BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 15 2010, pages 1913–1914
doi:10.1093/bioinformatics/btq288

Structural bioinformatics

Advance Access publication June 10, 2010

PESDserv: a server for high-throughput comparison of protein binding site surfaces

Sourav Das†, Michael P. Krein† and Curt M. Breneman*

Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180, USA

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Structure-based approaches complement ligand-based approaches for lead-discovery and cross-reactivity prediction. We present to the scientific community a web server for comparing the surface of a ligand bound site of a protein against a ligand bound site surface database of 106 796 sites. The web server implements the property encoded shape distributions (PESD) algorithm for surface comparison. A typical virtual screen takes 5 min to complete. The output provides a ranked list of sites (by site similarity), hyperlinked to the corresponding entries in the PDB and PDBeChem databases.

Availability: The server is freely accessible at http://reccr.chem.rpi.edu/Software/pesdserv/

Contact: brenec@rpi.edu

Received on March 5, 2010; revised on May 3, 2010; accepted on May 27, 2010

1 INTRODUCTION

The numerous genome sequencing projects have 'revealed that proteins involved in entirely different biochemical pathways, and even residing in different tissues and organs, may possess functional binding pockets with similar shapes and physiochemical properties' (Kinnings *et al.*, 2009; Weber *et al.*, 2004). The need to compare protein binding sites and establish structure–function relationships across protein families has become ever more important. Three direct outcomes of such an effort are: (i) identification of lead compounds for 'target-hopping', (ii) repositioning an existing drug and (iii) predicting adverse side-effects (Rognan, 2007; Shulman-Peleg *et al.*, 2004). The discovery of low molecular weight somatostatin receptor subtype 5 (hSST5R) antagonists involved use of astemizole as the lead structure whose original target was H1, a histamine receptor. H1 has binding site amino acid composition similar to hSST5R's (Martin *et al.*, 2007). The work by Kinnings *et al.* (2009) utilizing the SOIPPA algorithm (Xie and Bourne, 2008) is an example of how a catechol-*O*-methyltransferase inhibitor used for treating Parkinson's disease can be repositioned as a lead compound for treating multi-drug resistant and extensively drug resistant tuberculosis. Since the compound has an excellent safety profile, repositioning it to treat drug-resistant tuberculosis offers significant incentives. Side effects are a serious cause for concern in chemotherapy. A study by Paolini *et al.* (2006) found that out of 276 122 active compounds, over 96 000 had activity for more than one target (Kinnings *et al.*, 2009). While ligand-based approaches can point to such cross-reactivity, they may not be sufficient in cases where ligands with low apparent similarity bind on to the same target (Kinnings *et al.*, 2009; Das *et al.*, 2009).

The property encoded shape distributions (PESD) algorithm (Das *et al.*, 2009) offers high-throughput binding site comparison that makes it suitable for mining large databases such as the Protein Data Bank (PDB; Berman *et al.*, 2000). There are several web-based services and stand-alone programs for comparison of ligand binding sites in proteins. The SitesBase (Gold and Jackson, 2006) server uses a geometric hashing algorithm on an all-atom-based representation of the binding site. The SiteEngine and Multibind (Shulman-Peleg *et al.*, 2004, 2008) servers also utilize a geometric hashing based search protocol that compares pseudo-center representations of binding sites. The Catalytic Site Search server (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/CSS/makeEbiHtml.cgi?file=form.html) based on the Jess algorithm (Barker and Thornton, 2003) searches for specific groups of residues. PDBSiteScan (Ivanisenko *et al.*, 2004) works on a reduced representation of atoms (utilizing N, C and C_α) and implements a combinatorial extension (Shindyalov and Bourne, 1998) algorithm followed by alignment for similarity detection. The SuMo server (Jambon *et al.*, 2005) matches graphs representing triplets of functional groups such as unbound hydrogen bond donors or acceptors, accessible sides of aromatic rings and carboxylate, primary amide, etc. The FuzCav program utilizes a generic cavity fingerprint generated from a labeled C_α representation of the binding site (Weill and Rognan, 2010).

In Das *et al.* 2009, we have shown that binding sites can have similar shape and property distributions on the surface and yet show low conservation in underlying residue, pseudo-center or atom composition. In such cases, methods based on residue, pseudo-center or atom representations may not be able to detect the similarity that is evident from the surface. The eF-site server (Kinoshita *et al.*, 2002) can work directly on electrostatic potential-mapped surfaces; however, it uses a clique-detection-based algorithm that is NP-hard and slow for high-throughput global similarity searches. The PESD algorithm implemented in the PESDserv binding site comparison server presented here overcomes this limitation and quickly returns a list of globally similar ligand-bound site surfaces from the PDB in a few minutes.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 METHODS AND FEATURES

All X-ray crystal structures obtained from the PDB as of October 30, 2009, were separated into protein and ligand parts. The protein side chains were protonated with PROPKA (Li *et al*., 2005) at pH 7.0 and Gauss–Connolly surfaces were generated in Molecular Operating Environment (MOE Version 2007.09, Chemical Computing Group, Inc.). For any ligand (identified by the HETATM keyword in a PDB file) not a heme and having more than five non-hydrogen atoms, property-mapped Gauss–Connolly binding site surfaces were generated on the protein. A 4.5-Å cutoff distance from the protein surface to any atom of the ligand was chosen for defining the binding site surface. The maps were for electrostatic potential and mildly polar, hydrophobic and hydrogen bonding regions (MOE Active LP maps). The Amber99 force field was used for generating the electrostatic potential mapped surface. A total of 106 796 binding sites were finally available for PESD signature generation. The signatures were generated as described in Das *et al.* (2009). A bin width of 1 Å was employed for the first 24 bins whereas the 25th bin recorded all distances >24 Å. 100 000 pairs of points were sampled from each surface. This formed the 'pre-computed database' of binding sites obtained from the entire PDB.

The server requires a user to upload a protein file and a ligand file in PDB format. The binding site surface of the uploaded complex is generated as described previously. The user has choices of L1 (Manhattan), L2 (Euclidean) or χ^2 distances for calculating the similarity between the PESD signatures of the uploaded site and the signatures in the database. Lower similarity is indicated by a greater distance. The results are output as a sorted list of sites, sorting being based on distance option chosen. Links to downloadable tab-delimited results file are also provided in the output page. The output page has seven columns: (i) rank of a site from the pre-computed database; (ii) PDB accession code of the protein complex; (iii) ligand identifier of the ligand bound to the site; (iv) PESD distance; (v) ligand chain ID extracted from the PDB file; (vi) the PDB title extracted from the PDB file; and (vii) the *Z*-score. The *Z*-scores are computed from the mean and SD of scores of all 106 796 sites compared. The PDB accession codes are linked to the PDB database and the ligand accession codes are linked to PDBeChem database (http://www.ebi.ac.uk/msd-srv/msdchem/cgi-bin/cgi.pl) for convenience. The entire process of screening the pre-computed database typically takes 5 min.

The true-positive recall rate of the PESD method was determined in our previous work (Das *et al*., 2009). In a screening experiment involving 881 queries against corresponding 1256 member test sets, PESD was able to retrieve a binding site with identical Enzyme Commission (E.C.) numbers as the query in Rank 1 in 79.5% of cases.

3 CONCLUSIONS AND FUTURE WORK

A publicly accessible server for fast web-based global similarity analysis of surface shape and property distributions of ligand binding sites in proteins was created. Since the matching is global, functionally relevant sites differing in size will have a larger PESD distance than similar-sized sites. Binding sites of significantly different shapes but that bind to the same ligand will also result in a higher PESD distance compared to similarly shaped sites. We are presently working on a fragment-based approach, where surfaces around smaller and less flexible ligand fragments would be generated and matched to ensure local similarity-based matching of sites to overcome current limitations.

Funding: National Institutes of Health (grant number 1P20 HG003899) and Rensselaer Polytechnic Institute (institute fellowship to S.D.).

Conflict of Interest: none declared.

REFERENCES

Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Das,S. *et al*. (2009) Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model.*, **49**, 2863–2872.

Gold,N.D. and Jackson,R.M. (2006) SitesBase: a database for structure-based protein–ligand binding site comparisons. *Nucleic Acids Res.*, **34**, 231–234.

Ivanisenko,V.A. *et al*. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res.*, **32**, W549–W554.

Jambon,M. *et al*. (2005) The SuMo server: 3D search for protein functional sites. *Bioinformatics*, **21**, 3929–3930.

Kinnings,S.L. *et. al.* (2009) Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.*, **5**, e1000423.

Kinoshita,K. *et al*. (2002) Identification of proteins functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics*, **2**, 9–22.

Li,H. *et al*. (2005) Very fast empirical prediction and rationalization of protein pKa values. *Proteins*, **61**, 704–721.

Martin,R.E. *et al*. (2007) Discovery of the first nonpeptidic, small-molecule, highly selective somatostatin receptor subtype 5 antagonists: a chemogenomics approach. *J. Med. Chem.*, **50**, 6291–6294.

Paolini,G.V. *et al*. (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.

Rognan,D. (2007) Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.*, **152**, 38–52.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Shulman-Peleg,A. *et al*. (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.*, **339**, 607–633.

Shulman-Peleg,A. *et al*. (2008) MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.*, **36**, W260–W264.

Weber,A. *et al.* (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.*, **47**, 550–557.

Weill,A. and Rognan,D. (2010) Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.*, **50**, 123–135.

Xie,L. and Bourne,P.E. (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl Acad. Sci. USA*, **105**, 5441–5446.