Accurate Classification of Homodimeric vs Other Homooligomeric Proteins Using a New Measure of Information Discrepancy

Jie Song*[†,‡] and Huanwen Tang[†]

Institute of Computational Biology and Bioinformatics, Dalian University of Technology,
Dalian 116024, People's Republic of China, and Department of Mathematics, Shaoguan University,
Shaoguan 512005, People's Republic of China

Received December 10, 2003

It has been shown that protein primary sequence encodes quaternary structure information. In this present work, function of degree of disagreement (FDOD), a new measure of information discrepancy, is applied to discriminating between homodimers and other homooligomeric proteins from the primary structure. This new approach is based on subsequence distributions of the primary sequences, so the effect of residue order on protein structure is taken into account. When the length of subsequence is 4, the overall accuracy of the 10-fold cross-validation test attains to 82.5%, which is much better than that of the previous method on the same data set. Our tests demonstrate that the residue order along protein sequences plays an important role in the prediction of homooligomers. In addition, our results suggest that FDOD measure is a simple and powerful tool for the prediction of protein multimeric states.

INTRODUCTION

In general, the function of a gene relies on the protein for which it encodes; in turn, the function of a protein relies on its structure, that is, the conformation in which the protein folds. Thus, protein structure plays a key role in cell biology, biochemistry, and molecular biology. Protein structure is organized hierarchically from so-called primary structure toquaternary structure. Primary structure is the sequence of residues in the polypeptide chain. Secondary structure refers to regular, repeated patterns of folding of the protein backbone. Tertiary structure is the full three-dimensional folded structure of the polypeptide chain. Quaternary structure, the focus of this paper, only exists, if there is more than one polypeptide chain present in a complex protein. With multiple polypeptide chains, quaternary structure is their interconnections and organization.

The structure of a protein can be determined by physical methods. But this is a slow and expensive process. Owing to the dramatic increase in the numbers of protein sequences sent to the public data bank during the past few years, it is highly desirable to develop some effective computational methods to predict the structure of new proteins from the primary sequences. The primary structure is unique for each protein. It is generally accepted that a protein's primary structure is enough to determine how it will fold and combine with other proteins to make the appropriate secondary, tertiary, and quaternary structure.[1,2] Prediction of protein structure from amino acid sequences remains one of the most important problems in molecular biology.

Proteins with quaternary structure are said to be oligomeric (or multimeric), and the individual chains are called subunits. Oligomeric proteins are either homooligomeric, consisting of identical subunits, or heterooligomeric, consisting of different subunits. A considerable range of oligomers is found in proteins from dimeric creatine kinase to octomeric tryptophanase, and ribulose diphosphate carboxylase, which has 16 subunits. Arrangement of subunits in the oligomeric structure can also vary. An oligomeric protein is more than the sum of its parts and will have important properties not shared with its separated subunits. A variety of bonding interactions including hydrogen bonding, salt bridges, and disulfide bonds hold the various subunits into a particular geometry. Klotz et al. reviewed a number of quaternary structure properties such as stoichiometric constitution, the geometric arrangements of the subunits, the assembly energetics, intersubunit communication, and their functional aspects.[3] Some recent works have paid attention to analyzing protein−protein interactions and predicting interactions sites.[4−8] R. Garian found a rule-based classifier by using decision tree model and amino acid indices to discriminate between the primary sequences of homodimers and other homooligomeric proteins.[9] His results confirmed that protein primary sequence encodes quaternary structure information.

In this paper, we use function of degree of disagreement (FDOD),[10−13] a new measure of information discrepancy, to discriminate between homodimers and other homooligomeric proteins from their primary structure. FDOD has been successfully used to measure the discrepancy between DNA sequences and amino acids sequences from different species in the study of phylogeny and the prediction of protein structural classes.[13−15] Different from the method based on amino acid index, this method is based on the comparison of subsequence distributions. The concept of subsequence distribution of a protein primary sequence is a generalization of its amino acid composition. The amino acid composition of a sequence is a 20-D unit vector, which represents the occurrence frequencies of the 20 amino acids. The concept of amino acid composition has been used in many studies

* Corresponding author phone: +86-411-4700927; fax: +86-411-4709304; e-mail: jiesong@sgu.edu.cn.

[†] Dalian University of Technology.

[‡] Shaoguan University.

10.1021/ci034288y CCC: $27.50 © 2004 American Chemical Society
Published on Web 05/18/2004

of protein.[16−20] However, the methods based on amino acid composition could not sufficiently utilize the information of a sequence.[21,22] Subsequence distributions of a protein sequence involve the residue order along the sequence; therefore, they incorporate more information of the sequence than its amino acid composition does. The new overall accuracy is higher than that of the previous approach based on the decision tree and amino acid index in a 10-fold cross-validation test for the same data set.[9] In addition, the new method is very simple and easy to realize computationally.

DATA AND METHODS

Data Sets. Robert Garian selected a data set of homooligomeric protein sequences from Release 34 of the SWISS-PROT database.[9] It was limited to the prokaryotic, cytosolic subset of homooligomers in the database in order to eliminate membrane proteins and other specialized proteins. The data set consisted of 1639 homooligomeric protein sequences, 914 of which were homodimers and 725 other homooligomeric proteins. In the present work, we use FDOD measure to discriminate between homodimers and other homooligomeric proteins from their primary sequences and compare the new method with the method of decision tree through this data set.

Methods. As a new measure of information discrepancy, FDOD has a close connection with Shannon entropy and has many good characteristics, such as nonnegativity, identity, symmetry, boundedness, triangle inequality, absolute continuity, symmetric recursiveness, monotonicity, effectiveness in singular case, convexity, and so on.[10−13] When it is applied to measure discrepancies among sequences, the constructive information of a sequence is transformed into the set of subsequence distributions called the complete information set. In the following, we give the concept of subsequence distribution and introduce the new method.

For protein sequences, the 20 amino acids form the alphabet $\Sigma = \{$A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y$\}$. The number of all different contiguous sequences formed from Σ with length of l equals 20^l. Suppose S is a sequence of L residues and given the length of subsequence l ($l \leq L$), define p_i^l, $i = 1,...,20^l$, as a ratio between the number of times the ith subsequence of the length l occurs in S and the total number of sequences of the same length l found in S. So we obtain a subsequence distribution:

$$U_S^{\,l} = (p_1^{\,l}, p_2^{\,l}, ..., p_{20^l}^{\,l}) \tag{1}$$

Thus, for each sequence S of length L, there is a unique set of distributions

$$\{U_S^{\,1}, U_S^{\,2}, ..., U_S^{\,L}\} \tag{2}$$

which contains all the composition information of sequence S; it is called a complete information set of the sequence S. Any different sequences have different complete information sets and vice versa.

When $l = 1$, $U_S^1 = (p_1^1, p_2^1, ..., p_{20}^1)$ becomes the conventional amino acid composition. When $l = 2$, $U_S^2 = (p_1^2, p_2^2, ..., p_{400}^2)$ includes all information of the first-order coupled composition introduced by Liu and Chou in secondary structure content prediction.[18] According to the construction of complete information sets, the longer the subsequence is, the more information it includes. When $l = 1$, no information about the residue order is considered, and there are a great number of sequences with the same amino acid composition. But when $l \geq 2$, the residue order along a sequence is contained in its subsequence distribution set, and the occurrence of different sequences with the same subsequence distribution is largely confined because the adjoining subsequences have to overlap each other at $l - 1$ residues. Obviously, any sequence can be uniquely recognized by increasing the length of subsequence.

Given t sequences $S_1, S_2, \cdots, S_t$ and their distributions:

$$\begin{aligned} U_{S1}^{\,l} &= (p_{11}^{\,l}, p_{21}^{\,l}, ..., p_{20^l1}^{\,l}) \\ U_{S2}^{\,l} &:= (p_{12}^{\,l}, p_{22}^{\,l}, ..., p_{20^l2}^{\,l}) \\ &\cdots\ \cdots \\ U_{St}^{\,l} &= (p_{1t}^{\,l}, p_{2t}^{\,l}, ..., p_{20^lt}^{\,l}) \end{aligned} \tag{3}$$

The FDOD measure is defined as

$$B_k(U_{S1}^{\,l}, U_{S2}^{\,l}, ..., U_{St}^{\,l}) = \sum_{i=1}^{20^l} p_{ik}^{\,l} \log \frac{p_{ik}^{\,l}}{\sum_{j=1}^{t} p_{ij}^{\,l}/t} \tag{4}$$

where 0log 0 = 0 and 0log(0/0) = 0 are defined.

$B_k(U_{S1}^l, U_{S2}^l, ..., U_{St}^l)$ denotes a measurement of discrepancy between the kth sequence and all other sequences in the group. Since FDOD measure also satisfies the measurement conditions of complete information sets, such as the Augment and Heredity property, it is not necessary to consider all distributions of the complete information set.[13]

FDOD measure can be applied to multicategory classification. Here we only use it to discriminate between two classes. We suppose that a training set T is the union of two subsets of sequences $T = T^+ \cup T^-$, where T^+ consists of all training homodimers, T^- consists of all training nonhomodimers. Function (4) is used to calculate the discrepancy between a query protein X and T^+ or T^-. Denote the two discrepancy by B_X^+ and B_X^-, respectively. Accordingly, if $B_X^+ < B_X^-$, then X will be assigned to the class of homodimers, else X will be considered as a nonhomodimer protein.

RESULTS AND DISCUSSION

To examine the predictive quality of the present prediction method, we carry out both the resubstitution test and the 10-fold cross-validation test with different subsequence lengths from 1 to 4. The so-called resubstitution test is an examination for the self-consistency of a prediction method, and the method is not considered as a good one if its self-consistency is poor. When the resubstitution test is performed, each data in the data set is in turn identified using the same data set as the training set. k-fold cross-validation is a popular and honest technique for estimating generalization ability of a classifier. For a 10-fold cross-validation test, the data set is randomly partitioned into 10 blocks of examples of approximately equal size. A classifier is then trained on 9 blocks. The remaining block is set aside as a test block. This

Table 1. Results of Resubstitution Tests of the FDOD Method with Different Subsequence Lengths

confusion matrix	FDOD ($l = 1$)	FDOD ($l = 2$)	FDOD ($l = 3$)	FDOD ($l = 4$)
TP	520	639	858	914
TN	475	563	701	725
FP	250	162	24	0
FN	394	275	56	0

performance measures	FDOD ($l = 1$)	FDOD ($l = 2$)	FDOD ($l = 3$)	FDOD ($l = 4$)
Q	0.607	0.733	0.951	1
TPR	0.569	0.699	0.939	1
TNR	0.655	0.777	0.967	1
MCC	0.223	0.473	0.902	1

Table 2. Results of 10-Fold Cross-Validation Tests of the FDOD Method with Different Subsequence Lengths and Comparison with the Method of Decision Tree[a]

confusion matrix	decision tree	FDOD ($l = 1$)	FDOD ($l = 2$)	FDOD ($l = 3$)	FDOD ($l = 4$)
TP	714	513	596	695	651
TN	433	474	538	617	701
FP	292	251	187	108	23
FN	200	401	318	219	251

performance measures	decision tree	FDOD ($l = 1$)	FDOD ($l = 2$)	FDOD ($l = 3$)	FDOD ($l = 4$)
Q	0.699	0.602	0.692	0.801	0.825
TPR	0.781	0.561	0.651	0.760	0.712
TNR	0.597	0.653	0.743	0.851	0.967
MCC	0.386	0.214	0.392	0.608	0.685

[a] The decision tree results are taken from ref 9.

process is repeated for 10 iterations, each time setting aside a different test block.

Suppose that homodimers and other homooligomeric proteins are labeled "positive" and "negative", respectively. Four performance measures are used to assess the ability of the new method for the testing data, which are the overall accuracy (Q), true positive rate (TPR), true negative rate (TNR), and Matthews correlation coefficient (MCC), respectively, and defined as

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives. The four frequencies constitute the confusion matrix of a classifier. The Matthews coefficient is used to reflect the correlation between the predicted and the observed result.

Table 1 gives the results of resubstitution tests of the new method. It is observed that the four performance measures improve quickly when the length of subsequence increases from 1 to 4. When $l = 4$, the new method performs best and predicts totally correctly the homodimers and other homooligomeric proteins. So all performance measures attain to 100%.

Table 2 gives the results of 10-fold cross-validation tests of the new method with different subsequence lengths from 1 to 4 and compares them with the result of the decision tree method. We can see that the predictive accuracies also improve quickly when the length of subsequence increases from 1 to 3. When $l = 3$, the new method performs very well, and all performance measures are higher than that of the method of decision tree, except that the true positive rate is slightly lower than the later. When $l = 4$, the overall accuracy, the true negative rate, and the Matthews correlation coefficient of the FDOD method are up to 82.5%, 96.7%, and 0.685, respectively, which are 12.6%, 37.0%, and 0.299, respectively, higher than those of the method of the decision tree; only the true positive rate is lower than that of the method of decision tree. Obviously, the FDOD method is much more sensitive to other homooligomeric proteins instead of homodimers than the decision tree method. We also notice that the increasing extent begins to reduce when $l = 4$, and the true positive rate is lower than the one with $l = 3$.

Classification results of the decision tree method depend on the choice of the amino acid index. For the method in this paper, the coupling effect of the closest residues is taken into account by decomposing a sequence into subsequences. Moreover, the effect of the residue order along the sequence is involved through the overlaps of subsequences; the longer the subsequence is, the more information of the original sequence it includes.

We notice that when the length of subsequence increases from 1 to 4, almost all performance measures have the tendency to ascend for two kinds of tests. This trend suggests that prediction quality can be improved by increasing the length of subsequence. However, this fact does not mean that the longer the subsequence is, the better the result is. The accuracy of resubstitution test indeed improves with the increase of subsequence length, but that of the 10-fold cross-validation test does only for some lengths. When the length of subsequence increases to some degree, the discrepancies between different sequences tend to be the maximum value, even for the sequences in the same class. Therefore, it is harder to distinguish the classes, and computation error might effect the classification. It is usually appropriate to take $l =$ 3,4 when compared protein sequences have no more than 10 000 residues.[13]

When the dimension of input vector increases, another algorithm such as neural network may be infeasible when the size of the problem increases to some degree. So it is not practicable to use such an algorithm in the space of subsequence distribution. However, the FDOD measure is very convenient and fast for computation; its implementation is as simple as that of the Hamming distance but more powerful to measure the discrepancy of subsequence distributions. Furthermore, the FDOD measure is not only a measure of discrepancy between two distributions but also a measure of discrepancy among multiple distributions.

CONCLUSION

In this paper, a new measure of information discrepancy is applied to the discrimination between homodimers and other homooligomeric proteins from the primary structure, by which the effects of residue order along sequences are taken into account. The tests on the same data set show that the new method has a better predictive quality than that of the previous method. It also confirms that protein primary sequence encodes quaternary structure information. It is observed that the residue order plays important roles in the classification of homooligomers. The most important is that the new method is very simple and efficient, which does not need to choose any man-made parameter. In addition, the present method can be extended to multicategory classification of protein sequences. It is anticipated that FDOD measure may be combined with other prediction methods to become a very useful tool for predicting protein multimeric states.

ACKNOWLEDGMENT

This work has been supported by the Chinese NSF under the Grants 90103033. The authors would like to thank Dr. Robert Garian (School of Computational Sciences, George Mason University, U.S.A.) for providing the data set.

REFERENCES AND NOTES

(1) Anfinsen, C. B.; Haber, E.; Sela M.; White, F. H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309−1314.

(2) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181*, 223−230.

(3) Klotz, I. M.; Darnall, D. M.; Langerman, N. R. Quaternary structure of proteins. In *The Proteins*; Neurath, H., Hill, R. L., Eds.; Academic Press: New York, 1975; Vol. 1, pp 293−411.

(4) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates T. O.; Eisenberg, D. Detecting protein function and protein−protein interactions from genome sequences. *Science* **1999**, *285*, 751−753.

(5) Bock, J. R.; Gough, D. A. Predicting protein−protein interactions from primary structure. *Bioinformatics* **2001**, *17*, 455−460.

(6) Glaser, F.; Steinberg, D. M.; Vakser, I. A.; Ben-Tal, N. Residue frequencies and pairing preference at protein−protein interfaces. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 89−102.

(7) Nooren, I. M. A.; Thornton, J. M. Structural characterisation and functional significance of transient protein−protein interactions. *J. Mol. Biol.* **2003**, *325*, 991−1018.

(8) Ofran, Y.; Rost, B. Analysing six types of protein−protein interfaces. *J. Mol. Biol*. **2003**, *325*, 377−387.

(9) Garian, R. Prediction of quaternary structure from primary structure. *Bioinformatics* **2001**, *17*, 551−556.

(10) Fang, W. The disagreement degree of multi-person judgments in additive structure. *Math. Social Sci.* **1994**, *28*(2), 85−111.

(11) Fang, W. On a global optimization problem in the study of information discrepancy. *J. Global Optim.* **1997**, *11*, 387−408.

(12) Fang, W. The characterization of a measure of information discrepancy. *Inf. Sci.* **2000**, *125*, 207−232.

(13) Fang, W.; Roberts, F. S.; Ma, Z. A measure of discrepancy of multiple sequences. *Inf. Sci*. **2001**, *137*, 75−102.

(14) Wang, J.; Fang, W.; Ling, L.; Chen, R. Gene's functional arrangement as a measure of the phylogenetic relationships of microorganisms. *J. Biol. Phys.* **2002**, *28*, 55−62.

(15) Jin, L.; Fang, W.; Tang, H. Prediction of protein structural classes by a new measure of information discrepancy. *CBAC* **2003**, *27*, 373−380.

(16) Zhang, C. T.; Chou, K. C. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.* **1992**, *1*, 401−408.

(17) Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S. H. Recognition of a protein fold in the context of the SCOP classification. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 401−407.

(18) Liu, W. M.; Chou, K. C. Prediction of protein secondary structure content, *Protein Eng.* **1999**, *12*(12), 1041−1050.

(19) Hua, S.; Sun, Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.* **2001**, *308*, 397−407.

(20) Chou, K. C. Prediction of cellular attributes using pseudo-amino acid composition. *Proteins, Struct., Funct., Genet.* **2001**, *43*, 246−255.

(21) Bu, W. S.; Feng Z. P.; Zhang Z.; Zhang C. T. Prediction of protein (domain) structural classes based on amino acid index. *Eur. J. Biochem.* **1999**, *266*, 1043−1049.

(22) Grigoriev, I. V.; Kim, S. H. Detection of protein fold similarity based on correlation of amino acid properties. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14318−14323.

CI034288Y