Functional Annotation of Proteins through Substructure Matching

Mark Moll
Department of Computer Science
Rice University
Houston, TX 77005
Email: mmoll@rice.edu

Drew H. Bryant
Department of Computer Science
Rice University
Houston, TX 77005
Email: drew.h.bryant@gmail.com

Lydia E. Kavraki
Department of Computer Science
Rice University
Houston, TX 77005
Email: kavraki@rice.edu

***Abstract*—The number of known protein structures is rapidly increasing. The function of most proteins is still poorly understood or even completely unknown. At the same time, there are also many large protein families for which many structural variants are available. We present a substructure-based approach called LabelHash that can be used to annotate proteins with unknown function. We also describe a new method that uses LabelHash as a tool to help understand the structural variations within classes of proteins with known function. This structural variation within a family of related proteins can be exploited to design drugs with very high specificity.**

Determining the function of unannotated proteins would have a significant impact on understanding diseases and designing new therapeutics. However, experimental protein function determination is expensive and very time-consuming. Computational methods can facilitate function determination by identifying proteins that have high structural and chemical similarity. Our focus is on methods that determine binding site similarity. Although several such methods exist, it still remains a challenging problem to quickly find *all* functionally-related matches for structural motifs in the entire Protein Data Bank (PDB) with high specificity. In this context, a structural motif is a set of 3D points annotated with physicochemical information that characterize a molecular function. We have developed a method called LabelHash that creates a hash table of n-tuples of residues for all structures in the PDB [2]. The method is inspired by geometric hashing, a technique that originated in computer vision but which has also been applied to matching structural motifs [6, 5]. The key advantage of LabelHash over geometric hashing is that it uses much less space and scales more easily to very large data sets such as the entire PDB. Using the LabelHash hash tables, we can quickly look up partial matches to a motif and expand those matches to complete matches. We show that by applying only very mild geometric constraints we can find statistically significant matches with extremely high sensitivity and specificity for very general structural motifs (see Figure 1). The LabelHash method is also extremely fast; it can match motifs ranging in size from 3 to 11 residues in a matter of seconds to minutes to all structures in the 95% sequence identity filtered non-redundant PDB. A web server front-end for LabelHash as well as a command line version are available at http://labelhash.kavrakilab.org [3].

The LabelHash method is sufficiently fast that it can be used to perform a detailed analysis of the structural variability within large protein families or even superfamilies. Structural variations caused by a wide range of physicochemical and biological sources directly influence the function of a protein. Comparative analysis of drug-receptor substructures across and within species has been used for lead evaluation. Substructure-level similarity between the binding sites of functionally similar proteins has also been used to identify instances of convergent evolution among proteins. The Family-wise Analysis of Sub-Structural Templates (FASST) method uses LabelHash for all-against-all substructure comparison to determine substructural clusters [1]. Substructural clusters characterize the binding site substructural variation within a protein family. We focus on examples of automatically determined substructural clusters that can be linked to phylogenetic distance between family members (see Figure 2), segregation by conformation, and organization by homology among convergent protein lineages. The Motif Ensemble Statistical Hypothesis (MESH) framework constructs a representative motif for each protein cluster among the substructural clusters determined by FASST to build *motif*

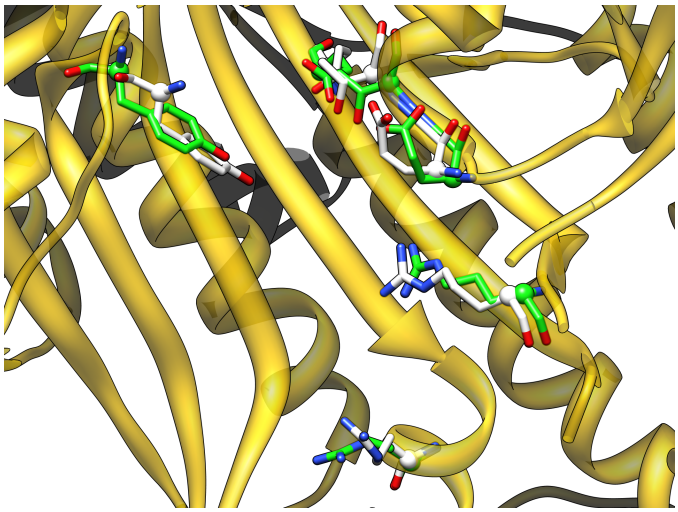

Fig. 1. A substructure match (in green) shown superimposed with a motif (in white), while the rest of the matching protein is shown in ribbon representation. *Figure reproduced from [2].*

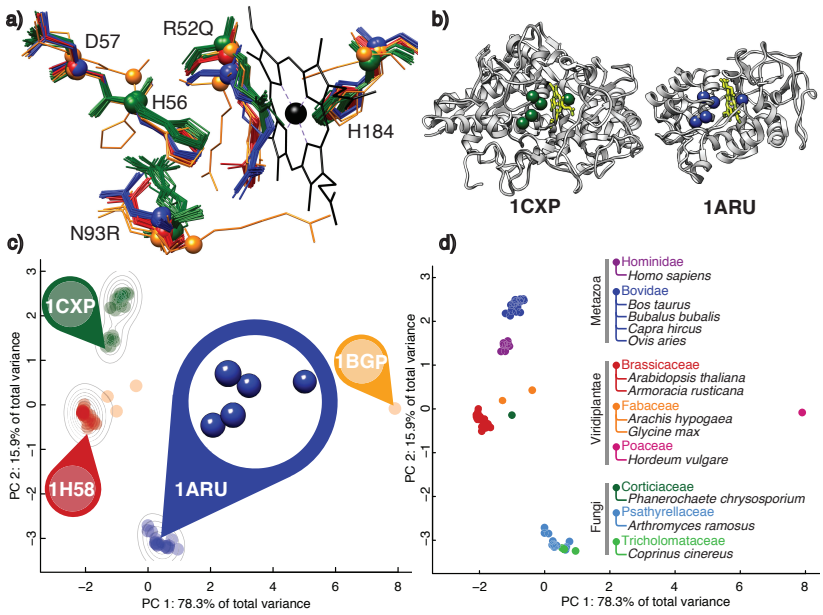

Fig. 2. **Substructural Clusters (SCs) for the heme-dependent peroxidases. (a)** Superposition of the propagated motifs for the animal and non-animal heme-dependent peroxidases of EC 1.11.1.7 demonstrates geometric diversity. The color of each aligned substructure corresponds to its cluster assignment in (c), and it can be seen that closely aligned substructures in (a) correspond to co-located points in the SCs shown in (c). **(b)** When the backbones of a class II fungal peroxidase [PDB:1ARU] and human myeloperoxidase [PDB:1CXP] are compared, substructural similarity within the heme-binding catalytic site region is evident, but the remainder of the enzyme structures can be seen to have significant topological differences and are assigned to separate topological classes within the CATH structural ontology [4]. **(c)** Applying FASST to the family of peroxidases yields a family-wise geometric feature vector for each catalytic substructure in the family, reducing each substructure shown in (a) to a point in the SCs. Gaussian mixture model (GMM) clustering of geometric feature vectors, projected onto a space of reduced dimension, identifies four clusters denoted by color. The gray isocontours show the smoothed density of substructures in each part of the SCs. **(d)** Substructure positions in the SCs colored by Family-level taxonomic classification reveal that phylogenetic distance between proteins is the main source of substructural diversity among the heme-dependent peroxidase binding sites. The open/closed plot characters correspond to apo/holo structures, respectively. *Figure reproduced from [1].*

ensembles that are shown through a series of function prediction experiments to improve the function prediction power of existing motifs [1]. FASST contributes a critical feedback and assessment step to existing binding site substructure identification methods and can be used for the thorough investigation of structure-function relationships. The application of MESH allows for an automated, statistically rigorous procedure for incorporating structural variation data into protein function prediction pipelines. Our work provides an unbiased, automated assessment of the structural variability of identified binding site substructures among protein structure families and a technique for exploring the relation of substructural variation to protein function. As available proteomic data continues to expand, the techniques proposed will be indispensable for the large-scale analysis and interpretation of structural data. We suspect that techniques similar to LabelHash and FASST can also be applied to other large, partially annotated, spatial datasets for the purpose of object recognition.

ACKNOWLEDGMENTS

This work has been supported in part by NSF Graduate Research Fellowship grant DGE-0237081 to DHB, NSF ABI grant ABI-0960612, the John and Ann Doerr Fund for Computational Biomedicine at Rice University, and the Texas Higher Education Coordinating Board NHARP 01907. Equipment used to run the experiments presented in this abstract is part of the Shared University Grid at Rice which is funded in part by NSF under Grant EIA-0216467, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.

REFERENCES

[1] Drew H. Bryant, Mark Moll, Brian Y. Chen, Viacheslav Y. Fofanov, and Lydia E. Kavraki. Analysis of substructural variation in families of enzymatic proteins with applications to protein function prediction. *BMC Bioinformatics*, 11 (242), May 2010. doi:10.1186/1471-2105-11-242.

[2] Mark Moll, Drew H. Bryant, and Lydia E. Kavraki. The LabelHash algorithm for substructure matching. *BMC Bioinformatics*, 11(555), November 2010. doi:10.1186/1471-2105-11-555.

[3] Mark Moll, Drew H. Bryant, and Lydia E. Kavraki. The LabelHash server and tools for substructure-based functional annotation. *Bioinformatics*, 27(15):2161–2162, June 2011. doi:10.1093/bioinformatics/btr343.

[4] C A Orengo, A D Michie, S Jones, D T Jones, M B Swindells, and J M Thornton. CATH—a hierarchic classification of protein domain structures. *Structure*, 5(8): 1093–1108, 1997. doi:10.1016/S0969-2126(97)00260-8.

[5] A Shulman-Peleg, R Nussinov, and H J Wolfson. Recognition of functional sites in protein structures. *J Mol Biol*, 339 (3):607–633, June 2004. doi:10.1016/j.jmb.2004.04.012.

[6] H. J. Wolfson and I. Rigoutsos. Geometric hashing: an overview. *IEEE Computational Science and Engineering*, 4(4):10–21, 1997. doi:10.1109/99.641604.