PDBinders: a new method for binding site prediction in protein structures

Bianchi V, Gherardini PF, Helmer-Citterich M, Ausiello G

Motivation

The identification of ligand-binding sites in protein structures is an open problem in functional annotation and can help increasing the efficiency of molecular docking and de novo drug design. The aim of binding site prediction is the detection of specific regions in uncharacterized proteins that can be associated to the ability of binding ligands of biological interest (i.e.: nucleotides, metals, other compounds). The major problem encountered by available methods is the identification of binding pockets in protein structures analyzed in their unbound conformation.

Methods

Here we describe a new method for ligand-binding site identification in protein structures. The method assigns a propensity value to each aminoacid in a protein structure based on the local similarity of the residue and its neighbors with portions of other binding pockets of known structure, independently of the type of bound ligand. In order to calculate this value we created a non-redundant dataset of all the protein-ligand binding pockets present in the PDB [1] and a complementary nr-dataset of all residues not belonging to any binding pocket. The structure to be analyzed is then used as a query in a search for small structural motifs using the Query3D local comparison program [2] in both datasets. The propensity value of each aminoacid is calculated as the ratio between its occurrences in structural motifs matching the binding dataset vs. the non-binding one.

Results

PDBinders was trained on a set composed of 1301 high quality non-redundant protein chains derived from the entire PDB and was able to discriminate binding residues with an average MCC of 0.27 and an AUC of 0.765. The method was validated on 337 protein structures in their holo and apo forms obtained from the LigASite dataset [3]. The results show that PDBinders is able to work with comparable efficacy on both bound and unbound protein structures, achieving an average MCC of 0.23 (sensitivity 0.24, specificity 0.98) for the holo set and an average MCC of 0.22 for the apo set. We compared PDBinders with the well-established Q-SiteFinder method [4] using the available server on the same set of reference structures. PDBinders and Q-SiteFinder gave similar results in the case

Centre for Molecular Bioinformatics, Dept of Biology, University of Rome Tor Vergata

of bound structures, but our method performed sensibly better in the apo set of structures.

Contact e-mail

valerio.bianchi@uniroma2.it

Supplementary information

References

[1] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. 28, 235-242. [2] Ausiello, G., Via, A., and Helmer-Citterich, M. (2005). Query3d: a new method for high throughput analysis of functional residues in protein structures. BMC Bioinformatics 6, Suppl 4:S5. [3] Dessailly, B.H., Lensink, M.F., Orengo, C.A., and Wodak, S.J (2008). LigASite - a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Research 36, 667-673 [4] Laurie, A.T.R., and Jackson, R.M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21, 1908-1916.