Structural Classification of Phosphate Binding Sites in Protein-Nucleotide Complexes: An Automated All-Against-All Structural Comparison Using Geometric Matching

A. Brakoulas and R.M. Jackson

School of Biochemistry and Molecular Biology
University of Leeds

Analysis of completed genomes from a number of organisms reveal that about half of all gene products are classified as functionally unknown hypothetical proteins, and structures are already being solved prior to any knowledge of function (1,2). This will make the structure-based prediction of molecular function of increasing importance in the future. The ability to detect and classify local atomic level similarity will allow the development of rules for the prediction of function directly from structure (the premise of Structural Genomics).

The determination of protein function from sequence and structure is major goal of bioinformatics techniques. In many cases sequence patterns are sufficient to identify function. The idea of using sequence patterns and profiles underlie the PROSITE(3), PRINTS(4) and Pfam(5) methodologies. However, structure can often provide a more powerful evolutionary link than sequence alone. The structural classification schemes of protein families (e.g. the CATH, SCOP, DALI databases) have proved invaluable tools for structural biologists frequently allowing functional inferences to be made between members of a family even when sequence similarity is statistically low. Furthermore, the incorporation of both sequence and structure into functional pattern recognition has been used to identify the recurrence of short motifs (often in different folds) that infer function, such as the phosphate binding P-loop (6) or the DNA binding helix-hairpin-helix (7). More recently Kasuya & Thornton (8) identified maintenance of 3D structure in PROSITE patterns. This suggests that many sequence motifs do infer common function.

Recently, we have identified a conserved main-chain pattern in canonical serine protease inhibitors where there is no sequence conservation (9). This main-chain fragment was used to search the structural database for other potential canonical serine protease inhibitors. Clusters of hits were found in several extracellular proteins including hydrolases, toxins, cytokines and viral proteins as well as the known canonical serine protease inhibitors.

Here we investigate the construction of a structural classification scheme for protein-ligand binding sites in the large class of nucleotide phosphate-binding proteins in an attempt to quantify common structural features that confer specific ligand binding properties in different binding sites. The method we have used exploits common atomic features through identification of the maximum clique. These features are not necessarily evident from sequence or from global structural similarity giving additional insight into molecular recognition not evident from current sequence or structural classification schemes. Another motivation for this work is that such a classification scheme could be used as a resource for creating structural templates that can be used to screen for binding sites in structurally determined yet functionally unknown hypothetical proteins that will arise from structural genomics initiatives.

Initially, proteins that bind ligands containing the phosphate moiety are being studied. In particular, the large class of nucleotide ligands (ATP/ADP, GTP/GDP, NAD, FAD etc.).

Figure 1. The local protein-ligand atomic environments are extracted from the PDB for proteins that bind phosphate containing ligands.

We have developed a method that identifies the extent of the similarity between the structure of different protein-ligand binding sites allowing their structural classification (10) based on the method of geometric hashing (11).

The Structural P-loop (Nucleotide Mono-Di-Tri Phosphate)

Cluster output (K=6.0, there are 4 clusters)

- Cluster has 32 separate group representatives (from total of 476)
- Ligand ATP > ADP > GDP > GTP

CATH/SCOP-fold (domain) PROSITE:

39 P-loop nucleotide triphosphate hydrolases 30 PS0017 - consensus pattern G-K-[T/S]

Figure 2. Following an all-against-all comparison of nucleotide phosphate binding sites group representatives are clustered to identify similar sites. This figure shows one such cluster (main-chain atoms only) which corresponds to the structural P-loop. The residues are coloured according to residue type. It can be seen that members of the cluster contain an identifiable sequence pattern identified in the PROSITE database (GK-[T/S]). The nucleotide moieties are also shown in the right hand picture.

The classification procedure allows the generation of structural binding site “templates”. These could be used to identify similarity between the binding site of proteins in the database (for which an enzyme mechanism or binding site is well determined) and a newly determined protein of unknown function. Clearly, knowledge that a protein has a nucleotide binding site as well as information on the function of other proteins in the database that share this site will be valuable.

References

- Hoh F., Yang Y.S., Guignard L., Padilla A., Stern M.H., Lhoste J.M., van Tilbeurgh H. (1998). *Structure*, **6**, 147-155.
- Colovos C., Cascio D., Yeates T.O. (1998). *Structure*, **6**, 1329-1337.
- Hofmann K., Bucher P., Falquet L., Bairoch A. (1999). *Nucl. Acids Res.*, **27**, 215-219.
- Attwood T.K., Beck M.E., Bleasby A.J., Parry-Smith D.J. (1994). *Nucl. Acids Res.*, **22**, 3590-3596.
- Bateman A., Birney E., Durbin R., Eddy S.R., Howe K.L., Sonnhammer E.L. (2000). *Nucl. Acids Res.*, **28**, 263-266
- Swindells M.B. (1993). *Protein Sci.*, **2**, 2146-2153.
- Doherty A.J., Serpell L.C., Ponting C.P. (1996). *Nucl. Acids Res.*, **24**, 2488-2497.
- Kasuya A., Thornton J.M. (1999). *J. Mol. Biol.*, **286**, 1673-1691.
- Jackson RM, Russell R.B. (2000). *J. Mol. Biol.*, **296**, 325-334.
- Brakoulas A., Jackson R.M. (2003). Structural comparison of phosphate binding sites: All-against-all comparison (manuscript in preparation).
- Bachar O., Fischer D., Nussinov R., Wolfson H. (1993). *Protein Eng.*, **6**, 279-288.