Geometry-inspired Optimization Methods for Structural Motifs for Protein Function Prediction

Brian Y. Chen[1], Drew H. Bryant[2], Viacheslav Y. Fofanov[3], David M. Kristensen[4],
Mark Moll[1], Marek Kimmel[3], Olivier Lichtarge[4], Lydia E. Kavraki[1]*

[1] Department of Computer Science, Rice University, Houston, TX 77005, USA
[2] Department of Bioengineering, Rice University. Houston, TX 77005, USA
[3] Department of Statistics, Rice University, Houston, TX 77005, USA
[4] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston TX 77030, USA

*To whom correspondence should be addressed: kavraki@cs.rice.edu

1. INTRODUCTION

One strategy for function prediction is to search the structures of “target” proteins with unknown function for sites which are geometrically and chemically similar to “motifs” representing a known active site. Like all function prediction strategies, the above strategy may have some inaccuracies, such as in the design of the motifs, which may have geometric and chemical dissimilarities to functionally related proteins (not sensitive), or similarities to functionally unrelated proteins (not specific). In this abstract we describe two techniques to optimizing structural motifs so as to increase specificity while maintaining sensitivity. Both techniques are based on a general principle called “Motif Profiling”. It is assumed that a reasonably designed motif will be optimized. The presented methods can be used as a post-processing step with many motif design methods.

2. METHODOLOGY AND RESULTS

Improving motif specificity requires the elimination of false positive matches which occur by random chance between a given motif and any large set of target structures. We have developed a general technique called Motif Profiling which provides a measure which seems to be useful for improving specificity in several applications. A “Motif Profile” is a frequency distribution which indicates the frequency of matches observed between a given motif at a specific degree of similarity. In our work the Least Root Mean Squared Distance (LRMSD) is used as a measure of geometric similarity for two sets of atoms that are chemically similar, and motif profiles are explicitly computed by matching a given motif to a representative subset of the PDB (1,2). Due to chance, functionally unrelated sites can still have a low LRMSD to a motif. During Motif Profiling, we aim to skew the distribution of matches such that this is less likely to happen. The threshold distance that best separates functionally related matched from unrelated matches varies per motif. So instead we use the p-value to determine the statistically significant matches (1,2).

The first method we have developed for optimizing, or refining, a given reasonable motif is called Geometric Sieving (GS). The goal of GS is to select a subset from a given motif that has increased specificity while maintaining the sensitivity of the original motif (3). GS takes as input a motif and k, an expected number of motif points in the output motif. It then finds a motif of size k that maximizes the median of the LRMSD of the matches between the motif and a representative subset of the PDB, or in other words a motif with the greatest overall dissimilarity to the PDB. We have shown that this increases the LRMSD of negative matches significantly more than the LRMSD of the positive matches, thus improving the specificity of the motif. To compute the median it is not necessary to compute the full distribution of LRMSD values. Instead, a narrow range for the median can be computed with high confidence with a relatively small number of samples. In (3) we showed that candidate motifs from six well-studied proteins, including a-Chymotrypsin, Dihydrofolate Reductase, and Lysozyme, can be optimized with GS to motifs that are among the most sensitive and specific motifs possible for the candidate motifs.

We also applied Motif Profiling towards the refinement of cavity-aware motifs. The later motifs employ C-spheres to eliminate false positive matching with targets that have atoms occupying volumes essential for protein function. C-spheres are spheres that are rigidly associated with some of the motif points. For a valid match, the C-spheres do not intersect any protein atoms. One difficulty in the design of cavity-aware motifs, in addition to the selection of points for the motif, is the desire to select C-spheres which eliminate many false positive matches. Our method, Cavity Scaling (CS), measures the change in motif profiles as C-spheres expand (4). We observed that some C-spheres, called high-impact C-spheres hereafter, eliminate many false positive matches, and cause the motif profile to shift dramatically towards higher LRMSDs, while low-impact C-spheres do not. In (4) we demonstrated that CS can be applied to identify high-impact C-spheres leading us to believe that in the absence of expert knowledge, CS can guide the design of cavity-aware motifs to eliminate many false positive matches.

3. DISCUSSION

Multiple studies have established the difficulty of designing sensitive and specific motifs for protein function annotation. Our work suggests that because of the speed of current matching algorithms and availability of computational power, it is possible to further refine exist motifs produced by recent motif design efforts by examining the motif profiles when matched to a representative subset of the PDB. Modifying the design of a motif by changing the way an active site is represented, affects the shape and position of the motif profile. In particular, changes to the motif design which cause the median of the profile to shift towards more dissimilar ranges identify changes which reduce the similarity of the motif to the space of known protein structures. In the two separate instances of GS and CS, we have been able to show that Motif Profiling, which is essentially a geometry-inspired approach, is effective in refining given motifs. Our work suggests that geometric criteria may have a role to play in the design of sensitive and specific motifs for protein function prediction.

5. ACKNOWLEDGMENT

This work has been supported in part NSF DBI-0318415 and DBI 0547695 under a subcontract to Rice University. Additional support is gratefully acknowledged from training fellowships of the W.M. Keck Center (NLM Grant No. 5T15LM07093) to B.C. and D.K. and from a VIGRE Training in Bioinformatics Grant (NSF DMS 0240058) to V.F. Experiments were run on equipment funded by NSF CNS-0454333, CNS-042119 in partnership with Rice University, AMD and Cray.

4. REFERENCES

1. B.Y. Chen, V.Y. Fofanov, D.M. Kristensen, M. Kimmel, O. Lichtarge, and L.E. Kavraki, Algorithms for Structural Comparison and Statistical Analysis of 3D Protein Motifs, *Pacific Symposium on Biocomputing 2005*, World Scientific, Hawaii, January, 2005, 334-345.

2. D. M. Kristensen, B.Y. Chen, V.Y. Fofanov, R.M. Ward, A.M. Lisewski, M. Kimmel, L.E. Kavraki, and O. Lichtarge. (2006). Recurrent Use of Evolutionary Importance for Functional Annotation of Proteins Based on Local Structural Similarity. Protein Science special section on Automated Function Prediction. 15:1530–1536.

3. B.Y. Chen, V.Y. Fofanov, D.H. Bryant, B.D. Dodson, D.M. Kristensen, A.M. Lisewski, M. Kimmel, O.Lichtarge, and L.E. Kavraki, Geometric Sieving: Automated Distributed Optimization of 3D Motifs for Protein Function Prediction, *Research in Computational Biology: 10th Annual International Conference (RECOMB)*, Venice, Italy, April 2006. Published in Lecture Notes in Bioinformatics, A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds), Springer, LNBI 3909/2006, 500-515.

4. B.Y. Chen, D.H. Bryant, V.Y. Fofanov, D.M. Kristensen, A.E. Cruess, M. Kimmel, O. Lichtarge, and L.E. Kavraki, Cavity-Aware Motifs Reduce False Positives in Protein Function Prediction, *Computational Systems Bioinformatics (CSB)*, Stanford, CA, August 2006, Series on Advances in Bioinformatics and Computational Biology, Vol 4, 311-323, Imperial College Press.