BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 10 2005, pages 2541–2543
doi:10.1093/bioinformatics/bti366

Structural bioinformatics

PSIbase: a database of Protein Structural Interactome map (PSIMAP)

Sungsam Gong[1], Giseok Yoon[2], Insoo Jang[3], Dan Bolser[4], Panos Dafas[5], Michael Schroeder[6], Hansol Choi[1], Yoobok Cho[2], Kyungsook Han[7], Sunghoon Lee[3], Hwanho Choi[1], Michael Lappe[8], Liisa Holm[9], Sangsoo Kim[3], Donghoon Oh[2] and Jonghwa Bhak[1,2,3,10,*]

[1]Biomatics Lab, Department of BioSystems, KAIST, Daejeon, Korea, [2]OITEK, Daejeon, Korea, [3]NGIC, KRIBB, Daejeon, Korea, [4]MRC-DUNN, Cambridge, UK, [5]City University, London, UK, [6]Biotechnologisches Zentrum, TU Dresden, Germany, [7]Inha University, Incheon, Korea, [8]Max Planck Institute for Molecular Genetics, Berlin, Germany, [9]Helsinki University, Finland and [10]BiO Centre, KAIST, Daejeon, Korea

Received on November 1, 2004; revised on January 27, 2005; accepted on February 28, 2005
Advance Access publication March 3, 2005

ABSTRACT

Summary: Protein Structural Interactome map (PSIMAP) is a global interaction map that describes domain–domain and protein–protein interaction information for known Protein Data Bank structures. It calculates the Euclidean distance to determine interactions between possible pairs of structural domains in proteins. PSIbase is a database and file server for protein structural interaction information calculated by the PSIMAP algorithm. PSIbase also provides an easy-to-use protein domain assignment module, interaction navigation and visual tools. Users can retrieve possible interaction partners of their proteins of interests if a significant homology assignment is made with their query sequences.

Availability: http://psimap.org and http://psibase.kaist.ac.kr/

Contact: biopark@kaist.ac.kr

Supplementary information: Supplementary material is available at http://psibase.kaist.ac.kr/Doc/supplementary_material.htm

INTRODUCTION

Most proteins function by interacting with other molecules. Therefore, it is important to investigate the interaction partners of proteins. Recently, high-throughput experiments, such as yeast (Uetz *et al*., 2000) and fly (Giot *et al*., 2003) proteomes, have enabled us to elucidate the interaction networks on a large scale. These large-scale experiment results are collected and well-curated into interaction databases such as the Database of Interacting Proteins (DIP) (Salwinski *et al*., 2000), Biomolecular Interaction Network Database (BIND) (Bader *et al*., 2003) and Molecular INTeraction database (MINT) (Zanzoni *et al*., 2002). There have also been computational approaches to map and predict the protein interactome in a genomic context using gene fusion and gene neighborhood methods (Huynen *et al*., 2000).

In parallel with the above methods, PSIMAP (Protein Structural Interactome map) has introduced a new mapping protocol in protein structural interactome study. An underlying concept of PSIMAP is homologous interaction: the interaction among protein structures is conserved as closely as the protein structures themselves (Park *et al*., 2001; Aloy and Russell, 2002; Aloy *et al*., 2003). With PSIMAP, we can view protein interactions in terms of family–family interactions, as well as individual protein–protein interactions. PSIMAP covers interaction information from both gene fusion style protein sequence level interaction and physical interaction within complexes or multi-domain proteins.

Here, we introduce PSIbase: the PSIMAP web server and database. It contains (1) domain–domain and protein–protein interaction information from proteins whose 3D-structures are identified, (2) a protein interaction map and its viewer at protein superfamily and family levels, (3) protein interaction interface viewers and (4) structural domain prediction tools for possible interactions by detecting homologous matches in the Protein Data Bank (PDB) from query sequences. Structural interaction data, in flat file format, can be downloaded from PSIbase (http://psibase.kaist.ac.kr/Download/download.shtml) for further analyses. It contains the smallest distance between two domains and the number of residue pairs that is within the threshold distance according to the PSIMAP algorithm. It not only provides raw data files, but it also serves biologists who need to look up the interaction partners of their proteins of interest. Simply putting a protein sequence is enough to search for possible interaction partners (interlogs). As the possible predicted domains of query sequence are based on a structural assignment protocol, users can see the interlogs' 3D structures if they accept the prediction made by PSIbase. For structural domain assignment, we used two databases and two algorithms. They were the SCOP (http://scop.kaist.ac.kr/scop, Murzin *et al*., 1995) database with an intermediate sequence library ISL, (Teichmann *et al*., 2000), and PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) with a hidden Markov model package (HMMER, http://hmmer.wustl.edu/). We believe that PSIbase is useful for those in the fields of structural bioinformatics and molecular biology.

*To whom correspondence should be addressed.

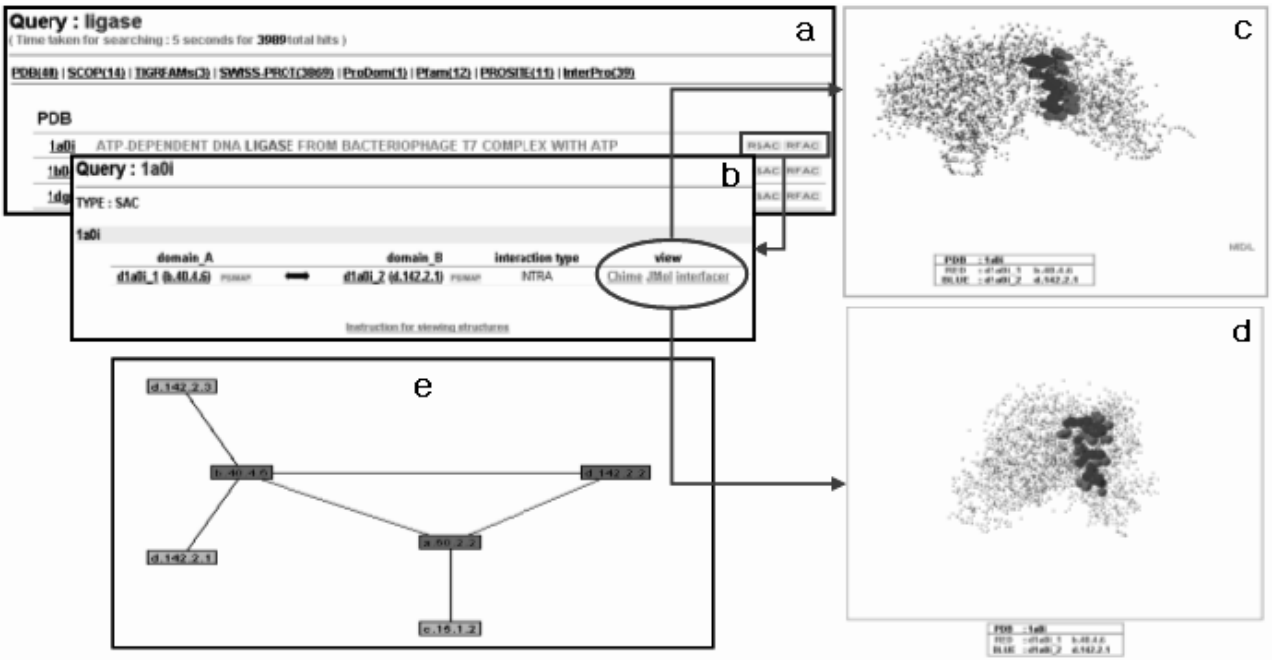

Fig. 1. The visualization of interaction information. Search results for 'ligase' as a query. 'RSAC' button in box *a* leads to a page *b* that shows interactions among domains which are defined by SCOP. *c* and *d* boxes show the interacting interfaces with different viewers. The interaction network is shown in *e*.

PSIMAP ALGORITHM

The basic mechanism to check interactions between any two domains or proteins is the calculation of the Euclidean distance in order to see if they are within a certain distance threshold. PSIMAP checks every possible pair of structural domains in a protein to see if there are at least five residue contacts within a 5 Å distance (5–5 rule). The current PSIMAP protocol has three methods. They are the Full Atom Contact (FAC) PSIMAP, Sampled Atom Contact (SAC) PSIMAP and Bounding Box Contact (BBC) PSIMAP (Dafas *et al*., 2004). (The supplementary material provides in-depth information about the three different PSIMAP algorithms.)

The FAC calculates all the atomic contacts among two or more protein structural domains. FAC PSIMAP is the most accurate of the three, as we take into account all the atoms in domain pairs.

The SAC and BBC algorithms are approximations of FAC. Their main purpose is to reduce the time taken in constructing PSIMAP. The BBC algorithm is a radically different approach, using a bounding box algorithm to dramatically reduce the time of computation. Dafas *et al*. introduced a bounding box and convex hull algorithm that can reduce the search space.

DATABASE ACCESS

The PSIbase server is available at http://psibase.kaist.ac.kr/. There are three different query interfaces to access the PSIbase. All queries are funneled into a web page that shows protein domain interactions with their partners.

First, PSIbase provides a simple search interface that looks up keywords or database accession IDs. Figure 1 shows the search result of 'ligase' as a query against 12 annotated DB resources (listed on the PSIbase webpage). Out of the 12, multiple matches for the query 'ligase' are listed up from the following databases: PDB, SCOP, TIGRFAMs, Swiss-Prot, ProDom, Pfam, Prosite and Interpro. There are three tools to view interaction interface structures: Chime (http://www.mdli.com), Jmol (http://jmol.sourceforge.net) and Interfacer (http://www.interfacer.org). Interfacer is a slow but advanced protein interface viewer with surface representation capability.

The second PSIbase query interface is a protein structural domain assignment utility that accepts protein sequences from users. There are two domain assignment algorithms available in PSIbase. One is a homology-based sequence search by PSI-BLAST utilizing the ISL (see Introduction) and the other is the HMMER profile search algorithm. These two are complementary in terms of the coverage in the assignment.

The last PSIbase query interface accepts specific domain IDs at SCOP family or superfamily levels. There are several levels to determine interactions among query domains. For example, interacting partners of a specific query domain can be identified within a specified interaction depth (the maximum depth limit is 4). Interactions between two or more input query domains can also be identified. Additionally, PSIbase is equipped with a simple open-source Java applet program that shows the interaction network of each query.

CONCLUDING REMARKS

There are 1294 superfamilies and 2327 families in SCOP 1.65. On average, PSIbase covers 87% (1136/1294) of SCOP superfamily interactions, indicating that the majority of SCOP superfamilies have interacting partner information. In the supplementary material, Table 2 shows the 20 most interactive superfamilies in PSIbase. These can be regarded as the most central interaction components in interactomes, so we call them the 'interactome core'. This core contains proteins with energy metabolism, RNA and DNA binding, and other key biological processes that have existed since the very early

days of interaction networks (Bolser *et al*., 2003). The interactions of non-protein molecules in cells are critical in biological functions. In the next version, PSIbase and PSIMAP will cover interactions between proteins and non-proteins such as nucleic acids and small molecules.

ACKNOWLEDGEMENTS

We thank Mr. Chung MoonSoul for donating $25 million to the Department of Biosystems at KAIST. This project was funded by IMT-2000 C3-4 grants from the Ministry of Information and Communication of Korea and a grant from KRIBB Research Initiative Program. J.B. is supported by Biogreen21. We thank Maryana Bhak for editing and commenting on this manuscript. We also send our loving gratitude to the anonymous reviewers for their precious comments.

REFERENCES

Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.

Aloy,P. *et al*. (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, **332**, 989–998.

Bader,G.D. *et al*. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 48–50.

Bolser,D.M. *et al*. (2003) Visualisation and graph-theoretic analysis of a large-scale protein structural interactome. *BMC Bioinform.*, **4**, 1471–2105.

Dafas,P. *et al*. (2004) Using convex hulls to compute protein interactions from known structures. *Bioinformatics*, **20**, 1–5.

Giot,L. *et al*. (2003) A Protein Interaction Map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Huynen,M. *et al*. (2000) Exploitation of gene context. *Curr. Opin. Struct. Biol.*, **10**, 366–370.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J.H. *et al*. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.*, **307**, 929–938.

Salwinski,L. *et al*. (2000) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Teichmann,S.A. *et al*. (2000) Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics*, **16**, 117–124.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Zanzoni,A. *et al*. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.