Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

A Fast Algorithm for Low-Resolution Protein Structure Prediction

Rajkumar Bondugula, *Student Member, IEEE*, Dong Xu, *Member, IEEE* and Yi Shang, *Member, IEEE*

Abstract— We propose a new approach for the protein tertiary structure prediction based on the concept of mini-threading. The method identifies useful fragments in Protein Data Bank (PDB) with variable lengths and retrieves spatial restraints. The multidimensional scaling method and least-squares minimization are used to build coarse-grain structural models. Our method uses the information in the PDB efficiently and the prediction time is in minutes when compared to hours and days required by existing methods.

I. Introduction

The 3-D structure of a protein holds key in understanding its function at the molecular level. The ability to predict the structure of a protein has a proven impact in the pharmaceutical industry by helping the researchers in the rational drug design [1]. In spite of research efforts for past three decades, the problem of accurately predicting the tertiary structure of a protein remains challenging [2] [3]. The study of the structure prediction algorithms is especially timely in this post-genomic era as the experimental methods cannot keep up with the large number of proteins from the genome sequencing projects, whose structures need to be characterized.

Existing protein structure prediction methods fall into three main categories. They are homology modeling [4] [5], *ab initio* structure prediction [6], and mini-threading [7-11]. Homology modeling relies on recognizing a native-like fold of a query protein in the Protein Data Bank (PDB) [12] of the experimentally determined structures. Homology modeling has limited performance if the query protein has no close homolog in the PDB. *Ab initio* methods predict the structure of a protein by optimizing the energy function that includes physical or statistical properties of the amino acids. This method of predicting the structure from the first principles does not produce accurate structure in general and requires long computing times. A new method called mini-threading, though still in its infancy has shown some promising results [13-15]. Mini-threading methods obtain matches between the query sequence and the short fragments of the proteins in the PDB for building local structures. These local structures are then assembled into a global structure through optimizing an objective function.

The current mini-threading process involves four steps [7]: 1) Search for compatible fragments of short sequences in the query protein against PDB. Typically, lengths of 9-amino acids are aligned using gapless alignment with the proteins in the database to search for the fragments. 2) The obtained fragments are used to build the Φ/Ψ angle distributions. The conformation of the protein backbone is defined by the Φ/Ψ angle on the protein backbone. 3) The 9-mer fragments are assembled in to global structures. The Φ/Ψ distributions are used as the soft constraints, together with other energy functions, for assembling the structures, typically using genetic algorithms 4) A large ensemble of possible structures is generated and these structures are grouped into clusters. The best cluster is selected based on the average energy function as the final prediction.

II. Methods and Materials

In this paper we present a novel approach that is similar to the concept of mini-threading but different from it in the following aspects. First, we do not restrict ourselves with the arbitrary choice of 9-mer fragments. We allow the fragments of various lengths. Second, traditional mini-threading methods treat all the retrieved fragments equally. We use different weights to different hits based on the statistical significance of the hit. Third, we use distances as restraints instead of Φ/Ψ angles. This enables us to use the information in the PDB more effectively. Finally, instead of using time-consuming optimization methods such as genetic algorithms and Monte Carlo simulations, we use simple and computationally efficient Classical Multi Dimensional Scaling (CMDS) [16] algorithm to generate the initial structure and use the fast Least Squares Minimization (LSM) [17] algorithm to perform local optimization. Our method predicts one possible tertiary structure in less than a minute on a desktop computer.

In the first step towards the tertiary structure prediction, we generate and retrieve the compatible protein fragments. In the second step, the structural fragments are transformed into in to distance restraints. In the third step, the distance restraints are used to generate the candidate tertiary structures of the protein using the CMDS method. In the fourth step, the structures are refined using LSM local optimization method. We employ a simplified representation of the protein by using only the C-alpha atom to represent each amino acid. During the development of the method, we compare the generated structure to the native (correct) structure. We generate 10,000 candidate structures and choose the configuration with the lowest cRMSD (square

Rajkumar Bondugula (e-mail: raj@mizzou.edu, Ph: +1 573 8825994, Fax +1 573 8828318) and Dong Xu are with Department of Computer Science and C.S. Bond Life Sciences Center, University of Missouri-Columbia, MO 65211, USA.

Yi Shang is with the Department of Department of Computer Science, University of Missouri-Columbia, MO 65211, USA.

This project was partially funded by the University of Missouri Research Board Grant

1-4244-0033-3/06/$20.00 ©2006 IEEE.

root of the mean squared deviation) of the predicted C-alpha atom of the predicted structure with the native structure) as the predicted structure.

A. Identification of Protein Fragments

The first task in the process of protein tertiary structure prediction is to identify the compatible fragments from which the structure of the query protein is predicted. In [18] we presented a protein secondary prediction using a nearest neighbor algorithm. The system first searches for protein fragments that are similar to the subsequences of the query protein. The information from these fragments is used to predict the secondary structure. In this paper, we show the results of the quantitative assessment of the information present in these fragments; specifically we look at average the Φ/Ψ angle deviation and the secondary structure similarity.

For this we prepare a database that contains the representative protein set (RPS). We used the July 2005 release of the PDBSelect [19] database that contains the protein chains such that the sequence identity between any two given proteins is at most 25%. The original database of 2810 chains was filtered to eliminate low-quality structures. Only the chains whose structures were determined using the X-Ray crystallography and are of resolution of less than or equal to 3 Å were selected. Further, if the more than 10% of a protein is composed of the non-standard amino acids, it is discarded too. Finally, the 1695 proteins that remain after the filtration make up our RPS. We randomly choose 200 proteins from our RPS to perform the analysis. Each of these 200 proteins is used to search for the compatible fragments against the RPS using the following procedure. In the first step, the Position Specific Scoring Matrix (PSSM) of the query protein is constructed using PSI-BLAST [20] with the *nr* database (the non-redundant sequence database at *http://www.ncbi.nlm.nih.gov*). The following parameters were used to construct the PSSM: *j*(number of iterations to construct the profile)=3, *e*(expectation value threshold) = 0.001, and *M*(scoring matrix)=BLOSUM90. The generated PSSM is then used to search for the compatible fragments in RPS by running the PSI-BLAST again. While running the PSI-BLAST second time, the value of the '*e'* was set to 10,000. We remove the query protein from the RPS during the search for compatible fragments.

The search with the 200 proteins resulted in 150,074 hits. We then analyzed these hits to examine, on average, by how much do the Φ/Ψ angles of the hit fragments vary from the query protein fragment. We also wanted to know how the secondary structure similarity varies with the E-Value (statistical significance value) [20] of the hit. The results are presented in Figure 1. From the plots in the Figure 1, we can conclude that if we have a hit with an associated E-value of 1e-4 or less, and if we simple copy that the secondary structure of the hit, on average, it is 80% accurate and if we copy the Φ angle, it is off by little more than 30°.

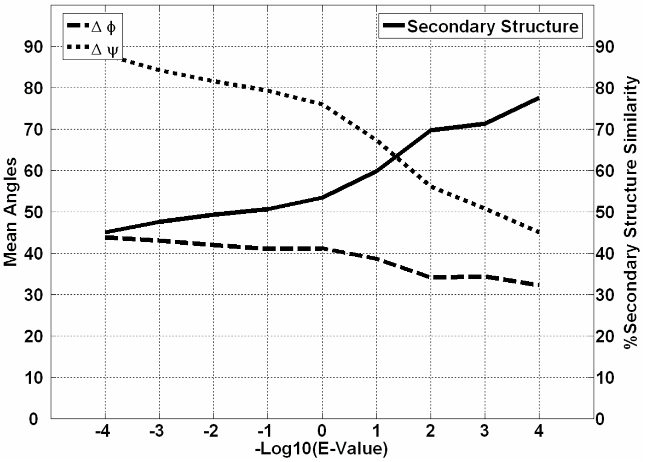

Figure 1: Information content in the hit regions using 200 proteins that resulted in 150,074 hits. The plots with the x-axis and the y-axis on the left side depict the relationship between the negative logarithm of the E-Value and mean absolute difference in the Φ/Ψ angles. The plot with the x-axis and the y-axis on the right side depicts the relationship between the negative logarithm of the E-Value and percentage of the secondary structure identity in the hit region.

B. Tertiary Structure Prediction

The protein tertiary structure prediction problem can be formulated as a graph realization problem. If there are *n* points (each representing the C-alpha atom of an amino acid) $x_j \in R^3$, $j=1,\ldots,n$, in a 3-D space. Suppose we know the Euclidean distances between some pairs of points (partial restraints), the task is to predict the distances of the remaining restraints from the partial distance matrix $D = (d_{ij})$. The restraint d_{ij} is the Euclidean distance between the nodes *i* and *j* on the aligned positions of the structure template derived from the alignment of the query protein sequence profile against the RPS. The realization problem can be formulated as the following error minimization problem:

$$\min_{x_1,\ldots x_n \in R^3} \sum_{(i,j)\in N} \left(\left\| x_i - x_j \right\| - d_{ij} \right)^2 \qquad (1)$$

For any pair of residues, there will be none to several restraints to choose from. In the query protein, for every possible pair of residues (i,j, $i\neq j$), we search for hits that have residues aligned with current residues *i* and *j*. To generate a population of structures for further clustering and selection, we choose one of the constraints d_{ij} with a probability that is proportional to the value of the score *S* associated with each hit. The score *S* is calculated as follows:

$$S = 4 - Log_{10}\left(E\ Value\right) \qquad (2)$$

Where, 4 is an empirical value based on training. Using this equation, the more similar between the two compared sequences, the smaller the E-value, and the larger the score *S*.

First, for the pairs for which the restraints are available they are filled in to D, the rest of the restraints are extrapolated using Floyd's shortest path algorithm [21]. Once the D matrix is completely filled, we use CMDS to construct a set of coordinates that satisfy the constraints in the distance matrix D. CMDS is a set of data analysis techniques that display the structure of the distance-like data as a geometrical picture [16]. Specifically, we choose the CMDS algorithm due to its simplicity and computational efficiency. In the CMDS method, the data is quantitative and the proximities of objects are treated as distances in the Euclidean space. The goal of the CMDS algorithm is to find a configuration of points in multidimensional space that satisfy the given restraints. If the distances matrices are accurate, the CMDS algorithm will be able to recreate the exact C-alpha coordinates that satisfy the restraints. Since, our problem is over-constrained (multiple restraints for a given pair of residues (i,j)), CMDS outputs the configuration of the points in more than three dimensions. We select only the first three dimensions of the constructed coordinates and treat them as the coordinates of the C-alpha atoms.

C. Refinement using the LMS algorithm

After the initial structure is generated using CMDS as described in the section II *B*, it is further refined using the non-linear LSM algorithm. LSM is effective at reducing the effect of the estimated distances (in our case, the distances estimated using the Floyd's algorithm) to improve the quality of the predicted structure. During the LSM, we used S (Equation 2) as the weight and the revised objective function is given in Equation 3.

$$\min_{x_1,\ldots x_n \in R^3} \sum_{(i,j)\in N} w_{ij} \left(\left\| x_i - x_j \right\| - d_{ij} \right)^2 \qquad (3)$$

Weighted objective functions are difficult to handle in the CMDS algorithm but are straight forward to when used with LSM. This procedure, i.e., using the CMDS for generating the initial structure using objective function and using the weighted objective function during LSM results in an implementation that takes less than a minute for a protein with a hundred residues on a desktop PC while most of the *ab initio* and mini-threading methods require a few hours to few days for structure optimization.

III. Results

We implemented the algorithms and the simulations in Matlab. The Statistical Toolbox of Matlab has an implementation of CMDS, called '*mdscale*' that supports metric CMDS. For the local optimization in refinement, we used the Levenberg-Marquardt method as implemented in the '*lsqnonlin*' function in the Optimization Toolbox for minimizing sum-of-squared-errors objective functions. We randomly picked 50 proteins from our RPS to test the algorithm; we present the prediction results of the four proteins with the lowest the cRMSD. The PDB codes of the selected proteins are: 1rb9 (an iron sulfur protein), 1vf6 chain D (a binding/transport protein), 1th7 chain A (an archaeal Sm protein) and 2bh1 chain y (a secretion pathway complex). The prediction results are given in Table 1.

Table 1: Protein tertiary structure prediction results on the four proteins with lowest cRMSD

Protein Name	cRMSD (Å)
1rb9	5.4
1vf6 chain D	6.9
1th7 chain A	6.2
2bh1 chain Y	6.6

cRMSD: square root of the mean squared deviation of c-alpha coordinates of the predicted structure to the native structure.

Another way of assessing the quality of the predicted structure is through the Hubbard plot. In Hubbard plot, segments of window size 'W' of the predicted structure are matched with the native structure to compute the minimum and average cRMSD. Usually, the starting with W=5, the cRMSD is calculated at uniform increments, until the value of W is equal to the length of the protein. The results are presented in Figure 2.

IV. Discussion

Our study provides an effective and efficient method to generate low-resolution protein structural models, which can be used as input for structure refinement. While useful, our method requires substantial further developments. One of the drawbacks of our system is that we consider the restraints at the amino acid level and not at the secondary structure segment level. Considering the restraints at the amino acid level poses at least one obvious disadvantage of inconsistent hits. For example, for a short length of query protein, few hits might be in the helix configuration, while few others might be in the sheet configuration. Since the restraints are choose one amino acid at a time based on the probability that is proportional to the E-value of the hit, there is a possibility that the for position i, the restraint(s) might come from a fragment in helix configuration and restraint(s) for position i+1 might come from a fragment that is in the sheet configuration, resulting in a unrealistic tertiary structure at the secondary structure level. Considering the restraints at the secondary structure segment level will alleviate from this problem. We plan to use the secondary structure prediction system we proposed in [18] to first predict the secondary structure of the query protein and then use the predicted segments as the basis to select the restraints.

Work is underway to explore various objective functions and optimization schemes to further refine the predicted structure, both at the local and global levels. We are also working on various possibilities to automatically pick the set of possible structures that are likely to be close to actual structure of the query protein.

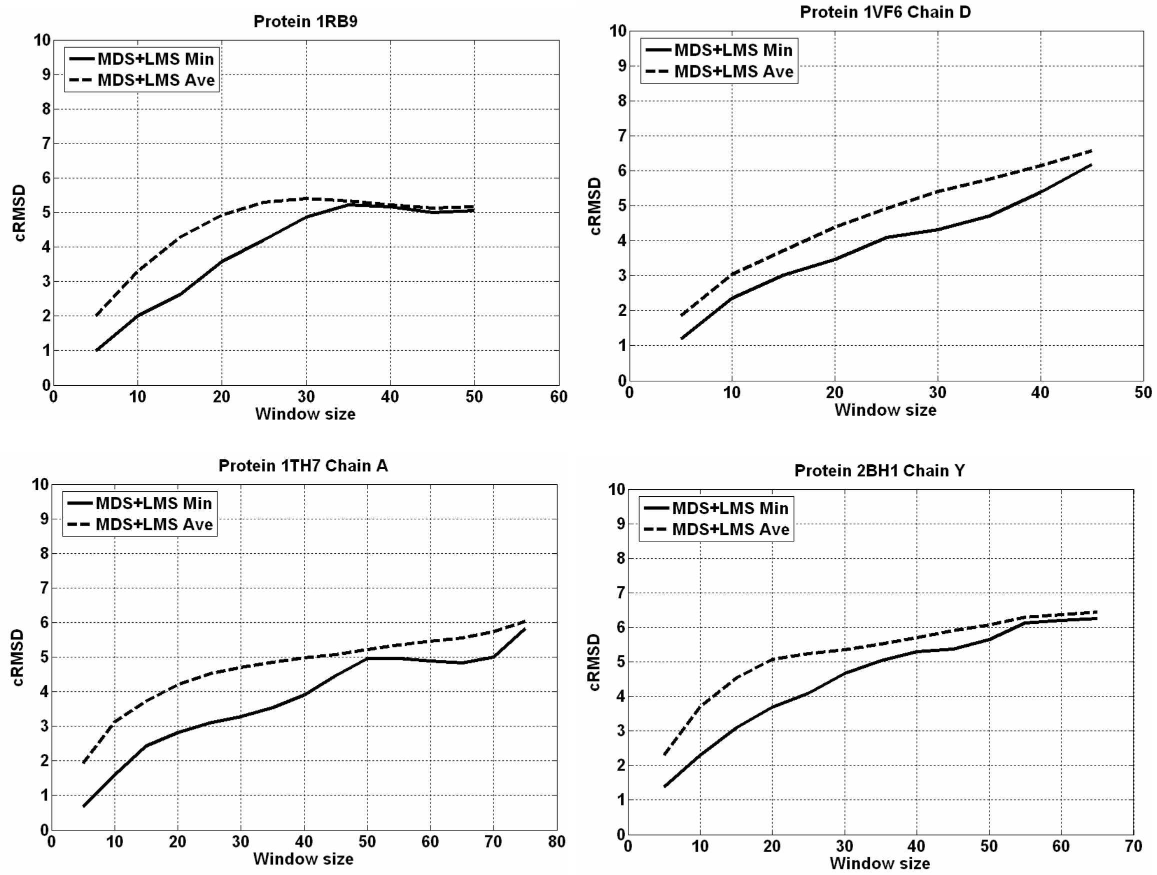

Figure 2: Average and minimum cRMSD vs. the window size of the protein for the predicted structure using the CMDS followed by the LSM on the four test proteins. Segments of the window sizes 5 and up of the predicted structures are matched with the native structure to compute the minimum and average cRMSD.

REFERENCE

[1] Kitchen DB, Decornez H, Furr JR, Bajorath J (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 3(11):935-949.

[2] Ginalski K, Grishin NV, Godzik A, Rychlewski L (2005). Practical lessons from protein structure prediction. Nucleic Acids Res. 33(6):1874-91.

[3] Moult J (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.

[4] Bowie JU, Luthy R, Eisenberg D (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science. 253:164-170.

[5] Ring CS, Cohen FE (1993). Modeling protein structures: construction and their applications. FASEB J. 7(9):783-790.

[6] Li Z, Scheraga HA (1987). Monte carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA. 84:6611-6615.

[7] Simons KT, Kooperberg C, Huang E, Baker D (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268(1):209-225.

[8] Bystroff C, Baker D (1999). Prediction of local structure in proteins using a library of sequence-structure motifs. J. Mol. Biol. 281:565-577.

[9] Inbar Y, Benyamini H, Nussinov R, Wolfson HJ (2003). Protein structure prediction via combinatorial assembly of sub-structural units. Bioinformatics. 19 Suppl 1:i158-i168.

[10] Chikenji G, Fujitsuka Y, Takada S (2003). A reversible fragment assembly method for de novo protein structure prediction. Journal of Chemical Physics. 119:6895-6903.

[11] Lee J, Kim SY, Joo K, Kim I, Lee J (2004). Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins. 1;56(4):704-714.

[12] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. Nucleic Acids Res. 28:235-242.

[13] Venclovas C, Zemla A, Fidelis K, Moult J (2003). Assessment of progress over the CASP experiments. Proteins. 53 Suppl 6:585-595.

[14] Li X, Jacobson MP, Friesner RA (2004). High-resolution prediction of protein helix positions and orientations. Proteins. 55(2):368-82.

[15] Bradley P, Misura KM, Baker D (2005). Toward high-resolution de novo structure prediction for small proteins. Science. 309(5742):1868-71.

[16] Borg I, Groenen P (1997). Modern Multidimensional Scaling, Theory and Applications. Springer-Verlag, New York.

[17] Luenburger DG (1984). Linear and Nonlinear Programming: Addison-Wesley Publishing Company, Reading, Mass

[18] Xu D, Bondugula R, Popescu M, Keller J (2006). Fuzzy Logic and Bioinformatics, Proceedings of the Fuzz-IEEE 2006 (IEEE International Conference on Fuzzy Systems), Vancouver, BC, Canada, July 2006.

[19] Hobohm U, Sander C (1994). Enlarged Representative set of protein structures, Protein Science 3: 522.

[20] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-3402.

[21] Floyd RW (1962). Algorithm 97: Shortest Path. Communications of the ACM 5 (6): 345.