International Journal of Bio-Science and Bio-Technology
Vol.6, No.3 (2014), pp.49-54
http://dx.doi.org/10.14257/ijbsbt.2014.6.3.06

Comparison of Two Academic Software Packages For Protein Structure Prediction

Bayumy.A.Youssef[1], Shaheera Rashwan[1] and ElRashdy M. Redwan[2]

[1]Informatics Research Institute,City for Scientific Research and Technological Applications, Borg ElArab, Alexandria, Egypt
[2]Genetic Engineering and BiotechnologyResearch Institute, City for Scientific Research and Technological Applications, Borg ElArab, Alexandria, Egypt
bbayumy@yahoo.com, rashwan.shaheera@gmail.com,redwan1961@yahoo.com

Abstract

Protein structure prediction has matured over the past few years to the point that even fully automated methods can provide reasonably accurate three-dimensional models of protein structures. However, until now it has not been possible to develop programs able to perform as well as human experts, who are still capable of systematically producing better models than automated servers. In this paper, we review and compare two recently developed and publicly available software packages, LOMETS and I-TASSER, for predicting protein structure in comparison with the commercial software package (Hyper Chem Release 8.0). These two software packages share some common features and also have some fundamental difference. Based on our experience, I-TASSER is more accurate in predicting protein structure. In contrast, LOMETS shows less performance than I-TASSER.

Keywords: *Protein structure prediction, TM-score, Amino-acid sequence*

1. Introduction

While it has been known for over 40 years that the three dimensional structures of proteins are determined by their amino acid sequences, protein structure prediction remains a largely unsolved problem for all but the smallest protein domains. The state-of-the-art Rosetta [3] structure prediction methodology, for example, is limited primarily by conformational sampling; the native structure almost always has lower energy than any non-native conformation, but the free energy landscape that must be searched is extremely large—even small proteins have on the order of 1000 degrees of freedom—and rugged due to unfavorable atom-atom repulsion which can dominate the energy even quite close to the native state. To search this landscape, Rosetta uses a combination of stochastic and deterministic algorithms: rebuilding all or a portion of the chain from fragments, random perturbation to a subset of the backbone torsion angles, combinatorial optimization of protein side chain conformations, gradient based energy minimization, and energy-dependent acceptance or rejection of structure changes [1].

The most successful general approach for predicting the structure of proteins involves the detection of homo logs of known three-dimensional (3D) structure—the so-called template-based homology modeling or fold-recognition. These methods rely on the observation that the number of folds in nature appears to be limited and that many different remotely homologous protein sequences adopt remarkably similar structures. Thus, given a protein sequence of interest, one may compare this sequence with the sequences of proteins with experimentally determined structures. If a homolog can be found, an alignment of the two sequences can be

ISSN: 2233-7849 IJBSBT
Copyright ⓒ 2014 SERSC

generated and used directly to build a 3D model of the sequence of interest. The practical applications of protein structure prediction are many and varied, including guiding the development of functional hypotheses about hypothetical proteins, improving phasing signals in crystallography, selecting sites for mutagenesis and the rational design of drugs [2].

The lack of comparisons between algorithms and software packages in this research area, which is the protein structure prediction, causes the difficulty and confusion. In this paper, we compare the performance of the two software packages in the aspects of predicting structure of protein and statistical analysis. The two academic software packages, LOMETS and I-TASSER, for predicting protein structures, are recently developed and publicly available [14, 15].

This paper is organized as follows: section 1 presents the introduction. Section 2 summarizes a background. Section 3 introduces the software for comparison. Section 4 shows the software results of the comparison. Section 5 concludes and discusses the software results of the comparison. Finally a list of references is given.

2. Background

Protein structure prediction is usually divided into three categories: ab initio (or de nono) prediction, fold recognition (or threading) and homology modelling, based on to which extent the homology information in sequence and structure databases has been used to construct the structural model.

Bonneau and Baker [5] reviewed the features of recent ab initio protocols in an attempt to highlight the foundations of recent progress in the field and suggest promising directions for future work. In [3], Rohl *et al.*, showed the Rosetta method where short fragments of known proteins are assembled by a Monte Carlo strategy to yield native-like protein conformations.

In [4], Fourrier *et al*., defined a structural alphabet composed of 16 average protein fragments, which they called Protein Blocks (PBs). Those PBs allow an accurate description of every region of 3D protein backbones and have been used in local structure prediction. They use this structural alphabet to analyze and predict the loops connecting two repetitive structures.

Sitao Wu and Yang Zhang [10] developed LOMETS, a local threading meta-server, for quick and automated predictions of protein tertiary structures and spatial constraints. Nine state-of-the-art threading programs are installed and run in a local computer cluster, which ensure the quick generation of initial threading alignments compared with traditional remote server- based meta-servers.

Moult *et al.*, [9] described the conduct of the experiment, the categories of prediction included, and outlines the evaluation and assessment procedures. Highlights are the first blind assessment of model refinement methods showing that under some circumstances substantial model improvements are possible; improvements in the performance of methods for determining the accuracy of a model; and some progress in the accuracy of comparative models in regions not present in a principal template.

Zhang, Y. [11] developed I-TASSER, a protein structure modelling approach based on the secondary-structure enhanced Profile-Profile threading Alignment (PPA) and the iterative implementation of the Threading ASSEmbly Refinement (TASSER) program.

Cooper *et. al.* [1] described Foldit, a multiplayer online game that engages non-scientists in solving hard prediction problems. Foldit players interact with protein structures using direct manipulation tools and user-friendly versions of algorithms from the Rosetta structure prediction methodology, while they compete and collaborate to optimize computed energy.

3. Software for Comparison

3.1. First Academic Software Package: LOMETS

LOMETS server, developed by Sitao *et al.*, [10], takes predictions from nine different servers that represent a diverse set of state-of-the-art threading algorithms, *i.e.*, FUGUE, HHSEARCH, PROSPECT2, SAM-T02, SPARKS2, SP3, PAINT, PPA-I and PPA-II. Models in LOMETS are selected from individual servers purely based on consensus, i.e. the structure similarity of the considered model with other threading alignments.

For the best performance, 30 models are taken from the top predictions of the nine servers sequentially, where the order of the servers are based on their performance on independent test runs. For each protein, threading models are categorized as ‘good’ or ‘bad’ depending on whether the inherent Z-score (the energy in standard deviation units relative to mean) of the alignment is above or below a threshold Z-scorecut. Regarding this, we choose the protein structure model with the highest Z-score as being the best predicted model.

3.2. Second Academic Software Package: I-TASSER

I-TASSER, developed by Zhang in [11], is a hierarchical protein structure modelling approach based on the secondary-structure enhanced Profile-Profile threading Alignment (PPA) [10] and the iterative implementation of the Threading ASSEmbly Refinement (TASSER) program.

The target sequences are first threaded through a representative PDB structure library (with a pair-wise sequence identity cut-off of 70%) to search for the possible folds by four simple variants of PPA methods. The continuous fragments are then excised from the threading aligned regions which are used to reassemble full-length models while the threading unaligned regions (mainly loops) are built by ab initio modeling. The cluster centroids are obtained by the averaging the coordinates of all clustered structures.

To rule out the steric clashes on the centroid structures and to refine the models further, we implement the fragment assembly simulation again, which starts from the cluster centroid of the first round simulation. Spatial restraints are extracted from the centroids and the PDBstructures searched by the structure alignment program TM-align, which are used to guide the second round simulation. Finally, the structure decoys are clustered and the lowest energy structure in each cluster is selected, which has the Cα atoms and the side-chain centers of mass specified.

4. Software Results of Comparison

4.1. Data Used for Comparison

In our work, we use four amino acids sequences of proteins (input as FASTA format) extracted from camel milk in order to design a drug for Hepatitis C virus. For more information about data used, see [17].

4.2. Software Results and Performance Evaluation

The Template Modeling Score or TM-score is defined to assess the topological similarity of two protein structures. The TM-score is intended as a more accurate measure of the quality of full-length protein structures. The equation [12, 13] is as follows:

$$\text{TM - score} = \frac{1}{L}\left[\sum_{i=1}^{L_{ali}} \frac{1}{1 + d_i^2 / d_0^2}\right]_{\max} \quad (1)$$

where L is the length of the target protein, and L_{ali} is the number of the equivalent residues in two proteins. d_i is the distance of the i-th pair of the equivalent residues between the two structures, which depends on the superposition matrix; the 'max' means the procedure to identify the optimal superposition matrix that maximizes the sum in equation (1). The scale $d_0 = \sqrt[3]{L - 15} - 1.8$ is defined to normalize the TM-score in a way that the magnitude of the average TM-score for random protein pairs is independent on the size of the proteins. TM-score stays in (0, 1] with a higher value indicating a stronger similarity, where 1 indicates a perfect match between two structures. The TM-score is designed to be independent of protein lengths.

In our work, we used the TM-score to measure the similarity between the protein structure produced by the commercial software package (HyperChem 8.0) and those produced by the two academic software packages (LOMETS and I-TASSER). Scores below 0.17 corresponds to randomly chosen unrelated proteins whereas structures with a score higher than 0.5 assume that the protein structures are quite typical and show that the LOMETS or I-TASSER succeeded in retrieving and predicting the protein structure.

Table 1. The TM-Scores for LOMETS and I-TASSER for the Four Protein Sequences

Protein Sequence	TM-score: LOMETS	TM-score: I-TASSER
1	0.1297	0.1309
2	0.116	0.12
3	0.1815	0.19
4	0.1422	0.1419
Average	0.14235	0.1457

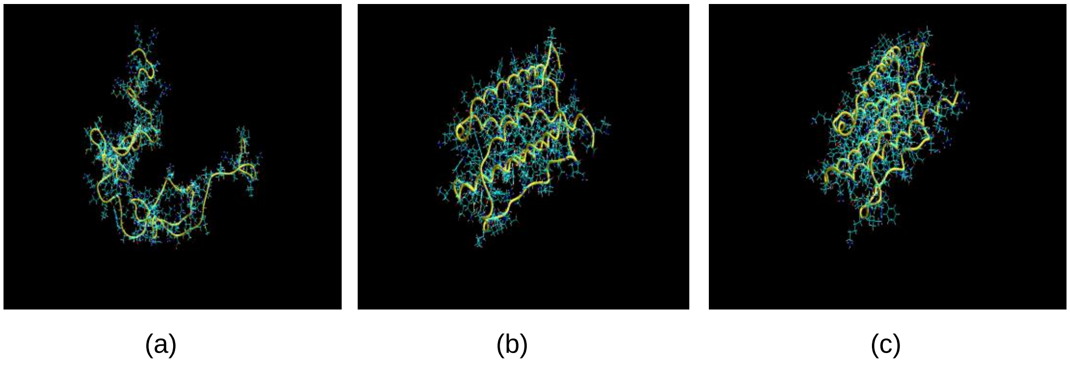
(a) (b) (c)

Figure 1. The Protein Structure Prediction Model of the First Protein Sequence by (a) HyperChem v. 8.0, (b) LOMETS and (c) I-TASSER

5. Conclusions and Discussion

We can see from Table 1 that the software package (I-TASSER) was more accurate in three models over four models and the average TM-score is higher than that of LOMETS. We acknowledged that our limited experience with the I-TASSER and LOMETS software may not fully reveal their performance in practice. Nevertheless, both free software packages benefit the biological research significantly and maximize the impact of protein sequences to a large extent. The contributions to the proteomic research community are of major significance. As future work, we can perform the comparison between more than two software packages, also with sequences of known proteins models.

References

[1] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee and M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, "Predicting Protein Structures with a Multiplayer Online Game", Nature, vol. 466, no. 7307, **(2010)**, pp. 756-760.

[2] L. A. Kelley and M. J. E. Sternberg, "Protein Structure Prediction on the Web: A Case Study Using the Phyre Server", Nature Protocols, vol. 4, no. 3, **(2009)**, pp.363-370.

[3] C. Rohl, C. Strauss, K. Misura and D. Baker, "Protein Structure Prediction Using Rosetta", Methods in Enzymology, vol. 383, **(2004)**, p. 66-93.

[4] L. Fourrier, C. Benros and A.G. de Brevern, "Use of a Structural Alphabet for Analysis of Short Loops Connecting Repetitive Structures", BMC Bioinformatics, vol. 5, no.58, **(2004)**.

[5] R. Bonneau and D Baker. "Ab initio Protein Structure Prediction: Progress and Prospects", Annual Review of Biophysics and Biomolecular Structure, vol. 30, **(2001)**, pp. 173-189.

[6] A.T.R. Laurie and R.M. Jackson, "Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening", Current Protein & Peptide Science, vol. 7, **(2006)**, pp. 395-406.

[7] R. Jauch, H.C. Yeo, P.R. Kolatkar, N.D. Clarke, "Assessment of CASP7 Structure Predictions for Template free Targets", Proteins, vol. 69, no. 8, **(2000)**, pp.57-67.

[8] Y. Zhang, "Progress and Challenges in Protein Structure Prediction", Current Opinion in Structural Biology, vol. 18, **(2008)**, pp. 342-348.

[9] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost and A. Tramontano, "Critical Assessment of Methods of Protein Structure Prediction -Round VIII", Proteins, vol. 77, no. 9, **(2009)**, pp. 1-4.

[10] W. Sitao and Y. Zhang, "LOMETS: A Local Meta-threading-server for Protein Structure Prediction" Nucleic Acids Research, vol. 35, no. 10, **(2007)**, pp. 3375–3382.

[11] Y. Zhang, "I-TASSER: Fully Automated Protein Structure Prediction in CASP8" Proteins, vol. 77, no. 9, **(2009)**, pp. 100–113.

[12] Zhang and Skolnick, "Scoring Function for Automated Assessment of Protein Structure Template Quality", PROTEINS: Structure, Function, and Bioinformatics, vol. 57, (2004), pp.702–710.

[13] J. Xu and Y. Zhang, "How Significant is a Protein Structure Similarity with TM-Score = 0.5?", Bioinformatics, vol. 26, no.7, **(2010)**, pp. 889-895.

[14] LOMETS Meta Server Based Protein Fold Recognitions, Zhang Lab, University of Michigan. (http://zhanglab.ccmb.med.umich.edu/LOMETS/)

[15] I-TASSER ONLINE Protein Structure & Function Predictions, Zhang Lab, University of Michigan. (http://zhanglab.ccmb.med.umich.edu/I-TASSER/)

[16] TM-score: A Quantitative Assesment of Protein Structure Similarity,Zhang Lab, University of Michigan. (http://zhanglab.ccmb.med.umich.edu/TM-score/)

[17] O. Almahdy, E.M. EL-Fakharany, E. EL-Dabaa, Ng TB, E.M. Redwan, "Examination of the Activity of Camel Milk Casein Against Hepatitis C Virus (Genotype-4a) and Its Apoptotic Potential in Hepatoma and HeLa Cell Lines", Hepat Mon., vol.11, no. 9, **(2011)**, pp. 724-730.

Copyright ⓒ 2014 SERSC