W320–W324 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl267

PreBI: prediction of biological interfaces of proteins in crystals

Yuko Tsuchiya[1], Kengo Kinoshita[2,3,*], Nobutoshi Ito[4] and Haruki Nakamura[1]

[1]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka, 565-0871, Japan,
[2]Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo, 108-8639, Japan,
[3]Structure and Function of Biomolecules, SORST, JST, 4-1-8 Honcho, Kawaguchi, Saitama,
332-0012, Japan and [4]School of Biomedical Science, Tokyo Medical and Dental University, 1-5-45 Yushima,
Bunkyo-ku, Tokyo, 113-8510, Japan

Received February 14, 2006; Revised March 1, 2006; Accepted March 31, 2006

ABSTRACT

PreBI is a server that predicts biological interfaces in protein crystal structures, according to the complementarity and the area of the interface. The server accepts a coordinate file in the PDB format, and all of the possible interfaces are generated automatically, according to the symmetry operations given in the coordinate file. For all of the interfaces generated, the complementarities of the electrostatic potential, hydrophobicity and shape of the interfaces are analyzed, and the most probable biological interface is identified according to the combination of the degree of complementarity derived from the database analyses and the area of the interface. The results can be checked through an interactive viewer, and the most probable complex can be downloaded as atomic coordinates in the PDB format. PreBI is available at http://pre-s.protein.osaka-u.ac.jp/~prebi/.

INTRODUCTION

X-ray crystallography is a powerful tool to determine the 3D structures of proteins. It is especially effective for analyzing the large protein complexes, and many protein complexes have been determined according to the recent progress of structural genomics projects (1,2). A problem with crystallography, however, is that crystals contain both crystallographic contacts and biologically relevant contacts. Therefore, it is necessary to discriminate between the biological and crystallographic contacts, for the structural information to be useful for understanding the functions of proteins.

The biological interfaces in crystal structures are usually identified by relying on biological information obtained by experiments, such as site-directed mutagenesis, alanine scanning and/or information inferred from that obtained for their homologous proteins. However, when no such information is available, one has to find the answer solely from the structural information.

To address this problem, some methods for biological interface identification have been developed. They are usually based on the observation that the interface with the largest contact area tends to be the biologically relevant interface, and they search for the interface with the maximum contact area among all of the possible contacts in the crystal or seek for the optimum value of the score function strongly related with the contact area. The assurance of the methods is relatively high (around 85% accuracy) (3–5), but the high performance can introduce the bias that the interface with the largest interface is selected as the possible biological interface, even when no experimental support is available. However, the largest interface is not always the biological interface, as in the case of human telomeric protein TRF2 (6), as described later. Therefore, we tried to develop another method to discriminate the biological interface from the crystal contacts.

Our indicator is made based on statistical analyses of the homo-interfaces within the PDB. The details of the analysis will be described elsewhere, but here we will describe it briefly. The analysis was done for 393 and 344 non-redundantly selected homo-interfaces for biological interfaces and crystallographic interfaces, respectively, by focusing on the complementarity of electrostatic potential, hydrophobicity and shape of the molecular surfaces. Therefore, the current version of our server is limited to the analysis of homo-interfaces, and the interfaces of ligand–protein and hetero complexes will be considered in the next version. These interfaces were selected by gathering all of the homo-oligomeric proteins in the PDB and choosing one from each SCOP family (7). The electrostatic potential was obtained by solving the Poisson–Boltzmann equation numerically with the program SCB (8), and the hydrophobicity was calculated by the

*To whom correspondence should be addressed at Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo, 108-8639, Japan. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: kino@ims.u-tokyo.ac.jp

Ooi–Oobatake method (9). These physicochemical features were mapped to every vertex of the molecular surface obtained by Connolly's algorithm (10), and the shape of the molecular surface was described using the curvature for each vertex (11). Then, the complementarity was evaluated by counting the number of complementary pairs of vertices. The pairs of vertices coming from different surfaces within less than a 1.0 Å distance are considered as complementary pairs of vertices if they have the opposite signs of electrostatic potential, the same sign of hydrophobicity or the opposite signs of curvature. The number of complementary pairs in each interface is converted to the ratio by dividing it by the number of all pairs in the interface, and the ratio is further divided by the median value of the ratios among all of the non-redundant homo-interfaces. This value is calculated for each property, and the sum of the values from all of the properties is used as a measure to find the biological interface, and we refer to it as the *degree of complementarity*.

INPUTTING DATA AND ACCESSING RESULTS

The server requires the coordinates of the protein's 3D structure in the PDB format (12), along with a chain identifier to specify the protomer to search for the biological interface, and the user's e-mail address for notification of the completion of the calculation. The PDB file can either be uploaded, or specified as the PDB-ID if the structure has already been registered in the PDB, and the coordinates appeared in ATOM record are used without any modification. It should be noted that the coordinate file must include information about the unit cell parameters and the space group symbol. More precisely, the symmetry operators and the scale matrix in the REMARK290 and SCALE records are needed for the calculation. However, if the records are not available, the server can generate the operators and the matrix using the CRYST1 record, if it is available. The existence of these records is checked automatically when the query is submitted, and the user will be required to confirm the submission, if the check is finished successfully.

A typical calculation will take several hours on a single CPU system. The calculation time largely depends on the sizes of the proteins and the degree of symmetry. When the degree of symmetry is high, the number of possible interfaces is large, and thus it will take more time to find the biological interface.

When the calculation is finished successfully, an e-mail will be sent. The e-mail contains two URLs of the result web pages, one for the prediction result based on our complementarity analyses and the other for that based on the maximum area of contact. In addition, our suggestion about which result is more probable is also included in the e-mail. In the result web pages, the user can access the interactive view of the calculation results using pdbjviewer (13), and the coordinate file of the most probable complex, in the PDB format, can be downloaded.

EXAMPLE OF A RESULT

The human telomeric protein, TRF2, is known to form a dimer under physiological conditions, and there are four different interfaces in the crystal structure [PDB: 1h6p (6)]. One of the interfaces is used as a dimerization domain, consisting of a four-helix bundle, and its contact area (464.7 $Å^2$) is smaller than the largest interface that is formed by two long and one short helices (617.0 $Å^2$). The other two interfaces were neglected, because they are too small.

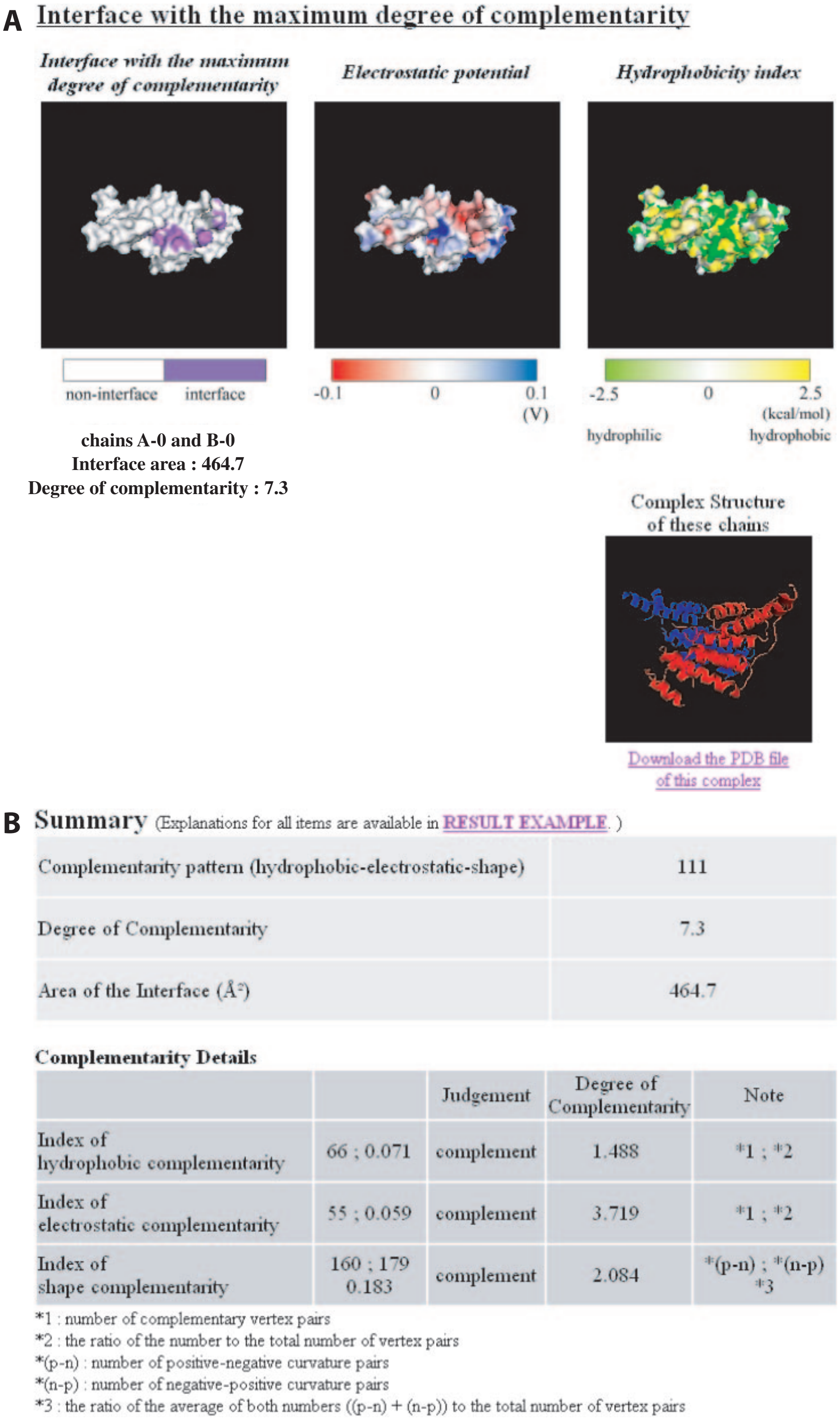
Since the PDB file of this protein contains two identical chains, A- and B-chains, the user should specify the A or B chain as the *representative* chain, for which all of the possible interfaces are considered. As this protein is a homo-dimer, the selection of the chain will not greatly affect the results in most cases, and thus we used the A-chain as the representative chain, as an example. However, it should be noted that proteins in crystal structures are sometimes missing some residues, which could change the result, depending on the selection of the representative chain.

When the user submits a job using PDB-ID (1h6p), the server can use the REMARK 290 and SCALE matrix data that appeared in the original PDB file, and then a confirmation page will appear immediately. If the uploaded coordinate file lacks this information and if it does not have the CRYST1 record, then an error page will be shown just after the submission. If the submission is successful, the user can start the job by pressing the 'START CALCULATION' button on the confirmation page, and then the results will be sent via e-mail in about 2 h for this protein, if no other jobs are being processed.

In the returned e-mail, the user will find two URLs, one for the prediction based on our method, and the other for the results obtained as the maximum contact surface. An example page for the result based on our prediction is shown in Figure 1, and that based on the maximum area is shown at http://pre-s.protein.osaka-u.ac.jp/~prebi/result.html.

The result page consists of two parts. In the upper part, the viewer part (Figure 1A), four viewers are attached to show the contact area, the electrostatic surface, the hydrophobicity of the molecular surface and the most probable contact pair with the ribbon model. The molecules in the four viewers can be translated and rotated interactively and synchronously. In addition, the most probable complex can also be downloaded by following the link located just below the cartoon model. In the lower part (Figure 1B), the details of the complementarity analysis are shown in two tables, the summary table and the complementarity details table.

The server generates all possible interfaces with other protomers in the crystal, as described later, and will add the chain identifier for all of the protomers as follows. For example, if two chains, A and B, are found in the PDB, then the protomers in the asymmetric unit (ASU) will be called A-0 and B-0. The other protomers in the unit cell will be called A-1, B-1, A-2, B-2, . . . , A-*n* and B-*n*, where the *n* is the number of each chain in the unit cell. When we consider the adjacent cells to find the possible interface, two adjacent cells for each x-, y- and z-direction can exist. Each protomer in the adjacent cell will be called A-0+x, A-0-x, A-1+x, A-2-x and so on. In the case of 1h6p, the space group of the crystal is C2 (C1 2 1) and there are two molecules in the ASU, and thus eight molecules, A-0 to A-3 and B-0 to B-3, can be generated. In addition, there are six cells adjacent to the center unit cell, and therefore $6 * 8 + 8 = 56$ protomers are generated in order to find the possible interface. It should be noted that there are 26 adjacent cells around the center unit cell, however, only the six adjacent cells (adjacent in each x-, y-, z-direction) are


A
Interface with the maximum degree of complementarity
Electrostatic potential
Hydrophobicity index
non-interface
interface
-0.1
0
0.1
(V)
-2.5
0
2.5
(kcal/mol)
hydrophilic
hydrophobic
Complex Structure of these chains
Download the PDB file of this complex
B

Interface with the maximum degree of complementarity

chains A-0 and B-0
Interface area : 464.7
Degree of complementarity : 7.3

Summary (Explanations for all items are available in RESULT EXAMPLE.)

Complementarity pattern (hydrophobic-electrostatic-shape)	111
Degree of Complementarity	7.3
Area of the Interface (Å^2)	464.7

Complementarity Details

		Judgement	Degree of Complementarity	Note
Index of hydrophobic complementarity	66 ; 0.071	complement	1.488	*1 ; *2
Index of electrostatic complementarity	55 ; 0.059	complement	3.719	*1 ; *2
Index of shape complementarity	160 ; 179 0.183	complement	2.084	*(p-n) ; *(n-p) *3

*1 : number of complementary vertex pairs
*2 : the ratio of the number to the total number of vertex pairs
*(p-n) : number of positive-negative curvature pairs
*(n-p) : number of negative-positive curvature pairs
*3 : the ratio of the average of both numbers ((p-n) + (n-p)) to the total number of vertex pairs

Figure 1. An example of the result page for PDB: 1h6p. (**A**) In the viewer part (upper half of the result page), four viewers are attached. The surface views from the left to right at the top of the page show the possible interface (purple), the electrostatic potential [from red (negative) to blue (positive)], and the hydrophobicity [from green (hydrophilic) to yellow (hydrophobic)]. The remaining view shows the complex with the maximum degree of complementarity. (**B**) In the table part (lower half of the result page), a summary table with the complementarity details is shown. The summary table describes the complementarity pattern, the degree of complementarity (3.6, see Calculation flow section), and the area of the interface in Å^2 unit. The complementarity pattern indicates whether the interface is complementary (1) or not (0) for each property (hydrophobicity, electrostatic potential and shape, in this order). For example, 111 indicates that the interface is complementary for all properties and 101 means the interface is complementary in hydrophobicity and shape, but not in electrostatic potential. In the complementarity details table, the raw data of the number of complementary pairs of vertices are shown.

considered in order to reduce the computation time in default by assuming that the contact area of the other 20 cells are usually small compared with the six *main* adjacent cells. The user can choose the full calculation using the 26 adjacent cells by enabling the check box in the submission page.

Among the 56 protomers, all possible combinations were checked, and the interface made from the A-0 and B-0 protomers was found to have the highest degree of complementarity, and that from the A-0 and A-1+z protomers was the interface with the maximum area. In many cases, the interface with the maximum degree of complementarity and that with the maximum area are the same, but in this example, they were different. According to the primary reference of the 1h6p structure (6), the interface with the maximum degree of complementarity is the biological interface, and the other interface with the maximum area of contact is the crystallographic one. The server's prediction of the most probable interface is included in the e-mail.

CALCULATION FLOW

Step 1: Generation of symmetry-related protomers

When the submission is successful, the amino acid sequence of the specified protomer in the input step will be compared with all of the other protomers in the ASU by using FASTA (14), and the protomers with sequence identity >85% are chosen for the next step. Then, the symmetry-related protomers in the unit cell and the adjacent six cells (two each for the x-, y-, and z-directions) are generated according to the symmetry operations appearing in the coordinate file, as described. It may be noteworthy that we used the amino acid sequence within the ATOM record in the PDB, in which some flexible loops are missing, and that even the same protein can sometimes appear to have a different sequence in the ATOM record. As described in the Example of a result section, although TRF2 is a homodimeric protein, both chains lack some residues in the crystal structure, and thus the sequence identity is 97.4%. Therefore, we adopted the 85% threshold of sequence identity, as the safe criteria to obtain the protomers that should be checked for contact between the specified protomer.

Step 2: Determination of the contacting pairs of protomers and identification of all possible interfaces

The distances between all pairs of protomers that were obtained in the previous step are calculated, where the distance between a pair of protomers is defined as the minimum distance of the pairs of atoms belonging to the different protomers. When the distance between a pair of protomers is <4 Å, the protomer pair is regarded as the contact-protomer pair. For each protomer, a molecular surface is generated by Connolly's algorithm (10), and the pairs of vertices that belong to different molecular surfaces and that have a distance <1.0 Å are defined as the vertices in the interface of the contact-protomer pair.

Step 3: Complementarity analysis

The degree of complementarity is calculated for each interface, and it is the sum of the complementarity of the three properties, that is, electrostatic potential, hydrophobicity and shape complementarity. The complementarity for a property is calculated as the ratio of the percentage of complementary pairs of vertices among the vertices in each interface to its median value in the learning dataset (393 non-redundantly selected proteins from each family in the SCOP database (7), which are shown at http://pre-s.protein.osaka-u.ac.jp/~prebi/393entry.html).

Step 4: Selection of the most probable interface

According to the complementarity analyses in step 3, the most probable interface is selected as the interface with the largest degree of complementarity or with the maximum contact area from all of the contact-protomer pairs picked up in step 2. The selection is performed as follows: (i) the interface with the largest degree of complementarity and that with the maximum area are selected, but the interface whose contact area <100.0 $Å^2$ is not considered. (ii) If their contact areas and their degrees of complementarity do not meet the criteria of $\geq$290.0 $Å^2$ and $\geq$1.25, respectively, then the interface is not considered as a possible interface. If both of the interfaces exceed the threshold, then the interface with the largest degree of complementarity is judged as the most probable interface. If the both of the interfaces do not meet the criteria, the interfaces are considered as non-biological. These criteria were determined by optimizing the Matthew's correlation coefficient in the learning dataset, where the optimum values were 0.78 for the contact area and 0.48 for the degree of complementarity, respectively. And finally, we obtained the performance that the sensitivity was 0.95 and the specificity was 0.79 for the 367 homo-oligomer interfaces and 2640 crystal contacts created according to the symmetry operation appeared in PDB, where 26 homo-interfaces were omitted because they have no crystal contacts. These 26 cases are such entries with non-identical protein chains and/or RNA molecules (e.g. 1e6t) and entries where only some of the protomers in the ASU have non-biological contacts (e.g. 1gtz).

ACKNOWLEDGEMENTS

This work was partially supported by a Research Fellowship from the Japan Society for the Promotion of Science for Young Scientists to Y.T. K.K. was supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (No. 15 710 150), and by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No. 17 081 003). H.N. was supported by Grant-in-Aid for Scientific Research on Priority Areas (No. 17 017 024) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Funding to pay the Open Access publication charges for this article was provided by Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Todd,A.E., Marsden,R.L., Thornton,J.M. and Orengo,C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.*, **348**, 1235–1260.

2. Zhang,C. and Kim,S.H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.*, **7**, 28–32.
3. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
4. Ponstingl,H., Henrick,K. and Thornton,J.M. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.
5. Valdar,W.S. and Thornton,J.M. (2001) Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, **313**, 399–416.
6. Fairall,L., Chapman,L., Moss,H., de Lange,T. and Rhodes,D. (2001) Structure of the TRFH dimerization domain of the human telomeric proteins TRF1 and TRF2. *Mol. Cell*, **8**, 351–361.
7. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structural and sequence family data. *Nucleic Acids Res.*, **30**, 264–267.
8. Nakamura,H. and Nishida,S. (1987) Numerical calculations of electrostatic potentials of protein-solvent systems by the self consistent boundary method. *J. Phys. Soc. Jpn.*, **56**, 1609–1622.
9. Ooi,T., Oobatake,M., Nemethy,G. and Scheraga,H.A. (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl Acad. Sci. USA*, **84**, 3086–3090.
10. Connolly,M.L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.
11. Tsuchiya,Y., Kinoshita,K. and Nakamura,H. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.
12. 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.
13. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
14. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Met. Mol. Biol.*, **24**, 307–331.