Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W337–W341***
doi:10.1093/nar/gki482

SiteEngines: recognition and comparison of binding sites and protein–protein interfaces

Alexandra Shulman-Peleg*, Ruth Nussinov[1,2] and Haim J. Wolfson

School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences and [1]Sackler Institute of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel and [2]Basic Research Program, SAIC-Frederick Inc., Laboratory of Experimental and Computational Biology NCI-Frederick, Building 469, Room 151, Frederick, MD 21702, USA

Received February 14, 2005; Revised April 4, 2005; Accepted April 18, 2005

ABSTRACT

Protein surface regions with similar physicochemical properties and shapes may perform similar functions and bind similar binding partners. Here we present two web servers and software packages for recognition of the similarity of binding sites and interfaces. Both methods recognize local geometrical and physicochemical similarity, which can be present even in the absence of overall sequence or fold similarity. The first method, SiteEngine (http:/bioinfo3d.cs.tau.ac.il/SiteEngine), receives as an input two protein structures and searches the complete surface of one protein for regions similar to the binding site of the other. The second, Interface-to-Interface (I2I)-SiteEngine (http:/bioinfo3d.cs.tau.ac.il/I2I-SiteEngine), compares protein–protein interfaces, which are regions of interaction between two protein molecules. It receives as an input two structures of protein–protein complexes, extracts the interfaces and finds the three-dimensional transformation that maximizes the similarity between two pairs of interacting binding sites. The output of both servers consists of a superimposition in PDB file format and a list of physicochemical properties shared by the compared entities. The methods are highly efficient and the freely available software packages are suitable for large-scale database searches of the entire PDB.

INTRODUCTION

Recognition and comparison of regions through which protein molecules function and interact are crucial for the prediction of molecular interactions, which govern practically all cellular processes. Consequently, a broad range of tools for sequence and overall structural alignment are routinely used by the scientific community in the analysis of biological processes and the prediction of function (1). However, the overall similarity of the sequences and folds does not necessarily imply similarity of biological function (2,3). It has been shown that proteins with the same fold can have different functions (4), and that proteins with different folds, such as serine proteases or zinc-binding proteins, can share the same function. Since proteins function by interacting with other molecules, similarity in their biological function is related to the similarity of their corresponding binding regions. These may be sequentially non-continuous regions with no common patterns of amino acids (5–7) but sharing a set of physicochemical properties which create similar surface regions. Several approaches have been proposed for the recognition of such functional sites (8–12). These are important for drug design (13) as well as functional annotation (2,14) and biological classification (15–17).

With the progress of the Structural Genomics project, the number of determined structures of protein–protein complexes is growing rapidly. These complexes contain valuable information regarding the functional groups that actually interact with each other. We define a *protein–protein interface* as a pair of regions of two interacting protein molecules that are linked by non-covalent bonds. Analysis and classification of protein–protein interfaces (17–19) is the first step in the recognition of preferred binding organizations, which may shed light on the driving forces stabilizing molecular interactions.

In this paper, we present two web servers and freely available software packages for the comparison of protein binding sites and protein–protein interfaces. The first method, SiteEngine (7), searches the complete structure of one protein for a region similar to the binding site of another. The second method, Interface-to-Interface (I2I)-SiteEngine (20), utilizes the information regarding patterns of interactions present in protein–protein complexes and performs a simultaneous alignment of pairs of interacting binding sites. In contrast to the methods that compare the locations of the backbone atoms or the identity of the amino acids, the methods presented here

*To whom correspondence should be addressed. Tel: +972 3 6405395; Fax: +972 3 6406476; Email: shulmana@tau.ac.il

consider the physico chemical properties of both backbone and side-chains and assume no similarity of the overall sequences or folds.

SiteEngine: FUNCTIONAL SITES STRUCTURAL SEARCH ENGINE

SiteEngine is an efficient method for the recognition of functional sites in protein structures (7). It is motivated by several goals. First, analysis of compounds bound to proteins with similar functional sites may suggest chemical groups and scaffolds that can be used in drug design and optimization. SiteEngine can also assist in the recognition of proteins with similar binding sites that can potentially cause side-effects. In addition, it can be applied to recognize regions on the surface of a novel protein that are similar to functional sites of known proteins. This may contribute to a better understanding of the novel proteins's function and activation mechanism. Furthermore, classification of proteins according to their functional site may facilitate the development of more efficient database organizations and search schemes.

Method

Following Schmitt *et al*. (13), for each amino acid we group atoms with similar physicochemical properties into functional groups, which are represented by three-dimensional points in space, denoted as *pseudocenters*. Each pseudocenter represents one of the following properties important for protein–ligand interactions (13): *hydrogen-bond donor (DON)*, *hydrogen-bond acceptor (ACC)*, *mixed donor/acceptor (DAC)*, *hydrophobic aliphatic (ALI) and aromatic*, *pi interactions (PII)*. We construct a smooth molecular surface as implemented by Connolly (21) and retain only pseudocenters that represent at least one surface exposed atom. When considering binding sites, we refer only to the surface regions that are within 4 Å of the binding partner.

Given the representation described, we calculate all possible transformations that superimpose the input binding site on a similar surface region of the other molecule. The algorithm is based on efficient hashing and matching of almost congruent triangles defined by triplets of pseudocenters. The hashing of the triangles is done with a key that consists of the three parameters of side lengths of a triangle and of an additional physicochemical index, which encodes the properties of its nodes. Each pair of matched triangles defines a candidate transformation which can superimpose the input binding site on a certain region of the complete protein. Similarity of the physicochemical properties and shapes aligned by each transformation is scored using a set of hierarchically applied scoring functions (7) and a list of top ranking solutions is selected.

Input

The input to the web server consist of two structures (defined by the PDB codes or uploaded files, see Figure 1a). The SiteEngine method will search the complete surface of the first molecule for a region similar to the binding site of the second. Although the first structure can be either bound or unbound, the second is required to contain a ligand in the binding site of interest. The definition of the binding site is done through the web form presented in Figure 1b, which provides a list of ligands bound to the input structure. (Only ligands listed as HETATM records of size more than 7 atoms are considered) The binding site is determined by the surface region located within a distance of 4 Å of the selected ligand.

Output

The output of SiteEngine is a three-dimensional transformation that can superimpose the binding site of interest on the regions that resemble it in the complete structure. The web server presents the details of the 10 top ranking solutions. For each solution, the calculated transformation, the score and the match list are presented (see Figure 1c). [The exact details of the score calculation can be found on the webserver and in the Supplementary Materials of Ref. (7)] The match list provides the details of the 1:1 correspondence between the pseudocenters of the two molecules. The first four columns correspond to the first molecule and the next four columns correspond to the second (query) molecule. For each molecule, the first column provides the chain identifier and the residue number. The second column contains the residue name. The third column gives an abbreviation of the physico chemical property (see Method), and the fourth provides information regarding the source of the property: backbone (b) or side-chain (s). The last two columns denote the distance between the matched pseudocenters and the conservation of the identity of the amino acid that originated the property. For each solution we provide a PDB file (aligned.pdb) with the superimposition of the input molecules (see Figure 1d).

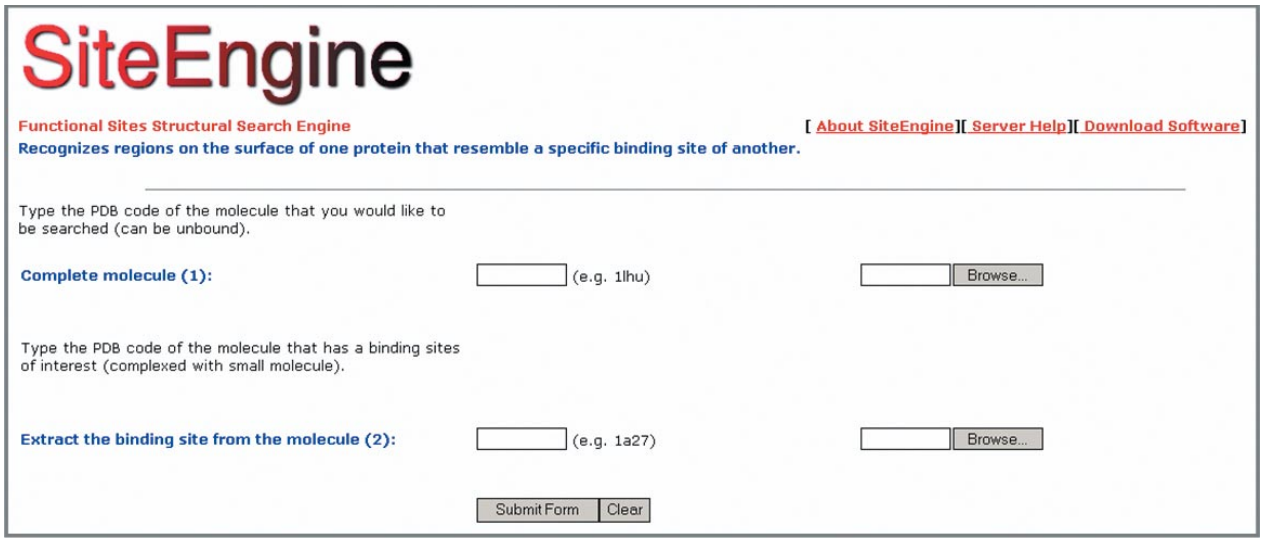
I2I-SiteEngine: ALIGNMENT OF PROTEIN–PROTEIN INTERFACES

I2I-SiteEngine is a method for the simultaneous structural alignment of two protein–protein interfaces (20). It utilizes information regarding patterns of interacting functional groups to increase the speed and the quality of the alignment. Similarly to SiteEngine, the method can be useful in drug discovery and the prediction of side-effects. However, in the case of protein–protein interfaces the drug design strategy may differ, since we are interested in prevention of association or dissociation of the protein molecules. The analysis and classification of protein–protein interfaces with methods such as I2I-SiteEngine allows the recognition of certain binding organizations shared by different protein families that might be important for the formation and stability of the protein–protein complex. Their recognition may assist in the discovery and optimization of drug leads that target these centers of interaction. In addition, given a structure of a novel complex with an unknown function, the method can be applied to recognize complexes with similar binding organizations and biological functions (17).

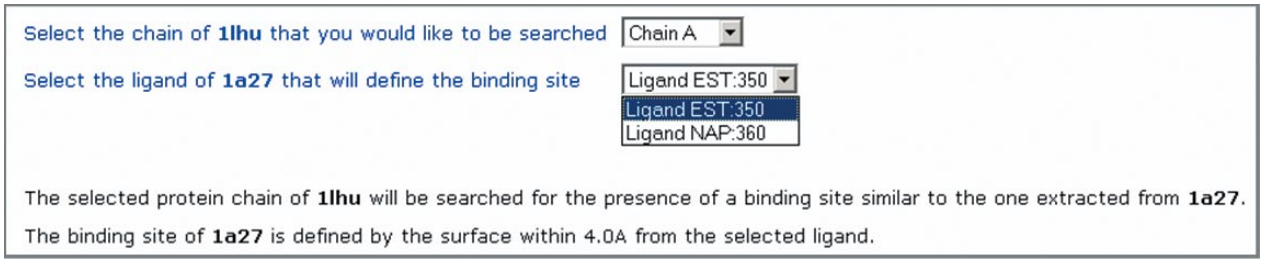
Method

We define an *interface* as an unordered pair of interacting binding sites (*A* and *B*) that belong to two non-covalently linked protein molecules. An interface is represented by a pair of interacting surfaces and a set of pseudocenters that create them. The interacting surfaces are defined by a set of solvent accessible surface points (21) that are located within

(a)



(b)



(c)

SiteEngine results for searching **EST** 350 binding site of **1a27** on the complete surface of **1lhu** (chain A).
Save all results: all_results.zip

Solution : 1 aligned.pdb

Rigid Transformation: -2.91659 0.496356 -1.54859 31.825 19.2983 17.844

Similarity Score: 3970.68 rmsd: 0.202234 Match Size: 12

Chain.ID	Amino Acid	Property	Source	Chain.ID	Amino Acid	Property	Source	Dist.	Conserved AA
A.40	Thr	DON	b	.142	Ser	DAC	b	1.6	
A.40	Thr	DAC	s	.185	Cys	ACC	b	2	
A.40	Thr	ALI	s	.143	Val	ALI	b	1	
A.41	Ser	DAC	s	.155	Tyr	DAC	b	0.43	
A.42	Ser	DAC	s	.152	Asn	DON	b	1.5	
A.56	Phe	PII	s	.218	Tyr	PII	b	1.6	
A.107	Met	ACC	b	.186	Gly	ACC	b	0.79	
A.107	Met	ALI	s	.187	Pro	ALI	b	1.6	
A.129	Gly	DON	b	.258	Arg	DON	b	0.81	
A.130	Pro	ACC	b	.258	Arg	ACC	b	1.5	
A.139	Met	ALI	s	.279	Met	ALI	b	2	*
A.171	Leu	ALI	s	.149	Leu	ALI	s	2	*

(d)

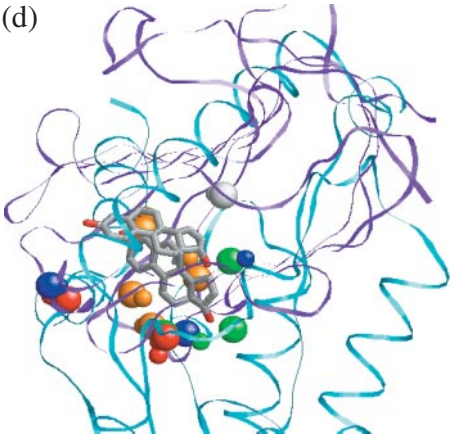


Figure 1. Web interface of SiteEngine. (**a**) Main web server interface. (**b**) Chains and binding site selection form. (**c**) Output page example. (**d**) Visualization of the superimposition defined by the output PDB file (aligned.pdb) for the example in (c). The web server provides a rasmol script (pha.rsm provided in all_results.zip) which colors the physicochemical properties shared by the aligned binding sites. Hydrogen-bond donors (DON) are colored blue; acceptors (ACC), red; donors/acceptors (DAC), green; hydrophobic aliphatic (ALI), orange; and aromatic (PII), white.

4 Å from the surface of the binding partner. The pseudocenters are extracted as described above, according to the definition of Schmitt *et al*. (13)

Given the representation described, we compute the candidate transformations that superimpose one interface on the other. To increase the speed and the quality of the alignment, we construct transformations only on triplets of *interacting* pseudocenters which have a *complementary* physicochemical property (with which it can interact) at the other binding site. Specifically, hydrogen-bond donors are complementary to acceptors, while hydrophobic aliphatic and aromatic pseudocenters can interact only with similar ones (17,20). Each candidate transformation is then scored by a set of scoring functions, which are an extension of those described for SiteEngine (7), to simultaneously compare two pairs of binding sites. The top ranking solutions are the transformations that, when applied, maximize the similarity of the physicochemical properties and shapes of the interfaces.

Input

The input to I2I-SiteEngine consists of two protein–protein complexes, which are defined either by the PDB codes or by

the uploaded structures. Since many protein–protein interfaces are part of multimolecular ensembles, we recognize the protein chains that interact with each other and prompt the user to select the interface of interest. (Two protein chains are considered to be interacting if there are at least five atoms of one chain that are within a distance of 6.0 Å of the other chain.) The interface is automatically extracted and the I2I-SiteEngine method is applied.

Output

Given two input interfaces $I = (A, B)$ and $I' = (A', B')$ we assume that the correspondence between the binding sites of the two complexes is unknown, that is, that binding site A can be aligned either to A' or to B'. Both alignments are considered and the solution that provides the highest score is selected. Assume that the highest scoring correspondence is obtained in the alignments of binding sites A to A' and B to B'. Then, the output of I2I-SiteEngine will first present the details of the alignment of the pair of binding sites (A, A') and then of the pair (B, B') in a format similar to the one described for SiteEngine. For each solution we provide an output PDB file (aligned.pdb) with the superimposition of the two complexes. Additional details can be found at http://bioinfo3d.cs.tau.ac.il/I2I-SiteEngine/help.html.

PERFORMANCE AND AVAILABILITY

The web servers of SiteEngine and I2I-SiteEngine are available from http:/bioinfo3d.cs.tau.ac.il/. Given the input PDB structures, the software is immediately invoked and the results are presented to the user. The response time is a matter of minutes and the main bottlenecks of the server are its general overload and the construction of the surfaces and grids, which is done before the algorithm invocation. Users who are interested in performing large-scale database searches (7,17) are advised to download the software packages, which are freely available for download from the web servers. The packages contain the software as well as additional scripts for database construction, screening and ranking. The user manual is also provided.

CONCLUSIONS AND FUTURE WORK

We have presented two web servers for online recognition and comparison of binding sites and protein–protein interfaces. The algorithms behind the web servers are highly efficient and have been applied to perform large-scale database searches of the entire PDB (7). Recently, Mintz *et al*. (17) applied I2I-SiteEngine to perform 5 million comparisons in order to classify all the protein–protein complexes currently available in the PDB. By downloading the software packages the user can perform an offline construction and search of any database of interest. Next we intend to develop an efficient method of preprocessing the available structural data, which will allow online searches of the entire PDB.

ACKNOWLEDGEMENTS

We would like to thank Maxim Shatsky, Dina Schneidman and Shira Mintz for useful discussions and technical help. We would like to thank Dr Shuo Liang Lin for valuable suggestions. This research has been supported in part by the Center of Excellence in Geometric Computing and its Applications funded by the Israel Science Foundation (administered by the Israel Academy of Sciences) and by the Tel Aviv University Adams Brain Center. The research of H.J.W. is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University. The research of R.N. has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the view or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Funding to pay the Open Access publication charges for this article was provided by SAIC-Frederick, Inc.

Conflict of interest statement. None declared.

REFERENCES

1. Wolfson,H.J., Shatsky,M., Schneidman-Duhovny,D., Dror,O., Shulman-Peleg,A., Ma,B. and Nussinov,R. (2005) From structure to function: methods and applications. *Curr. Prot. Pept. Sci*., **6**, 171–183.
2. Kinoshita,K. and Nakamura,H. (2004) Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci*., **14**, 711–718.
3. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol*., **307**, 1113–1143.
4. Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol*., **321**, 741–765.
5. Moodie,S.L., Mitchell,J.B.O. and Thornton,J.M. (1996) Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol*., **263**, 486–500.
6. Denessiouk,K.A., Rantanen,V.S. and Johnson,M.S. (2001) Adenine Recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins*, **44**, 282–291.
7. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2004) Recognition of functional sites in protein structures. *J. Mol. Biol*., **339**, 607–633.
8. Artymiuk,P.J., Poirrette,A.R., Grindley,H.M., Rice,D.W. and Willett,P. (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol*., **243**, 327–344.
9. Russell,R.B. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol*., **279**, 1211–1227.
10. Spriggs,R.V., Artymiuk,P.J. and Willett,P. (2003) Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci*., **43**, 412–421.
11. Binkowski,T.A., Adamian,L. and Liang,J. (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol*., **332**, 505–526.
12. 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.
13. 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.
14. Kinoshita,K. and Nakamura,H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci*., **12**, 1589–1595.
15. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci*., **5**, 1001–1013.

16. 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.
17. Mintz,S., Shulman-Peleg,A., Wolfson,H.J. and Nussinov,R. (2005) Generation and analysis of a protein–protein interface dataset with similar chemical and spatial patterns of interactions. *Proteins*, in press.
18. Valdar,W.S. and Thornton,J.M. (2001) Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–124.
19. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
20. Shulman-Peleg,A., Mintz,S., Nussinov,R. and Wolfson,H.J. (2004) Protein–protein interfaces: recognition of similar spatial and chemical organizations. In Jonassen,I. and Kim,J. (eds), *Proceedings of the Fourth International Workshop on Algorithms in Bioinformatics, LNCS 3240*. Springer-Verlag GmbH, pp.194–205.
21. Connolly,M.L. (1983) Analytical molecular surface calculation. *J. Appl. Crystallogr.*, **16**, 548–558.