Matching Volumetric Models of Protein Active Sites

Thomas Funkhouser, Roman Laskowski, and Janet Thornton

SHORT ABSTRACT:

This paper investigates the use of a knowledge-based algorithm to build volumetric models of active site cavities from protein structures, a fast rotational matching algorithm to detect similarities in the volumetric models, and a nearest neighbor classifier to predict the type of ligand bound based on similarities in volumetric models.

LONG ABSTRACT:

The goal of our project is to detect functional similarities between protein active sites. Given the 3D atomic coordinates for a protein, along with the location of an active site, we employ a knowledge-based algorithm based on X-Site [Laskowski96] to build a volumetric model of the chemical and geometric properties inside its cavity. We then use a fast rotational matching algorithm [Kovacs02] to find the correlation between pairs of volumetric models at the optimal rotational alignment. Finally, we use a nearest neighbor classifier to transfer functional annotations between similar active sites.

To test the proposed methods, we performed a leave-one-out classification study aimed at predicting the type of ligand bound in active sites of proteins in the PDB. Our test set comprised 105 active sites of non-homologous proteins partitioned into 6 groups (ATP, FMN, BOG, NAD, FAD, and HEM). We constructed a volumetric model for each site, matched all pairs, and then measured how often the best match for each active site binds the same type of ligand. For comparison sake, we also computed the classification rate achieved with FASTA [Pearson90], (a sequence-alignment program), ICP [Besl92] (a method for aligning atoms in a 15Å sphere surrounding the active site), CE [Shindyalov98] (a structure alignment program), SCOP [Murzin95] (a structural classification), and random (a random classifier).

We found that volumetric matching correctly predicts the bound ligand type for 61% of the active sites, as compared to 50% with ICP, 50% with SCOP, 47% with CE, 46% with FASTA, and 34% with a random classifier. These results suggest that volumetric matching of binding sites is useful for predicting gross categories of molecular function (e.g., bound ligand type). Further study is required to determine if they are also effective at distinguishing more specific molecular functions (e.g., EC number) and whether they can be used to discover the functions of novel proteins for which no function is currently known.

REFERENCES

[Kovacs02] J.A. Kovacs and W. Wriggers (2002). Fast rotational matching, *Acta Cryst.*, D58:1282-1286.
[Laskowski96] R.A. Laskowski et al. (1996). X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins, *J. Mol. Biol*, 259:175-201.
[Murzin95] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol*, 247:536-540.
[Pearson90] W.R. Pearson (1990). Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol*, 183:63-98.
[Shindyalov98] I.N. Shindyalov and P.E. Bourne (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng*, 11:739-747.