Protein Function Inference Using Structure-Based Fingerprints

Department of Computer Science University of North Carolina at Chapel Hill March 2005

Introduction

Proteins are long chains of amino acids that perform all the functions required for life. In the biological environment, proteins fold to form unique three-dimensional structures that enable them to carry out their function. The function is usually dependent on a few residues coming together in a specific geometric arrangements.

We are developing tools to infer the function of protein structures, using a graph representation of protein structure[1,2] and subgraph mining[1] on families of proteins from classifications such as SCOP, EC and GO. We can find subgraphs that occur in almost all members of the family, and are rare in the rest of the Protein Data Bank, called *the background*. We call these subgraphs *family-specific fingerprints*, since taken together they uniquely identify the family.

Structural Genomics projects have produced many new structures with unknown function from proteins encoded in the fully-sequenced genomes. Fast automated methods are needed to infer their function.

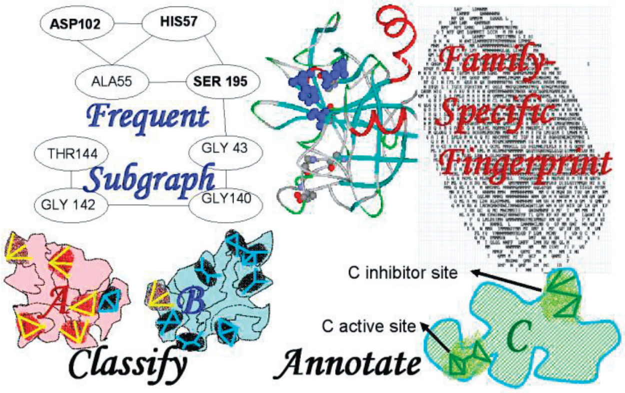

Figure 1: Schematic overview of function inference

Method

We present a method for functional inference based on protein family fingerprints (spatial motifs) and a fast graph search method. Given the structure of a protein to be annotated, we search within it for occurrences of the family fingerprints of some protein families that it is suspected to belong to, and make a family assignment based on occurrences found. Our fingerprints can identify the function where sequence alignment and overall structural alignment with known proteins are ambiguous or absent.

Highlights

- **Method to identify function of proteins with unknown function**
- **Based on local structural features that characterize a protein family, known as fingerprints**
- **Fast fingerprint computation using frequent subgraph mining, and fast retrieval using index of graph similarity**

Our function inference method has the following steps:

1. Select families of non-redundant proteins from any level of a classification such as SCOP or EC, or manually.
2. Represent protein structures as graphs, with nodes at each residue, and contact edges defined between residues using the almost-Delaunay[2] edges, which are sparse and robust enough to find complex patterns quickly in the presence of coordinate perturbations.
3. Mine family-specific fingerprints using the Fast Frequent Subgraph Mining method[1]. Fingerprints are defined to occur in at least 80% of the family (support), and at most 5% of the background (background occurrence).
4. Search for fingerprints in a new structure, using an index of graph similarity to speed up Ullman's subgraph isomorphism algorithm.
5. Assign a significance to the function inference by counting the fingerprints found; more fingerprints indicate a higher probability that the protein belongs to the family.

Results : Identifying and Distinguishing SCOP Families

We searched the background (6,500 proteins) using 79 trypsin-like Serine Protease fingerprints, the largest of which is shown in upper left corner of Figure 1. Based on the number of fingerprints found, we annotated several proteins as serine proteases that were not annotated in SCOP. Our inference was validated by the literature.

New annotations (missing from SCOP 1.65):
Trypsin-like Serine Proteases: (from 79 FP)
1op0A (73); *1os8A*(73); *1p57B*(73);
1s83 (73); *1ssx* (46); *1md8* (45).
Triosephosphate Isomerase: (from 1920 FP)
1r2r (1885)

In another study, we were able to use fingerprints mined from triosephosphate isomerase, xylose isomerase, alcohol dehydrogenase and amylase families within the TIM fold to mutually distinguish these families, which have strong overall structural overlap. Also, we could distinguish all the proteins with TIM fold from each one of these families.

Results: Function Inference for Structural Genomics

Next, we applied the method to suggest function assignments for Structural Genomics proteins in the PDB with no known function. For example, the Ycdx protein (PDB: 1m65, CASP5 target T0147) has a rare seven-stranded $\beta\alpha$-barrel fold, and no significant sequence or overall structure similarity with proteins of known function. We inferred this protein to have a Metallo-Dependent Hydrolase (MDH) function, as shown in Figure 2.

We mined 49 fingerprints from 17 non-redundant members of the MDH family in SCOP, which adopt the 8-stranded $\beta\alpha$ TIM barrel fold. These fingerprints were concentrated in the metal binding site and an adjoining region, as shown in the top row (subgraph view) and middle row (residue surface view). We found 30 out of 49 fingerprints in the Ycdx protein, suggesting that it is a Metallo-Dependent Hydrolase (one fingerprint shown in bottom row). Our inference was validated by active site and ligand template matches, but still needs final validation by experimental determination of function.

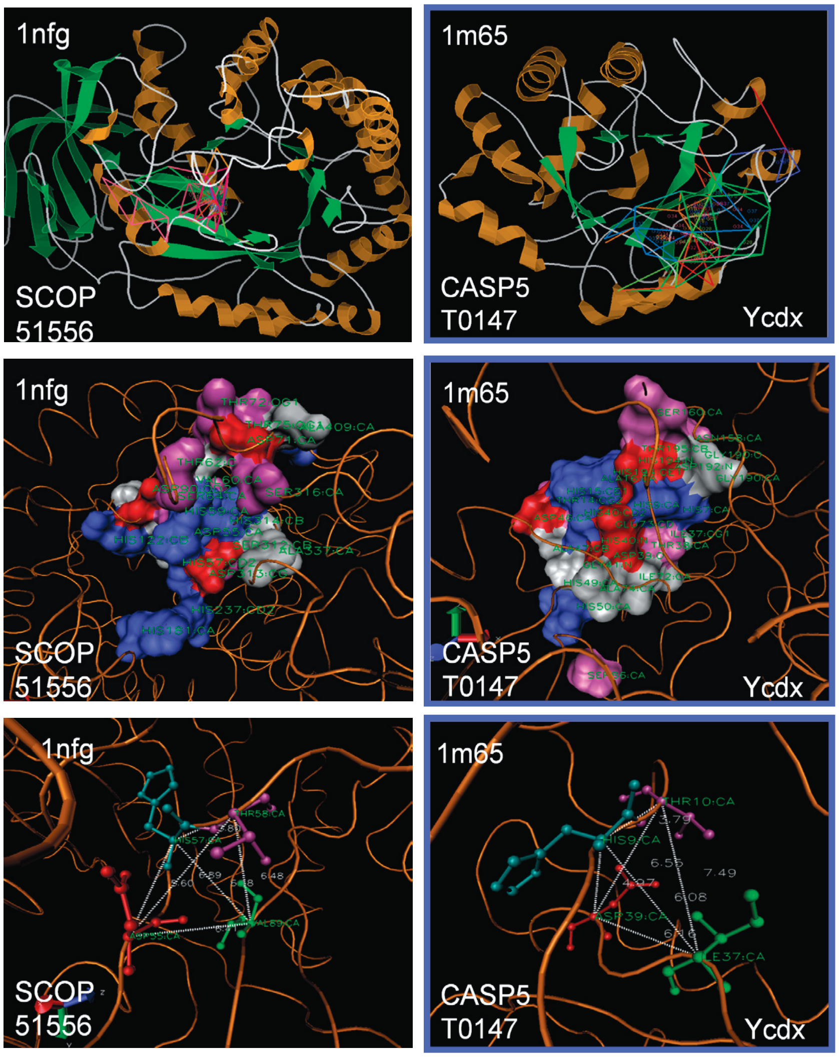

Figure 2: Identifying function of the Ycdx protein as Metallo-Dependent Hydrolase using fingerprints

Ongoing Research

Our method is robust enough to infer the functional family of predicted structures. In particular, we can check for membership of a homology modeled structure in the template structure's family, to determine if a different template should be chosen.

By searching for fingerprints within the background, and clustering the proteins found based on their enrichment in certain nodes of the SCOP or GO hierarchies, we can find *functional neighbors* of a family, i.e. families that are structurally dissimilar but may have related function. Functional neighbors may help characterize entire families of proteins with unknown function.

Our method is also applicable to function prediction at the sequence level, based on predicted structures and on sequence patterns derived from structural fingerprints that are conserved in sequence.

Project Leaders

Jack Snoeyink, professor (Computer Science)
Jan Prins, professor (Computer Science)
Wei Wang, professor (Computer Science)
Alexander Tropsha, professor (Pharmacy)

Graduate Research Assistants

Deepak Bandyopadhyay
Jun (Luke) Huan
Jinze Liu

Research Sponsors

National Science Foundation (Biogeometry)

References

[1] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, A. Tropsha. "Mining Spatial Motifs from Protein Structure Graphs." RECOMB 2004. Invited to Journal of Computational Biology, 2005.

[2] D. Bandyopadhyay and J. Snoeyink. "Almost-Delaunay Simplices : Nearest Neighbor Relations for Imprecise Points." ACM-SIAM SODA 2004.

For More Information

http://www.cs.unc.edu/~debug/papers/RECOMB/subgraph.html