BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 2 2012, pages 286–287
doi:10.1093/bioinformatics/btr651

Structural bioinformatics

Advance Access publication November 22, 2011

FTSite: high accuracy detection of ligand binding sites on unbound protein structures

Chi-Ho Ngan[1], David R. Hall[1], Brandon Zerbe[1], Laurie E. Grove[2], Dima Kozakov[1,*] and Sandor Vajda[1,*]

[1]Department of Biomedical Engineering, Boston University and [2]Department of Sciences, Wentworth Institute of Technology, Boston, MA 02115, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Binding site identification is a classical problem that is important for a range of applications, including the structure-based prediction of function, the elucidation of functional relationships among proteins, protein engineering and drug design. We describe an accurate method of binding site identification, namely FTSite. This method is based on experimental evidence that ligand binding sites also bind small organic molecules of various shapes and polarity. The FTSite algorithm does not rely on any evolutionary or statistical information, but achieves near experimental accuracy: it is capable of identifying the binding sites in over 94% of *apo* proteins from established test sets that have been used to evaluate many other binding site prediction methods.

Availability: FTSite is freely available as a web-based server at http://ftsite.bu.edu.

Contact: vajda@bu.edu; midas@bu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on September 28, 2011; accepted on November 13, 2011

1 INTRODUCTION

Locating ligand binding sites is of fundamental importance for a range of applications, such as the structure-based prediction of function, the elucidation of functional relationships among proteins, protein engineering and drug design. A number of methods aimed at ligand binding site identification have been developed; these include geometric analyses, energy calculations, evolutionary considerations, machine learning and various combinations of these approaches. Over the years, improvements in methodology have led to an increase in the success rate of binding site detection as the top prediction from 52% by SURFNET (Laskowski, 1995) to 83% by VICE (Tripathi and Kellogg, 2010) for the LIGSITE^{CSC} test set of 48 unbound protein structures (Huang and Schroeder, 2006), which has been used to evaluate many binding site detection methods. The VICE algorithm (Tripathi and Kellogg, 2010) has not been implemented as a server, and the best server currently available is MetaPocket 2.0 (Zhang *et al.*, 2011), which seeks consensus among eight different methods and reaches 80% accuracy for the unbound LIGSITE^{CSC} test set. Here we describe the energy-based method FTSite, which is capable of identifying the binding sites with 94% success rate as the top prediction for the same LIGSITE^{CSC} test set, and has been implemented as a server.

*To whom correspondence should be addressed.

2 METHODS

The premise behind FTSite is that ligand binding sites also bind small organic molecules of various shapes and polarity, as observed by nucleic magnetic resonance (NMR) (Hajduk *et al.*, 2005) and X-ray crystallography experiments (Mattos and Ringe, 1996). The solvent mapping algorithm (Brenke *et al.*, 2009) in FTSite is a direct computational analog of these screening techniques. Briefly, computational mapping individually places each of 16 different small molecular probes (Supplementary Fig. S1) on a dense grid around the protein and finds favorable positions using empirical free energy functions. For each probe type, the individual probes are then clustered and the clusters are ranked on the basis of the average free energy (Supplementary Material). Next, consensus clusters are identified as sites in which different probe clusters overlap. The consensus clusters are ranked on the basis of the total number of non-bonded interactions between the protein and all probes in the cluster. The consensus cluster with the highest number of contacts is ranked first; nearby consensus clusters are also joined with this cluster. The amino acid residues in contact with the probes of this newly defined cluster constitute the top ranked predicted ligand binding site (Fig. 1a). Clusters with fewer contacts define lower ranked predictions.

3 RESULTS

FTSite was evaluated using *apo* structures from two test sets of proteins that have been previously used in the evaluation of other binding site prediction methods. These test sets are the LIGSITE^{CSC} set (Huang and Schroeder, 2006) and QSiteFinder set (Laurie and Jackson, 2005). FTSite achieves accuracy rates of 94 and 97% for the LIGSITE^{CSC} and QSiteFinder sets, respectively, when determining the binding site as the highest ranked consensus site. For each test set, we employed the same assessment criteria that were used in the respective previous studies (Supplementary Material). A comparison to other methods is shown in Figure 1c.

FTSite was able to identify the ligand binding site using only the top ranked prediction in a number of difficult cases. Two examples are shown in Figure 1b (see the Supplementary Material for other examples); the ligand from the *holo* structures is superimposed on the results of the apo structures for comparison. In both cases, conformational differences between the *holo* and *apo* structures make binding site detection difficult. In the case of β-amylase, the loop formed by residues V99, G100 and D101 closes down on the ligand in the *holo* form, yielding a better defined ligand binding site. Similarly, in the *apo* form of HIV-2 protease the ligand binding site

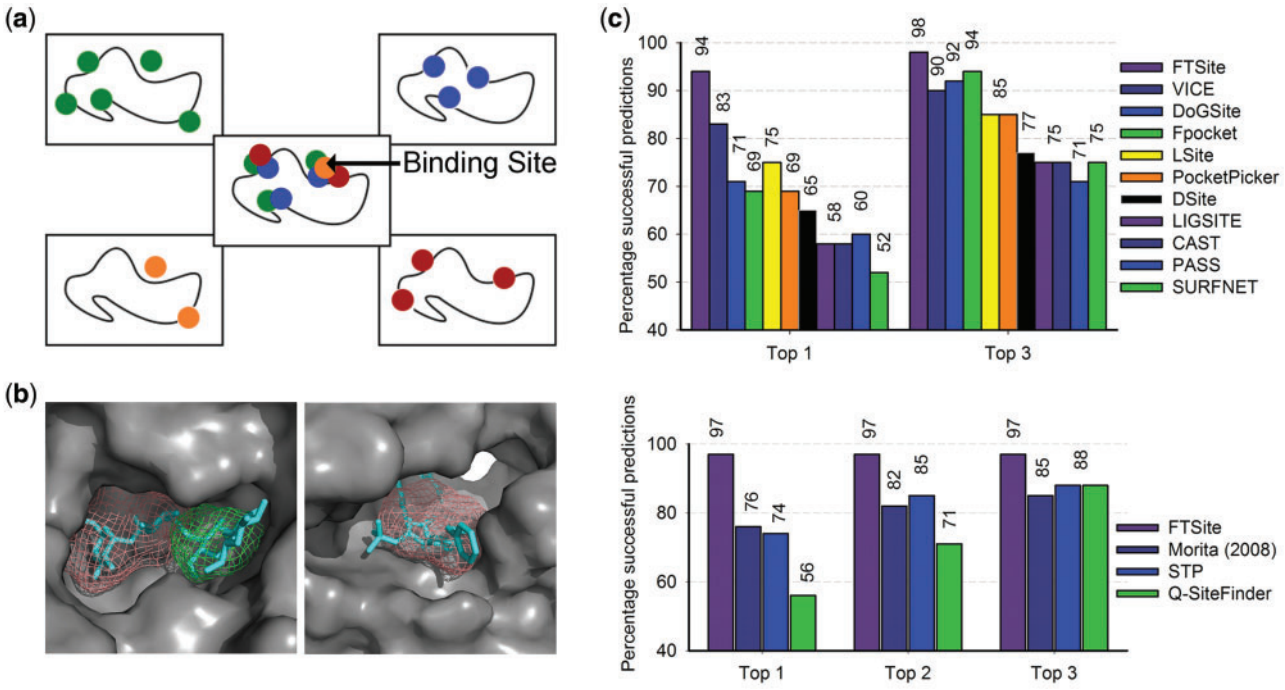

Fig. 1. Methodology of FTSite and its performance on two sets of test proteins. (**a**) FTSite identifies regions that have the highest number of non-bonded interactions with overlapping low energy clusters of several small molecular probes. (**b**) Two successful examples are shown. The ligands of each respective target are shown as cyan sticks and the putative ligand binding sites are colored salmon and green for the first and second highest ranked sites, respectively. Left: β-amylase (PDB ID: 1BYA (*apo*) and 1BYB (*holo*)). Right: HIV-2 protease (PDB ID 1HSI (*apo*) and 1IDA (*holo*)). (**c**) Top: on the LIGSITE^{CSC} test set, FTSite has an accuracy of 94% using only the top ranked prediction of the ligand binding site, and 98% using the top three. Bottom: on the QSiteFinder test set, FTSite has an accuracy of 97% using only the top ranked prediction.

is very open, and becomes well formed only upon ligand binding. Despite the poorly defined binding sites present in the apo structures of these two proteins, FTSite identifies the binding site as the top ranked prediction.

4 DISCUSSION

Key to the success of FTSite is the use of multiple molecular probes rather than a single probe as implemented in most other energy-based methods. Screening by NMR and X-ray crystallography shows that the binding sites of proteins possess a tendency to bind small organic compounds that vary in size, shape and polarity, thus improving the robustness of FTSite to conformational changes. Although individual probes may bind to other cavities, the largest clusters of multiple probes occur in ligand binding sites (Hajduk *et al.*, 2005). Thus, FTSite does not rely on surrogate measures of ligand-binding propensity such as pocket volume, cavity depth or the ability of binding non-polar spheres. Due to its strong biophysical basis, the method provides high accuracy without evolutionary considerations. We note that solvent mapping also had success in identifying druggable binding sites in protein–protein interfaces (Kozakov *et al.*, 2011).

Funding: National Institute of General Medical Sciences (grant GM064700).

Conflict of Interest: none declared.

REFERENCES

Brenke,R. *et al.* (2009) Fragment-based identification of druggable "hot spots" of proteins using Fourier domain correlation techniques. *Bioinformatics*, **25**, 621–627.

Hajduk,P.J. *et al.* (2005) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.*, **48**, 2518–2525.

Huang,B. and Schroeder,M. (2006) LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19–29.

Kozakov,D. *et al.* (2011) Structural conservation of druggable hot spots in protein-protein interfaces. *Proc. Natl Acad. Sci. USA*, **108**, 13528–13533.

Laskowski,R. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330.

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.

Mattos,C. and Ringe, D. (1996) Locating and characterizing binding sites on proteins. *Nat. Biotechnol.*, **14**, 595–599.

Tripathi,A. and Kellogg,G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins*, **78**, 825–842.

Zhang,Z. *et al.* (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **26**, 2920–2921.