Poster I-45

Profiling Tertiary Protein Motifs Using PHMM

Authors:
Jalal Mahmud (*Department of Computer Science, Stony Brook University, Stony Brook , NY 11794, USA*)
Chang Zhao (*Department of Computer Science, Stony Brook University, Stony Brook , NY 11794, USA*)
I.V. Ramakrishnan (*Department of Computer Science, Stony Brook University, Stony Brook , NY 11794, USA*)

Short Abstract: Structural motif comparison can detect homology when there is no sufficient sequence similarity or overall structural similarity. Profile based methods have been proved to be more effective in recognizing remote homologs. We adapt PHMM to profile the tertiary structure of protein motifs. Our experiments demonstrate reasonable result.

Long Abstract:

Homology search is a well known method of predicting protein functions. Structural motif based methods [1] can detect homology when there is neither sufficient sequence similarity nor overall structural similarity between proteins. Most existing structural motif based methods compare individual motifs. Such pair-wise comparison methods do not capture the conserved part of similar motifs from the same protein family. Just as for sequence homology search where profile base methods can detect more remote homologues than pair-wise comparison, we can build a profile for a set of similar structural motifs and use that for remote homology search. PHMM [2] have been shown to be an effective model for sequence homology search. We adapt PHMM to profile the tertiary structure of protein motifs. More specifically, we are interested in protein active sites which are tertiary motifs in proteins that dictates how a protein interacts with others.

Traditional PHMM methods requires sequences of observations but atoms in 3D space do not form sequences. Therefore we need to introduce an ordering of the atoms to serialize them. In particular we use ordering of all the atoms according to the distance from their center of mass. We need to fix the emission symbols associated with the states of the PHMM. Note that they need to capture not only atom types, which correspond to residue types in traditional PHMMs, but also spatial characteristics of the atoms. However, directly using 3D coordinates of the atoms makes the quality of profiling and search heavily dependent on correct superimposition. We used the distance of an atom to the center of mass to approximate the spatial feature of the atom. Therefore, each emission symbol is a pair (T, D) where T is the atom type and D is the distance to the center of mass. The alphabet for the atom types is adopted from [6] which contains 40 atom types defined for all heavy atoms of 20 amino acids according to the atom location, connectivity and chemical nature.

To learn the parameters, we follow the same approach as used in traditional PHMMs: The training sequences are generated from the protein family, one sequence per motif. Since

emission symbols are pairs we need to compute the joint distribution of atom types and distance at each state. For computational tractability purposes, we make the standard independence assumption. Hence, the probabilities of the atom types and the distances for a state are computed separately. Since distance from center of mass is a continuous feature, we represented them as Multivariate Gaussians. We estimated parameters of the model by Baum-Welch algorithm, a widely used techniques for HMM parameter estimation.

To use the adapted PHMM for searching, we proceed as follows: using tools such as Active Site Finder [5] we find potential substructures of the input protein and generate their observation sequences. These sequences are run through the adapted PHMM and a modified Viterbi algorithm determines their closeness to the family. This modification was requires the emissions be pairs instead of single symbols used in traditional PHMMs.

We developed profiles for protein families based on the ideas sketched above and performed experiments. The dataset is taken from [3]. Here we listed precision and recall values (precision/recall) for protein classification of the 5 families using our methods:

Name Precision/Recall

Ribonuclease A (92%/90%)
Ribonuclease T1 (86%/88%)
Eukaryotic Lysozyme (91%/90%)
Prokaryotic Lysozyme (88%/89%)
Nu:-His-Elec catalytic triad (86.5%/88%).

Profiling of structural motifs is not extensively studied. Wallace et.al [3] used template of motifs to search for similar motifs in protein structures. They manually generate template for each family. They assume that domain experts specify the set of atoms that are conserved among the members of that family. Clearly it is not suitable for a large number of families because of this manual process. In contrast, we do not require such domain knowledge and learn the profiles. Thus our approach is scalable.

Gernstein et.al [4], incorporated 3D co-ordinates as emission symbols. However they use RMSD criteria to align the protein tertiary sequences. Their approach is dependent on the performance of superimposing tool. Our approach does not depend on such tool. Moreover, some proteins with similar functions don't resemble each other at the whole structure level. Only their active sites are similar. Their approach is not applicable for those cases since they build profile from the entire structure of the proteins.

Our experiments demonstrate that it is feasible to develop computational models for profiling active sites. There are several avenues for further work: In our experimentation, we utilized one geometric feature, namely distance to the center of mass. We can incorporate other geometric features such as pair-wise distances between atoms as well as non-geometric features such as stereochemical and charge constraints of the active site motif, etc. Adding these features will yield richer profiles and may improve the accuracy of prediction. Finally, HMMs assume that transitions are Markovian, i.e., they depend only on the previous state. However, active sites are localized substructures of a protein. To accommodate such localized dependencies, a state transition have to be influenced by a set of neighboring

states. Hidden Markov Random Fields (HMRF) address such kinds of dependencies [7]. It is possible to develop techniques for profiling active sites of protein families using HMRF with a rich set of active site features.

References:

[1] alpha2.bmc.uu.se/usf/spasm.html
[2]Sean Eddy, Anders Krogh, and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids , Cambridge University Press, 1998.
[3] A. C. Wallace, N. Borkakoti And J. M. Thornton, TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites, Protein Sci., 1997 6: 2308-2323
[4] V. Alexandrov and M. Gerstein. Using 3D hidden Markov models that explicitly represent spatial coordinates to model and compare protein structures. BMC Bioinformatics, 5:2, 2004
[5] http://www.chemcomp.com/journal/sitefind.htm
[6] F. Melo and E. Feytmans. Novel knowledge-based mean force potential at atomic level. J. Mol. Biol., 267:207-222, 1997.
[7] H. Kuensch, S. Geman, and A. Kehagias. Hidden Markov random fields. Annals of Applied Probability, 5:577-602, 1995.