Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W331–W336***
doi:10.1093/nar/gki585

PRISM: protein interactions by structural matching

Utkan Ogmen, Ozlem Keskin*, A. Selim Aytuna, Ruth Nussinov[1,2] and Attila Gursoy

Koc University, Center for Computational Biology and Bioinformatics and College of Engineering, Rumelifeneri Yolu, Sariyer, Istanbul 34450, Turkey, [1]Basic Research Program, Science Applications International Corporation (SAIC)–Frederick, Inc., LECB, NCI–Frederick, Frederick, MD 21702, USA and [2]Department of Human Genetics and Molecular Medicine, Sackler Institute of Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Received February 16, 2005; Revised April 18, 2005; Accepted May 2, 2005

ABSTRACT

Prism (http://gordion.hpc.eng.ku.edu.tr/prism) is a website for protein interface analysis and prediction of putative protein–protein interactions. It is composed of a database holding protein interface structures derived from the Protein Data Bank (PDB). The server also includes summary information about related proteins and an interactive protein interface viewer. A list of putative protein–protein interactions obtained by running our prediction algorithm can also be accessed. These results are applied to a set of protein structures obtained from the PDB at the time of algorithm execution (January 2004). Users can browse through the non-redundant dataset of representative interfaces on which the prediction algorithm depends, retrieve the list of similar structures to these interfaces or see the results of interaction predictions for a particular protein. Another service provided is interactive prediction. This is done by running the algorithm for user input structures.

INTRODUCTION

Most molecular and cellular operations are largely sustained by interactions between proteins. Identifying interaction sites of proteins and knowing which proteins interact with which others are crucial for a better understanding of the bases of many biological processes. Despite the ongoing effort to decipher the complex nature of protein interactions, they are still not entirely understood (1–4). Protein binding sites have been thoroughly analyzed for the presence of certain physicochemical and geometric properties that can be used to distinguish these regions from the non-interacting surface regions. Notable differences have been found in both the chemical composition and the geometric properties of these sites (5–9). Prediction of binding sites using these specific properties can be used to improve docking algorithms. Alongside experimental methods for detecting and analyzing protein–protein interactions (9–11), computational approaches are becoming increasingly important as large amounts of data become available. Development of predictive methods is a major goal in computational biology that will lead to protein engineering and drug discovery (7,8,12). Hence, an efficient computational technique with acceptable error rates that can be utilized to predict the binding sites and binding partners in proteins will surely be of great use.

Here, we present the results of our novel, high-performance and efficient algorithm to predict protein–protein interactions (13). We have implemented PRISM (Protein Interactions by Structural Matching), a web server that can be used to explore protein interfaces and predict protein–protein interactions. Our algorithm principally seeks pairs of proteins that may interact in a dataset of protein structures (target dataset) by comparing them with a dataset of interfaces (template dataset) which is a structurally and evolutionarily representative subset of biological and crystal interactions present in the Protein Data Bank (PDB) (14). PRISM consists of a web interface to the dataset of interfaces and target structures including a summary of the proteins the interface belongs to (with cross-references to other biological databases where available), similarity matching results, solvent accessible surface area calculation results on a residue-level scale, interface visualization of the protein using both static images and an interactive interface viewer implemented using a browser plug-in.

METHODOLOGY AND RESULTS

The rationale of our protein–protein interaction prediction algorithm is that, if any two structures contain particular regions on their surfaces that resemble the complementary partners of a known interface, they 'possibly interact' through these regions. In other words, if *A* is known to interact with *B*,

*To whom correspondence should be addressed. Tel: +90 212 338 1538; Fax: +90 212 338 1548; Email: okeskin@ku.edu.tr
Correspondence may also be addressed to Attila Gursoy. Tel: +90 212 338 1720; Fax: +90 212 338 1548; Email: agursoy@ku.edu.tr

a shares similarity with the binding site of *A* and *b* shares similarity with the binding site of *B*, then we predict that *a* interacts with *b*. This resemblance indicates the ability of these structures structurally and evolutionarily to complement each other along an interface, as chains of any template interface might do. The algorithm requires a 'template' dataset, i.e. the representative dataset of 'available' interfaces, and a 'target' dataset to seek every potential binary interaction between its members (13).

Interface dataset

The interface dataset is a non-redundant dataset of protein–protein interfaces. Interfaces were defined as the set of residues representing a region through which two polypeptide chains bind to each other through non-covalent interactions. This set consisted of contacting residues between the chains (interacting residues) and those that are in their vicinity within a certain distance threshold (neighboring residues), representing the scaffold of the interface. Two residues from the opposite chains were marked as interacting if there was at least a pair of atoms, one from each residue, at a distance smaller than the sum of their van der Waals radii plus a threshold of 0.5 Å. If the C-α of a non-interacting residue lay at a distance of ⩽6.0 Å from a C-α of an already assigned interface residue in the same chain, it was flagged as a neighboring residue. All interfaces between two protein chains obtained from higher complexes of proteins available in the PDB were extracted (15). As a result, 21 684 two-chain interfaces were obtained. These interfaces were compared structurally using a sequence order-independent computer vision-based algorithm (16). Interfaces sharing similar architectures were grouped into clusters. At the end of the iterative structural clustering procedure, we obtained 3799 interface clusters. Each cluster includes a representative interface structure and members similar to the representative interface. The list of all clusters is available in our web server as the 'Interface dataset'. Figure 1 is a screenshot of the sample search form where users can enter their queries.

Template interface dataset

The evolutionary conservation of certain residues at protein interfaces is another characteristic of binding sites. For this purpose, we used a dataset of computational hotspots, consisting of the critical residues for binding on representative interfaces. The members of the 3799 interface clusters were put through a filtering process which eliminated the redundant sequences from the clusters. A cluster was defined as non-redundant if it contained at least five non-homologous sequences. Then, simultaneous structural alignments among the non-homologous members of each cluster were performed (17). If a residue was conserved at a particular spot among interfaces of similar architectures with ⩾50% frequency, it was flagged as a computational hotspot (18). As a result, we could detect the hotspots of 67 clusters out of 3799 since most of the clusters did not pass the non-homologous filtering. (A flowchart of the template preparation procedure is given in Supplementary Material Figure S1A.) The prediction algorithm serviced by PRISM uses only these 67 template interfaces for similarity matching. Hence, PRISM considers both shape complementarities and evolutionary conservation while searching for binding sites on the surface of a target protein.

Target dataset

The target dataset is a sequentially non-redundant subset (with a sequence identity upper limit of 50%) of all the polypeptide

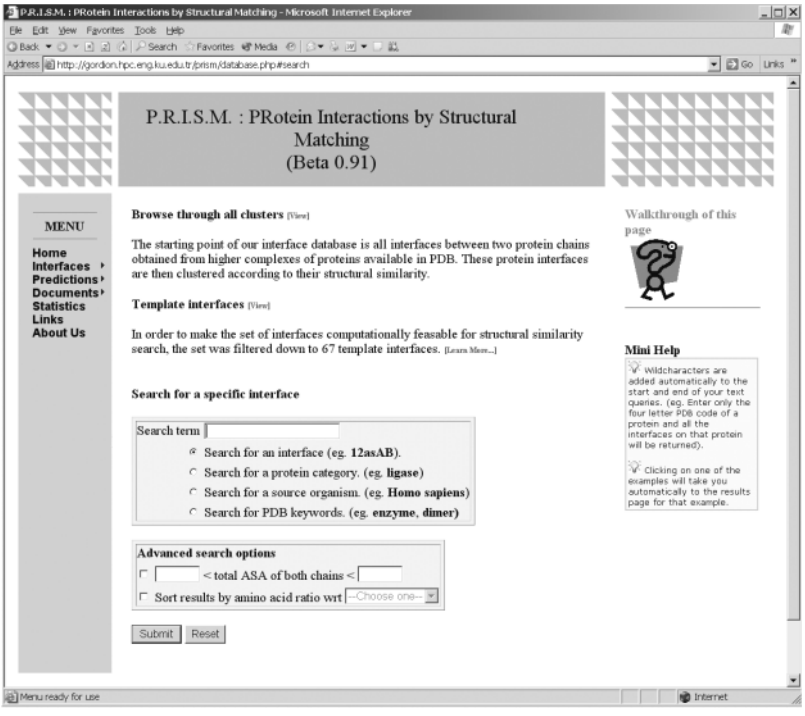


Figure 1. Sample search form where users can enter their queries.

chains and complexes existing in the PDB. Every pair of member structures in this dataset is checked for potential interactions. The protein chains may be in the form of monomers or in the form of isolated chains from multimeric complexes. As of January 27, 2004, the target dataset contained 6170 structures (13). The generation of this dataset is a two-step process. The first is a preprocessing step that involved downloading the set of proteins obtained by applying a sequence identity filter of 50% to all existing protein structures in the PDB. This resulted in a list containing 5427 proteins. Then, the multimeric proteins were split into constituent chains, where homologous chains are counted only once. The target dataset consists of 6170 structures, of which 1981 are multimeric and 4189 are monomeric. Of the monomeric structures, 2483 are derived from complexes. All these structures are on our web server as the 'Target structure dataset'. (A flowchart of the target preparation procedure is given in Supplementary Material Figure S1B.)

Prediction of protein–protein interactions

To find every possible binary interaction between pairs of structures in the target dataset, we need a method to measure the similarity between partners of these representative interfaces and surfaces of target proteins. To do this, we extract surfaces of target proteins and perform successive structural alignments between these surfaces and the partner chains of interfaces in the template interface dataset, in an all-against-all manner. This enables us to measure the 'structural similarity' of a target structure to a template binding site. If the surfaces of two target proteins (*A* and *B*) contain regions 'similar' to complementary partner chains of a template interface, we say that *A* and *B* may interact through these 'similar' regions. Further, we check for the presence of hotspots on the target structure. The hotspot match ratio is used for the calculation of an 'evolutionary similarity score', whereas the structural match ratio is used for a 'structural similarity score'. Combination of these scores contributes to the overall prediction score. A simplified flowchart of the algorithm is given in Figure 2. We have run our algorithm using the template interface set and target structure set, which resulted in a total of 62 616 protein–protein interactions. These can be accessed from our website under 'Predictions'.

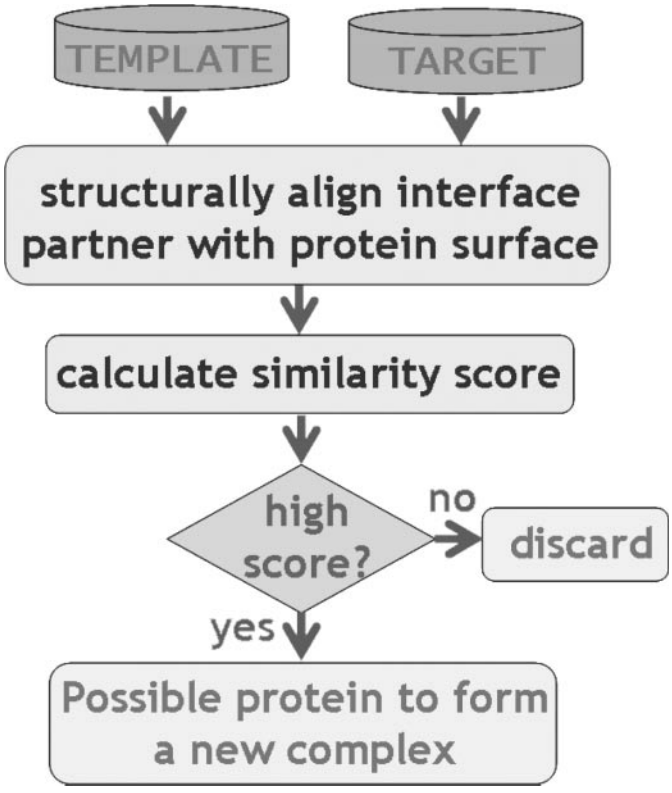


Figure 2. Flowchart summary of the prediction algorithm. The surfaces of the target proteins are compared with the template interface dataset.

Services provided by PRISM

The PRISM web server provides its users with a front end to the datasets used in our prediction algorithm, an interface to the offline results of our calculations based on the most recent run of our algorithm and the ability to run our algorithm for a user input protein. Services provided to the user and the input types differ accordingly, so they are discussed separately.

In the interfaces section we make our interface dataset available to the scientific community. A total of 21 684 interfaces are stored, clustered into 3799 clusters according to their structural similarity. Users are provided with a search facility using which they can find specific interfaces in the interfaces dataset that match a set of search criteria. Their inputs can be a simple search string, which is searched for in the corresponding records in the title section of the PDB file of the protein which the template interface belongs to. For example, a user might be interested in interfaces that are extracted from proteins that play a role in apoptosis or might want to see only interfaces that are extracted from enzymes. In addition to this basic search functionality, some advanced search options can be used, enabling the user to search for interfaces of a certain size—in terms of solvent accessible surface areas (ASAs) measured in $Å^2$—or interfaces that have the highest frequency for a certain type of amino acid. Once the user clicks on an interface, an output containing the following data is provided.

(i) A summary of the proteins the interface is extracted from, including cross-references to other biological databases where available.
(ii) Details about the interface in question such as the names of the constituent chains, interface size (in terms of number of residues), solvent ASAs buried upon complexation, polar and non-polar ASA and a listing of all interface residues with their respective interface ASA. (Supplementary Material Figure S2 shows a snapshot of the web server's results for the summary of the proteins, i.e. name of the protein, number of atoms of the protein, ASA of the interfaces, etc.)
(iii) A visualization of the interface as static images where the interface is highlighted on the protein. These are dynamically generated by running RasMol scripts. The whole protein is represented using stick representation, whereas the interface atoms are shown using spheres. Constituent binding sites of the interface are distinguished using a coloring scheme. (Supplementary Material Figure S3 shows an example of the static images.)

Another service provided by PRISM is an interactive visualization tool. The interface viewer is implemented using the MDL® Chime software from Elsevier MDL. The viewer window is divided into two frames. In the left frame is a 3D model of the protein where the interface in question is shown as part of the protein. The right frame contains control buttons used for manipulating the 3D model. These buttons can be used to rotate and zoom in/out of the model, show/hide constituent binding sites and change certain aspects of the display representation. (Supplementary Material Figure S4 shows an example.)

In the targets section (under 'Prediction'), users are provided with a search facility to find specific structures in the target dataset that match a set of search criteria. The input can be a simple search string which is searched for in the corresponding records in the title section of the PDB file of the protein. In addition, using advanced search options, specific sets of target structures can be returned, for instance, target structures of a predefined size (size defined as number of residues) or type (monomer, complex, split chain). Once the user clicks on a certain target protein, the following data are provided: a summary of the target protein, a list of template interfaces that the target structure is found to match and several dynamically generated static images visualizing the target structure.

In the 'Predictions' section of PRISM, we provide users with an interface to our prediction results. Users can search our results in two different ways. One possibility is directly to search for the presence of similarities between a template interface and a target structure. Alternatively, a user can input either the PDB ID or the sequence of a protein [whose sequence is then aligned to the target dataset using BLAST (19)], which is then checked for any predicted protein–protein interactions that the input protein participates in. The combinatorial search space of target structures that match different partner chains of a template interface is then displayed to the user as a list of proteins that are candidates for an interaction. This is done by first checking to see whether the input protein has a binding site similar to any one of the template interfaces, as explained above. All the target structures that are a priori found to have a binding site similar to the partner of the matched interface are listed as predicted interacting proteins.

Figure 3 is a screenshot of the prediction results. The left column lists the possible binding partners for the protein with PDB code 1mr8. The corresponding entries in the middle column show which template interfaces were used in the prediction phase. The third column gives the prediction score. Detailed information about the predictions is given in related pages. Figure 4 is an example of the output. Here one of the putative binding partners of 1mr8, 1e8a, is detailed. The template is 1mr8AB (in the template dataset, the A chain of 1mr8 interacts with the 1mr8B chain). The target is 1e8aA. Each row in the figure displays which residue in the template dataset is structurally aligned with those of the target protein. The red residues are the computational hotspots of the template interface. These are also invariant for the target protein.

The PRISM website can also be used to perform online calculations to predict binding partners of input proteins not covered by our datasets. At the moment we have implemented a preliminary service in which users can ask to see with which of the proteins in our datasets their input protein interacts. PRISM accepts an input protein either by its PDB code or by file upload. The online calculations build on our previous results. First the target dataset is replaced with the structure in question. Then the algorithm is run using the original template

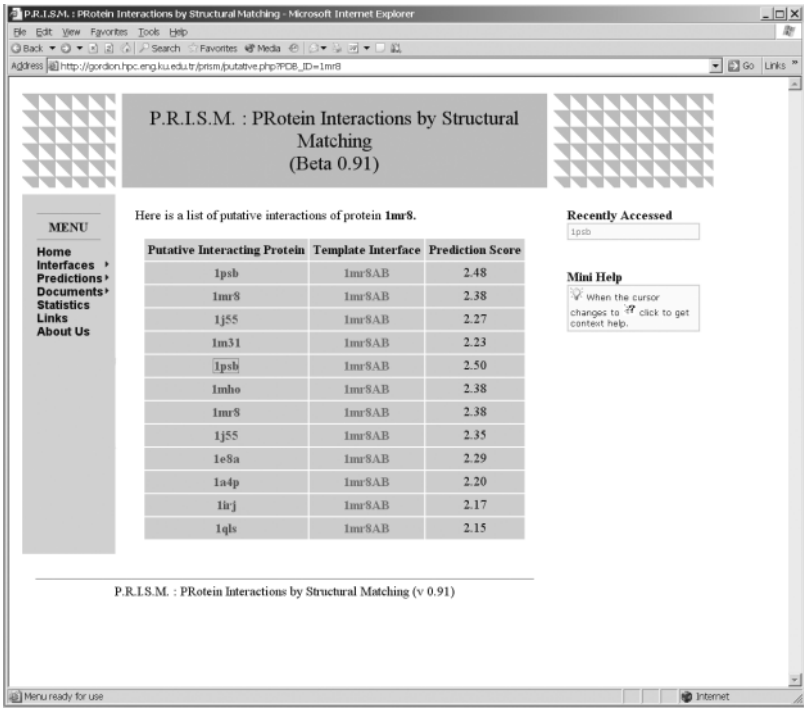
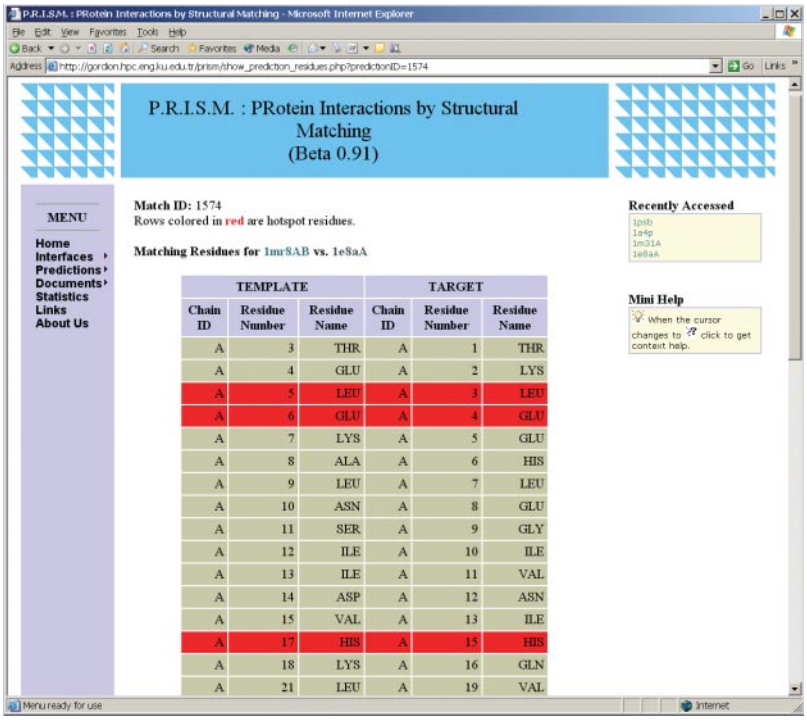


Figure 3. List of putative interacting proteins for an input protein.



TEMPLATE			TARGET		
Chain ID	Residue Number	Residue Name	Chain ID	Residue Number	Residue Name
A	3	THR	A	1	THR
A	4	GLU	A	2	LYS
A	5	LEU	A	3	LEU
A	6	GLU	A	4	GLU
A	7	LYS	A	5	GLU
A	8	ALA	A	6	HIS
A	9	LEU	A	7	LEU
A	10	ASN	A	8	GLU
A	11	SER	A	9	GLY
A	12	ILE	A	10	ILE
A	13	ILE	A	11	VAL
A	14	ASP	A	12	ASN
A	15	VAL	A	13	ILE
A	17	HIS	A	15	HIS
A	18	LYS	A	16	GLN
A	21	LEU	A	19	VAL

Figure 4. A screenshot displaying the results of the list of residues of one side of the predicted interface (target columns). The template columns are the residue listing of the template interface through which the interface was predicted. Red is used for the computational hotspots of the template interface.

set and the user input structure. Upon completion of the algorithm we know which of the template interface partners are structurally similar to the surface of the structure in question. The algorithm then finds the original structures in our target set that are similar to the partner of the template interface. These structures are output as the proteins with which the input protein is predicted to interact.

CONCLUSIONS AND FUTURE WORK

In this paper, a web server designed for the analysis of existing protein–protein interfaces is introduced. This includes a non-redundant dataset of 3799 interface clusters. Of these clusters, 67 have structurally conserved residues, computational hotspots, along their interfaces. This set is the template dataset of interfaces. A non-homologous dataset of protein structures is provided as the target dataset. The web server includes putative protein–protein interaction predictions based on our pre-calculated results. These predictions include every possible binary interaction between the target proteins. The predictions are calculated using the structure and sequence information of the template interfaces. Currently, our predictions are derived from only a subset of the known interfaces, since only 67 of the 3799 interfaces have hotspots. Therefore, finding the computational hotspots for the whole set of 3799 clusters would certainly improve and enlarge our existing predictions.

Another service provided is interactive prediction. This is done by running the algorithm for user input structures. At the moment the online prediction of an interaction between a user input protein and all the structures in our target dataset is possible. In the future, we intend to provide different types of online calculations using our prediction algorithm. For example, in a different query scenario, our prediction algorithm can also be used to see whether two specific structures not included in our datasets interact with each other.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The projects described here have been funded in part by Federal Funds from the National Cancer Institute, under contract number NO1-CO-12400. The contents of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.
2. Chakrabarti,P. and Janin,J. (2002) Dissecting protein–protein recognition sites. *Proteins*, **47**, 334–343.
3. LoConte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
4. Jones,S. and Thornton,J. (1997) Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

5. Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.
6. Jones,S. and Thornton,J.M. (1995) Protein–protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, **63**, 31–65.
7. Zhou,H.X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.
8. Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *J. Mol. Biol.*, **338**, 181–199.
9. Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T. *et al.* (2000) Toward a protein–protein interaction map of budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.
10. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
11. Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.
12. Kortemme,T. and Baker,D. (2004) Computational design of protein–protein interactions. *Curr. Opin. Chem. Biol.*, **8**, 91–97.
13. Aytuna,A.S. (2004) A high performance algorithm for automated prediction of protein–protein interactions. Master thesis, Graduate School of Engineering, Koc University, Istanbul,Turkey.
14. Berman,H.M., Westbrook,J.Z., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shyndalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
15. Keskin,O., Tsai,C.J., Wolfson,H. and Nussinov,R. (2004) A new, structurally non-redundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.*, **13** (4), 1043–1055.
16. Nussinov,R. and Wolfson,H.J. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci.*, **88**, 10495–10499.
17. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2002) Multiprot—a multiple protein structural alignment algorithm. *LNCS*, **2452**, 235–250, Springer-Verlag.
18. Keskin,O., Ma,B. and Nussinov,R. (2004) Hot regions in protein–protein interactions: The organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, **345**, 1281–1294.
19. 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 protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.