DMP Final Report: Geometric Analysis of Multiple Protein Structures for the Design of Optimized 3D Protein Motifs

Amanda Cruess

I. Introduction

Understanding the function of proteins continues to be a fundamental problem of biology[7]. Functional annotation of proteins through biological experimentation, however, is time-consuming and expensive. Many computational methods for prediction of protein function have been developed. Some tools, such as PSI-BLAST[2], EMATRIX[13], and PROSITE[6], use sequence similarity to help predict function, while others, such as JESS[12], PINTS[1], webFEATURE[10], Geometric Hashing[11], and Match Augmentation[3], predict function using comparisons of geometric structure. Although these computational techniques are accurate and efficient in determining geometric similarity, the choice of protein components to compare is important as well. These components must be both functionally significant and a geometrically distinct representation of that protein relative to other protein structures to prevent matches with proteins that are not functionally similar. Here we explore methods for choosing protein components that will increase the sensitivity and specificity of searching.

II. Previous Work

Some algorithms that search for matches with greatest geometric similarity, such as Geometric Hashing[11] and Match Augmentation[3], use least root mean square distance (LRMSD) as a measure of geometric similarity. A statistically significant LRMSD between two protein structures can suggest similar function[3]. In their work on Match Augmentation, Chen et. al. define a *motif* as a known protein component, usually an active site, that is used to search for structural *matches* in a set of functionally uncharacterized proteins, known as *targets*. A motif is composed of points in three-dimensional space, usually chosen from the area surrounding the active site of the protein. Figure1 shows an example of a set of motif points (left) and the same motif superimposed on a target protein structure (right). Each motif point is defined by its geometric configuration and its chemical makeup. In previous work, motifs have been designed from a single protein structure. However, many proteins occur in a variety of states due to conformational changes that may occur during ligand binding or catalysis (Figure2). For example, the unliganded form of glutathione transferase (1gsd), its complex with an ethacrynic acid glutathione conjugate (1gse), and its complex with ethacrynic acid alone (1gsf) are among the 153 structures of glutathione transferases contained in the Protein Data Bank[5] (PDB). Here we present a process for using this knowledge of multiple structures to improve our method of motif design.

III. Data and Methods

Our primary data consisted of ten distinct protein structures selected from the Protein Data Bank. A data set appropriate for studying the development of motifs based on multiple structures

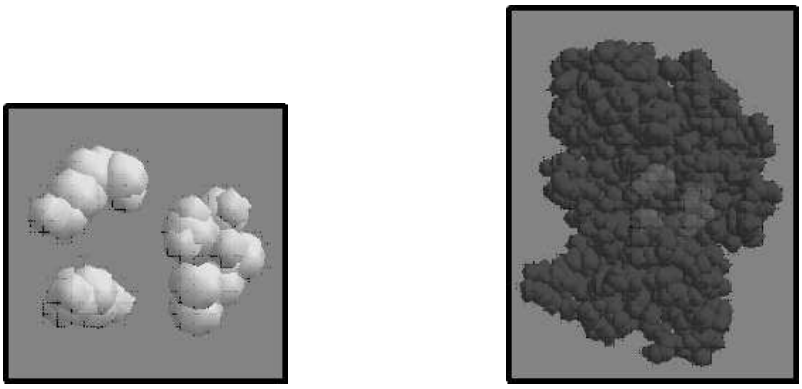

Fig. 1. An example motif (left) and a sample target with motif (right)

needed to consist of a variety of proteins, each having multiple structures in the PDB and a set of functionally homologous proteins. Functional homology was determined using the Enzyme Commission (EC) classification, which, although imperfect, is standard and useful for our purposes. For each of the ten protein structures, a motif consisting of four to ten residue points was designed based on documentation of functionally significant amino acids and sequence analysis information. For example, peroxidase from the fungus Arthromyces ramosus (1aru) is a heme protein belonging to EC family 1.11.1.7. Five points were selected for this motif, including histidine 184, which binds the heme iron[8], and the distal arginine (Arg-52 in this structure[4]), which has been suggested to play a role in substrate binding and stabilization of the product of the first step of the enzyme reaction[9]. Also included was histidine 56, which is suggested to be responsible for proton translocation in the hydrogen peroxide substrate and has been shown to undergo conformational change in complexes with both cyanide and triiodide[4]. Asparagine 93 and glutamic acid 87 were chosen because they form a hydrogn bond network with histidine 56[4].

Using the motif designed for each single protein structure, Match Augmentation[3] was then used to compare the geometric similarity of the motif with that of a set of target proteins composed of those structures in the functional homolog family. Match Augmentation consists of two parts: seed matching and augmentation. Seed matching takes advantage of a prioritization assigned to the motif points based on sequence analysis information, searching for targets that match the three highest priority motif points. The k targets that match this *seed* with lowest LRMSD are then passed to the Augmentation phase where the matches are iteratively expanded, adding the remaining motif points in order of decreasing priority. We use $k = 30$. Multiple complete matches may be found, but we use only the match with the lowest LRMSD. For all protein structures, approximately 80% of the functional homologs returned a match for the initial motif. This indicates that a motif designed using the conventional method is not sensitive or specific enough to match all functionally homologous proteins. The LRMSD values obtained for each target structure returning a match were then used to separate the targets into subgroups. Targets with a low pairwise LRMSD have similar structure, while those with high pairwise LRMSD have less similar structure. For example, the functional homolog family containing the peroxidase from the fungus Arthromyces ramosus was divided into three distinct structural subgroups.

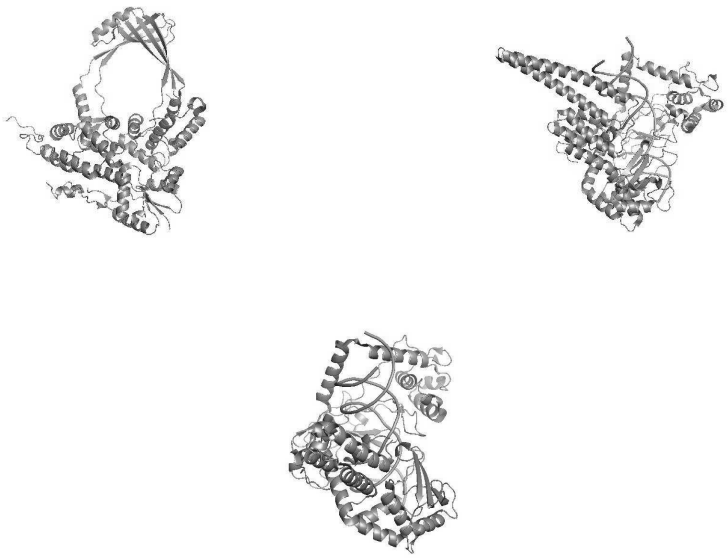

Fig. 2. Three crystallized structures of Human DNA Topoisomerase I

IV. Results, Conclusions, and Future Work

For each of the protein structures in our dataset, we were able to divide the set of functionally homologous structures into subgroups, although the number and size of these subgroups vary widely. These results indicate that groups of proteins with similar function can be subdivided into more specific subgroups based on structural similarity. The results from the segmentation of these protein families provides information that can be used to develop an ensemble of motifs which describe the various states in which target protein structures can be found more effectively than any motif defined by conventional means. We hypothesize that motifs developed based on multiple protein structures will have higher sensitivity and specificity than single structure optimized motifs. The next step in the project will be to use this information to create the new motifs and then compare the motifs designed based on this multiple structure information with motifs developed from a single structure. In the future, we are also interested in developing a means of automating this motif design based on multiple structures.

A. Acknowledgements

This work was supported by the CRA-W Distributed Mentor Project award for summer 2005.

References

[1] Stark A., Sunyaev S., and Russell R.B. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.

[2] Altschul S.F. et. al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids. Res.*, 25(17):3389–3402, Sept 1997.

[3] Chen et. al. Algorithms for structural comparison and statistical analysis of 3d protein motifs. *To appear in Pacific Symposium on Biocomputing*, 2005.

[4] Fukuyama K. et. al. Crystal structures of cyanide- and triiodide-bound forms of *Arthromyces ramosus* peroxidase at different ph values: perturbations of active site residues and their implication in enzyme catalysis. *J. Biol. Chem.*, 270(37):21884–21892, September 1995.

[5] H.M. Berman et. al. The protein data bank. *Nucleic Acids Research*, 28:235–242, Sept 2000.

[6] Hulo N. et. al. Recent improvements to the PROSITE database. *Nucl. Acids. Res.*, 32:D134–D137, 2004.
[7] Jones S. et. al. Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, 8(1):3–7, 2004.
[8] Kunishima N. et. al. Crystal structure of the fungal peroxidase from arthromyces ramosus at 1.9 a resolution. structural comparison with the lignin and cytochrome c peroxidases. *J. Mol. Biol.*, 235(1):331–344, January 1994.
[9] Vitello L.B. et. al. Effect of arginine-48 replacement on the reaction between cytochrome c peroxidase and hydrogen peroxide. *Biochemistry*, 32(37):9807–9818, September 1993.
[10] Laing M.P. et.al. Webfeature: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucl. Acids Res.*, 31(13):3324–7, 2003.
[11] Wolfson H.J. and Rigoutsos I. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.
[12] Barker J.A. and Thornton J.M. An algorithm for constraint-based structural template matching: application to 3D templates. *Bioinf.*, 19(13):1644–1649, 2003.
[13] Wu T.D, Nevill-Manning C.G, and Brutlag D.L. Fast probabilistic analysis of sequence function using scoring matrices. *Bioinf.*, 16(3):233–44, 2000.