CLASSIFICATION OF PROTEIN 3D FOLDS BY HIDDEN MARKOV LEARNING ON SEQUENCES OF STRUCTURAL ALPHABETS

SHIOU-LING WANG
Institute of Biomedical Engineering, Taiwan University, Taipei, Taiwan
Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

CHUNG-MING CHEN
Institute of Biomedical Engineering, Taiwan University, Taipei, Taiwan

MING-JING HWANG
Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

Fragment-based analysis of protein three-dimensional (3D) structures has received increased attention in recent years. Here, we used a set of pentamer local structure alphabets (LSAs) recently derived in our laboratory to represent protein structures, i.e. we transformed the 3D structures into one-dimensional (1D) sequences of LSAs. We then applied Hidden Markov Model training to these LSA sequences to assess their ability to capture features characteristic of 43 populated protein folds. In the size range of LSAs examined (5 to 41 alphabets), the performance was optimal using 20 alphabets, giving an accuracy of fold classification of 82% in a 5-fold cross-validation on training-set structures sharing < 40% pairwise sequence identity at the amino acid level. For test-set structures, the accuracy was as high as for the training set, but fell to 65% for those sharing no more than 25% amino acid sequence identity with the training-set structures. These results suggest that sufficient 3D information can be retained during the drastic 3D->1D transformation for use as a framework for developing efficient and useful structural bioinformatics tools.

1. Introduction

Ever since Anfinsen's experiments in the 1950's demonstrating that the factors determining the three-dimensional (3D) structure of a protein is encoded in its one-dimensional (1D) sequence of amino acids [1], protein structure prediction has been a central interest in computational biology. Among the markedly diverse approaches used, the success of the Rosetta method developed by Baker and co-workers in Critical Assessment of Structure Prediction competitions [2] has highlighted the practicality of using short structural motifs for protein 3D prediction and has stimulated many fragment-based studies in recent years [3-10]. The strategy used in these studies is to cut known protein structures into short overlapping fragments, which are then collected and clustered based on measures of geometric similarity, each cluster being represented by a central fragment, called a centroid; these centroids are then used to construct or analyze protein 3D structures.

We have recently derived a library of structural centroids (http://gln.ibms.sinica.edu.tw/jccs) for protein fragments of 5 amino acids, and shown its performance in approximating protein 3D structures to compare favorably with several others reported in the literature [10]. In the present work, we assigned each centroid of the fragment library an alphabet and used these alphabets to represent protein 3D structures: i.e. with a certain loss of resolution, the protein 3D structure was conveniently transformed into a 1D sequence string of local structural alphabets (LSAs). Using Hidden Markov Model (HMM) machine learning [11], we then evaluated the possibility of using this 3D->1D transformation to assign the Structure Classification Of Proteins (SCOP) fold [12] of a given protein structure, and determined the size of the alphabet set required for optimal performance.

2. Materials and Methods

2.1. Derivation of structural alphabets

The details of our method for deriving LSAs have been described [10]. Briefly, we employed a two-stage procedure to cluster a total of 136,765 pentamer fragments cut from 1,059 randomly selected protein chains of a non-redundant (sequence identity < 25%) Protein Data Bank (PDB) [13] set. The first stage involved the application of an Expectation-Maximization (EM) algorithm [14] using six intra-fragment distances of non-adjacent Cα atoms as feature vectors for clustering. In the second stage, the EM clusters were refined by splitting and merging iteratively to achieve high conformational homogeneity among within-cluster fragments. The results showed that half of the fragment database could be approximated within 0.65 Å by the centroids of the top (i.e. those with most members) 5 clusters. At the same level of approximation, the top 20 clusters covered 80% of the database and the top 40 clusters 90%, but 264 clusters were required to cover the entire database. The root-mean-square (rmsd) error to fit residual fragments (those that cannot be approximated within 0.65 Å of any of the centroids) using 20 and 40 clusters was 0.43 Å and 0.38 Å, respectively [10].

2.2. Hidden Markov training and fold classification

Using the alphabets, we can approximate a protein 3D structure by converting it into a 1D character string, or sequence, of the alphabets (i.e. LSAs). To evaluate to what extent the LSA sequence representation can capture the essence of a protein 3D fold, we examined the fold classification performance by HMM training on 43 of the most populated SCOP folds (release 1.61), each containing at least 20 domains. We employed the ASTRAL Compendium database [15] to choose those domains in the 43 folds sharing less than 40% sequence identity. In all, we used 2,041 domains (~10% of all the SCOP domains belonging to the 43 folds) for training. For HMM training, we followed the procedures and model architecture of HMMER [16]. For each fold selected, we identified a reference structure as the one with the largest number of structurally similar domains

within its own fold and aligned onto it all the other structures of the same fold using the fast structure comparison algorithm FLASH [17]. The resulting multiple structural alignment was represented in the form of a multiple LSA sequence alignment, and this representation was used to train HMM models. Two groups of HMM models were derived with or without the use of an alphabet substitution matrix to estimate a prior relationship between alphabets. The substitution matrix contained probabilities transformed from the rmsd values computed for all pairs of the structural alphabets using the formula of Altschul et al. [18]. Using the HMM models, a given protein structure was then assigned to one of the 43 SCOP folds, with the HMM model of this fold scoring highest for the given structure. Figure 1 outlines the procedures involved in HMM training and fold classification. To evaluate fold classification performance, we performed a 5-fold cross-validation on the training set. For a further evaluation, we tested the trained HMMs on a second set of SCOP domains. This second set, which contained 17,959 structures from the 43 folds, was selected from a newer SCOP release (1.63), excluding those already used in the training set.

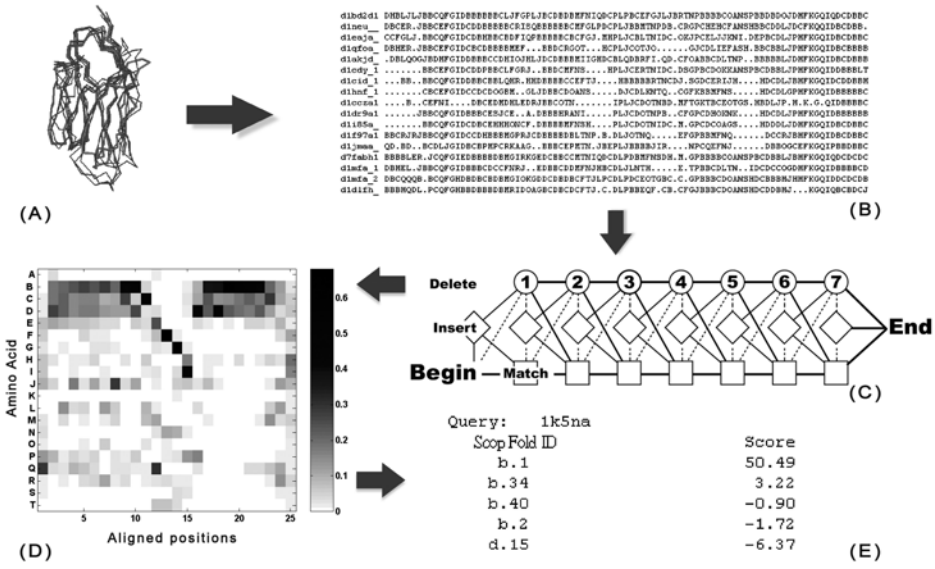

Figure 1. Schematic diagram of the HMM training and fold classification procedures used in this work. For each fold, we (A) selected a suitable reference structure (thick) and aligned the other structures (thin) with it using FLASH, a structural comparison program [17], (B) transformed all these 3D structures into 1D sequence strings of LSAs and produced a multiple LSA sequence alignment based on the multiple structure alignment of (A), (C) trained HMMs iteratively using the multiple LSA sequence alignment of (B) as the initial input, (D) produced HMM profiles showing the emission distribution of LSAs at aligned positions (only part of the aligned positions shown), (E) ranked each fold according to the HMM score for a given structure.

3. Results

3.1. Number of alphabets for optimal performance

The HMM was run on different sets of LSAs containing 5, 10, 15, 20, 25, 33, or 41 alphabets. The 5-fold cross-validation results showed that the performance, as measured by the TP-rate (the fraction of correctly assigned domains), reached a plateau at 20 alphabets, beyond which improvement was negligible (Figure 2). Furthermore, the use of a substitution matrix to take into account different degrees of similarity among the alphabets increased the classification accuracy by ~7% for all alphabet sets, the TP-rate being maximal at 82% for the set of 20 alphabets.

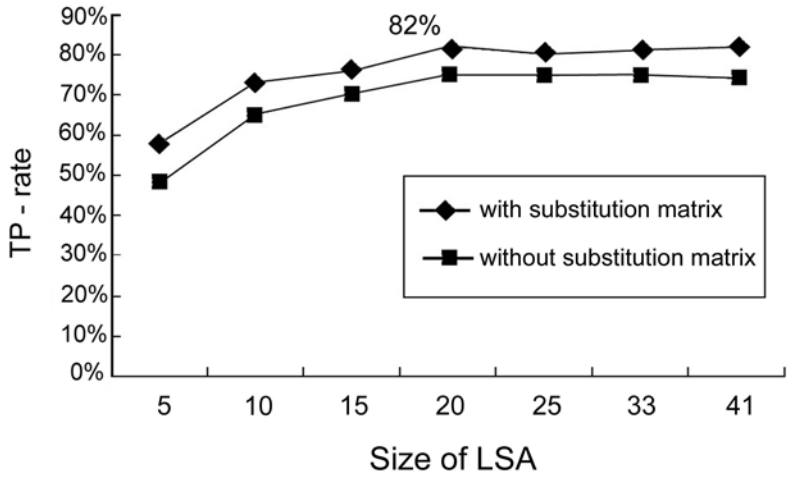

Figure 2. Results of the 5-fold cross-validation using different number of LSAs, with and without a substitution matrix (see Methods). The performance of the fold classification was measured by the TP-rate[a], the fraction of test domains that were correctly assigned.

$$^{a}TP-rate=\sum_{i=1}^{43} TP_i \Big/ (TP_i + FP_i) \ ,$$

TP_i: true positives for fold i; FP_i: false positives for fold i.

3.2. Comparison with the results of Cootes et al. [19]

Having determined the number of LSAs for optimal fold assignment, we then retrained the HMM on the entire training set using 20 alphabets plus the substitution matrix, and compared the results with those reported by Cootes et al. [19], who used inductive logic programming, a machine learning algorithm, to capture signatures of 45 SCOP folds expressed in rules such as “has a parallel sheet of eight strands for a TIM barrel fold”. As

shown in Table 1, our method performed considerably better for three of the four major protein classes. Furthermore, the poorer result for the α+β structures was due to a gross misalignment in the form of LSA sequence for a particular SCOP fold (SCOP fold ID d92, the Zincin-like fold). This misalignment resulted from the difficulty in aligning two domains which differ greatly in size, especially for alignments involving many helices, which, as represented by LSA, are rather featureless. Discounting this fold, our results for α+β structures were 94% for precision, 82% for recall, and 88% for the F-measure (Table 1).

Table 1. Comparison of fold classification performance in this work (first value) and the study of Cootes et al. [19] (second value)

	Precision [a] (%)	Recall [b] (%)	F-measure [c] (%)
All-α	78 / 76	72 / 53	75 / 62
All-β	91 / 64	83 / 45	87 / 53
α/β	85 / 78	74 / 54	79 / 64
$\alpha+\beta$ (-d.92) [d]	82 (94) / 93	72 (82) / 71	77 (88) / 81
Total	84 (87) / 77	75 (78) / 55	79 (82) / 65

According to [20]:

[a] $\mathrm{Precision} = TP_i / (TP_i + FP_i)$

[b] $\mathrm{Recall} = TP_i / (TP_i + FN_i)$

[c] $\mathrm{F-measure} = (2 \times \mathrm{Precision} \times \mathrm{Recall}) / (\mathrm{Precision} + \mathrm{Recall})$

where TP_i :True positive for fold i.
FP_i :False positive for fold i.
FN_i:False negative for fold i.

[d] The data in parenthesis are the results discounting the Zincin-like fold. (fold id d92)

3.3. Test results at different levels of amino acid sequence identity

The trained HMM was then tested on structures that were not included in the training set. These test structures were grouped in different ranges of amino acid sequence identity, which, for any given test structure, was taken to be the highest sequence identity when compared to the training set structures of the same fold. As shown in Figure 3, for structures with sequence identity greater than 30%, the test accuracy was as good as that for the training set (Figure 2 and Table 1). Below this level of sequence identity, performance degraded because of increasing assignment difficulty or decreasing sequence identity. However, the correct assignment was generally within the top ranked folds (87% accuracy within the top 5 folds) even for structures with low sequence identity (Figure 3).

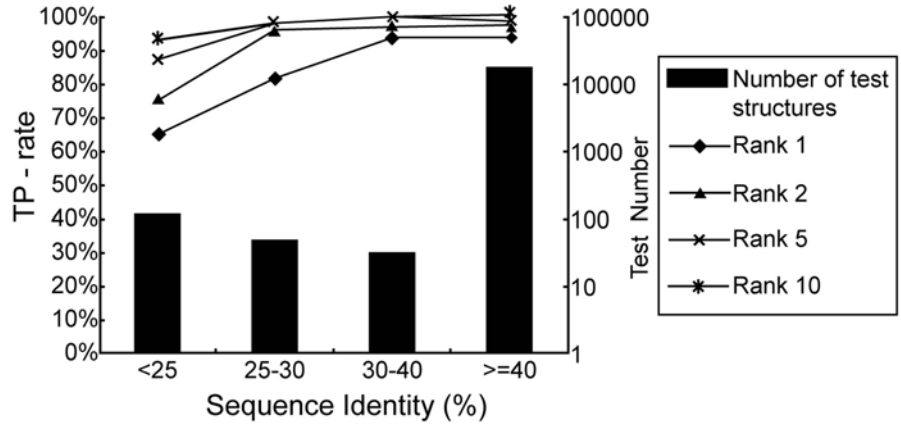

Figure 3.Results of fold assignment for test structures grouped by sequence identity.

4. Discussion

In recent years, many studies have used clustering techniques to identify and characterize short structural motifs of proteins [3-10,21-27]. These studies showed that, although only a few motifs, i.e. a few structural alphabets, are sufficient to represent a large proportion of protein local structures, many more alphabets – hundreds or thousands, depending on the desired resolution – would be required to cover the rest. Consequently, to make use of these structural alphabets, one must evaluate the trade-off between computational efficiency and accuracy. In this work, we showed that, using pentamer fragments, 20 alphabets were optimal for capturing fold-specific features. Using 20 structural alphabets, many bioinformatics tools developed for analyzing amino acid sequences, which, coincidentally, have the same number of alphabets, may now be adopted to analyze protein 3D structures.

Protein fold classification is one such application demonstrated in this work. As the number of new structure entries in the database is increased rapidly by structural genomics projects, there is a need for the accurate and fast classification of protein structures. The baseline accuracy of fold classification using amino acid sequence information alone was recently established as 69.6% for proteins in the 27 most populated folds with sequence identity < 35% [28]. Other methods have been shown to be much more accurate, but require more detailed geometric information, such as the spatial relationship between secondary structure elements [29]. The data presented here demonstrate that LSAs can capture specific features of a protein fold, provided that it is sufficiently populated, with an accuracy intermediate between that of methods requiring no 3D information whatsoever (e.g. [28]) and those requiring detailed knowledge (e.g.

[29]). By removing artifacts that distort HMM training, such as that identified for the Zincin-like fold, our method can be significantly improved.

Acknowledgments

We thank Ta-Tsen Soong and Edward S. C. Shih for helpful discussions. This work was supported by an Academia Sinica program project.

References

1. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181: 223-230, 1973.
2. R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C Rohl, C. Strauss, and D. Baker. Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*, Suppl 5: 119-126, 2001.
3. A. S. Yang and L. Y. Wang. Local structure prediction with local structure-based sequence profiles. *Bioinformatics*, 19: 1267-1274, 2003.
4. C. G. Hunter and S. Subramaniam. Protein fragment clustering and canonical local shapes. *Proteins*, 50: 580-588, 2003.
5. C. G. Hunter and S. Subramaniam. Protein local structure prediction from sequence. *Proteins*, 50: 572-579, 2003.
6. T. R. Hvidsten, A. Kryshtafovych, J. Komorowski, and K. Fidelis. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics*, 19 Suppl 2: II81-II91, 2004.
7. J. B. Holmes and J. Tsai. Some fundamental aspects of building protein structures from fragment libraries. *Protein Sci.*, 13: 1636-1650, 2004.
8. A. C. Camproux, R. Gautier, and P. Tuffery. A Hidden Markov Model derived structural alphabet for proteins. *J .Mol. Biol.*, 339: 591-605, 2004.
9. A. V. Tendulkar, A. A. Joshi, M. A. Sohoni, and P. P. Wangikar . Clustering of protein structural fragments reveals modular building block approach of nature. *J .Mol .Biol.*, 338: 611-629, 2004.
10. T. T. Soong, M. J. Hwang, and C. M. Chen. Discovery of recurrent structural motifs for approximating three-dimensional protein structures. (to appear in *Journal of the Chinese Chemical Society*)
11. L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77: 257-286, 1989.
12. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol .Biol.*, 247: 536-540, 1995.
13. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28: 235-242, 2000.
14. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *J R Stat Soc*, Series B 39: 1-38, 1977.
15. J. M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL compendium in 2004. *Nucleic Acids Res.*, 32: D189-D192, 2004.

16. S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14: 755-763, 1998.
17. E. S. Shih and M. J. Hwang. Protein structure comparison by probability-based matching of secondary structure elements. *Bioinformatics*, 19: 735-741, 2003.
18. S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, 219: 555-565, 1991.
19. A. P. Cootes, S. H. Muggleton, and M. J. Sternberg. The automatic discovery of structural principles describing protein fold space. *J. Mol. Biol.*, 330: 839-850, 2003.
20. I. H. Witten and E Frank. Data mining: practical machine learning tools and techniques with JAVA implementations. 145-146, 1999. CA, U.S.A., Morgan Kaufmann.
21. R. Unger, D. Harel, S. Wherland, and J. L. Sussman. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5: 355-373, 1989.
22. M. J. Rooman, J. Rodriguez, and S. J. Wodak. Automatic definition of recurrent local structure motifs in proteins. *J. Mol. Biol.*, 213: 327-336, 1990.
23. B. Oliva, P. A. Bates, E. Querol, F. X. Aviles, and M. J. E. Sternberg. An automated classification of the structure of protein loops. *J. Mol. Biol.*, 266: 814-830, 1997.
24. J. S. Fetrow, M. J. Palumbo, and G. Berg. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins*, 27: 249-271, 1997.
25. C. Bystroff and D. Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, 281: 565-577, 1998.
26. C. Micheletti, F. Seno, and A. Martin. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins*, 40: 662-674, 2000.
27. A. G. de Brevern, C. Etchebest, and S. Hazout. Bayesian Probabilistic Approach for Predicting Backbone Structures in Terms of Protein Blocks. *Proteins*, 41: 271-287, 2000.
28. C. S. Yu, J. Y. Wang, P. C. Lyu, C. J. Lin, and J. K. Hwang. Fine-grained protein fold assignment by support vector machines using generalized *n*peptide coding schemes and jury voting from multiple-parameter sets. *Proteins*, 50: 531-536, 2003.
29. A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton, and C. Orengo. Recognizing the fold of a protein structure. *Bioinformatics*, 19: 1748-1759, 2003.