The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 324–331

Assessing Distance Measures for Protein Structure Comparison*

Zikai Wu[1,4,†] Yong Wang[2,3] Enmin Feng[1]
Luonan Chen[3,4]

[1] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
[2] Institute of Applied Mathematics, Academy of Mathematics and Systems Science, CAS, Beijing 100080, China
[3] Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka 574-8530, Japan
[4] Institute of Systems Biology, Shanghai University, Shanghai 200444, China

Abstract Distance measure is a key component for fast protein structure comparison. The identical representation of protein structure combining with different distance measures results in different structure comparison methods. In this paper, we provide and further analyze three structure comparison methods with contact vector representation based on three different distance measures respectively, i.e. Euclidean distance, cross entropy and FDOD function, in an empirical manner. Relying on a public data set, we evaluate the ability for detecting function similarity of the three methods. The comparison results reveal that two information-based measures outperform the Euclidean distance, in particular the method based on FDOD function is the best in the three distance measures. We show that the information-based measure can really identify the subtle difference given abstract representations of protein structures. Moreover, the FDOD for protein comparison is also superior to cross entropy in terms of mathematical consistency.

Keywords Protein structure comparison; contact vector; euclidean distance; cross entropy; FDOD function; function prediction.

1 Introduction

Proteins are macromolecules that regulate all biological processes in a living organism and their structures are generally better conserved than sequences. Thus, identifying similarity of structures by comparing proteins could yield valuable clues to their functions, and can be employed for fold family classification, motif finding, phylogenetic tree reconstruction and even protein docking[1].

So far, a number of automatic protein structure comparison methods have been proposed [1-17]. Generally, these works can be roughly classified into two categories, namely, structural alignment methods and alignment-free methods. The traditional alignment methods mainly focus on finding the optimal rigid-body superposition of

*Zikai Wu and Yong Wang equally contributed to this paper
†Email:wuzikai1981@163.com

two structures such that the root mean square deviation (RMSD) between the aligned atoms is minimized [2]. They all use the element-based representation for a structure, such as atoms, residues, and secondary structure elements (SSEs), and adopt the RMSD scoring scheme to measure the similarity [2,5]. Generally, these methods require intensive computation and are too slow to scan large databases.

In contrast, recently alignment-free methods have been developed by adopting a different strategy to speed up the computation[2-6,12-17]. They all accord with such a framework that the relevant features are extracted and represented in structure descriptions, and the equivalence is obtained by the specific distance measure [2,4,5,7,12]. In other words, alignment-free methods are mainly composed of two key components: the abstract representation of protein structure and a distance measure used to compare different structures' representations. For example, in GSM method, a protein structure is represented as a 30 dimensional vector, then Euclidean distance is used to measure the difference between two 30 dimensional vectors [14,15]. In contact metric method, a contact vector is used to represent a protein structure and the difference between two vectors is measured by Euclidean distance.

Besides the protein structure representation, the choice of distance measures is also very important for developing a high-accuracy comparison method. It is obvious that the identical representation of protein structure combining with different distance measures generally results in different structure comparison methods. A distance measure may be as crucial as the choice of the representation itself. Despite the important role in structure comparison method, the performance of the distance measures has not been systematically explored yet, and these measures are currently used on an ad-hoc basis.

In this paper we provide and further analyze three structure comparison methods based on distance measures, in order to gain insight or hints on structure comparison problems. Since protein structure can be abstracted as a vector, frequency distribution and so on in the existing alignment-free methods, Euclidean distance, cross entropy and FDOD function are commonly used to evaluate the difference between two vector representations or frequency distribution representations. Thus we will focus on these three distance measures to assess their ability in measuring the difference between two vectors or frequency distributions.

Specifically, a procedure is designed in this paper to compare Euclidean distance, cross entropy and FDOD function. Firstly, contact vector [17] is applied to represent protein structure, then it is combined with each of the three distance measures to form three structure comparison methods. Finally, relying on a public data set, these three structure comparison methods are compared in terms of the ability detecting functional similarity.

The remainder is organized as follows. In Section 2, we give the details of contact vector representation and the three distance measures respectively. Then, in Section 3 we present functional prediction test on public data sets. The Section 4 discusses the results with several general remarks. Finally, Section 5 shows some

directions needed to be explored in future.

2 Methods

2.1 Contact vector representation of a protein structure

On the process of protein folding, the amino acids along the polypeptide chain interact with each other in a cooperative manner to form a stable native structure. Some residues that are spatially neighboring can contact each other. These contact patterns can reflect the overall fold topology. The contact vector representation of protein structures just expresses the contact pattern, where each of its components records the number of residue-residue contacts as a function of their separation along the sequence.

In a given protein, every pair of residues with C_α backbone atoms closer than 9 is recorded in a contact matrix. Summing up diagonally all the contacts of the contact matrix among residues i and j such that $i-j=k$ with $k \geq 2$, leads to a histogram that enumerates all structural contacts among residues that are k positions apart in the sequence[17]. Typically many contact lengths k are short ($k = 2,3,4$), and are consistent with the local secondary structure constraints of helices and turns[17]. But other contact lengths are longer, almost equal to the length of the chain, and carries important information of the entire fold[17]. Thus the contact vector representation of a tertiary structure is $(q_2, q_3, \cdots q_T)$, where q_k is the absolute number of contact lengths $2 \leq k \leq T$. The cut-off T is often set to $T = 399$ which reflects that the two amino acids seldom interact if they have a gap larger than 399 in sequence position.

2.2 Three distance measures

2.2.1 Euclidean distance

Given two proteins A and B, their structures are represented as

$$Q_A = (q_2^A, q_3^A, \cdots q_{399}^A)$$

and

$$Q_B = (q_2^B, q_3^B, \cdots q_{399}^B)$$

respectively. Then the difference between them can be measured by Euclidean distance as:

$$Ed(Q_A, Q_B) = \sum_{k=2}^{399} |q_k^A - q_k^B| \tag{1}$$

However, different lengths in protein chains bias (raise) the contact metric distances of longer chains. It is known that $dl(X,Y) < c(LX+LY)$, where L_A and L_B denote A and B's chain lengths respectively. This can be corrected by normalizing the contact metric with the factor $1/[c(LX+LY)]$ to yield the length-corrected contact metric. The length-corrected contact metric can then be written by the simple

formula[17]:

$$LCM(Q_A, Q_B) = \frac{\sum_{k=2}^{399} |q_k^A - q_k^B|}{\sum_{k=2}^{399} (q_k^A + q_k^B)} \tag{2}$$

2.2.2 Cross Entropy

Cross entropy is a tool based on information theory to measure the discrepancy between two distributions [18]. It has been successfully applied to phylogenetic tree reconstruction and so on[19].

Cross entropy is used to compare two distributions, so the contact vector representation should be transformed into frequency distribution. Finally, the difference between A and B can be measured by cross entropy as:

$$CVS(Q_A, Q_B) = \sum_{k=2}^{399} \frac{q_k^A}{\sum_{i=2}^{399} q_i^A} \log \frac{\frac{q_k^A}{\sum_{i=2}^{399} q_i^A}}{\frac{q_k^B}{\sum_{i=2}^{399} q_i^B}} + \sum_{k=2}^{399} \frac{q_k^B}{\sum_{i=2}^{399} q_i^B} \log \frac{\frac{q_k^B}{\sum_{i=2}^{399} q_i^B}}{\frac{q_k^A}{\sum_{i=2}^{399} q_i^A}} \tag{3}$$

2.2.3 FDOD Function

Function of Degree of Disagreement (FDOD) is another measure based on information theory for information discrepancy [20] and it has a close connection with Shannon entropy. Comparing with cross entropy, it has many good mathematical characteristics, such as symmetry, boundedness, triangle inequality, and so on. FDOD has been successfully applied to the study of phylogeny, multiple sequence alignment, discrimination of homodimeric proteins and protein structure comparison and so on [21-24].

FDOD function is also a tool to compare two distributions, provided that the contact vector representation is transformed into frequency distribution. Finally, the difference between A and B can be measured by FDOD function as:

$$\begin{aligned} FDODVS(Q_A, Q_B) = \sum_{k=2}^{399} \Bigg(& \frac{q_k^A}{\sum_{i=2}^{399} q_i^A} \log \frac{2 \frac{q_k^A}{\sum_{i=2}^{399} q_i^A}}{\frac{q_k^A}{\sum_{i=2}^{399} q_i^A} + \frac{q_k^B}{\sum_{i=2}^{399} q_i^B}} \\ & + \frac{q_k^B}{\sum_{i=2}^{399} q_i^B} \log \frac{2 \frac{q_k^B}{\sum_{i=2}^{399} q_i^B}}{\frac{q_k^A}{\sum_{i=2}^{399} q_i^A} + \frac{q_k^B}{\sum_{i=2}^{399} q_i^B}} \Bigg) \end{aligned} \tag{4}$$

3 Experiment design and Results

The characterization of biological function among newly determined protein structures is a central challenge in structural genomics. One class of computational solutions to this problem is based on the similarity of protein structure. At present, rapid detection of similarity in protein function through protein structure comparison has become one of the most important applications. To test whether these three protein comparison methods formulated above can detect functional similarity, as defined by Gene Ontology(GO), they are compared in large-scale computational experiment.

3.1 Dataset

To make the comparison objectively, we employ the data set used originally by contact metric [17]. The data set is composed of 1662 non-redundant protein structures with $< 25\%$ mutual sequence identity from PDBselect25, March 2006, that also had at least one available GO annotation term recorded (version GOA 28.0). The 1662 proteins's GO terms represent 261 molecular functions, 216 biological processes and 75 cellular components. It is widely accepted that two proteins with common GO term can be viewed as functionally similar.

3.2 Function predictor

Next we focus on the assessment of the three methods on the ability of function similarity detection. The Prediction procedure is designed as follows [17]:

Step 1. Computation of discrepancy: Each pair of structures is compared using the three methods respectively.

Step 2. Ranking: These $\frac{1662\times1661}{2} = 1380291$ scores are sorted ascendingly.

Step 3. Performance assessment: We define two criteria to evaluate performance: *sensitivity* and *specificity*. They are defined as

$$sensitivity(S) = \frac{N_{tp}(S)}{N_t}$$

$$specificity(S) = \frac{N_{tn}(S)}{1380291 - N_t},$$

where $N_{tp}(S)$ denotes the number of protein pairs whose discrepancy score is $\leq S$ with a common GO term, N_t denotes the number of protein pairs with a common GO term. $N_{tn}(S)$ denotes the number of protein pairs whose discrepancy score is $> S$ without a common GO term.

It is obvious that *sensitivity* and *specificity* are all functions dependent on S. Thus, many pairs of *sensitivity* and *specificity* can be obtained. At present, the test is limited to specificities of no less than 90%. Finally, the Receiver Operating Characteristic (ROC) curve that combines *sensitivity* and *specificity* is applied to assess the three structure comparison methods in terms of the ability of function similarity detection. The resulting ROC graphs are shown in Fig 1.

4 Discussion and Conclusion

From Fig.1, we can see that the three methods are all have satisfactory functional similarity detection ability. These three methods are all based on contact vector representation, so their good performance in detecting function similarity demonstrates that contact vector representation is an outstanding abstraction of protein structure. It also hints that contact vector may be a intrinsical attribute.

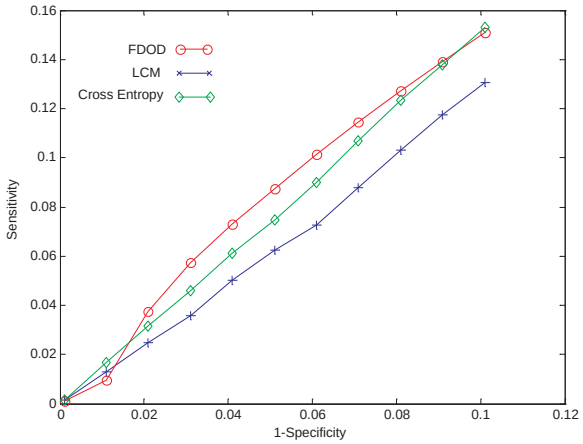

Figure 1: Gene Ontology standardized ROC curves for 1.38 million pairs of PDB chains with$<$ 25% sequence identity. The FDOD function based method is most sensitive among all methods tested.

From Fig.1, we also can see that the method based on FDOD function has the best performance, while the method based on Euclidean distance has the worst performance. This fact shows that FDOD function is more suited for measuring the difference between contact vectors than Cross Entropy and Euclidean distance to some extent. The two information-based measures outperform Euclidean distance, thereby indicating that they can detect the difference of distribution in a more subtle way. The improvement FDOD over Cross Entropy reveals that the specific mathematical properties such as non-negative, symmetric, continuous, identical and recursive conditions lead to a more reasonable measurement on protein similarity. As a conclusion we can apply the FDOD function to improve contact metric method.

The difference in the performance of function similarity detection also confirms that the identical representation of protein structure combining with different distance measures results in different structure comparison methods. In view of the results, further attention should be paid to the selection of a proper distance measure for comparing abstract representations of protein structures.

5 Future works

Nowadays, there are many abstract representations for protein structures, which can be adopted for protein structure comparison by combining with a specific similarity measurement or criterion. In order to evaluate the three distance measures more objectively and extensively, our future work will focus on applying distance measures to a variety of representations, and further conducting extensive application experiments. Besides, Kolmogorov-Smirnov test and contingency table analysis also

have been applied to measure some representations of protein structure. The evaluation of Euclidean distance, cross entropy, FDOD function, Kolmogorov-Smirnov test and contingency table analysis in the framework of protein structure comparison is another future work.

6 Acknowledgements

This work is supported by National Natural Science Foundation of China under grant No. 10471014. The authors would like to thank Dr. Andreas Martin Lisewski (Department of Molecular and Human Genetics, Baylor College of Medicine, U.S.A.) for providing the data set.

References

[1] L. Chen, L. Wu, Y. Wang, S. Zhang, X. Zhang, BMC Struct Biol, doi:10.1186/1472-6807-6-18, (2006).

[2] K. Mizuguchi, N. Go, Curr. Opin. Struct. Biol. 5 (1995) 377.

[3] Z. Michalewicz, I. Eidhammer, I. Jonassen, W.R. Taylor, J. Comput Biol. 7 (2000) 685.

[4] O. Carugo, S. Pongor, Curr. Protein Pept. Sci. 3 (2002) 441.

[5] C. Guerra, S. Istrail, Springer, Berlin, 2003.

[6] O.Carugo, Curr. Bioinf. 1 (2006) 75.

[7] L. Holm, C. Sander, Science 273 (1996) 595.

[8] I.N. Shindyalov, P. E. Bourne, Protein Eng. 11 (1998) 739.

[9] T. Zhou, L. Chen, Y. Tang, X. S. Zhang, J. Bioinf. Comput. Biol. 3 (2005) 837.

[10] M. M. Young, A. G. Skillman, I. D. Kuntz, Proteins 34 (1999) 317.

[11] D. Gilbert, D. Westhead, N. Nagano, J. thornton, Bioinf. 15 (1999) 317.

[12] T. Kawabata, K. Nishikawa, Proteins 41 (2000) 108.

[13] O. Carugo, S. Pongor, J. Mol. Biol. 315 (2002) 887.

[14] P. Rogen, H. Bohr, Math Biosci. 182 (2003) 167.

[15] P. Rogen, B. Fain, PNAS. 100(1) (2003) 119.

[16] D. Bostick, I. I. Vaisman, Biochem. Biophys. Res. Commun. 304 (2003) 320.

[17] A. M. Lisewski, O. Lichtarge, Nucleic. Acids. Res. doi:10.1093/nar/gkl788, (2006).

[18] S. Kullback, Wiley, New York, 1959.

[19] J. Liu, F. Xu, Acta Scientiarum Naturalium Universitatis Pekinensis. 39 (2003) 76.

[20] W. W. Fang, F. S. Roberts, Z. Ma, Inform. Sciences. 137 (2001) 75.

[21] J Wang, W Fang, L ling, J. Biol. Phys. 28 (2002) 55.

[22] Min Zhang, Weiwu Fang, Junhua Zhang, Comput. Biol. Chem. 29 (2005) 175.

[23] Jie Song, Huanwen Tang, J. Chem. Inf. Comp. Sci. 44 (2004) 1324.

[24] Zikai Wu, Yong Wang, Enmin Feng, Luonan Chen, Chem. Phys. Lett. 433 (2007) 432.