

Mining 3D-Motifs using physical-chemical constraints: application to Cardiolipin binding sites

Dmitrii Polshakov^{1,2}, Keith Marsolo² and Srinivasan Parthasarathy^{2*}

¹Department of Chemistry,

²Department of Computer Science,

The Ohio State University, Columbus, OH, USA

Motivation:

Biologically-relevant structural motifs can provide important clues to a protein's functionality. The detection of such motifs is an open problem. Here we report a new approach toward the discovery of biologically-meaningful structural motifs in proteins.

Approach:

Our work is an extension of the MotifMiner algorithm, which was designed to find frequent structural motifs in macromolecules. It is an extensible toolkit and allows for the detection of motifs at varying resolutions, ranging from the atomic scale to the secondary structure level. The interactions between objects (atoms, residues, secondary structures, etc.) are represented as a set of *mining bonds*. Substructures (or motifs) with the same set of mining bonds are said to be equivalent. By determining the count of all equivalent substructures, MotifMiner can return the set of motifs that occur above a certain user-specified frequency threshold. Rather than finding *all* frequent motifs, here we limit our search for those motifs that are biologically-meaningful. We use the principles of the MotifMiner algorithm, but add several extensions to achieve our task. Those extensions are as follows:

1. The definition of the mining bond is extended to take into account the distance between amino acids in the physical-chemical space. As a result, we do not need to use specific labels to represent the physical-chemical properties of the amino acid side chains and are not affected by the problems inherent in clustering those properties. This extension also serves to reduce the search space, in some cases by more than two orders of magnitude over the previous version.
2. The *Recursive Fuzzy Hashing* technique, which was designed to handle the noise inherent in X-ray crystallization data, is replaced with a flexible resolution scheme that is more appropriate for motifs at the amino acid level.

3. A novel template-based support-counting scheme that provides a new approach to constrained structural mining is implemented. This scheme is particularly important for discovering local motifs such as protein-lipid binding sites.

Results:

1) Zinc Finger Proteins: We tested the ability of our algorithm to find specific structural motifs in the set of 36 zinc finger proteins (obtained from the Structural Classification of Proteins [SCOP] Database). All of these proteins contain the C2H2 zinc finger. This motif has very specific physical-chemical and structural properties that facilitate the binding of the zinc ion. Typically, zinc finger proteins contain multiple zinc binding sites. Therefore, our algorithm should be able to detect such a location. The key features of this motif include the presence of a pair of Cys residues separated from a pair of His residues. Our algorithm found such motifs to be frequent across all proteins in dataset. These motifs contain both the Cys and His residues found in the classical zinc finger.

2) Binding Site of Cardiolipin (CL) Head Group: We also performed the first structural search on a subset of the membrane proteins containing the lipid Cardiolipin (CL), in an attempt to discover specific protein-CL binding sites. This set consists of six structures: 1KB1, 1KQF, 1M3X, 1OKC, 1V54, 1OGV. Several of them contain more than one CL molecule.

In order to find the structural motifs that can serve as a binding site for a Cardiolipin head group, we used one of the structures from our dataset and extracted the local space region in the vicinity of each head group. Since CL is a dianionic phospholipid, we selected the region containing all the residues within 10 Å of the corresponding phosphorus atoms. We then used this selected region as the basis for a template structure, counting the support of motifs in the remaining 5 structures. If a motif was found in the template structure and one other protein, we denoted the motif as frequent.

We found that the frequent structural motifs located within 10 Å of the CL head groups primarily contain a combination

*To whom correspondence should be addressed

of polar (N, T, S), charged (R, E) and aromatic (Y, W) residues. This agrees with previous results, where a short motif (2-3 residues) that contained a combination of one polar and two charged residues was proposed to be the head group-binding motif. It has also been noted that a structural search would yield a more adequate motif architecture since the residues involved in the interactions with the head groups are not sequentially connected. Even more, it is possible for residues from different subunits to contribute to the composition of the protein-lipid binding sites. Since our structural approach does not require the residues in the motif to be in sequence order, many of the motifs we find are non-sequential.

Using our approach we found that the nearest neighbors to the lipid head group always contain polar residues while some motifs are altered with charged residues. The motifs that were found to have an identical set of mining bonds to the motifs from the template structure always appeared at the potential lipids binding sites. Cardiophilin molecules resolved

by X-ray crystallography commonly orient their head groups toward either the positive or negative leaflet of the lipid bilayer, which favors stabilization by the polar and charged residues commonly found at the membrane boundary. In addition, such orientation of the lipid head groups favors the access of water, which provides additional stabilization due to hydrogen bond interaction.

Although all of the proteins from our dataset contained resolved Cardiophilin molecules, not all of the motifs we found had a CL head groups in the vicinity. Since the membrane protein crystallization process involves severe delipidation in order to make the protein soluble, some of the CL molecules can be lost. Nonetheless, the motifs that the algorithm finds can be used to predict the approximate location of CL in the membrane proteins where crystallization under harsh delipidation caused the lipids to be completely or partially eliminated.