Published online 19 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W441–W444***
doi:10.1093/nar/gkq400

SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison

Jingyuan Ren[1,2], Lei Xie[1,*], Wilfred W. Li[1,2,*] and Philip E. Bourne[1,3]

[1]San Diego Supercomputer Center, [2]National Biomedical Computation Resource and [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received March 9, 2010; Revised April 26, 2010; Accepted April 29, 2010

ABSTRACT

The proteome-wide characterization and analysis of protein ligand-binding sites and their interactions with ligands can provide pivotal information in understanding the structure, function and evolution of proteins and for designing safe and efficient therapeutics. The SMAP web service (SMAP-WS) meets this need through parallel computations designed for 3D ligand-binding site comparison and similarity searching on a structural proteome scale. SMAP-WS implements a shape descriptor (the Geometric Potential) that characterizes both local and global topological properties of the protein structure and which can be used to predict the likely ligand-binding pocket [Xie,L. and Bourne,P.E. (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand-binding sites. *BMC bioinformatics*, 8 (Suppl. 4.), S9.]. Subsequently a sequence order independent profile–profile alignment (SOIPPA) algorithm is used to detect and align similar pockets thereby finding protein functional and evolutionary relationships across fold space [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]. An extreme value distribution model estimates the statistical significance of the match [Xie, L., Xie, L. and Bourne, P.E. (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, 25, i305–i312.]. These algorithms have been extensively benchmarked and shown to outperform most existing algorithms. Moreover, several predictions resulting from SMAP-WS have been validated experimentally. Thus far SMAP-WS has been applied to predict drug side effects, and to repurpose existing drugs for new indications. SMAP-WS provides both a user-friendly web interface and programming API for scientists to address a wide range of compute intense questions in biology and drug discovery.

AVAILABILITY

SMAP-WS is available from the URL http://smap.nbcr.net.

INTRODUCTION

The 3D structure of a protein is an essential component in elucidating biological function(s) at the molecular level. Ligand-binding sites and their interactions with binding partners provides a strong correlation between structure and function and thus are critical to address a wide range of fundamental and practical problems—predicting functions for structural genomics targets, bridging knowledge derived from small molecules and proteins, correlating molecular functions to physiological processes, studying protein evolution and diversity, and designing safe and efficient therapeutics.

The SMAP web service (SMAP-WS) is distinct from the downloadable software and is designed for web accessible 3D ligand-binding site comparison and similarity searching on a structural proteome scale. The underlying algorithms comprising SMAP-WS and the standalone software, SMAP, are distinct from existing web servers SiteEngine (1), SitesBase (2,3), CavBase (4–6), SuMo (7), PdbSiteScan (8), eF-Site (9,10), pvSOAR (11), ProFunc (12), PevoSoar (13) and fPOP(14). First, SMAP represents protein structures using C-α atoms only and hence is tolerant to structural variation, meaning it can be

*To whom correspondence should be addressed. Tel: +1 858 822 3686; Fax: +1 858 822 0873; Email: lxie@ucsd.edu
*Correspondence may also be addressed to Wilfred W. Li. Tel: +1 858 534 0591; Fax: +1 858 822 1619; Email: wilfred@sdsc.edu

applied to homology models and low-resolution structures. Second, amino acid residues are characterized by surface orientation and a geometric potential (15) which provides a geometrical constraint to reduce the search space when undertaking ligand-binding site comparison. Third, two structures are compared using a sequence order independent profile–profile alignment (SOIPPA) algorithm (16). SOIPPA aligns two structures in the spirit of local sequence alignment, but independent of the sequence order. As a result the location and boundary of the ligand-binding site does not need to be pre-defined. This property is important for real world applications since information on the ligand-binding site may be unknown. Fourth, SMAP can compare two biological units that may include multiple chains. This is important since binding sites may be located in the homo- or hetero-dimer interface. For example, the binding site of the antibiotic myxopyronin to the bacterial RNA polymerase is located in the 'switch region' between the β and β′ chains. Finally, SMAP determines the similarity between two binding sites through the combination of geometrical fit, residue conservation and physiochemical similarity. The statistical significance of the similarity is estimated using an extreme value distribution model (17). Putting these features together within a parallel computing environment means that SMAP-WS is capable of an all-by-all comparison of binding sites for a complete structural proteome.

In benchmark studies, SOIPPA outperforms most existing ligand-binding site comparison algorithms (16). Around 30% of evolutionary and functional relationships across superfamilies are identified by SOIPPA with a false–positive ratio of 5%. Moreover, SOIPPA outperforms global structural alignment algorithms in detecting remote homologous that belong to the same superfamily. For a false–positive ratio of 5%, SOIPPA detects 15% more true positives than the global structural alignment. More important, several predictions from SMAP have been experimentally validated (16,18–20). Given the reliability of SMAP, it has been applied to constructing drug–target interaction networks on a structural proteome scale (17), predicting molecular mechanisms of drug side effects (21,22), repurposing old drugs for new medical usage (19), designing polypharmacology (dirty) drugs (18), and establishing evolutionary relationships across protein fold space (16). Thus, SMAP is useful for studying fundamental questions in protein structure, function and evolution, as well as for computer aided drug design based on polypharmacology. As standalone software, SMAP can be installed locally and executed from the command line. SMAP-WS has several improvements that make it more user-friendly and computationally efficient. SMAP-WS has a web-based interface for the input of PDB structures, the set of required parameters, a Jmol visualization plugin to analyze results, pre-computed databases to search against and a parallel implementation of SMAP accessible from a large compute cluster to improve database search speed. Thus SMAP-WS facilitates the application of comparative ligand-binding site analysis to address practical problems in biology and drug discovery.

METHODS

Opal powered SMAP web services

SMAP-WS is powered by Opal (23), a toolkit that enables scientists to easily wrap applications as web services that have user-friendly web forms by configuring simple XML files. Two SMAP-WS interfaces are implemented: (i) pair-wise comparison of two potential ligand-binding sites; and (ii) search using a query structure against a non-redundant structure database from the RCSB Protein Data Bank (PDB) (24). In the first application structures and their components can be chosen from the PDB or uploaded by the user. In the second application the user may either choose to enter a PDB structure id and its chain id(s) or upload a structure file in PDB format. The user can then choose to perform a search using this structure against several databases, including human homologous proteins and non-redundant PDB structures based on sequence identities of 30 and 90%, respectively. The user has the option to modify the appropriate SMAP-WS parameters for both applications. In order to improve the search speed, the database search has pre-cached the protein structures used for the SMAP comparison. The structure cached is characterized by geometric, evolutionary and physiochemical properties and uses default parameters. The pair-wise comparison interface provides the user with the ability to modify more parameters for comparing two protein structures based on the similarity of their potential ligand-binding sites.

In additional to the web input forms (http://smap.ncbr.net), SMAP-WS can be accessed through a programming API. The details of how to write a client program can be found on the web site.

Output of SMAP-WS

The hits from a database search are sorted by the similarity score of the match, along with *P*-values of the match, their PDB structure ids, chain ids and biological descriptions. The PDB id is linked to the structure summary page of the RCSB PDB (http://www.rcsb.org/pdb). For each of the hits, detailed information on the ligand-binding site similarity is presented (*P*-value, raw alignment score, RMSD and Tanimoto coefficient of overlap). The amino-acid residue alignment between two ligand-binding sites and the transformation matrix to superpose them are also displayed.

It is important to evaluate if the predicted residue cluster is a potential binding region. SMAP-WS relies on the geometric potential (15), which is a shape descriptor to characterize both local and global topological properties of each residue, to determine whether a residue is located in a pocket on the protein structure or not. However, in a real application where the binding region is unknown, additional information such as ligand-binding affinity may be required to determine if the predicted region is suited for ligand binding. Thus, a visualization tool that allows the user to inspect the protein–ligand complex structure was implemented. A Jmol plugin (Jmol: an open-source Java viewer for chemical structures in 3D.) that displays the superposition of two protein structures

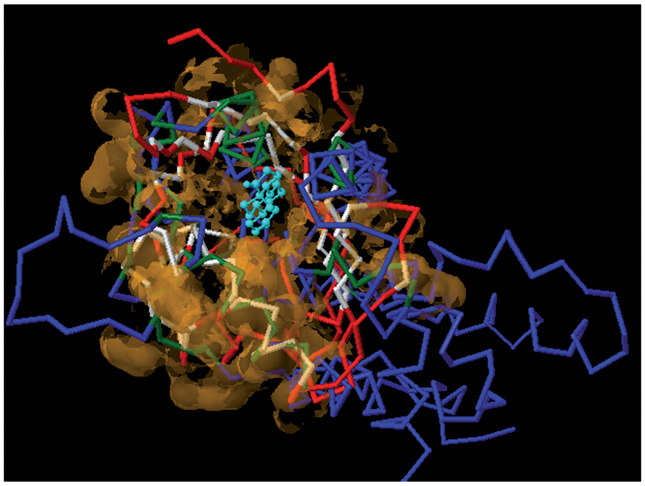

Figure 1. Two superposed structures of steroid delta-isomerase (PDB id: 1OHP, blue backbone) and estrogen receptor ligand-binding domain (PDB id 1QKT, red backbone) from an SMAP alignment. The co-crystalized estradiol in 1QKT is shown as a light blue stick model. The aligned residues between the two structures are highlighted.

with predicted and aligned ligand-binding site residues is provided. An example of such a superposition is shown in Figure 1. The estrogen receptor ligand-binding domain (PDB id: 1QKT) is compared with the steroid delta-isomerase (PDB id: 1OHP) without specifying the co-crystallized ligand-binding site (in the web interface, the option of 'search only co-crystal ligand sites' for both structures was set to false). A statistically significant similarity between the two estradiol ligand-binding sites is detected (P = 4.09e–6), although the two structures do not share global structure or sequence similarity [FATCAT (25) P-value 1.19e–1, and sequence identity %9.62], and the estradiol-binding sites are not pre-defined in both of the structures.

One of the major applications of SMAP is to predict off-targets given a known protein–ligand complex. The ligand-binding site similarity between two proteins alone is prerequisite, but not sufficient to determine their cross-reactivity for a specific ligand. The chemical nature of the ligand also contributes to the binding promiscuity. For example, staurosporine can bind to a large panel of kinases with $K_d < 100$ nm. However, compound VX-745 is a highly specific ATP competitive inhibitor of p38 MAP kinase (26). A more recent case is the chemical phylogenetic study of histone deacetylases (HDAC), where two chemical analogs have different binding profiles across multiple members of the HDAC superfamily (27). To computationally determine the potential off-target binding, it is necessary to calculate the binding free energy of the protein–ligand complex using techniques such as protein–ligand docking and molecular dynamics simulation. SMAP-WS narrows down the potential off-targets to a small subset of the whole structure proteome as well as provides an initial binding pose for the given ligand to the off-target it found in the query structure. The accuracy of predicted binding poses by SMAP-WS has been evaluated in a previous study (16). In a rigorous benchmark test, 6.5 and 25.9% of predicted binding poses fall within RMSD values of <2.0 and <5.0 Å, respectively, when compared with co-crystallized ligands that bind to proteins with different folds. Hence, the predicted protein–ligand complex from SMAP-WS could be used as a starting point for more computational intensive studies. The pipeline has been successfully applied to determine the polypharmacological targets of *Trypanosoma Brucei* RNA-ligase inhibitors (18). To facilitate such applications, SMAP-WS allows users to download the structure of potential off-targets with the superposed ligand. These complexes can then be subject to more computationally intensive studies such as protein–ligand docking and MD simulation.

Paralleled implementation of SMAP-WS

SMAP-WS database search is scheduled by the Sun Grid Engine (SGE), which allows SMAP pair-wise comparison to be executed concurrently on all available compute nodes. As a result, SMAP-WS significantly improves the speed of ligand-binding site database searching. Using SMAP on a single processor, sequential comparison of a query structure against a database of about 40 000 non-redundant structures from the PDB takes more than 20 days (17). Our solution to speed up this process was to set up a wrapper program that submits a SGE array job for the set of SMAP comparisons to allow these comparisons to run in parallel on the computer nodes in a cluster. The SMAP-WS server cluster has available up to 99 computer nodes with two processors on each node. Thus 198 SMAP-WS jobs can run in parallel, when all computer nodes are available, with a scan of the non-redundant PDB being done within one day.

CONCLUSION

We have developed a high performance computation environment SMAP-WS for protein ligand-binding site comparison and database searching. SMAP-WS provides both user-friendly interfaces and a programming API to help a wide spectrum of scientists to access the service. It is expected that the integration of SMAP-WS with other bioinformatics, molecular modeling and systems biology tools will facilitate the study of protein–ligand interactions on a structural proteome scale and drug design based on polypharmacology.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Li Xie for helpful discussions. The authors appreciate the constructive suggestions of the anonymous reviewers and the editor in improving the manuscript.

FUNDING

National Institutes of Health grant GM078596; and National Center for Research Resources, National Institutes of Health NIH P41RR08605 for the National

Biomedical Computation Resource (NBCR). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.*, **339**, 607–633.
2. Gold,N.D. and Jackson,R.M. (2006) SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res.*, **34**, D231–D234.
3. Gold,N.D. and Jackson,R.M. (2006) A searchable database for comparing protein-ligand binding sites for the analysis of structure-function relationships. *J. Chem. Inf. Modeling*, **46**, 736–742.
4. Kuhn,D., Weskamp,N., Schmitt,S., Hullermeier,E. and Klebe,G. (2006) From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.*, **359**, 1023–1044.
5. Schmitt,S., Kuhn,D. and Klebe,G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.
6. Weskamp,N., Kuhn,D., Hullermeier,E. and Klebe,G. (2004) Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics*, **20**, 1522–1526.
7. Jambon,M., Imberty,A., Deleage,G. and Geourjon,C. (2003) A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, **52**, 137–145.
8. Ivanisenko,V.A., Pintus,S.S., Grigorovich,D.A. and Kolchanov,N.A. (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.
9. Kinoshita,K., Furui,J. and Nakamura,H. (2002) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Func. Genomics*, **2**, 9–22.
10. Kinoshita,K. and Nakamura,H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci. Pub. Protein Soci.*, **12**, 1589–1595.
11. Binkowski,T.A., Freeman,P. and Liang,J. (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res.*, **32**, W555–W558.
12. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.
13. Tseng,Y.Y., Dundas,J. and Liang,J. (2009) Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. *J. Mol. Biol.*, **387**, 451–464.
14. Tseng,Y.Y., Chen,Z.J. and Li,W.H. fPOP: footprinting functional pockets of proteins by comparative spatial patterns. *Nucleic Acids Res.*, **38**, D288–D295.
15. Xie,L. and Bourne,P.E. (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics*, **8(Suppl. 4)**, S9.
16. 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.
17. Xie,L., Xie,L. and Bourne,P.E. (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, **25**, i305–i312.
18. Durrant,J.D., Amaro,R.E., Xie,L., Urbaniak,M.D., Ferguson,M.A., Haapalainen,A., Chen,Z., Di Guilmi,A.M., Wunder,F., Bourne,P.E. *et al.* (2010) A multidimensional strategy to detect polypharmacological targets in the absence of structural and sequence homology. *PLoS Comput. Biol.*, **6**, e1000648.
19. Kinnings,S.L., Liu,N., Buchmeier,N., Tonge,P.J., Xie,L. and Bourne,P.E. (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.
20. Miller,J.R., Dunham,S., Mochalkin,I., Banotai,C., Bowman,M., Buist,S., Dunkle,B., Hanna,D., Harwood,H.J., Huband,M.D. *et al.* (2009) A class of selective antibacterials derived from a protein kinase inhibitor pharmacophore. *Proc. Natl Acad. Sci. USA*, **106**, 1737–1742.
21. Xie,L., Li,J., Xie,L. and Bourne,P.E. (2009) Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.*, **5**, e1000387.
22. Xie,L., Wang,J. and Bourne,P.E. (2007) In silico elucidation of the molecular mechanism defining the adverse effect of selective estrogen receptor modulators. *PLoS Comput. Biol.*, **3**, e217.
23. Krishnan,S., Clementi,L., Ren,J., Papadopoulos,P. and Li,W. (2006) Design and Evaluation of Opal2: A Toolkit for Scientific Software as a Service. *The 2009 IEEE Congress on Services (SERVICES-1 2009), July, 2009*. IEEE Conference, Los Angeles, CA, USA.
24. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
25. Ye,Y. and Godzik,A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–W585.
26. Karaman,M.W., Herrgard,S., Treiber,D.K., Gallant,P., Atteridge,C.E., Campbell,B.T., Chan,K.W., Ciceri,P., Davis,M.I., Edeen,P.T. *et al.* (2008) A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnol.*, **26**, 127–132.
27. Bradner,J.E., West,N., Grachan,M.L., Greenberg,E.F., Haggarty,S.J., Warnow,T. and Mazitschek,R. (2010) Chemical phylogenetics of histone deacetylases. *Nature Chem. Biol.*, **6**, 238–243.