doi:10.1093/nar/gkp944

GWIDD: Genome-wide protein docking database

Petras J. Kundrotas[1], Zhengwei Zhu[1] and Ilya A. Vakser[1,2,*]

[1]Center for Bioinformatics and [2]Department of Molecular Biosciences, 2030 Becker Drive, The University of Kansas, Lawrence, KS 66047, USA

Received August 10, 2009; Revised September 27, 2009; Accepted October 12, 2009

ABSTRACT

Structural information on interacting proteins is important for understanding life processes at the molecular level. Genome-wide docking database is an integrated resource for structural studies of protein–protein interactions on the genome scale, which combines the available experimental data with models obtained by docking techniques. Current database version (August 2009) contains 25 559 experimental and modeled 3D structures for 771 organisms spanned over the entire universe of life from viruses to humans. Data are organized in a relational database with user-friendly search interface allowing exploration of the database content by a number of parameters. Search results can be interactively previewed and downloaded as PDB-formatted files, along with the information relevant to the specified interactions. The resource is freely available at http://gwidd.bioinformatics.ku.edu.

INTRODUCTION

Function of proteins in the living cell is determined by their ability to interact with other biologically relevant molecules (other proteins, DNA, RNA, small ligand, etc.). Thus understanding mechanisms of these interactions is critically important for studying life processes at the molecular level. Genome sequencing provided vast amount of information on proteins, spanning the entire universe of life from viruses to the highest eukaryotic organisms. In the post-genomic era, the efforts focus on the function assignment of the sequenced proteins based on their three-dimensional (3D) structures and/or participation in interactions. Because of the limitations of the experimental techniques for structural characterization, computational methods play a vital role (1).

Success in recreating maps of interactions for specific organisms and/or specific biochemical pathways emphasize the need for large-scale modeling efforts to deliver 3D structures of the protein complexes. Computational methods for structural modeling of the protein–protein interactions (PPIs) historically started with *ab initio* (or template free) methods based on shape complementarity and were later supplemented by the constraints derived from statistical analysis of properties of known protein complexes or from the experimentally acquired additional biochemical/biophysical knowledge (2). Most of the existing docking servers (3) employ constrained-based template-free approach. Despite the significant progress in development of the template-free algorithms, their accuracy in the high-throughput applications is limited.

Accumulation of experimental data in the last decade have caused paradigm shift in 3D modeling of individual proteins from *ab initio* to template-based techniques. A similar trend is underway in structural modeling of protein complexes (protein docking). Recently, several groups assessed quality of the models produced by the homology/threading docking techniques where a protein complex is modeled based on similarity to another protein complex with the known structure (4–8). It was demonstrated that the majority of the homology-docking models are of acceptable and medium quality, according to the CAPRI criteria (3). It was estimated that the homology docking can account for a significant part (15–20%) of known PPI (7). Structural alignment techniques were also benchmarked on various sets of protein complexes (9,10).

Success in developing the high-throughput modeling techniques makes it feasible to create a long-needed comprehensive resource, which would reflect large-scale efforts in structural modeling of known protein complexes. Genome-wide docking database (GWIDD) provides annotated collection of experimental and modeled 3D structures of protein–protein complexes from the entire universe of life spanning from viruses to humans. The database provides user-friendly interface for searching and browsing database content and downloading experimental and modeled structures of protein complexes.

DATABASE CONTENT AND DESCRIPTION

Source of PP1s data

PPIs are imported to GWIDD from external sources specialized in collecting and curating PPI. Currently they include BIND (http://www.bind.ca) (11) and DIP

[*] To whom correspondence should be addressed. Tel: +1 785 584 1057; Fax: +1 785 864 5558; Email: vakser@ku.edu

Table 1. Distribution of GWIDD entries for various categories of living organisms[a]

Living organisms	Number of species[b]	Number of interactions[c]	Number of model structures[d]	Number of experimental structures
Archaea	41	1128	369	723
Bacteria	288	13 871	3183	5488
Lower eukaryota[e]	80	29 289	2058	811
Plants	79	2055	365	399
Animals	136	72 395	7858	2746
Viruses	147	2080	802	757
Total	771	120 818	14 635	10 924

[a] The data is for protein–protein interactions where both partners are from the same organisms, except for the viruses where interactions are between a protein from the virus and a protein from the host organism.
[b] Number of species for which at least one protein–protein interaction is present in DIP and BIND databases.
[c] As in DIP and BIND, including interactions with no modeled structures.
[d] Modeled by homology docking.
[e] Includes primitive organisms and fungi.

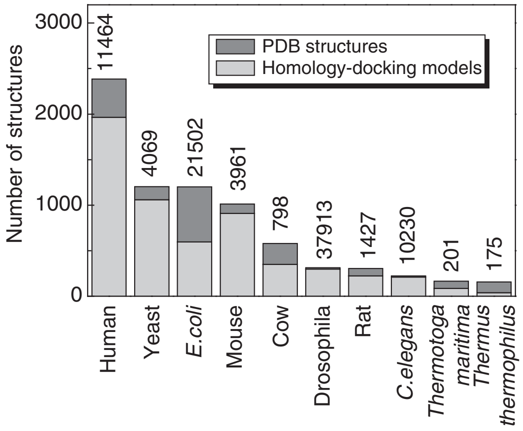


Figure 1. Number of experimental structures (dark gray bars) and structures modeled by homology docking (light gray bars) for 10 organisms with the largest structural coverage in GWIDD. Numbers at the bars indicate the total amount of non-identical interactions, including those with no structure, in DIP and BIND databases.

(http://dip.doe-mbi.ucla.edu) (12,13) databases. These databases were chosen because their content is not restricted to a single genome or group of genomes like in many other PPI databases [e.g. different flavors of MINT (14) or MIPS (15)]. They are also regularly updated providing up-to-date pool of the initial data. The interactions are obtained through either high-throughput discovery methods or small-scale experiments and thus are of diverse reliability. However, at the current stage, evaluation of credibility of the PPI data is outside the scope of GWIDD.

Current content of the database

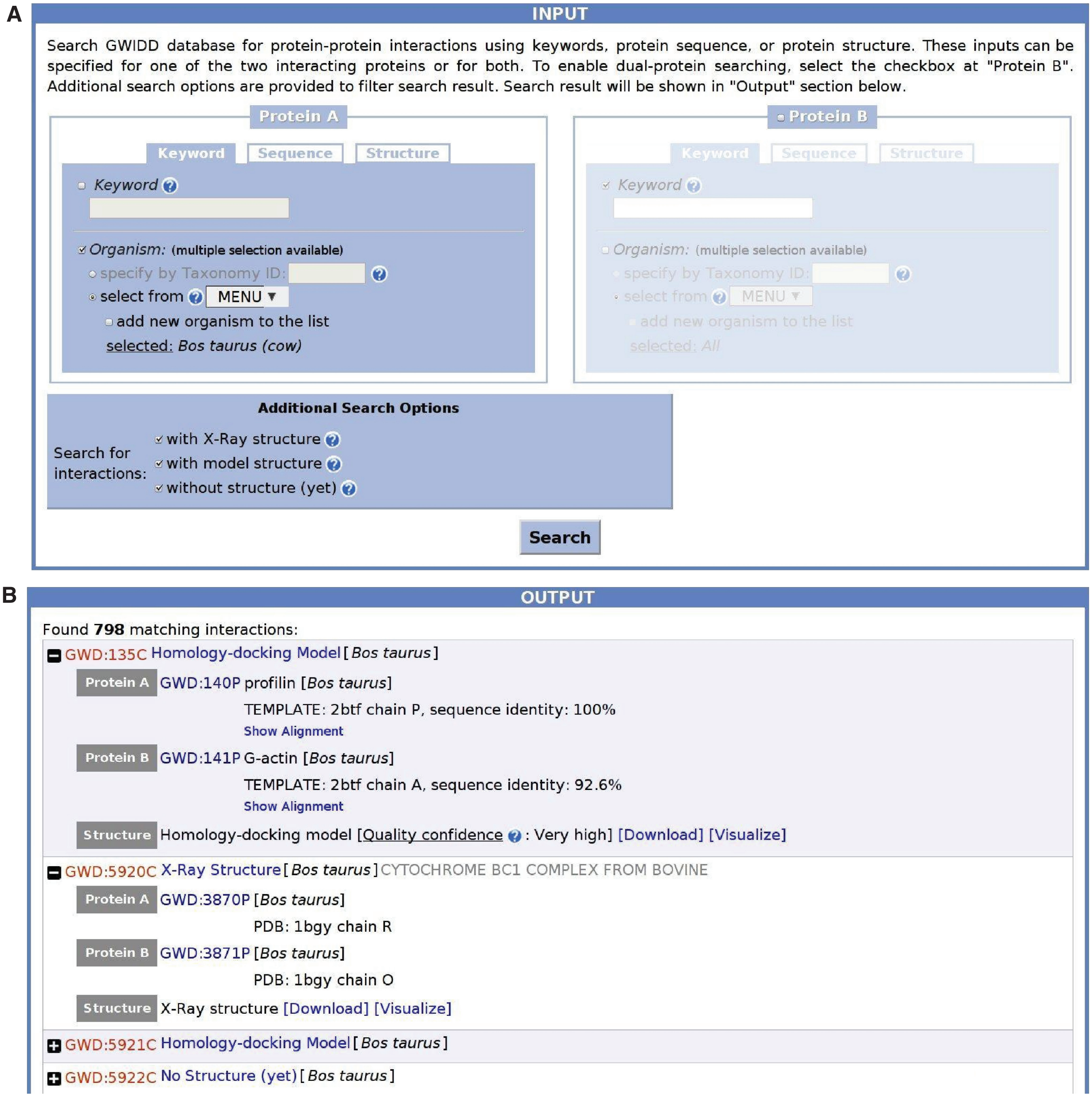
The ultimate goal of the GWIDD resource is to provide the 3D structures for all known PPI. At the current stage, the following steps are taken toward this goal. First, if an interaction is found in protein data bank (PDB), this structure is used and no modeling is performed. Otherwise, a search for a pair of homologous sequences from a complex with known structure is performed and the model is build by homology docking (6,7). We used earlier-described criteria for statistical significance of the sequence alignments (7) with an additional requirement that both alignments contain at least 80% of the target sequences. In the future, for the interactions not covered by these two steps, we will use other docking methods (e.g. structural alignment, template free docking), which will be incorporated in the upcoming GWIDD releases. However, even the current limited modeling approach provides structures for 14 635 PPI, which together with the available non-redundant X-ray structures (10 924) constitutes >20% of the currently known PPI. Summary of the GWIDD content is provided in Table 1. As of August 2009, GWIDD contained 126 897 binary interactions, involving 43 976 proteins from 771 different organisms spanning the entire universe of life (Table 1). Among those, 6079 entries are either cross-organism interactions or do not have organism annotation in the source data. Thus they are not present in Table 1, although will appear in search results. The distribution of available structures (X-ray and modeled) is shown in Figure 1 for 10 organisms with the largest numbers of structured GWIDD entries. The database is automatically updated every half year.

Implementation of the database and its user interface

The data from the external source databases have different formats and different levels of details. Thus such data are unified into a single dataset of PPI, removing redundancy and retaining common data fields for all the sources. Due to the large amount of data and complex data dependency as well as complex query requirement, all interaction data are stored in a relational database, except for large files, such as PDB ones, which are stored directly in the file system and are linked from the relational database. Implementation of the web interface is based on LAPP (Linux-Apache-PostgreSQL-PHP) software stack. Web user interface is built using PHP and jQuery library, where PHP is for web presentation and logic as well as back-end database access. jQuery is responsible for AJAX and other JavaScript-based dynamic features. Visualization of protein structures is implemented utilizing Jmol (http://www.jmol.org) technology. Homology docking was performed by NEST (16), BLAST (17) and in-house profile-to-profile alignment program used previously for the benchmarking of homology docking (7). The above parts are joined by a set of Python scripts.

User interface description

The database can be freely accessed at http://gwidd.bioinformatics.ku.edu. The default option offered to users is search of the database by keywords related to a single interaction partner ('Protein A' part in Figure 2A). Other search options are available by clicking tabs 'Sequence' (explicit input or upload of sequence in the FASTA format) or 'Structure' (upload of a



A

INPUT

Search GWIDD database for protein-protein interactions using keywords, protein sequence, or protein structure. These inputs can be specified for one of the two interacting proteins or for both. To enable dual-protein searching, select the checkbox at "Protein B". Additional search options are provided to filter search result. Search result will be shown in "Output" section below.

Protein A

Keyword | Sequence | Structure

□ Keyword

☑ Organism: (multiple selection available)

○ specify by Taxonomy ID:

● select from MENU ▾

□ add new organism to the list

selected: *Bos taurus (cow)*

□ Protein B

Keyword | Sequence | Structure

☑ Keyword

□ Organism: (multiple selection available)

○ specify by Taxonomy ID:

● select from MENU ▾

□ add new organism to the list

selected: *All*

Additional Search Options

Search for interactions:

☑ with X-Ray structure

☑ with model structure

☑ without structure (yet)

Search

B

OUTPUT

Found **798** matching interactions:

⊟ GWD:135C Homology-docking Model [*Bos taurus*]

Protein A GWD:140P profilin [*Bos taurus*]

TEMPLATE: 2btf chain P, sequence identity: 100%

Show Alignment

Protein B GWD:141P G-actin [*Bos taurus*]

TEMPLATE: 2btf chain A, sequence identity: 92.6%

Show Alignment

Structure Homology-docking model [Quality confidence : Very high] [Download] [Visualize]

⊟ GWD:5920C X-Ray Structure [*Bos taurus*] CYTOCHROME BC1 COMPLEX FROM BOVINE

Protein A GWD:3870P [*Bos taurus*]

PDB: 1bgy chain R

Protein B GWD:3871P [*Bos taurus*]

PDB: 1bgy chain O

Structure X-Ray structure [Download] [Visualize]

⊞ GWD:5921C Homology-docking Model [*Bos taurus*]

⊞ GWD:5922C No Structure (yet) [*Bos taurus*]

Figure 2. Example of a search by organism (**A**) and the results of this search (**B**).

PDB-format file). When searching by keywords, user can either enter any keyword in the protein description (name of organism, cellular location, biological function, etc.) or choose from the series of drop-down menus containing lists of all organisms currently in GWIDD. By repeating the selection with the box 'Add another organism to the list' checked, user can choose several organisms. When the box is unchecked, the search will clear the list of previously selected organisms. Also, in each submenu, user can select all listed organisms by a single click on the top 'Select All' position. An option to search by standard taxonomy ID with link to taxonomy database http://www.uniprot.org is also provided for convenience. Search results for the 'Keyword' tab can be, for example, all PPI related to a certain pathway (defined by the keyword) or all interactions within certain organism or group of organisms. Search results for the 'Sequence' and 'Structure' tabs contain all interactions with the input sequence as one of the interaction partners (in the case of input PDB file the sequence extracted from the SEQRES tags or, if the SEQRES part is not available, from ATOM tags for the C_α atoms). The amino acid sequences from different sources can differ in length even for the same protein (e.g. due to unresolved residues in the

X-ray structure). Thus advanced options are provided in the sequence search parts. An example of search by organism and its results is shown in Figure 2.

If information related to the other interaction partner is also known, user can enable the second part of the search interface ('Protein B,' see Figure 2A) by checking the corresponding box and input the information similarly to 'Protein A'. In addition, search results can be filtered by the availability of different types of GWIDD entries (experimental structures, modeled structures or interactions with no structures). Online help is provided in pop-up windows (question marks inside blue circles, see Figure 2). Search results screen (Figure 2B) displays all interactions in the database satisfying the input search criteria in the form of expandable list of GWIDD interaction IDs with minimum additional information. The expanded item in the list contains the name and the GWIDD IDs of the interacting partners along with information on the type of 3D structure available for this interaction (if applicable). For the homology-docking models, the alignments used to build the model are provided and the model quality is assessed by the sequence identity criteria (5). For the available structures, links are provided to download the PDB-format file along with the text file containing relevant information, as well as to the visualization screen where the structure is displayed in colored-by-chain space-filled interactive representation.

COMPARISON TO OTHER EXISTING RESOURCES

There are several resources available that are similar in spirit (genome-wide approach to PPI) to the GWIDD resource. Michigan molecular interactions (MiMIs), database (http://mimi.ncibi.org) (18) provides one cohesive view of molecules found in several popular interaction databases, including BIND, HPRD, IntAct, GRID and others, with complementary or conflicting data among the sites highlighted. POINT (http://point.bioinformatics.tw) (19) is a functional database for prediction of the human protein–protein interactome based on available orthologous interactome datasets with the emphasis on extraction of mouse, fruit fly, worm and yeast PPI datasets from DIP, followed by their conversion to predicted human interactome. 3D-GENOMICS (http://www.sbg.bio.ic.ac.uk/3dgenomics) (20) provides structural annotations for proteins from sequenced genomes and in August 2003 included data for 93 proteomes. NCBI Inferred Biomolecular Interactions Server (IBIS, http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi) reports physical interactions observed in experimentally determined structures for sequences homologous to the input amino acid sequence, thus inferring interacting partners and binding sites. However, none of the above resources provide single integrated and searchable pool of experimental and modeled 3D structures for all genomes for which at least one PPI is annotated. Recently developed ProtInfo PPC server (http://protinfo.compbio.washington.edu/ppc) (21) provides model structures for user's supplied sequences, but lacks the annotated database of 3D structures.

FUTURE DEVELOPMENT

The major direction in the future development of GWIDD is expanding the pool of available structures modeled by other modeling techniques, such as docking by structural alignment (to be submitted) and template-free docking by GRAMM methodology (22–24). To assess the applicability of these methods to the high-throughput, genome-wide modeling, large-scale benchmarking is currently underway. New sources of PPI will be incorporated as they become available.

ACKNOWLEDGEMENTS

Andrey Tovchigrechko and Tatiana Baronova made important contributions to GWIDD project at the earlier stages of development.

FUNDING

Funding for open access charge: National Institutes of Health (R01 GM074255).

Conflict of interest statement. None declared.

REFERENCES

1. Russell,R.B., Alber,F., Aloy,P., Davis,F.P., Korkin,D., Pichaud,M., Topf,M. and Sali,A. (2004) A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.*, **14**, 313–324.
2. Vakser,I.A. and Kundrotas,P. (2008) Predicting 3D structures of protein-protein complexes. *Curr. Pharm. Biotech.*, **9**, 57–66.
3. Lensink,M.F., Mendez,R. and Wodak,S.J. (2007) Docking and scoring protein complexes: CAPRI 3rd edn. *Proteins*, **69**, 704–718.
4. Aloy,P., Pichaud,M. and Russell,R.B. (2005) Protein complexes: Structure prediction challenges for the 21st century *Curr. Opin. Struct. Biol.*, **15**, 15–22.
5. Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
6. Kundrotas,P.J. and Alexov,E. (2006) Predicting 3D structures of transient protein-protein complexes by homology. *Bioch. Biophys. Acta.*, **1764**, 1498–1511.
7. Kundrotas,P.J., Lensink,M.F. and Alexov,E. (2008) Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. *Int. J. Biol. Macromol.*, **43**, 198–208.
8. Lu,L., Lu,H. and Skolnick,J. (2002) MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.
9. Gunther,S., May,P., Hoppe,A., Frommel,C. and Preissner,R. (2007) Docking without docking: ISEARCH - prediction of interactions using known interfaces. *Proteins*, **69**, 839–844.
10. Launay,G. and Simonson,T. (2008) Homology modelling of protein-protein complexes: A simple method and its possibilities and limitations. *BMC Bioinformatics*, **9**, 427.
11. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K. and Burgess,E.e.a. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
12. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

13. Xenarios,I., Rice,D.W., Salwinski,L., Baron,N.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: The Database of Interacting Proteins. *Nucleic Acids Res.*, **28**, 289–291.
14. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: A Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
15. Pagel,P., Kovac,S., Oesterheld,M., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Mark,P., Stümpflen,V., Mewes,H.W. *et al.* (2005) The MIPS mammalian protein–protein interaction database. *Bioinformatics*, **21**, 832–834.
16. Petrey,D., Xiang,Z.X., Tang,C.L., Xie,L., Gimpelev,M., Mitros,T., Soto,C.S., Goldsmith-Fischman,S., Kernytsky,A., Schlessinger,A. *et al.* (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53**, 430–435.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of database programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Tarcea,V.G., Weymouth,T., Ade,A., Bookvich,A., Gao,J., Mahavisno,V., Wright,Z., Chapman,A., Jayapandian,M., Ozgur,A. *et al.* (2009) Michigan molecular interactions r2: From interacting proteins to pathways. *Nucleic Acids Res.*, **37**, D642–D646.
19. Huang,T.W., Tien,A.C., Huang,W.S., Lee,Y.C.G., Peng,C.L., Huei-Hun Tseng,H.H., Kao,C.Y. and Huang,C.Y.F. (2004) POINT: A database for the prediction of protein–protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273–3276.
20. Fleming,K., Muller,A., MacCallum,R.M. and Sternberg,M.J.E. (2004) 3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. *Nucleic Acids Res.*, **32**, D245–D250.
21. Kittichotirat,W., Guerquin,M., Bumgarner,R.E. and Samudrala,R. (2009) Protinfo PPC: A web server for atomic level prediction of protein complexes. *Nucleic Acids Res.*, **37**, W519–W525.
22. Katchalski-Katzir,E., Shariv,I., Eisenstein,M., Friesem,A.A., Aflalo,C. and Vakser,I.A. (1992) Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.
23. Vakser,I.A., Matar,O.G. and Lam,C.F. (1999) A systematic study of low-resolution recognition in protein-protein complexes. *Proc. Natl Acad. Sci. USA*, **96**, 8477–8482.
24. Tovchigrechko,A., Wells,C.A. and Vakser,I.A. (2002) Docking of protein models. *Protein Sci.*, **11**, 1888–1896.