Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W503–W507***
DOI: 10.1093/nar/gkh413

BioInfo3D: a suite of tools for structural bioinformatics

Maxim Shatsky, Oranit Dror, Dina Schneidman-Duhovny, Ruth Nussinov[1,2] and Haim J. Wolfson*

School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences and [1]Sackler Institute of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel and [2]Basic Research Program, SAIC-Frederick, Inc., Laboratory of Experimental and Computational Biology, NCI-Frederick, Building 469, Room 151, Frederick, MD 21702, USA

Received February 26, 2004; Revised and Accepted April 2, 2004

ABSTRACT

Here, we describe BioInfo3D, a suite of freely available web services for protein structural analysis. The FlexProt method performs flexible structural alignment of protein molecules. FlexProt simultaneously detects the hinge regions and aligns the rigid subparts of the molecules. It does not require an a priori knowledge of the flexible hinge regions. MultiProt and MASS perform simultaneous comparison of multiple protein structures. PatchDock performs prediction of protein–protein and protein–small molecule interactions. The input to all services is either protein PDB codes or protein structures uploaded to the server. All the services are available at http://bioinfo3d.cs.tau.ac.il.

INTRODUCTION

Currently, major effort is focused on structural genomics initiatives. From the computational standpoint, the rapid increase in the number of structures presents a major challenge: how to best exploit the structural data to extract the biologically relevant features. Here, we present several recently available web services for structural analysis.

The importance of efficient protein structural comparison tools can hardly be overstated. Structural comparisons are essential for classification, for detection of conserved protein folding cores, for detection of similarities in functional binding sites, similarities in enzyme mechanisms, evolutionary conservation, construction of non-redundant databases of single chains and of protein–protein (protein–ligand) interfaces, detection of similarities between domains, identification of conserved residues, consensus motifs and pharmacophores. They are further used in fold recognition and in homology modeling. Programs for comparisons of protein structures are routinely run, existing in numerous packages.

Yet, despite the relatively large number of structural comparison algorithms, the majority perform *pairwise rigid structural comparison* [reviewed by Eidhammer *et al.* (1)]. There are very few *multiple structure comparison* algorithms. Here, we present two web services, MultiProt and MASS, that perform robust and efficient multiple protein structural alignment.

While the algorithms referred to above treat proteins as rigid bodies and carry out rigid structural comparisons, proteins are flexible molecules. Around their native state, there is an ensemble of conformational isomers separated by low-energy barriers. The movements reflect side-chain motions as well as large-scale hinge-bending movements, with the molecular parts rotating with respect to each other as relatively rigid bodies on a common hinge. Thus, two proteins with similar structures may appear different if one is hinge-bent with respect to the other. Nevertheless, despite the obvious recognition of molecular flexibility, very few algorithms have been designed to compare flexible molecules. The web server of the FlexProt method performs flexible protein alignment by simultaneous detection of the hinge regions and alignment of the rigid subparts of the molecules.

Despite the growth of the Protein Data Bank (PDB), the number of available protein–protein complexes is relatively small. Consequently, algorithms for docking are becoming integral tools of modern bioinformatics. Docking methods are used for prediction of protein–protein interactions. In addition, docking is very helpful in prediction of the the three-dimensional (3D) structures of large molecular complexes and detection of possible pharmacological targets.

Docking of drugs to receptor molecules is commonly applied in the rational drug design process. The most common application is the virtual screening of a compound library for discovery of new leads. This approach has been applied successfully, leading to a number of novel ligands (2). Docking is also applied for a partial prediction of ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties. Screening against different proteins can predict a drug's potential toxicity and its metabolites. Such computational filtering can save not only

*To whom correspondence should be addressed. Tel: +972 3 640 8040; Fax: +972 3 640 6476; Email: wolfson@cs.tau.ac.il

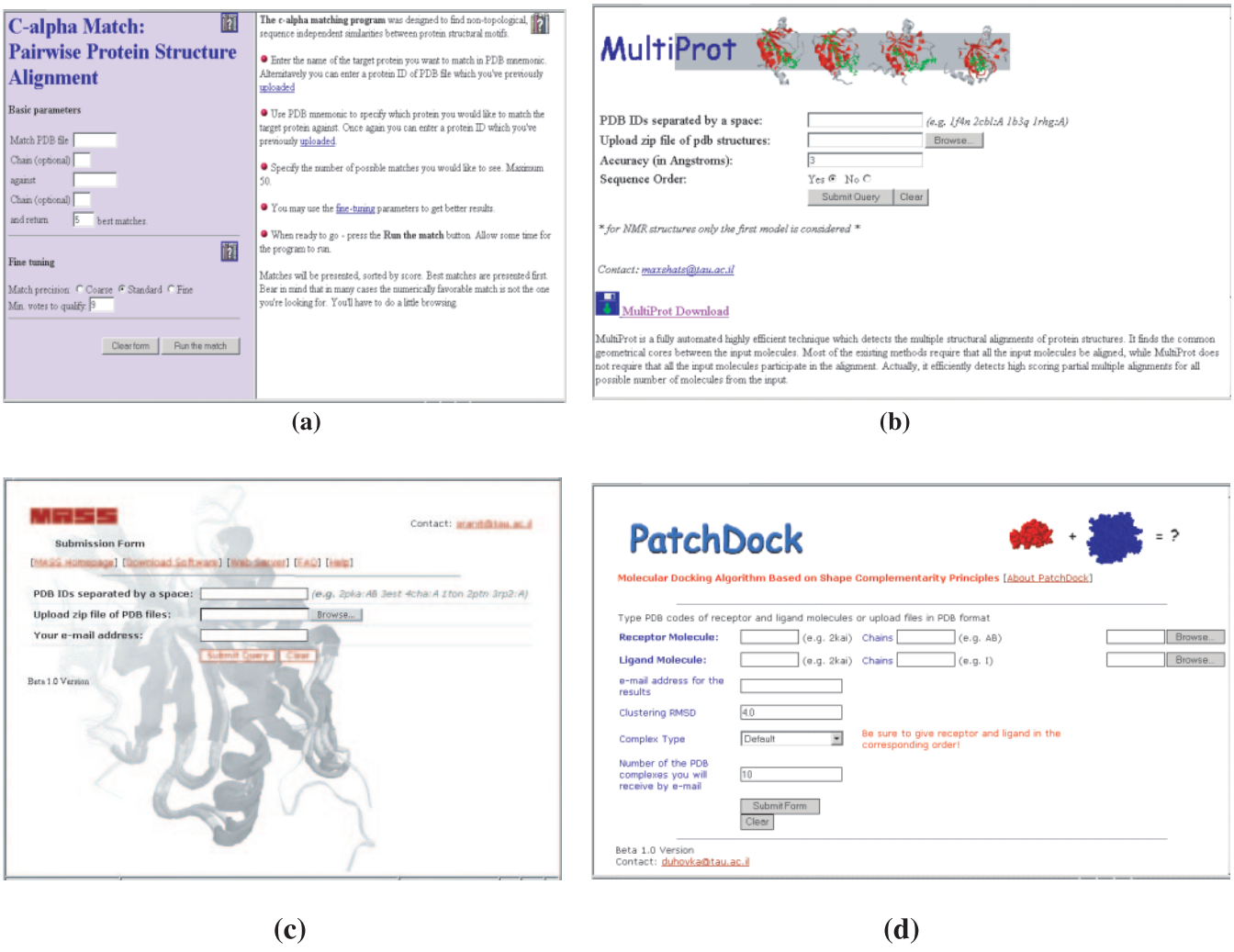

Figure 1. (**a**) C_α-Match server. (**b**) MultiProt server. (**c**) MASS server. (**d**) PatchDock server.

many *in vitro* tests, but also some of the expensive *in vivo* experiments.

The PatchDock web server presented here is among few available web services that perform prediction of protein–protein and protein–small molecule interactions.

PROTEIN STRUCTURAL ALIGNMENT: PAIRWISE, FLEXIBLE, MULTIPLE

C_α-Match: Pairwise protein structural alignment

One of the most basic operations in protein structural analysis is comparison of two protein structures. What differentiates the C_α-Match program from many other available structural alignment methods is its ability to detect order-independent alignments (3). There are numerous examples of proteins sharing similar 3D structures and functions but having different sequences and even different 3D topological order (e.g. C2 domain-like, four-helix bundle). Further, comparison of protein cores or binding sites may involve similar 3D configurations with different sequence order. Therefore, in these cases, sequence-order-independent methods should be used.

Use of the C_α-Match server (http://bioinfo3d.cs.tau.ac.il/c_alpha_match) is straightforward. The input is two protein structures either given as a four-letter PDB code or uploaded by a user in PDB format (Figure 1a). Only C_α atoms are considered. Therefore, for an uploaded file it is enough to include only C_α atoms. The *Match precision* parameter controls the maximum allowed deviation of the matched C_α atoms. The output page contains several high-scoring solutions. For each solution, there is a list of the matched amino acid pairs (the indices of the alignment do not necessarily follow the protein backbone order, i.e. it is a sequence-order-independent alignment). For each solution, there is a file in PDB format that contains two superimposed structures.

FlexProt: flexible protein alignment and hinge detection

FlexProt is a technique for the alignment of flexible proteins and has been described previously in Shatsky *et al*. (4,5). It does not require an a priori knowledge of the flexible hinge regions. FlexProt simultaneously detects the hinge regions and aligns the rigid subparts of the molecules. It is not sensitive to insertions and deletions. It is based on 3D pattern-matching algorithms combined with graph-theoretic techniques. Briefly, it works as follows. At the first stage, all structurally similar rigid fragments are detected between

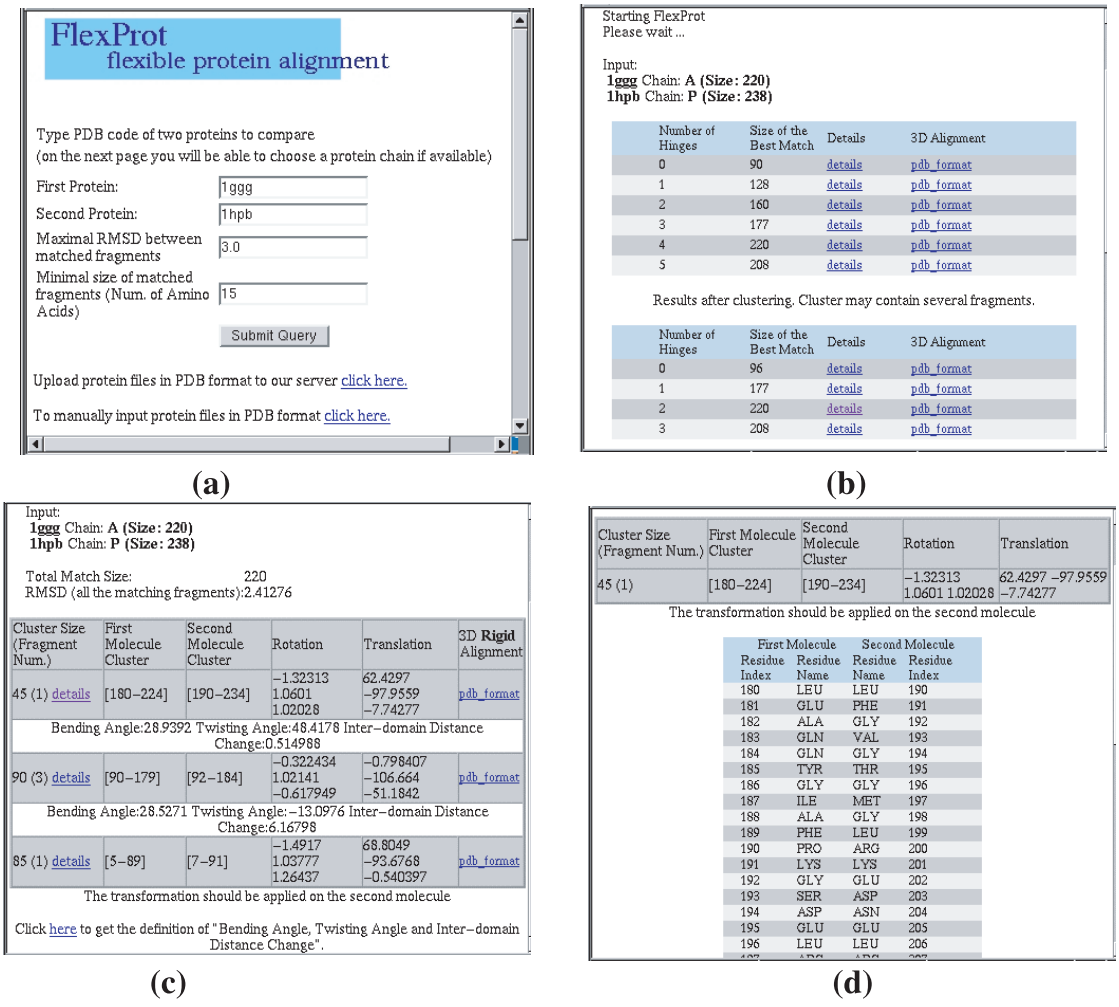

Figure 2. (**a**) FlexProt's entrance page. The user must enter two PDB protein codes or upload files in PDB format. (**b**) The results of the FlexProt algorithm. The first table contains the flexible alignments before the 'Clustering' stage is applied. The second table presents results extended by the clustering. For each result, the user can follow the link to the more specific details about flexible alignment. In addition, each flexible alignment is represented by a PDB file containing flexibly aligned molecules. (**c**) The flexible alignment is represented by a number of rigid fragment pairs. For each fragment from the second molecule, the 3D rigid transformation, which aligns this fragment with a matched fragment from the first molecule, is shown. (**d**) The alignment of the amino acid residues for the selected fragment pair.

the input molecules. Then, for each possible number of hinges an optimal combination of such fragments is computed. The third, 'clustering', stage identifies those consecutive fragments that have a similar 3D transformation, joining them into one rigid cluster.

Despite the fact that FlexProt automatically detects the hinge regions, the algorithm is as efficient as rigid structure alignment algorithms. Typical run times on proteins with a few hundred amino acids consisting of several rigid parts are ∼7 s on a 400 MHz desktop PC.

The input format to the FlexProt server (http://bioinfo3d.cs.tau.ac.il/FlexProt) is the same as for the C_α-Match server. It is either the four-letter PDB code or uploaded files in PDB format (Figure 2a). Only C_α atoms are considered. There is no restriction on the protein pair: proteins may have different lengths and even missing residues. In addition, two program parameters can be changed from the default values. Both parameters control the properties of the matched rigid fragment pairs. The first parameter, *Maximal RMSD between matched fragments*, specifies a maximum allowed root mean square deviation between the rigid fragments. Higher values allow detection of larger alignments and are more tolerant to structural changes. For example, one hinge motion may be detected with a 3 Å threshold, while for the same input with a 4 Å threshold two proteins may be structurally aligned without hinges. On the other hand, lowering the threshold to 1 Å may result in too many spurious hinges. The second parameter, *Minimal size of matched fragments*, defines a minimum length for the rigid fragment. Therefore, this parameter allows the omission of short rigid fragment pairs that can introduce noise.

Figure 2b shows the results of the program. The results are presented in two tables according to the number of flexible regions. The second table (which is recommended more for a regular user) displays the solutions processed by the 'Clustering' stage, i.e. each rigid part might contain several rigid fragment pairs which have almost the same 3D rigid transformation. Each solution can be downloaded as a file in PDB format ('pdb_format' link). The file contains the first molecule (chain A) and the matched rigid fragments (with different chain identifications) aligned with the first

molecule. Thus, only aligned rigid fragments from the second molecule are contained in the resulting PDB file. The details of each solution can be viewed by following the 'details' link from the solution table.

Figure 2c shows the details of the flexible alignment with two hinges (flexible regions). Each rigid fragment pair is presented with its 3D rigid transformation that aligns the fragment from the second molecule onto the matched fragment from the first molecule. One can view the match between the amino acid residues of each aligned fragment pair by following the 'details' link on the left (Figure 2d).

MultiProt: multiple structure alignment

MultiProt is a fully automated, highly efficient technique addressed at detecting multiple structural alignments of protein structures (6). MultiProt finds the common geometrical cores among protein molecules, aligning all molecules *simultaneously*. Further, MultiProt does not require that all molecules be in the match. The algorithm efficiently detects high-scoring partial multiple alignments for every possible number of molecules in the input. Thus, it enables detection of a number of (different) common structural motifs, and it is capable of distinguishing between similar and dissimilar molecules, not including the latter in the alignment. MultiProt is highly efficient running on tens of protein molecules. Depending on the application, it runs from a few seconds to a few minutes. For the seed alignment (a user-defined value, with a minimum of three residues), the residue-order has to be preserved. The seed extension is residue-order-independent. Therefore, like C_α-Match, it is capable of detecting non-topological alignments. The MultiProt server is available at http://bioinfo3d.cs.tau.ac.il/MultiProt (Figure 1b).

MASS: multiple 3D alignment by secondary structures

MASS is an efficient algorithm for aligning multiple protein structures and detecting 3D motifs that are common to two or more input proteins (7,8). The ability to find motifs shared by non-predefined subsets of the input molecules makes the method insensitive to structural outliers and may aid in distinguishing between subsets of similar and dissimilar protein structures. MASS is a two-tier algorithm, using both secondary structure and C_α atomic representation. In the first stage, the molecules are represented as sets of 3D line-segments, each representing a secondary structure element (SSE). Then, the Geometric Hashing paradigm is applied to obtain initial alignments between them. In the second stage, MASS uses the C_α coordinates of the protein structures to compute atomic superpositions based on the initial SSE alignments. Since proteins inherently consist of SSEs, using these in multiple structure alignments complements C_α-matching. Secondary structure configurations define the protein folds and provide the stabilizing protein scaffold, onto which the functional sites are grafted. As a result, they are evolutionarily highly conserved, while mutations frequently occur at the flexible loops, which are more difficult to match. On the practical side, since proteins are dense molecules, matching of atoms can be noisy, as there are many ways in which atoms can match. Secondary structures are more robust, practically without random, meaningless alignments. Since they are also fewer in number per protein (around 15 in a 300 residue protein), the time complexity is reduced. It should also be noted that MASS disregards the order of the SSEs along the polypeptide chain and thus it is able to detect non-topological alignments.

The input to the MASS server (http://bioinfo3d.cs.tau.ac.il/MASS) is a list of protein structures, either described by their PDB codes or uploaded to the server (Figure 1c). The output is a list of potential alignments shared by non-predefined subsets of the input proteins. The alignments are scored based on the core size and the number of participating proteins, and the highest scoring ones are sent to the user by email.

DOCKING

PatchDock: molecular docking algorithm based on shape complementarity principles (9)

The input to the PatchDock server (http://bioinfo3d.cs.tau.ac.il/PatchDock) is two molecules of any type: proteins, DNA, peptides or small drug-like molecules in PDB format (Figure 1d). The output is a list of potential complexes sorted by geometric shape complementarity score. The PatchDock algorithm is inspired by object-recognition and image-segmentation techniques used in Computer Vision. Docking can be compared to assembling a jigsaw puzzle. When solving the puzzle, we concentrate on the patterns that are unique to the puzzle element and look for the complementary patterns in the rest of the pieces. PatchDock employs a similar technique. Given two molecules, their surfaces are divided into patches according to the surface shape (concave, convex or flat). A patch is a set of neighboring critical points. These patches correspond to patterns that visually distinguish between puzzle pieces. PatchDock applies the Geometric Hashing algorithm to match concave patches with convex patches and flat patches with flat patches. Since surface patches are more stable features, the number of false positive docking configurations is reduced. The matching stage produces a list of candidate complexes. At the final scoring stage, a number of filtering scores are applied to reduce the size of the list. The scoring stage of the algorithm is enhanced by a multi-resolution surface representation, which contributes to its efficiency. The first filter checks the complexes for unacceptable steric clashes. Next, the geometric shape complementarity score is computed and low-scoring complexes are discarded. The desolvation score (10) is computed for the remaining complexes and can also be used as a filter. The final output is a list of complexes, including the shape complementarity score, penetration extent, interface area and desolvation score. This list together with complexes in PDB format is sent by the server to the email address specified by the user. The algorithm was successfully used and improved during the last four CAPRI rounds (http://capri.ebi.ac.uk), leading to acceptable predictions in several targets (11).

CONCLUSIONS

Here, we have presented BioInfo3D, a suite of tools for protein structure exploration. The methods which have

already been implemented in our server include (i) pairwise rigid structure comparison, independent of the order of the amino acids on the chain (C_α-Match); (ii) pairwise flexible structure comparison, without a predefinition of the hinges (FlexProt); (iii) multiple structure comparison (MultiProt); (iv) multiple structure comparison using secondary structure elements (MASS); and (v) docking (PatchDock). Combined with other software tools developed by our group, many of which will be soon available for download: SiteLight (12), for mapping phage display libraries onto the 3D protein surface; CombDock (13), for combinatorial docking of multiple molecules or domains for multimolecular assemblies and for folding, respectively; SiteEngine (14), for comparisons of binding sites or prediction of binding sites on protein surfaces using functional chemical groups and for comparisons of the patterns of functional groups in contact across protein–protein interfaces (I2I-SiteEngine); and Staccato, for multiple sequence alignment that is consistent with a multiple structural alignment. They form a suite of tools for structural bioinformatics. In particular, their synergistical integration according to the user's needs enhances their usefulness for various applications.

ACKNOWLEDGEMENTS

We thank our Structural Bioinformatics Group and the system team at Tel Aviv University. The research of R.N. and H.J.W. in Israel has been supported in part by the Center of Excellence in Geometric Computing and its Applications funded by the Israel Science Foundation (administered by the Israel Academy of Sciences) and by the Tel Aviv University Adams Brain Center. The research of H.J.W. is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the view 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.

REFERENCES

1. Eidhammer,I., Jonassen,I. and Taylor,W. (2000) Structure comparison and structure patterns. *J. Comput. Biol.*, **7**, 685–716.
2. Schneider,G. and Böhm,H. (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today*, **7**, 64–70.
3. Bachar,O., Fischer,D., Nussinov,R., and Wolfson,H. (1993) A Computer Vision based technique for 3-D sequence independent structural comparison. *Protein Eng.*, **6**, 279–288.
4. Shatsky,M., Nussinov,R. and Wolfson,H., (2002) Flexible protein alignment and hinge detection. *Proteins*, **48**, 242–256.
5. Shatsky,M., Nussinov,R., and Wolfson,H. (2004) FlexProt: alignment of flexible protein structures *without* a pre-definition of hinge regions. *J. Comput. Biol.*, **11**, 83–106.
6. Shatsky,M., Nussinov,R. and Wolfson,H. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, in press.
7. Dror,O., Benyamini,H., Nussinov,R. and Wolfson,H. (2003) MASS: multiple structural alignment by secondary structures. *Bioinformatics*, **19**(Suppl. 1), i95–i104.
8. Dror,O., Benyamini,H., Nussinov,R. and Wolfson,H. (2003) Multiple structural alignment by secondary structures: algorithm and applications. *Protein Sci.*, **12**, 2492–2507.
9. Duhovny,D., Nussinov,R. and Wolfson,H. (2002) Efficient unbound docking of rigid molecules. In Guigo,R. and Gusfield,D. (eds), *Workshop on Algorithms in Bioinformatics*, Rome, Italy, LNCS 2452, Springer-Verlag, pp. 185–200.
10. Zhang,C., Vasmatzis,G., Cornette,J.L., and DeLisi,C. (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, **267**(3), 707–726.
11. Schneidman-Duhovny,D., Inbar,Y., Polak,V., Shatsky,M., Halperin,I., Benyamini,H., Barzilai,A., Dror,O., Haspel,N., Nussinov,R. and Wolfson,H. (2003) Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins*, **52**(1), 107–112.
12. Halperin,I., Wolfson,H., and Nussinov,R. (2003) SiteLight: binding-site prediction using phage display libraries. *Protein Sci.*, **12**, 1344–1359.
13. Inbar,Y., Benyamini,H., Nussinov,R. and Wolfson,H. (2003) Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics*, **19**(Suppl. 1), i158–i168.
14. Shulman-Peleg,A., Nussinor,R. and Wolfson,H.J. (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.*, in press.