Conclusion

This book has presented a new method for sequence-based protein homology detection that compares two proteins through alignment of two Markov Random Fields (MRFs), which model the multiple sequence alignment (MSA) of a protein set using an undirected general graph in a probabilistic way. The MRF representation is better than the extensively-used PSSM (position-specific scoring matrix) and HMM (Hidden Markov Model) representations in that the former can model long-range residue interactions while the latter two cannot. As such, MRF-based homology detection shall be much more sensitive than PSSM- and HMM-based methods. Our large-scale experimental tests show that MRF-MRF comparison can greatly improve alignment accuracy and remote homology detection over currently popular sequence-HMM, PSSM-PSSM, and HMM-HMM comparison methods. Our method also has a larger advantage over the others on mainly-beta proteins.

We build our MRF model from multiple sequence alignment (MSA) without using any native structures, so the accuracy of an MRF model depends on the accuracy of an MSA. Currently our MRF model is built upon the MSA generated by PSI-BLAST. In the future, we may explore better alignment methods for MSA building or even utilize a few solved structures to improve MSA. The accuracy of the MRF model parameter usually increases with respect to the number of non-redundant sequence homologs in the MSA. Along with more and more protein sequences are generated, very accurate MRFs will be available for more and more protein families and thus, their homologous relationship can be studied more accurately using MRFs.

An accurate scoring function is essential to MRF-MRF comparison. Although in this book we only present one scoring function, various scoring functions can be used to measure the similarity of two MRFs, just like quite a few scoring functions are developed to measure the similarity of two PSSMs or HMMs. It is computationally intractable to find the best alignment between two MRFs when long-range residue interaction is considered. This book presents an ADMM algorithm that can efficiently solve the MRF-MRF alignment problem to suboptimal. However, it is about 10 times slower than the dynamic programming algorithm for PSSM-PSSM alignment. Further tuning of this ADMM algorithm is needed for very large-scale homology detection on a laptop computer.

J. Xu et al., *Protein Homology Detection Through Alignment of Markov Random Fields*, SpringerBriefs in Computer Science,
DOI 10.1007/978-3-319-14914-1

Acknowledgements

This work is supported by National Institutes of Health grant R01GM089753, NSF CAREER Award CCF-1149811 and Alfred P. Sloan Research Fellowship. The authors are also grateful to the University of Chicago Beagle team and TeraGrid for their support of computational resources. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

J. Xu et al., *Protein Homology Detection Through Alignment of Markov Random Fields*, SpringerBriefs in Computer Science,
DOI 10.1007/978-3-319-14914-1