], an integrated package for protein structure prediction and functional design, is one of the leading *ab initio* performers in Critical Assessment of Protein Structure Prediction (CASP). ROSETTA uses fragments to model the protein backbone. Then, the model is refined and rotamers from the Dunbrack library are assembled to model the side chains. The fragment assembly phase is guided by Monte Carlo Simulated Annealing search [17]. Two energy functions are used in ROSETTA; the probability values used in the energy function are collected using Bayesian statistics from the PDB [17].

This work presents a fragment-based protein tertiary structure prediction method that yields good suboptimal structures. The method employs the CHARMM36 energy model [18] and is based on designing a two-phase scatter search metaheuristic that minimizes the energy function. Backbone fragments are extracted from the Robetta server and, later, side chains are extracted from the Dunbrack Library. The results of applying our method to three proteins are assessed by calculating their energy and root mean square deviation (RMSD) values and by visualizing them. The best structures generated are compared with structures generated by ROSETTA, I-TASSER, and previous work performed by Mansour et al. [19]. The adapted scatter search algorithm yields promising results.

The paper is organized as follows. Section 2 presents a protein structure model, the energy function used, and the assumptions made. Section 3 explains the design of the proposed scatter search algorithm. Section 4 discusses the experiments performed and the results obtained. Section 5 concludes the paper.

II. Protein Model and Energy Function

In the dihedral angles model of protein presentation, the backbone conformation is determined by three torsion angles, Phi φ, Psi ψ and Omega ω, and the conformation of side chains is determined by the Chi χ angles. Phi is formed by the C-N-Cα and N-Cα-C planes and rotating around the N-Cα bond. Psi is formed by the N-Cα-C and Cα-C-N planes and rotating around the Cα-C bond. Omega is formed by the Cα-C-N and C-N-Cα planes and rotating around the C-N bond.

The All-atom Chemistry at HARvard Macromolecular Mechanics (CHARMM36) protein force field function computes the potential energy of a protein structure. The potential energy is the sum of individual terms representing the internal and non-bonded contributions. Internal terms include bond, angle, Urey-Bradley, improper torsion, torsion, and backbone torsional correction energy values. The non-bonded terms include electrostatic, Van der Waals, and solvation values. The following equation represents the nine terms of the CHARMM36 energy function E as a function of the conformation c [18][20].

$$E(c)=\sum_{bonds}K_b(b-b_o)^2+\sum_{angles}K_\theta(\theta-\theta_o)^2+\sum_{impropers}K_{imp}(\varphi-\varphi_o)^2+$$
$$\sum_{torsions}K_x(1+\cos(n\chi-\delta))+\sum_{solvation}\sigma_i A_i+\sum_{electrostatic}\frac{q_i q_j}{r_{ij}}+$$
$$\sum_{vanderWaals}\varepsilon_{ij}\left(\left(\frac{R\min_{ij}}{r_{ij}}\right)^{12}-2\left(\frac{R\min_{ij}}{r_{ij}}\right)^{6}\right)+\sum_{Urey-Bradley}Ku(u-u_o)^2+\sum_{\alpha carbons}CMAP(\phi,\psi)$$

Assumptions have been made in this work to simplify the representation of a solution including: constant bond lengths and bond angles; and insignificant improper torsion, Urey-Bradley, and CMAP components. Also, Hydrogen atoms are combined with neighboring heavy atoms referred to as the "extended-atom representation”, which reduces the size of the problem.

III. Scatter Search Basic Algorithm

Scatter Search (SS) is a population based, evolutionary and stochastic meta-heuristic that generates and maintains high quality solutions by controlling the search space through randomization, recombination and diversification [21]. Scatter Search generates a random set of candidate solutions, improves them and selects 20% of these solutions and places them in the reference set. Half of the selected solutions are high quality and the other half diverse. Then, it iterates through a subset generation, solution combination, improvement and reference set update methods, where new subsets are generated, combined, improved and included in the reference set according to a certain criteria. In the following sections, we describe the design of the various methods that adapt scatter search to provide good suboptimal solutions for the protein structure prediction problem.

A. Solution Encoding

A PROTEIN candidate solution is represented as a list of consecutive objects, AMINO ACIDs. The position of an Amino Acid in a PROTEIN object list is consistent with its position in the protein chain. Consequently, the size of the PROTEIN object is equal to the number of amino acids of the protein. Each AMINO ACID consists of a name, Phi, Psi, omega, Chi1 to Chi4 angle values (if present), van der Waals, electrostatic, torsion, and ASA energy values, and a list of ATOM objects representing the atoms of that particular AMINO ACID. An ATOM has a name and a POSITION object, representing the Cartesian coordinates of that atom.

B. Diversification Generation Method

The Diversification Generation Method (DGM) generates random, diverse and valid initial solutions. These solutions are formed by randomly selecting a nine width window of consecutive amino acids from the protein chain, then randomly selecting a 9 width fragment (containing phi, psi and omega values) for this particular position and placing the torsion angles in their corresponding spots. Next, the Cartesian coordinates of the atoms of each amino acid are calculated and the energy of the solution is computed. These steps are repeated until the chain is full and the generated structure is valid, that is, each amino acid in the chain has phi, psi and omega values and there are no collisions between the atoms.

C. *Improvement Method*

The Improvement Method (IM) enhances the solutions generated by the DGM. After saving the existing values of the torsion angles, for every amino acid position in the chain a fragment is randomly selected for that position and torsion angle values are inserted into the solution. If the newly generated solution is feasible and its potential energy value is lower than the old solution, the move is accepted. The improvement method is run 25 times the protein size. Then, the same procedure is repeated with length 3 fragments, with number of moves being 50 * protein size.

D. *Reference Set Update Method*

The Reference Set Update Method (RSUM) constructs two reference sets (RefSet), high-quality and diverse solutions. The RefSet b contains b1 high-quality solutions, and b2 diverse solutions. Since b=20% of the population's size and PopSize=100, RefSet has 20 solutions. The b1 solutions are the top 10 minimum energy valued solutions generated from IM. The b2 solutions are the solutions having diverse energy values from the b1 high-quality solutions. After selecting the top 10 solutions of minimum energy and placing them in the RefSet (HQRefSet), for every solution not in the HQRefSet, the minimum distance between this solution and all solutions in the HQRefSet is computed and sorted in decreasing order of minimum distances. The first b2 (most diverse) solutions having the highest energy values are inserted into the RefSet (DivRefSet). The algorithm terminates when no new solutions are found to be inserted into the RefSet or when the number of added solutions reaches a limit.

E. *Subset Generation Method and Solution Combination Methods*

In the Subset Generation Method (SGM), subsets of the reference set are generated by using a method that groups every pair of elements in a subset. (b!/2!(b – 2)!) subsets are generated, where b is the size of the RefSet.

Then, the pairs generated by the SGM are combined to generate one candidate solution for each pair. For every amino acid, the dihedral angles from either candidate solution are used and the partial energy function, up to this amino acid, is calculated. The angle values that yield a lower energy value of the structure are chosen to be included in the combined candidate solution.

F. *Side chain Assembly*

After the termination of phase one, the solution with the lowest Cα-RMSD value in the final Reference Set is chosen to go through the side chain assembly phase. In this phase, fragments from the Dunbrack library are chosen and inserted into the solution. The method utilized is the same method utilized in the Improvement method of the Scatter Search algorithm in phase one, with 100 * protein size attempted moves. In this phase, the energy function includes the energy values produced by the side chain atoms and all-atom RMSD value is calculated.

IV. EXPERIMENTAL RESULTS

A. *Fragment-based SS and Mansour et al. Results*

In this section, we compare our results to the results generated by the pure ab initio results of Mansour et al. [19]. The generated structures are evaluated by computing the root mean square deviation expressed in Å.

Table 1 tabulates the minimum RMSD values generated by both algorithms for 1CRN, 1ROP and 1UTG proteins. Figures 1-3 display the tertiary structures of the three proteins in their native state (PDB) and generated by the two algorithms. Table I shows that the RMSD for 1CRN dropped from 9.01 Å to 8.05 Å, for 1ROP from 12.14 Å to 5.43 Å, and for 1UTG from 14.78 Å 12.34 Å. This shows that the approach utilized in this study significantly improves the three protein RMSD values. Furthermore, unlike [19], the structures generated, have no discontinuities in them.

TABLE1. FRAGMENT-BASED SS RMSD VALUES

Methodology Proteins	1CRN	1ROP	1UTG
Mansour et al. [19]	9.01 Å	12.14 Å	14.78 Å
Fragment based SS	8.05 Å	5.43 Å	12.34 Å

B. *Fragment-based SS, ROSETTA, and I-TASSER Results*

In these experiments, we compare the generated structures from our algorithm with those generated by I-TASSER and ROSETTA. Since I-TASSER and ROSETTA do not set the first amino acid coordinates of the structures to the coordinates of the corresponding PDB protein, after generating the structures from their servers we translated the coordinates to calculate the RMSDs and to visualize them.

As shown in Table II, the RMSD results generated by our code are the lowest for the three proteins. However, it seems that since RMSD is a global measure, a small disorientation in one part of a protein results in a large root mean square deviation increase. For all the three tested proteins, the visualized generated structures by I-TASEER and ROSETTA looked reasonable. Hence, their RMSDs should have been less. Figure 4 is a case in point. Consequently, the results of our fragment-based SS algorithm can be interpreted as comparable to those of I-TASSER and ROSETTA for these three proteins.

TABLE II. RMSD VALUES GENERATED BY FRAGMENT-BASED SS, I-TASSER AND ROSETTA

Method Proteins	1CRN	1ROP	1UTG
I-TASSER	12.14 Å	26.14 Å	19.94 Å
ROSETTA	11.35 Å	23.28 Å	18.20 Å
Fragment-based SS	8.05 Å	5.43 Å	12.34 Å

V. Conclusions

In this paper, an *ab initio* fragment-based protein structure prediction method is presented. This method is based on a scatter search metaheuristic. Given a protein sequence and its corresponding fragments, the algorithm first assembles the backbone of the candidate solutions then the side chains of the best generated solution. The RMSD values of the generated structures of three proteins show promising results that are comparable to those of well-recognized algorithms.

Major limitations of this work are presented by the inaccuracy of the energy function and the lack of accuracy of is the dihedral to Cartesian transformation method utilized that is not 100% accurate. Further future work can focus on including more terms in the energy functions. The CMAP term ignored for simplicity, should be added to the energy function. In addition, hydrogen atoms, can be added to the solution representation, thus adding the hydrogen bonding term to the energy function. This would require parallelizing the algorithm in order to speed up the processing and to explore areas of the search space.

Acknowledgment

This work was partially supported by the Lebanese American University and the National Council for Scientific Research.

References

[1] M. S. Abual-Rub and R. Abdullah, “A survey of protein fold recognition algorithms,” Journal of Computer Science, vol. 4, no. 9, pp. 768-776, 2008.

[2] L. Chen, G. Liu, Q. Wang, and W. Hou, “Homology modeling of the three-dimensional structure of bovine serum albumin” Proc. 3rd International Conference on Biomedical Engineering and Informatics, Yantai, Oct. 2009, pp. 2377-2381, 2009.

[3] L. Jaroszewski, Z. Li, X. H. Cai, C. Weber, and A. Godzik, “FFAS server: Novel features and applications,” Nucleic Acids Research, vol. 39, pp. 38-44, 2011.

[4] J. Kopp and T. Schwede, “Automated protein structure homology modeling: a progress report,” Pharmacogenomics, vol. 5, no. 4, pp. 405-416, 2004.

[5] C. A. Floudas, “Computational methods in protein structure prediction,” Biotechnology and Bioengineering, vol. 97, no. 2, pp. 207-213, 2007.

[6] P. Bradley, K. M. S. Misura, and D. Baker, “Toward high-resolution de novo structure prediction for small proteins,” Science, vol. 309, pp. 1868-1871, 2005.

[7] F. Liang and W. H. Wong, “Evolutionary Monte Carlo for protein folding simulations,” Journal of Chemical Physics, vol. 115, no. 7, pp. 3374-3381, 2001.

[8] C. B. Anfinsen, “Principles that govern the folding of protein chains,” Science, vol. 181, pp. 223-230, 1973.

[9] N. Mansour, F. Kanj, and H. Khachfe, “Particle swarm optimization approach for protein structure prediction in the 3D HP model,” Interdisciplinary Science: Computational Life Science, vol. 4, pp. 190–200, 2012.

[10] A. Liwo, J. Pillardy, C. Czaplewski, J. Lee, D. R. Ripoll, et al. “UNRES: A united-residue force field for energy-based prediction of protein structure-origin and significance of multibody terms,” Proc. 4th Int. Conf. on Computational Molecular Biology, Tokyo, Japan, April 2000, pp. 193-200.

[11] N. Mansour, C. Kehyayan, and H. Khachfe, “Scatter search algorithm for protein structure prediction,” Int. Journal of Bioinformatics Research and Applications, vol. 5, pp. 501–515, 2009.

[12] S. Schulze-Kremer, “Genetic algorithms and protein folding,” Methods in Molecular Biology, vol. 143, pp.175–222, 2000.

[13] A. Roy, J. Yang, and Y. Zhang, “COFACTOR: An accurate comparative algorithm for structure-based protein function annotation,” Nucleic Acids Research, vol. 40, pp. 471- 477, 2012. Doi:10.1093/nar/gks372.

[14] D. W. Buchan, S. M. Ward, A. E. Lobley, T. C. Nugent, K. Bryson, and D. T. Jones, “Protein annotation and modelling servers at University College London,” Nucleic Acids Research, vol. 38, pp. 563-568, 2010.

[15] D. T. Jones and L. TJ. McGuffin, “Assembling novel protein folds from super-secondary structural fragments,” Proteins, vol. 53(S6), pp. 480-485, 2003.

[16] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler, “ Practically useful: What the Rosetta protein modeling suite can do for you,” Biochemistry, vol. 49, no. 1, pp. 2987-2998, 2009.

[17] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, “Protein structure prediction using Rosetta,” Methods in Enzymology, vol. 383, pp. 66-93, 2004.

[18] J. Huang and A. D. MacKerell, “CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data,” J. Comput. Chem., vol. 34, pp. 2135–2145, 2013. DOI: 10.1002/jcc.23354.

[19] N. Mansour, I. Ghalayini, M. El-Sibai, and S. Rizk, “Evolutionary algorithm for predicting all-atom protein structure,” Proc. ISCA Third International Conference on Bioinformatics and Computational Biology, New Orleans, Louisiana, March 2011, pp. 7-12.

[20] B. R. Brooks, B.E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations,” Journal of Computational Chemistry, vol. 4, no. 2, pp. 187-217, 1983.

[21] F. Glover, “A template for scatter search and path relinking,” In J.K. Hao, E. Lutton, E. Ronald, M. Schoenauer, D. Snyers (Eds.), pp. 13-54, 1997, Springer-Verlag.

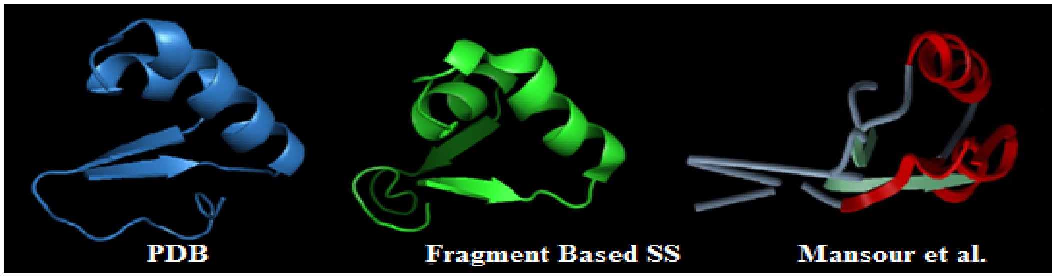

Figure 1. Structures generated by the two methods and the PDB structure for 1CRN.

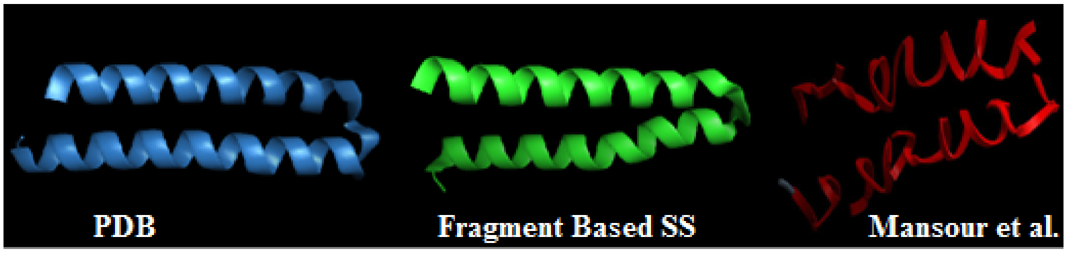

Figure 2. Structures generated by the two methods and the PDB structure for 1ROP.

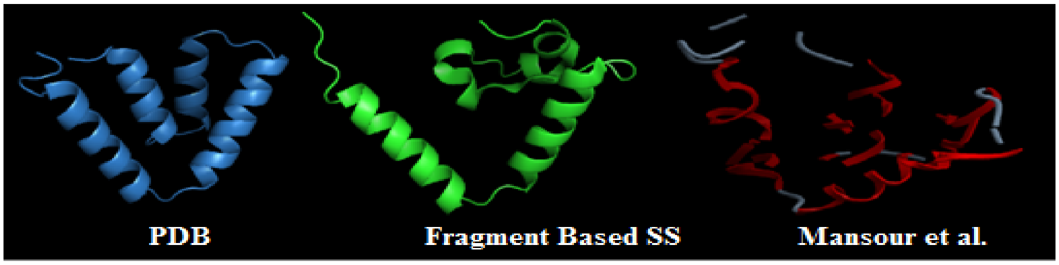

Figure 3. Structures generated by the two methods and the PDB structure for 1UTG.

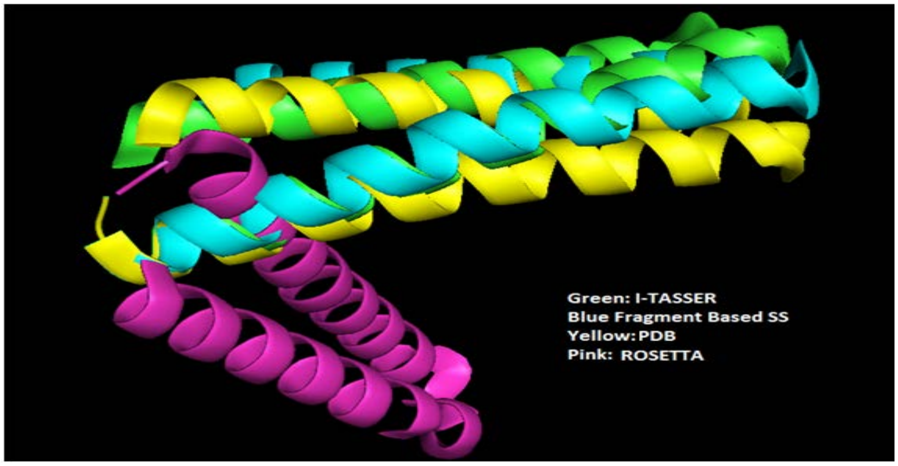

Figure 4. 1ROP Structures Generated.

ISBN: 978-1-61208-354-4