LFM-Pro: a tool for mining family-specific sites in protein structure databases

Ozgur Ozturk, Ahmet Sacan, Hakan Ferhatosmanoglu, Yusu Wang
The Ohio State University

Currently available structural comparison methods are both computationally expensive and fail to detect biologically significant local structural features. Developing better methods to generate highly representative and compact signatures is a crucial step in designing scalable and accurate data mining systems for proteins. We propose LFM-Pro (Local Feature Mining in Proteins) as a framework for automatically discovering family specific local sites and the features associated with these sites. Our method uses the distance field to protein backbone atoms to detect geometrically significant centers of the protein structure. A feature vector is generated from the geometrical and bio-chemical environment around these centers. These features are then tested for their ability to distinguish a family of proteins from a background set of unrelated proteins, and successful features are combined into a representative set of features for the protein family. The utility and success of LFM-Pro are demonstrated on Globins family and Serine/Threonine family of proteins.

Motivation

➢ Protein structure can provide valuable information about biochemical function or evolutionary relationship of proteins.

➢ The increasing size of structure databases presents a processing challenge. The classification and analysis of the new protein entries is still primarily a manual task, and there is need for automated methods of structural annotation.

➢ Proteins generally consist of a small, functional site that provides the specific biochemical function, and scaffold residues that form the structural environment.

➢ Structure is more conserved than sequence: functionally important sites resist changes due to selective advantage, whereas scaffold residues can accommodate changes.

➢ Automated identification of functionally important sites in proteins can have a great impact on protein classification, protein function prediction, and protein folding.

Background

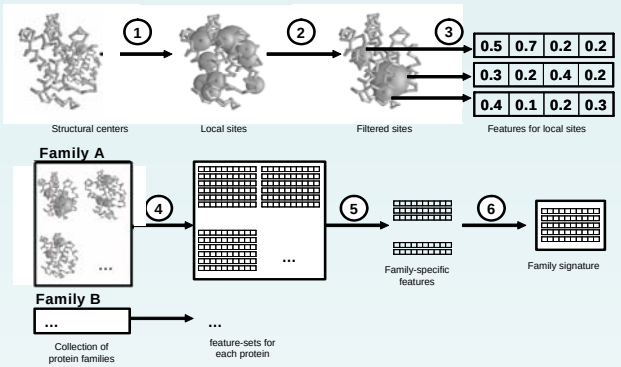
➢ Earlier studies have focused on dedicated cataloguing and analysis of specific protein families, like metal-binding proteins, or G-protein coupled receptors.

➢First attempts of structural motif extraction have assumed prior knowledge about the location or nature of the functional sites.

➢ Currently available methods rely only on a family classification, and do not require prior knowledge of the functional sites.

➢ Structural motif search is generally based on graph-theoretic methods or geometric-hashing. These methods are slow and do not allow large-scale motif search, and are sensitive to noise in the location or type of the protein residues.

Overall Framework



1. Local Structural Centers: location of the ***critical points*** of distance field to backbone atoms are identified,

2. Filtering: the critical points are filtered based on ***topological persistence*** and trivial secondary structures they capture,

3. Local Features: a feature vector that captures the ***topological and bio-chemical properties*** of its spatial neighborhood is associated with each critical point.

➢ Topological properties: persistence of the critical point, volume of the neighborhood, and writhing value of the contained backbone.

➢Bio-chemical properties: the density and center of mass for the side-chain Carbon, Nitrogen, Oxygen, and Sulfur atoms.

4. Feature vectors for the remaining critical points of each protein in the dataset are pooled and

5. Selection of discriminative features: Features that best discriminate the family members from the rest of the proteins are identified, and associated with a *degree of representativeness*.

6. Representative Feature Set: The representative feature set is obtained as a sum of the discriminative features, weighted by their degree of representativeness.

Critical Points

➢Critical points of distance function give four types of motifs: minima, maxima, and two types of saddle points.

➢In maxima motif (a), four pieces of protein backbone come close in space, forming a contact as indicated by the tetrahedron in the middle. In motif (b), the cross-point is a saddle point.

➢Local spatial patterns can be captured by taking a ball centered at these critical points.

Experiments

➢A representative ASTRAL [10] database with less than 40% sequence identity was used as the primary source of proteins.

➢ LFM-Pro was tested on two SCOP [11] families: Globins (a.1.1.2) and Serine/Threonine Kinases (d.144.1.1).

➢The dataset contained a total of 200 proteins: 10 proteins from each of these families, and an additional 180 randomly selected proteins to provide a background collection of structural features.

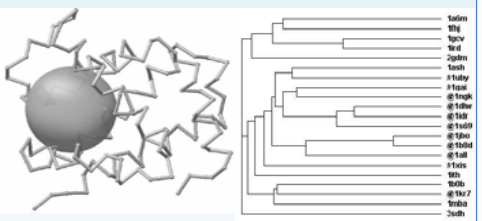
Results

➢ The sites discovered by LFM-Pro matched with the known functional sites for both protein families.

➢ The extracted signatures were tested for their ability to distinguish family members from among the rest of the proteins, using a cross-validation test:

Family	Precision	Recall
Globins	0.79	0.90
Globins (extended)	1.00	0.90
Ser/Thr Kinases	1.00	0.90
Negative Random Control	0.05	0.50

Case Study: Globins



➢ LFM-Pro's top scoring local site for the Globins family is located in the functional pocket of the protein responsible for binding the heme group.

➢The proteins 1uby, 1gai, and 1xis were not in the original SCOP family, but were detected to have the Globin family signature. Multiple sequence alignment reveals that these proteins are in fact close relatives of the Globin family proteins.

Selected References

1. Chakraborty, S., and Biswas, S. 1999. Approximation algorithms for 3-D common substructure identification in drug and protein molecules. Workshop on Algorithms and Data Structures, 253–264.
2. S. C. Bagley and R. B. Altman. Characterizing the microenvironment surrounding protein sites. Protein Sci, 4:622–635, 1995.
3. A. C. Wallace, R. A. Laskowski, and J. M. Thornton. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci, 5(6):1001–1013, 1996.
4. R. V. Spriggs, P. J. Argymiuk, and P. Willett. Searching for patterns of amino acids in 3D protein structures. J Chem Inf Comput Sci, 43(2):412-421, 2003.
5. J. Huan, et. al. Mining spatial motifs from protein structure graphs. In 8th International Conference on Research in Computational Molecular Biology (RECOMB), pages 308-315, 2004.