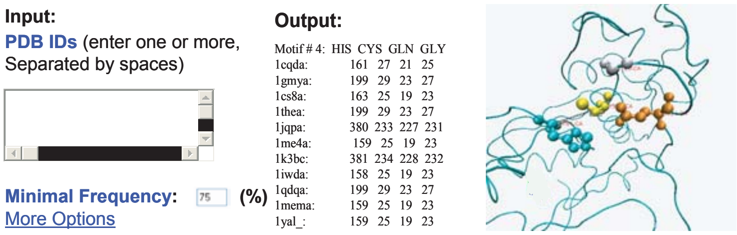
Rapid Determination of Local Structural Features Common to a Set of Proteins

Jun Huan[1], Deepak Bandyopadhyay[1], Jinze Liu[1], Jan Prins[1]
Jack Snoeyink [1], Alexander Tropsha [2], Wei Wang[1]

Traditionally, protein structure comparison has focused on *global similarity* between two structures. Recent research has focused on finding local structural features in common among a group of proteins. Such shared features are called *spatial motifs* and correspond to amino-acid packing patterns that may be implicated in function shared among the proteins in the group [4, 5, 6]. Searching for spatial motifs shared among multiple proteins yields fewer spurious results and improved statistical significance of features than those found using pairwise analysis. However, the basic techniques involved are considerably more complex.

We propose to demonstrate the current state of our efforts on this problem. Our most recent implementations locate shared spatial motifs among a group of several dozen protein structures in tens of seconds. The motifs are sequence order independent and may occur in every member of a group of proteins or a significant fraction of them (as specified by a threshold parameter). The spatial motif matching process accommodates variation inherent in structure determination. With our current software we expect to be able to provide real time responses to queries submitted by users in the ISMB demo session. Our command line and simple Graphics User Interface (GUI) shown in 1 is being extended. We intend to present a web-based interface with a fully integrated GUI to our server that implements our algorithms and provides access to PDB structures and previously determined spatial motifs.

The kernel of our software package is a subgraph mining algorithm that detects all frequent subgraphs from a graph database with a user specified minimal frequency. Our algorithm uses the pattern growth paradigm [3] with an efficient depth first enumeration scheme, searching through the graph space for frequent subgraphs. The recent algorithm incorporates several improvements by taking into account the properties of protein 3D structural graphs, searching only for maximal subgraphs, and incorporating constraints about interesting motifs [1, 2]. Using the tool, we are able to locate common functionally-correlated motifs from proteins with different global structures, such as a NAD binding motif from proteins with different folds [1], which are hard to be identified using sequence or global structure comparison. The algorithm FFSM [3] is written in C++ and compiled and tested in the Linux environment. This software is freely downloadable from http://www.cs.unc.edu/~huan/FFSM.html. We will demonstrate an improved version, CliqueHashing, which will soon be released at the same web site, with the improvements we discussed.



Input:

PDB IDs (enter one or more, Separated by spaces)

Minimal Frequency: 75 (%)

Output:

```
Motif # 4:  HIS  CYS  GLN  GLY
1cqda:      161   27   21   25
1gmya:      199   29   23   27
1cs8a:      163   25   19   23
1thea:      199   29   23   27
1jqpa:      380  233  227  231
1me4a:       159   25   19   23
1k3bc:      381  234  228  232
1iwda:      158   25   19   23
1qdqa:      199   29   23   27
1mema:      159   25   19   23
1yal_:      159   25   19   23
```

Figure 1. The input of the algorithm is a list of PDB IDs, (with chain numbers), and the minimal frequency required for spatial motifs. We output a text file recording the motifs and residues they may be mapped to in a protein. A visualization is generated to display a given motif as a collection of amino-acid residues highlighted within a representative protein containing the motif.

References

[1] J. Huan, D. Bandyopadhyay, J. Liu, J. Prins, J. Snoeyink, A. Tropsha, and W. Wang. Identify spatial motifs from protein families. *UNC Technical Report*, 2005.

[2] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. Mining protein family specific residue packing patterns from protein structure graphs. *RECOMB*, 2004.

[3] J. Huan, W. Wang, and J. Prins. Efficient mining of frequent subgraph in the presence of isomorphism. *ICDM*, 2003.

[4] M. Shatsky, A. Shulman-Peleg, R. Nussinov, and H. Wolfson. Recoginition of binding patterns common to a set of protein structures. *to appear in RECOMB*, 2005.

[5] A. Tropsha, C. Carter, S. Cammer, and I. Vaisman. Simplicial neighborhood analysis of protein packing (SNAPP) : a computational geometry approach to studying proteins. *Methods Enzymol.*, 374:509–544, 2003.

[6] P. Wangikar, A. Tendulkar, S. Ramya, D. Mali, and S. Sarawagi. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol*, 326(3):955–78, 2003.

[1] Computer Science Department, University of North Carolina, {huan, debug, liuj, prins, snoeyink,weiwang}@cs.unc.edu

[2] The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, University of North Carolina at Chapel Hill, tropsha@email.unc.edu