Finding and Extracting Active Site Cavities in Proteins

Thomas A. Funkhouser[1], Roman A. Laskowski[2], and Janet M. Thornton[2]
[1] Princeton University, Princeton, NJ, 08540, USA
[2] European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

SHORT ABSTRACT:

This paper investigates geometric algorithms for locating and extracting active site cavities in proteins. Given a 3D protein structure, volumetric methods estimate the probability of finding a ligand at every position in space, and then sampling and reconstruction algorithms produce shape representations suitable for visualization and matching.

LONG ABSTRACT:

The objective of this project is to develop and analyze algorithms that predict the locations, extract the shapes, and match the properties of protein active site models represented by volumetric models.

For every protein, a model is constructed that describes the shape of the protein with a distribution of the probability that a ligand will be found at every position in the space surrounding the protein. We consider several existing algorithms to build such a model from 3D atomic coordinates, including Ligsite [Hendlich97], Pocketfinder [An04], and Surfnet [Laskowski95], and we provide a new variants and algorithms based on mathematical morphology and consensus strategies. We also investigate the relative advantage of augmenting these algorithms with weights derived from residue conservation estimates [Glaser06].

Given such a model represented on a 3D grid, we: 1) predict the locations of ligand binding sites by sampling the probability distribution; 2) predict the shapes of bound ligands by extracting volumes or surfaces enclosing regions with highest probability; and, 3) predict functional similarities between cavities by matching the extracted shapes. In particular, we provide algorithms for sampling a probability distribution to estimate the locations of bound ligands and for estimating the volumetric extent of ligands in predicted active site cavities (as in [Kahraman07]).

We have evaluated these methods using two sets of proteins selected from the PDB. The first set contains PDB files with bound ligand(s), while the second set contains similar proteins without bound ligands (every protein in the first set has a counterpart in the second set with high sequence similarity and no bound ligands). For every protein in these two sets, we constructed a probabilistic model of the free-space around the protein structure and used it to estimate ligand binding site locations and shapes. We then measured the quality of the model by plotting precision vs. recall of the volume enclosed by bound ligands (using an alignment with proteins in the first set to estimate the positions of bound ligands in the second set). We also provided the extracted models as input to a shape matching algorithm to predict functional similarities (e.g., bound ligand types). The results show that the proposed model is effective at localization, shape extraction, and matching active sites in relation to previous work.

ACKNOWLEDGEMENTS

This work was funded by the National Science Foundation (IIS-0612231).

REFERENCES

[An04] An, J., Totrov, M., Abagyan, R. (2004) Comprehensive Identification of ``Druggable'' Protein Ligand Binding Sites. Genome Informatics. 15(2):31-41.

[Glaser06] Glaser F., Morris R.J., Najmanovich R.J., Laskowski R.A., Thornton J.M. (2006)
A method for localizing ligand binding pockets in protein structures. Proteins. 62(2): 479-88.

[Hendlich97] Hendlich, M. Rippman, F., and Barnickel, G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.” J. Mol. Graph., 15:359-363.

[Kahraman07] Kahraman, A., Morris, R.J., Laskowski, R.A. and Thornton, J.M. (2007).
Shape variation in protein binding pockets and their ligands. J Mol Biol 368:283-301.

[Laskowski95] Laskowski, R.A. (1995) Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph, 13:323-330.