

Structure Based Identification of Protein Family Signatures for Function Annotation

Ruchir Shah¹, Jun Huan², Wei Wang², Alexander Tropsha¹

Keywords: Structure based annotation of protein sequences.

1 Abstract

We present a novel approach to identifying recurrent structure-sequence motifs common to particular protein families. The approach employs Graph representation of proteins and frequent subgraph mining to obtain the motifs. We demonstrate the utility of these motifs for highly accurate annotation of several protein families.

2 Introduction

Protein sequences are known to evolve much faster than structures. Pairs of distantly homologous proteins with very low sequence similarity may still share very similar folds and functions. One example of well-preserved structural motifs is given by specific three-dimensional orientations of amino acid residues in the active site of a protein, which are responsible for protein's biochemical functions. However, most of the time active site residues are hard if not impossible to infer from the primary sequence since the active site residues are frequently quite distant from one another. Developing automated approaches to identifying protein family signatures can be very useful in protein family classification and function annotation.

We present a novel automated approach to identifying recurrent structure-sequence motifs common to particular protein families. We use a labeled graph to represent the structure of a protein where each amino acid residue is represented by a C α node. Two independent approaches were used to generate edges. In the first approach two nodes are connected by an edge if the two residues are either connected by a peptide bond or the distance between them was below a certain threshold in the absence of a peptide bond. Whereas in the second approach the edges result naturally from applying computational geometry approach, i.e., Delone tessellation to C-alpha nodes in 3D space. We then employ a fast and efficient frequent subgraph mining algorithm^{[1],[2]} developed in our group to extract common packing patterns (i.e., subgraphs) from a graph family corresponding to a protein family in the SCOP database. Finally, we examine these common substructural packing patterns for the presence of conserved sequence patterns which thereby represent family specific structure-sequence signatures similar to *PROSITE* motifs. These signatures are used to query protein sequence databases and the hits (i.e., sequences that contain the signatures) are predicted to belong to the same structural/functional class as the protein family from which the signatures were derived.

The Serine Protease and Nuclear Receptor Binding protein families were used as test cases. A diverse subset of structures belonging to a given family was selected from the SCOP database such that the pair-wise sequence identity between any two proteins in the set was below a certain threshold. Subgraph mining was performed to extract recurring structural motifs specific to the given family. Motifs were investigated for the location of residues in the three dimensional structures by superposing structures in SwissPDBviewer. Furthermore, amino acid residues in the motifs were

¹ The Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. E-mail: ruchir@email.unc.edu

² Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

analyzed for their order in the primary sequence and the lengths of the loops separating residues in the primary sequence were calculated.

Function-specific sequence signatures were derived and used to search the Swissprot sequence database. A protein sequence was called a *hit* if it contained the query functional signature and hits were analyzed for *precision* (True Positive/[True Positive + False Positive]) and *recall* (True Positive/[True Positive + False Negative]) values. Results for functional signatures derived from trypsin like serine protease are summarized below; for comparison, we have conducted similar searches using known *PROSITE* motifs for the same family.

In addition, we used two above motifs and two *PROSITE* motifs simultaneously for sequence searches. Here, a protein was considered a *hit* if it contained at least one of these motifs. Searching with the pair of motifs discovered by our approach yielded 93% precision and 91% recall; whereas the pair of *PROSITE* motifs gave 93% Precision and 95% Recall. It is interesting to note that, the query using our motifs identified five trypsin like serine proteases (Swissprot IDs: VSP2_TRIFL, VSP4_AGKAC, VSPA_LACMU, VSP1_AGKRH, VSP2_AGKRH) that were missed by the query using the *PROSITE* motifs.

The results indicate that our structure based approach is able to annotate protein sequences with accuracy comparable to that of other sequence-based methods such as *PROSITE*. However, our motifs appear more compact (see Table 1) and are derived from much smaller set of proteins with known structure. In some cases we were able to correctly predict the protein functions that *PROSITE* was unable to annotate. This structure-based approach is currently being extended to all major protein families in the SCOP and similar databases. The results of this study shall provide a diverse set of family specific signatures that can be used for the genome-wide annotation of protein functional families.

3 Figures and Tables

| Motif | Precision | Recall |
|-----------------------|-----------|--------|
| C-G-x{11}-A-A-H-C | 100% | 75% |
| D-S-G-G-P | 93% | 90% |
| <i>PROSITE</i> | | |
| PS00134* | 95% | 88% |
| PS00135** | 98% | 88% |

*PS00134: [LIVM]-[ST]-A-[STAG]-H-C

**PS00135:[DNSTAGC]-[GSTAPIMVQH]-x{2}-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH]

Table 1: Comparison of our structure based sequence annotation approach with PRSOTE motifs.

4 References

[1] Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J. and Tropsha, A. 2004. Mining spatial motifs from protein structure graphs. In: *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*, pp. 308-315.

[2] Huan, J., Wang, W., Washington, A., Prins, J., Shah, R. and Tropsha, A. 2004. Accurate classification of protein structural families using coherent subgraph analysis. In: *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pp. 411-422.