W398–W402 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm351

eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape

Kengo Kinoshita[1,2,*], Yoichi Murakami[3] and Haruki Nakamura[3]

[1]Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo, 108-8639, Japan, [2]Structure and Function of Biomolecules, SORST, Japan Science and Technology Corporation, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan and [3]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan

Received January 30, 2007; Revised April 22, 2007; Accepted April 23, 2007

ABSTRACT

We have developed a method to predict ligand-binding sites in a new protein structure by searching for similar binding sites in the Protein Data Bank (PDB). The similarities are measured according to the shapes of the molecular surfaces and their electrostatic potentials. A new web server, eF-seek, provides an interface to our search method. It simply requires a coordinate file in the PDB format, and generates a prediction result as a virtual complex structure, with the putative ligands in a PDB format file as the output. In addition, the predicted interacting interface is displayed to facilitate the examination of the virtual complex structure on our own applet viewer with the web browser (URL: http://eF-site.hgc.jp/eF-seek).

INTRODUCTION

In the post-genomic era, the functional identification of gene products, proteins, is one of the most important steps toward understanding living organisms at the molecular level. Progress in structural genomics and structural proteomics (1–4) is producing vast amount of structural information, and recently even structural information for hypothetical proteins, which are proteins without any functional annotations, is becoming available. The biochemical functions of proteins are strongly correlated with their 3D structures, and thus the structural information is increasingly useful to infer the function of the protein (5,6). However, the relationship between protein structure and function has not yet been elucidated and it seems to be very complicated. Therefore, the similarity searches against known proteins represent a practical approach to infer the function, as in a sequence similarity search, where the basic but naïve assumption is that proteins with similar structure should have similar functions (5,6).

Several approaches based on similarity searches against known proteins with structural information have been developed by many groups (5–9). In most cases, the function prediction based on the protein structure is focused on the prediction of the ligand-binding site as the first step toward determining the molecular function of proteins. We have also been developing our own method to predict the ligand-binding sites of proteins, which focuses on the protein molecular surfaces (10) and the electrostatic potential at the surfaces (11), where most interactions with small molecules will occur. Our method has been successfully used for over three years to infer the function of hypothetical proteins (12,13). One of the examples for a hypothetical protein is TT1576 from *Thermos thermophilus* (13). We predicted it might be involved in the sugar hydrolysis, and it is now strongly supported by the structural determination of the homologous protein MshB (14). Since surface similarity searches require more time-consuming calculations as compared to similarity searches of folds and/or the spatial arrangement of atoms in the functional sites, the implementation of a similarity search on the web was not an easy task. However, we have constructed a web server, eF-seek, which uses GRID technology to overcome these difficulties. We used the Sun Grid Engine to mange the jobs over two different architectures with four cluster computers in two universities. eF-seek provides a user-friendly interface to our search method against representative ligand-binding sites in the eF-site database, which is a database of the electrostatic surfaces of functional sites of proteins (15).

Flow of job submission to eF-seek

In the first step, eF-seek requires a coordinate file in the Protein Data Bank (PDB) (16) format and an e-mail address for reporting the job information. The user can

*To whom correspondence should be addressed. Tel: +81-3-5449-5131; Fax: +81 3 5449-5133; Email: kino@ims.u-tokyo.ac.jp

also specify a title for the job or make a note on the submitted job as an option. When the job is submitted, the server carries out the format check of the uploaded file. If no errors are found, the user receives an e-mail including the URL of the *job-start page*. The most frequent errors are caused by an incorrect coordinate file format. The coordinate file must be in the PDB format. Less frequent, but serious errors come from unusual atoms in the ATOM record in the coordinate file, when they are not the normal elements of the natural 20 amino acids in proteins. All of the atoms in HETATM records are basically ignored, but MSEs (selenomethionine) are used as methionine in the following calculations. To calculate the electrostatic potential, all of the hydrogen atoms are automatically generated by referring the stable covalent local conformations based on the AMBER force field (17), which also provide the charge on every atom in the query protein using program PRESTO (18). Thus, at the moment, our system cannot process a molecular system that includes such unusual atoms in the ATOM record, and it issues a failure notice through e-mail.

In the second step, the user confirms the title of the job and the file name of the upload file described in the *job-start page*. When the start button on the *job-start page* is clicked, the calculation is registered in the eF-seek server. This step serves to validate the e-mail address entered in the first step. The user can also cancel the calculation request at this stage.

After the registration of a job, the server checks the registered jobs in the queue every 5 min and starts the calculation with the first job in the queue, where a job queue is generated for each user, in order to prevent a calculation delay caused by multiple jobs from a single user. Although the calculation time largely depends on the number of other jobs and the size of the query protein, the typical calculation time will be from a few hours to one day if no other jobs are in the calculation queue at the same time. Large proteins may take more calculation time. It may be noteworthy that all jobs are prioritized for each user, thus multiple jobs by a single user will not affect the other users' jobs much, but it will take more time to finish the multiple jobs when compared with that of other users' jobs.

When the calculation is finished, the user receives an e-mail notification. In the e-mail, a URL to access the search result will be included, if no calculation errors occur. The typical error in this step is arising from the fact that the size of the protein is too large. In that case, cutting out a domain from the entire structure will be a useful approach.

Access to the result page

The search result is shown visually in a clickable *density-plot* implemented as a Java applet (Figure 1). The details of the density plot in the result page are described in Kinoshita and Nakamura (12), and we will briefly describe it here.

For each query, eF-seek carries out a comparison against all binding sites in the representative ligand database, in which about 17 500 binding sites have been registered. The similarity for each binding site is measured by two values, that is, Z-score and coverage. The Z-score assesses how similar each binding site is as compared with all other binding sites in the binding site database, and the coverage evaluates the breadth of a similar area of the binding site, i.e. the ratio of the number of the corresponding vertexes to that of all vertexes in each binding site. Thus, a binding site with a larger Z-score value and a larger coverage value should have more similar features to the query. The comparisons against the representative binding site database generate Z-score and coverage values, which are represented in the form of a 2D scatter plot between the Z-score and the coverage. Since the number of dots is quite large (about 17 500), a density plot is shown instead of a scatter plot in the *result page* (Figure 1).

In the density plot shown in Figure 1, the significantly similar binding sites [or the binding sites above the threshold line (12)] are represented as filled circles (Figure 1). Each filled circle is clickable, and the user can check the name of the binding site in the text box below the density plot (Figure 1). When some dots are overlapped in the plot, the names of the binding sites are shown in the text box. To examine the predicted complex structure interactively, the user can select one binding site in the text box by clicking the name, and push the view complex button, located just below the text box (Figure 1), and then the new window of the view structure page will appear (Figure 2). The ligand in the putative binding site is simply obtained by rotation and translation, according to the same rotation matrix and the same translation vector that is obtained by the superimposition between the query surface and the subsurface of the binding site. Therefore, some ligands from less similar binding sites can have collisions between the ligand atoms and protein atoms.

The complex structure with the predicted ligand is visualized using jV, which is the latest version of the PDBjViewer (15). The jV uses the JOGL library to render the object, such as a cartoon model of a protein and the surface structure, and thus it requires the installation of JOGL for the first use. The jV runs with typical browsers on Windows and Macintosh PCs. See the jV web page at http://www.pdbj.org/PDBjViewer for details.

In the *view structure page*, two jV panels appear (Figure 2). In the left panel, the putative complex structure with a cartoon model is shown, and the interacting residues are depicted by ball-and-stick models. In the right panel, a surface model with a putative ligand is provided. The molecular surface is colored according to the electrostatic potential at the surfaces in the same way as the eF-site database (15). The two views can be rotated, translated and zoomed simultaneously, according to the mouse operation, and thus the prediction results are easily examined intuitively.

The jV panels will be useful for a quick check of our prediction, but if the user wants to examine it in detail, it may be more convenient to download the structure to the user's own PC. The complex structure in the PDB format file can be downloaded by using the *download PDB*

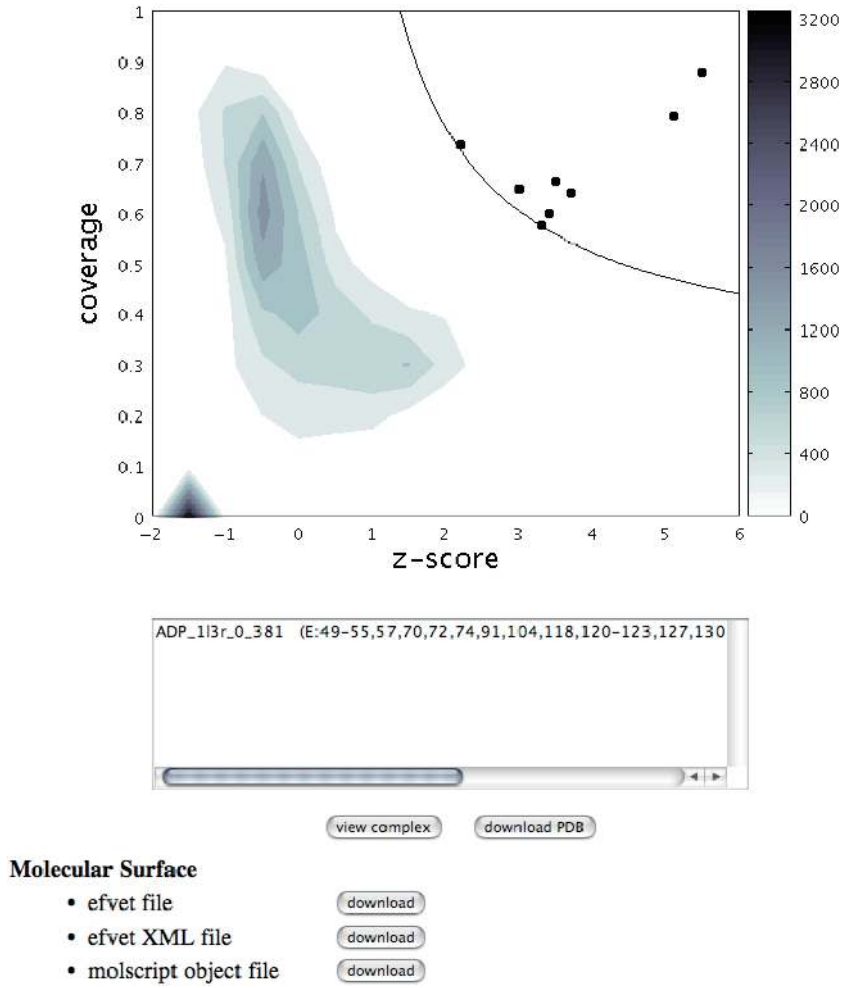

Figure 1. An example of a *result page*. In the result page, a density plot of the search result and a text box are shown, as described in the main text. In this example, c-AMP- dependent protein kinase (PDB: 1atp chain E) was used as a query, and the name of the top hit, which corresponds to the filled circle in the top right corner, is shown in the text box. The scale of the density plot is shown according to the color bar on the right side of the plot.

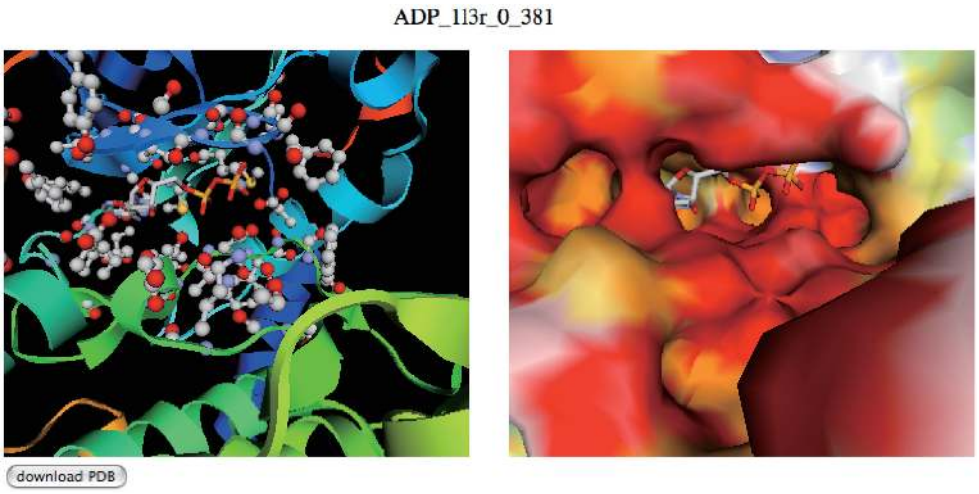

Figure 2. An example of a *view structure page*. In the view structure page, interactive view of the putative complex is provided with two jV [15] panels. The complex structure with ribbon model (left panel) and that with the surface model (right panel) are presented. In the right panel, the residues within 5.0 Å from the putative ligand are specified with ball and stick model. These viewers can be rotated and translated with the mouse operation synchronously. In this example, the binding mode of ADP with the query protein (PDB: 1atp, E chain), predicted from the binding site appearing in 1l3r in PDB (c-AMP-dependent protein kinase), is shown.

button next to the *view complex* button in the *result page* (Figure 1) or in the *view structure page* (Figure 2). The surface structure with the electrostatic potential in our own format and/or the Molscript (19) format can also be downloaded from the *result page*. The electrostatic potential and molecular surface are calculated under exactly the same conditions as those used in the eF-site. The calculation of the electrostatic potential is independently available at http://ef-site.hgc.jp/eF-surf.

Representative ligand-binding site

In the PDB in Aug 2006, there were 350 101 binding sites with 2.5 Å or better resolution, excluding the binding sites of HOH, WAT, PO3, SO4, SUL, MSE, PEG, EPE, IOH, PG4, DMS, TPO, SEP, PTR, HIP, PAS, ASQ, DOD, ACT, EDO, TRS, CL1, CIT, NO3 and FS4, because most of them are the results of crystallization additives. In addition, we excluded the sugar-binding sites (NAG, GLC, MAN, FUC and BGC), due to the weak interactions and the lower prediction accuracy. We also excluded the lipids (LDA and BOG) bound to membrane proteins, small ions (CA, ZN, NA and CL) and a large ligand (HEM), since the prediction of their binding sites is not very informative. The main purpose of the elimination is to reduce the number of false positive ligands as much as possible. The false positive hits are usually coming from the non-specific binding sites or the binding sites of the small molecules of crystallization. However, it is almost impossible to check all the binding sites of the small molecules in advance. Therefore, we now decided to take the strictest approach that is to eliminate all the ligands that can often be found in non-specific binding. Some of the eliminated ligands can actually be legitimate ligands, and that may weaken the detection power of eF-seek, but many of them are expected to be the artifacts of crystallization in the current system.

For the other ligands, we carried out all-against-all comparisons of the spatial arrangement of the surrounding atoms (20) with the same hetero compound code (3-letter code given by the PDB, such as ATP), and selected one representative binding site for each group that is identified by single linkage clustering, with a maximum r.m.s.d. of 1.0 Å and a minimum coverage of corresponding atoms of 70% (70% or more atoms of the surrounding atoms in each binding site should be corresponding). This selection aims to reduce the computation time and it is expected to significantly influence the detection power as discussed later. As a result, in the current database, 17 528 binding sites are registered. Now, we are planning to update the representative binding sites twice per year.

In the current version of the representative binding site, similarities are evaluated by the comparison of atomic configurations. As compared in Kinoshita *et al.* [21], similar atomic configurations usually indicate similar molecular surfaces. Therefore, the selection of representative binding site based on the similarity of atomic configurations does not influence much the detection power of the eF-seek.

ACKNOWLEDGEMENTS

The authors thank Dr Hiroyuki Sato (Information and Mathematical Science Laboratory, Inc.) for his technical support. Development of the eF-seek server has been supported by Grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Corporation (BIRD-JST) to HN, for Strategic Japan-UK Cooperative Program in the field of Structural Genomics and Proteomics, JST to KK, YM and HN, and for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan to KK. Funding to pay the Open Access publication charge was provided by BIRD-JST to KK.

Conflict of interest statement. None declared.

REFERENCES

1. Bateman,A. and Valencia,A. (2006) Structural genomics meets computational biology. *Bioinformatics*, **22**, 2319.
2. Lundstrom,K. (2006) Structural genomics: the ultimate approach for rational drug design. *Mol. Biotechnol.*, **34**, 205–212.
3. Rigden,D.J. (2006) Understanding the cell in terms of structure and function: insights from structural genomics. *Curr. Opin. Biotechnol.*, **17**, 457–464.
4. Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
5. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.*, **15**, 275–284.
6. Kinoshita,K. and Nakamura,H. (2003) Protein informatics towards function identification. *Curr. Opin. Struct. Biol.*, **13**, 396–400.
7. Powers,R., Copeland,J.C., Germer,K., Mercier,K.A., Ramanathan,V. and Revesz,P. (2006) Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins*, **65**, 124–135.
8. Gold,N.D. and Jackson,R.M. (2006) Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.*, **355**, 1112–1124.
9. Dawe,J.H., Porter,C.T., Thornton,J.M. and Tabor,A.B. (2003) A template search reveals mechanistic similarities and differences in beta-ketoacyl synthases (KAS) and related enzymes. *Proteins*, **52**, 427–435.
10. Connolly,M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.
11. 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.
12. Kinoshita,K. and Nakamura,H. (2005) Identification of the ligand binding sites on the molecular surface of proteins. *Protein Sci.*, **14**, 711–718.
13. Handa,N., Terada,T., Kamewari,Y., Hamana,H., Tame,J.R., Park,S.Y., Kinoshita,K., Ota,M., Nakamura,H. *et al.* (2003) Crystal structure of the conserved protein TT1542 from Thermus thermophilus HB8. *Protein Sci.*, **12**, 1621–1632.
14. McCarthy,A.A., Peterson,N.A., Knijff,R. and Baker,E.N. (2004) Crystal structure of MshB from Mycobacterium tuberculosis, a deacetylase involved in mycothiol biosynthesis. *J. Mol. Biol*, **335**, 1131–1141.
15. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
16. 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.
17. Case,D.A., Cheatham,T.E.,3rd, Darden,T., Gohlke,H., Luo,R., Merz,K.M., Jr., Onufriev,A., Simmerling,C., Wang,B. *et al.* (2005)

The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
18. Nakajima,K., Higo,J., Kidera,A. and Nakamura,H. (2000) Free energy landscapes of peptides by enhanced conformational sampling. *J. Mol. Biol.*, **296**, 197–216.
19. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of proteins structures. *J. Appl. Cryst.*, **24**, 946–950.
20. Kinoshita,K., Sadanami,K., Kidera,A. and Go,N. (1999) Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. *Protein Eng.*, **12**, 11–14.
21. Kinoshita,K., Furui,J. and Nakamura,H. (2002) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics*, **2**, 9–22.