Protein Function Prediction by Matching Volumetric Models of Active Sites

Thomas A. Funkhouser,*[1] Roman A. Laskowski,[2] and Janet M. Thornton[2]
[1] Princeton University, Princeton, New Jersey, 08540, USA
[2] European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK
*To whom correspondence should be addressed: funk@cs.princeton.edu

1. INTRODUCTION

The goal of the proposed project is to develop an algorithm for predicting the molecular function of a protein from a structural model of its active sites. Given the 3D atomic coordinates for a novel protein and the location of a ligand binding site, we build a model of the site cavity, match it against a database of active sites models having known molecular functions, and make predictions based on the functional annotations associated with the best matches.

This general strategy is common in structural bioinformatics. The most widely used methods represent protein active sites by sets of points representing atoms, residues, pseudo-centers, templates, and/or surface critical points and match them with algorithms based on exhaustive search, geometric hashing, or association graphs. However, these methods usually rely upon a close similarity in the geometric arrangement of several key residues, and they focus on properties of surface residues rather than of cavities, and thus they can produce false positives when similar arrangements of residues are found in proximity to very different cavities.

In this paper, we represent a protein active site by a set of 3D grids describing the volumetric properties of the void inside its cavity. Given the location of an active site, we employ a knowledge-based method to build a volumetric model of the chemical and geometric properties inside its cavity and a grid matching algorithm to detect similarities to models of other sites. Our motivation is to leverage the fact that properties in the interior of an active site cavity are more likely to be functionally preserved than the properties of any individual residue or property on the protein surface.

2. METHODS

Our method proceeds in two main steps: modeling and matching. During the modeling step, we employ an algorithm based on X-SITE [3] to analyze the 3D atomic structure of a protein and build a grid-based description its active site. During a training phase, the algorithm analyzes proteins from the Protein Data Bank (PDB) with bound ligands and stores the spatial distribution of ligand atoms for every element type (C, N, O, and P) with respect to every amino acid type in the coordinate systems defined by Singh and Thornton [7]. Then, for every test protein, those spatial distributions are resampled to build a volumetric model representing the likelihood of finding a ligand atom of each element type at every position within the active site cavity. The resulting volumetric model is stored on a regularly sampled 3D grid.

During the matching step, we compute the similarity between two active site models by finding the sum of the correlations of grids of the same element type at the optimal relative rotation. To accelerate this step, we use Fast Rotational Matching [2], a method that computes the correlation for a pair of spherical functions at all rotations in the frequency domain. Given two sets of grids, each representing the predicted spatial distribution of ligand atoms for a particular element type within an active site cavity, our method decomposes every grid into a set of concentric spherical shells and decomposes every shell into spherical harmonics. Then, the correlation between the spherical harmonic coefficients is computed for all pairs of shells and the Wigner-D^{-1} transform is used to map the correlations back to the space of rotations. Finally, the maximal correlation found for any rotation is used as our measure of active site similarity. This process finds correlations between two grids at all rotations in $O(N^4)$ time for grids with NxNxN resolution, whereas $O(N^6)$ would be required for a more naïve method. In practice, our implementation runs in less than a second when matching two sets of 64x64x64 grids (0.5Å resolution within a sphere of radius 16Å).

2. RESULTS

To test these volumetric modeling and matching methods for function prediction, we performed a leave-one-out classification study to predict the bound ligand type (e.g., ATP vs. NAD vs. …) of proteins found in the PDB. This task was chosen because it provides a first step towards prediction of molecular function and it is supported by enough data to perform a systematic study over a large number of proteins.

To build our test set, we scanned the PDB and selected all protein active sites with at most 3Å resolution containing at least one bound ligand having at least 20 hetero atoms. We then retained only one example within each homology family of the CATH hierarchy in order to minimize the bias due to evolutionary inter-dependence of our test set. Finally, we grouped active sites by bound ligand type and retained only the groups with at least five examples. This process yielded 105 active sites in 6 groups (ATP, FMN, BOG, NAD, FAD, and HEM). For all of the tested active sites, we constructed models of their active site cavities and matched them using the methods described in the previous section. The rank order of matches for each active site were used in a nearest neighbor classifier to predict the bound ligand type, and the percentage of correct classifications was used to evaluate the method. For comparison, we also computed the classification rate achieved with ranks computed using FASTA [5], (a sequence-alignment program), CE [6] (a structure alignment program), SCOP [4] (a structural classification), ICP [1] (a method for aligning point sets, in this case atoms within 15Å of the active site center), and random (a randomly generated rank order). We also compare to the results achieved by fast rotational matching (FRM) when given an ideal volumetric model derived from the ligands bound in the tested active sites (this test is for comparison only – it does not represent results obtainable in practice).

The table below reports the computational costs and classification rates achieved by the tested matching methods. The first result to note is that our methods for modeling and matching volumetric models of active sites (the top two rows) provide higher classification rates than the others (≥61% vs. ≤50%). However, this improvement comes at extra storage and compute costs. The second result to note is that the fast rotational matching algorithm can achieve very high classification rates (95%) when given an ideal volumetric model of every active site (the top row). This result suggests that developing and testing better methods for modeling the volumetric properties of active site cavities may be the best way to improve our results in the near term.

From this study, we conclude that methods for matching volumetric models of active site cavities can be useful for predicting the coarse molecular function (type of bound ligand) from the structure of a protein in its bound conformation. Further study is required to determine whether similar results can be achieved with proteins in unbound conformations and/or whether these methods can be used to predict the functions of proteins for which no function is currently known.

Matching Method	Algorithm	Storage (bytes)	Match Time (sec)	Classification Rate (%)
Our method (ideal model)	FRM	10^6	1	95%
Our method (X-SITE model)	FRM	10^6	1	61%
Point alignment	ICP	10^3	0.1	50%
Structural classification	SCOP	10^1	?	50%
Structural alignment	CE	10^3	10	47%
Sequence alignment	FASTA	10^2	0.1	46%
Random	-	0	0	34%

Table 1: Comparison of compute costs and classification rates for several protein matching methods.

ACKNOWLEDGEMENTS

The authors would like to thank the NSF, BBSRC, and Leverhulme Trust for funding this project.

REFERENCES

[1] P. J. Besl and N. D. McKay (1992). A method for registration of 3D shapes, *PAMI,* 14(2): 239-256.
[2] J.A. Kovacs and W. Wriggers (2002). Fast rotational matching, *Acta Cryst.,* D58:1282-1286.
[3] R.A. Laskowski, J.M. Thornton, C. Humblet, and J. Singh (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.
[4] 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.
[5] W.R. Pearson (1990). Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol,* 183:63-98.
[6] 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.
[7] J. Singh and J.M. Thornton (1992). *Protein Side-Chain Interactions*, Oxford University Press.