MotifSpace: Mining Patterns in Protein Structures

Department of Computer Science | University of North Carolina at Chapel Hill | March 2010

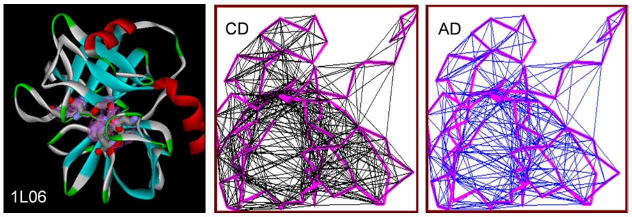
The Challenge

A protein is a long sequence of amino acids that fold into a unique 3D shape. A central tenet of modern biology is that a protein's function is determined by its structure. Often local substructures within the protein determine its function. These substructures, composed of a small number of amino acid residues, often have conserved spatial arrangements across a group of proteins of the same function, and are referred to as spatial motifs. The challenge is to develop automated techniques to identify spatial motifs in proteins.

The Approach

MotifSpace is a collection of computational methods for discovering, cataloging, querying, and visualizing spatial motifs. These tools can be used to study the locations of spatial motifs within protein structures, to build predictive models for protein function inference, and to construct hypotheses to guide the design of biological experiments.

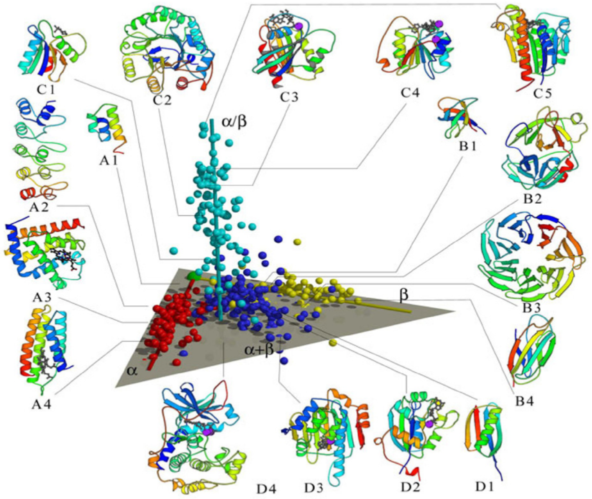
We represent a protein's structure using a *labeled multigraph* and detect spatial motifs by searching for common subgraphs among a group of protein graphs. In our representation, a node abstracts an amino acid residue in a protein structure with the amino acid type as the node label. An edge connects two amino acid residues and is labeled by (1) the discretized Euclidian distance between the two amino acid residues, and (2) the potential interaction between the two amino acid residues. A spatial motif corresponds to a subgraph where edges are labeled by distance intervals that allow some perturbation of the amino acids to account for dynamics and uncertainty in structure determination.



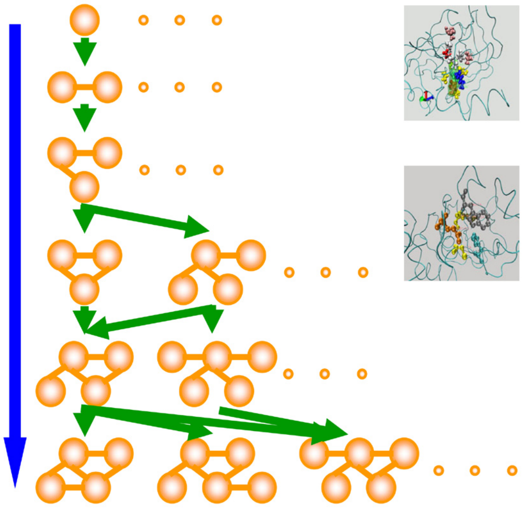
Protein structures can be represented by three different graph representations.

Highlights

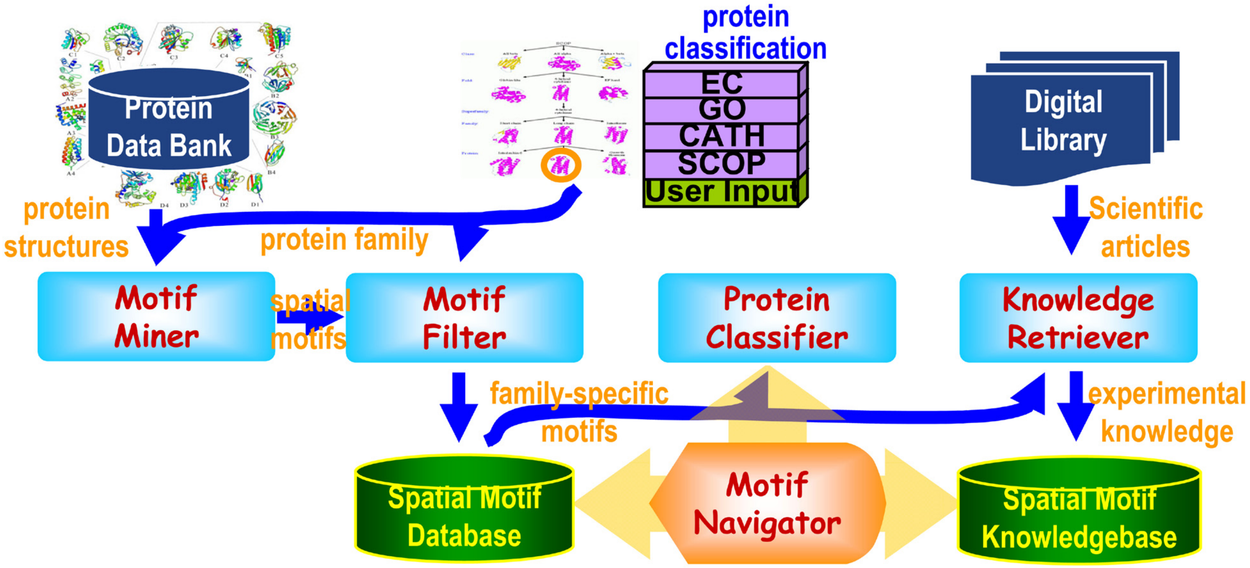
- **A database and user interface to organize, query, and visualize spatial motifs.**
- **A feature selection method to determine the most significant motifs.**
- **A classification system to predict protein function from motifs present in the protein structure.**
- **A knowledge retrieval agent to automatically retrieve information associated with spatial motifs from the literature.**



Proteins are classified into classes/folds/superfamilies/families by their global structure (figure from http://www.nigms.nih.gov/psi).



Left: an enumeration of sub-/super-graphs that form a lattice. Right: patterns identified from serine protease (top) and papain-like cysteine protease (bottom).



The architecture of MotifSpace

We have developed a **frequent subgraph mining** method to search all subgraphs that appear in at least a fraction of members in a group of graphs. For protein graphs, such subgraphs represent spatial motifs. As a proof of concept, we locate patterns with known biological functions such as the catalytic triad in serine protease, the catalytic dyad and the hydrophobic binding pocket in papain-like cysteine protease, the ligand binding sites in nuclear binding domains, and the co-factor binding sites in NADP binding proteins.

We identified more than six million spatial motifs from thousands of representative proteins in the Protein Databank (PDB). MotifSpace will provide an integrated view of information and knowledge of spatial motifs.

Participating Faculty

Wei Wang, Assistant Professor and Director
Jan Prins, Professor
Jack Snoeyink, Professor
Alex Tropsha, Professor (School of Pharmacy)

Graduate Research Assistant

Ning Jin

Undergraduate Research Assistant

Calvin Young

Research Sponsors

National Science Foundation, National Institute of Health, Microsoft

Selected Publications

J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha. "Comparing Graph Representations of Protein Structure for Mining Family-Specific Residue-Based Packing Motifs," *Journal of Computational Biology* (JCB), 2005.

J. Huan, W. Wang, J. Prins, and J. Yang. "SPIN: Mining Maximal Frequent Subgraphs from Graph Databases," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 581-586, 2004.

J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. "Mining Family Specific Residue Packing Patterns from Protein Structure Graphs," in *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology* (RECOMB), pp. 308-315, 2004.

J. Huan, W. Wang, A. Washington, J. Prins, R Shah, and A. Tropsha, "Accurate Classification of Protein Structural Families using Coherent Subgraph Analysis," in *Proceedings of the Pacific Symposium on Biocomputing* (PSB), pp. 411-422, 2004.

For More Information

Wei Wang
Phone: (919) 962-1744
Fax: (919) 962-1799
E-mail: weiwang@cs.unc.edu
http://www.cs.unc.edu/~weiwang