Training Hidden-Markov Models on Sequences of Local Structural Alphabets for Protein Fold Assignment

Shiou-Ling Wang[1,2], Chung-Ming Chen[1] and Ming-Jing Hwang[2]

Keywords: structural alphabet, protein fold assignment, HMM, structural alignment, fold signature

Introduction

Recurring local structures of proteins, which may be represented by a set of structural alphabets or libraries of structural motifs, are increasingly used to study the relationship between sequence and structure and to predict protein three-dimensional (3D) structures. We have recently derived a set of protein local structural alphabets (LSA) from clustering > 130,000 fragments, each of five residues in size, excised from ~1,000 non-redundant and diverse known protein structures [1]. In the present study, we employed the derived LSA for fold assignment, i.e. assigning the SCOP fold for a given protein structure, and evaluated the size of LSA required for optimal performance of the assignment.

Methods

With LSA, we can approximate a protein 3D structure and converted it into a 1D character string, or sequence, of LSA. To evaluate to what extent the LSA sequence representation can capture the essence of a protein 3D fold; we tested the fold assignment performance by training Hidden-Markov models (HMM) on 43 populated SCOP fold families, each having at least 20 member structures. For each fold family selected, we identified a reference structure, and aligned all the other member structures onto it using a fast structure comparison algorithm FLASH [2]. The HMM was trained on this multiple structural alignment, which was represented in the form of a multiple LSA sequence alignment. A protein structure can then be assigned to one of the 43 SCOP folds, i.e. the HMM having the maximal probability score. For evaluation of the fold assignment performance, we conducted a 5-fold cross-validation on a dataset with less than 40% pair wise sequence identity chosen according to the ASTRAL Compendium database.

Results

The HMM was run on different sets of LSA, in size of 5, 10, 15, 20, 25, 33 and 40 alphabets, respectively. The 5-fold cross-validation results showed that a performance plateau was reached at 20 alphabets, beyond which improvement was negligible. Furthermore, the use of a substitute matrix giving different substitution scores for different alphabets elevated the assignment accuracy by ~7% for all the different alphabet sets, and yielded an accuracy of 82% for the set of 20 alphabets. A comparison with the results of Cootes et al. [3], which used a very different approach to capture fold signatures, showed that our method performed better in three of the four major protein classes. The less-optimal results for $\alpha + \beta$ structures can be attributed in large part to a gross misalignment of long helices in the form of 1D LSA sequence for, particularly, the

[1] Institute of Biomedical Engineering, Taiwan University, Taipei, Taiwan.

[2] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. E-mail:mjhwang@ibms.sinica.edu.tw

Zincin-like fold. Our results suggested that protein fold signatures can be largely captured by local structures even if they are represented in the form of 1D alphabet sequences.

References

[1] Soong, T.T. 2002. Clustering and characterizing local protein structures by an Expectation-Maximization (EM)-assisted approach. *Master Thesis*, National Taiwan University.

[2] Shih, E. S. and M. J. Hwang. 2003. Protein structure comparison by probability-based matching of secondary structure elements. *Bioinformatics.* 19:735-741.

[3] Cootes, A. P., S. H. Muggleton, and M. J. Sternberg. 2003. The automatic discovery of structural principles describing protein fold space. *J.Mol.Biol.* 330:839-850.