PROTEINS: Structure, Function, and Genetics 52:107–112 (2003)

Taking Geometry to Its Edge: Fast Unbound Rigid (and Hinge-Bent) Docking

Dina Schneidman-Duhovny,[1‡] Yuval Inbar,[1‡] Vladimir Polak,[1] Maxim Shatsky,[1] Inbal Halperin,[3] Hadar Benyamini,[3] Adi Barzilai,[3] Oranit Dror,[1] Nurit Haspel,[1] Ruth Nussinov,[2,3] and Haim J. Wolfson[1†]

[1]*School of Computer Science, Beverly and Raymond Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[2]*Basic Research Program, SAIC-Frederick, Inc., Laboratory of Experimental and Computational Biology, NCI-Frederick, Frederick, Maryland*
[3]*Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*

ABSTRACT **We present a very efficient rigid "unbound" soft docking methodology, which is based on detection of geometric shape complementarity, allowing liberal steric clash at the interface. The method is based on local shape feature matching, avoiding the exhaustive search of the 6D transformation space. Our experiments at CAPRI rounds 1 and 2 show that although the method does not perform an exhaustive search of the 6D transformation space, the "correct" solution is never lost. However, such a solution might rank low for large proteins, because there are alternatives with significantly larger geometrically compatible interfaces. In many cases this problem can be resolved by successful a priori focusing on the vicinity of potential binding sites as well as the extension of the technique to flexible (hinge-bent) docking. This is demonstrated in the experiments performed as a lesson from our CAPRI experience. Proteins 2003;52:107–112.**

Key words: CAPRI; unbound docking; flexible docking; PatchDock; binding site focusing

INTRODUCTION

We report the results of applying geometric docking algorithms developed by our group to the targets of CAPRI rounds 1 and 2. We analyze our original submissions, single out the errors that could have been avoided, outline improved strategies to tackle these targets, and present the results of applying these improved strategies. The application of these strategies, which included flexible docking for target 1 and improved focusing on binding sites for targets 2–6, have resulted in the successful docking of all but two targets (targets 4 and 5). Even for targets 4 and 5, solutions as close as 2.67 Å and 1.82 Å root-mean-square deviation (RMSD) to the native have been obtained, alas ranking at places 169 and 153, respectively. This indicates, that if a successful energy-based reranking could have been applied to the top few hundred results of the geometric docking, one could obtain a top ranking solution for these targets as well.

Geometric docking algorithms can be classified into two broad categories: (i) exhaustive enumeration of the transformation space and (ii) local shape feature matching. Exhaustive enumeration algorithms search the entire six-dimensional (6D) transformation space of the ligand. Most of these methods follow Katchalski-Katzir et al.[1] by using brute force search for the three rotational parameters and the elegant fast Fourier transform (FFT) for fast enumeration of the translation space. Because of the exhaustive enumeration, such algorithms are always expected to detect a correct solution, if the sampling of the rotation space is fine enough. The need for fine sampling of the rotation space is also the major deficiency of these algorithms, because it results in large run times. Another algorithm that exhaustively enumerates the rotation space is the "soft docking" method.[2] There are also nondeterministic search methods that use genetic algorithms.[3]

Local shape feature matching algorithms have been pioneered by Kuntz et al.[4] Connolly[5] suggested a method to match quadruples of local curvature maxima and minima points to detect candidate transformations. Our group has further developed and improved this candidate transformation detection technique by matching pairs of points with their associated normals[6,7] using geometric hashing. This algorithm, named PPD,[8] was quite successful in unbound docking. Recently, we developed two additional local feature-based unbound docking algorithms, BUDDA,[9] and PatchDock,[10] which have been applied in the CAPRI rounds 1 and 2 docking experiments. Major advantages of local feature docking algorithms is their speed and relatively natural extension to flexible (hinge-bent) docking.[11] A potential disadvantage of such algorithms, which do not use an exhaustive search, is the (theoretical) possibility to lose the correct transformation. However, our experience shows that we always detect the correct solution, although it might not be ranked high enough by a purely geometric score. Obviously, reranking of the top few hundred results by an energetic score[12] has the potential of improving the rank of the "correct" result. In this article we also propose

Grant sponsor: National Cancer Institute, National Institutes of Health; Grant number: NO1-CO-12400.

‡Dina Schneidman-Duhovny and Yuval Inbar contributed equally to the work

†Correspondence to Haim J. Wolfson at School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: wolfson@post.tau.ac.il

Received 31 October 2002; Accepted 20 December 2002

different methods and approaches for finding potential binding sites. The docking algorithms we use focus on these sites.

MATERIALS AND METHODS

Because the three algorithms that we have applied in our docking experiments are based on local feature matching, we shortly outline the general scheme of these algorithms while focusing on the salient techniques of each algorithm.

Most of the local shape feature docking algorithms can be roughly divided into the following major steps:

1. *Molecular surface representation:* A popular representation is that of the solvent accessible surface as calculated by Connolly.[13] This representation is used to obtain densely sampled molecular surface points with associated normals. On the basis of Connolly's analysis, Lin et al.[14] extracted sparser critical points with normals, each defined by the projection of the gravity center of a Connolly face. This results in a significant reduction of the points, while roughly retaining similar information. An even sparser representation by "critical" points and their normals is achieved by focusing on "knobs" and "holes," which are local maxima and minima of a shape function, roughly representing local minima and maxima of the surface curvature.[5,7,10]
2. *Focusing on candidate binding (active) sites:* To significantly reduce the number of false positives and reduce computation time, it is desirable to focus a priori on the approximate areas of the molecular surface, where binding is likely to appear. Such candidate binding sites are usually detected by biological and shape criteria. An excellent example of "biologically" defined binding regions are the complementarity determining regions (CDRs) in antibodies. Additional focusing can be achieved by preferring areas with high hot spot concentration.[15,16] An example of a shape criterion is the binding of drugs and small ligands in large cavities of a receptor. There one might restrict the receptor surface to be explored to such cavities.
3. *Complementary spatial pattern detection:* This is the heart of geometric docking algorithms and usually matches triplets or pairs of critical points with associated normals, which should align in roughly opposite directions. After proper clustering, the output of this step is a set of candidate rigid transformations, which dock one molecule to the other.
4. *Geometric complementarity scoring and ranking:* Because molecules cannot penetrate into each other, candidate transformations from the previous step are discarded if they cause a significant penetration. Minor penetrations are allowed to reflect conformational changes of the molecular surface on docking. In this step one also calculates a geometric compatibility score based on the size of the computed interface, while penalizing the allowed minor penetrations as a function of the penetration depth and size.
5. *Biological scoring and reranking:* In this step one would like to accept the high enough scoring hypotheses of the previous step and rerank them according to a free-energy function, which could discriminate between the biologically valid hypotheses and geometrically compatible false positives. This step was not applied in our algorithms (there is a limited application of energy terms in PPD).

Within the above mentioned scheme, PPD[8] represents the molecular surface by dense Connolly points, matches pairs of sparse knobs/holes with their associated normals, scores interfaces by geometry (based on the dense Connolly representation), electrostatics and propensity of aromatic residues. BUDDA[9] represents the molecular surface by the sparser Lin et al.[14] caps/pits/belts in addition to a distance transform grid, matches triplets of critical features based on one knob/hole and a pair of caps/pits in their vicinity, scores interfaces by geometric complementarity (based on the distance transform grid), and focuses on binding sites in the matching step. PatchDock[10] uses a multiresolution representation of the molecular surface by Connolly points of different densities in addition to a distance transform grid. On the basis of the shape function, which measures approximate curvature, it partitions the molecular surfaces into convex, concave, and flat local patches of almost equal size. Patches with higher probability of belonging to the binding site (e.g., hot spot propensity) are considered, and complementary configurations of pairs of knobs/holes with associated normals within the patches are detected. Alignment of such pairs induce rigid transformations, which are subsequently tested for shape penetration and scored by geometric complementarity by using the multiresolution ligand surface and distance transform grid of the receptor molecule. The use of surface patches reduces the number of potential docking hypotheses, while still (in almost all tested cases) retaining the correct transformation.

RESULTS

In the CAPRI rounds 1 and 2 docking experiments we have applied the PPD[8] and BUDDA[9] algorithms for target 1, BUDDA for targets 2 and 3, and BUDDA and PatchDock[10] for targets 4–7. As a result of the lessons, we have learned from the analysis of our submissions after the publication of the complexes, we rerun target 1 with a new BUDDA-based flexible docking algorithm and rerun targets 2–6 with BUDDA and PatchDock, applying an improved focusing on potential binding sites of the antigens. Below we report the highlights of the results for each target. All the reported run times are on a 1.8 GHz Pentium IV PC.

Target 1: Lactobacillus HPr Kinase-*B. subtilis* HPr

Submitted results

In this docking experiment we have a priori restricted the distance between the side-chain oxygen of Ser(Asp)-46 and the closest phosphate oxygen to 10.0 Å. We have also restricted the algorithm conformational search to the area of the HPr kinase P-loop. Our best result within the top 10

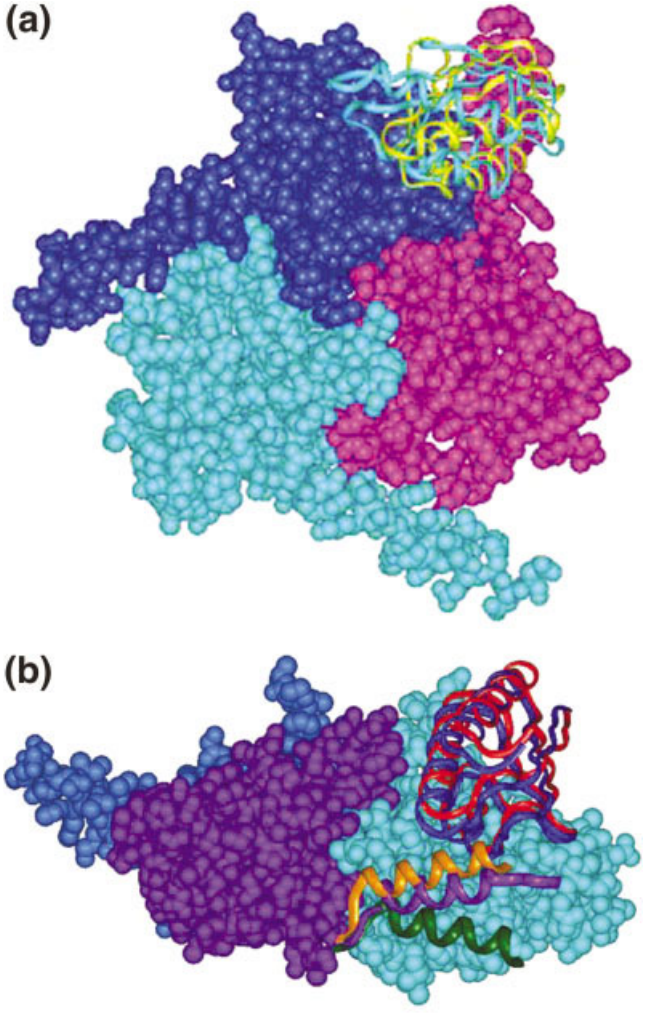

Fig. 1. Target 1. **a:** Best rigid docking result within the top 10, 8.0 Å from the native. The enzyme is shown in spacefill. **b:** The flexible helix at the bottom of the figure: green: position of the helix in the native uncomplexed enzyme; purple: position in the complex; orange: position in our best flexible docking solution. The hinge-based flexibility improved the geometrical complimentarity of the near-native solutions: best hinge-bent docking result (ranked 2nd), 3.0 from the native; blue: docking solution; red: crystal.

was about 8.0 Å RMSD close to the native and ranked 7 th by the geometric score [Fig. 1(a)]. The main reason for this mediocre performance in a relatively easy target is due to the flexibility of the enzyme [see Fig. 1(b)]. This strongly affects the geometric docking score, because a considerable part of the interface surface area is between the HPr and this flexible, helix of the enzyme.

Lessons learned

Because the major problem was lack of flexibility, we have implemented a new flexible (hinge-bent) docking algorithm based on the local feature matching of BUDDA and flexibility handling in the spirit of our previous approach.[11] The enzyme was divided into two rigid parts: the helix of chain C and the body of chain A without that helix. The results of this algorithm have been significantly better than the rigid version. The second best scoring result was 3.0 Å RMSD from the native [see Fig. 1(b)]. These results have been achieved without using the 10.0 Å distance constraint. The run time of the algorithm was 2 min.

Target 2: Bovine rotavirus VP6-Fab

Submitted results

In all the antibody-antigen docking examples, we have restricted the potential antibody binding site to the CDRs. The CDRs were detected by alignment of the antibody sequence with a "CDR sequence template." More precisely, a CDRs detection procedure was developed. This procedure receives as input a sequence of an antibody (light chain and/or heavy chain) and outputs the most likely residues that the CDRs consist of. The likelihood calculations are based on a multiple alignment of thousands of known antibodies sequences. We used statistical data available at http://home.ust.hk/hxue/igprfs and the CDRs were set according to the union of the Kabat and Wu[17] and Chothia and Lesk[18] CDR definitions.

In target 2 the antigen VP6 potential binding site was restricted to the β-domain. We selected solutions with interfaces that include at least four CDRs of the antibody with high TYR, TRP concentration and at least two chains of the antigen. Clustering of the solutions obtained for the different chains of the trimer was performed. Our best solution was 15 Å RMSD from the native and ranked 7th.

Lessons learned

Our main conclusion from the experiment was that we should focus better on the binding site of the antigen. Thus, we introduced several restrictions that seem general enough and biologically justifiable. In particular, we limited the search for the antigen-binding site to the loop regions of the antigen. We further restricted it to the exposed part of the virus capsid. In addition, we discarded results that cause steric clash of the three (symmetric) antibodies binding to the antigen trimer. Under these restrictions we received among the first 10 a result with 5.54 Å RMSD from the native (ranked 9) in 7-min runtime. It should be noted that although the area of the interface of the native complex is about 400 $Å^2$, the interface area of our (geometrically) highest ranked solution is 600 $Å^2$. In our calculations, the interface area is equivalent to half of the surface area of both molecules, which becomes buried on binding. In the highest ranked solution, the light chain of the antibody is shifted toward the center of the virus capsid, enlarging shape complementarity. The heavy chain is very close to its original location (Fig. 2).

Target 3: Influenza Hemagglutinin-Fab HC63

Submitted results

We selected solutions with interfaces that included the following restrictions: (i) at least four CDRs of the antibody with high TYR, TRP concentration should participate in the interface, (ii) only one chain of the antigen should participate in the interface (this was an erroneous assump-

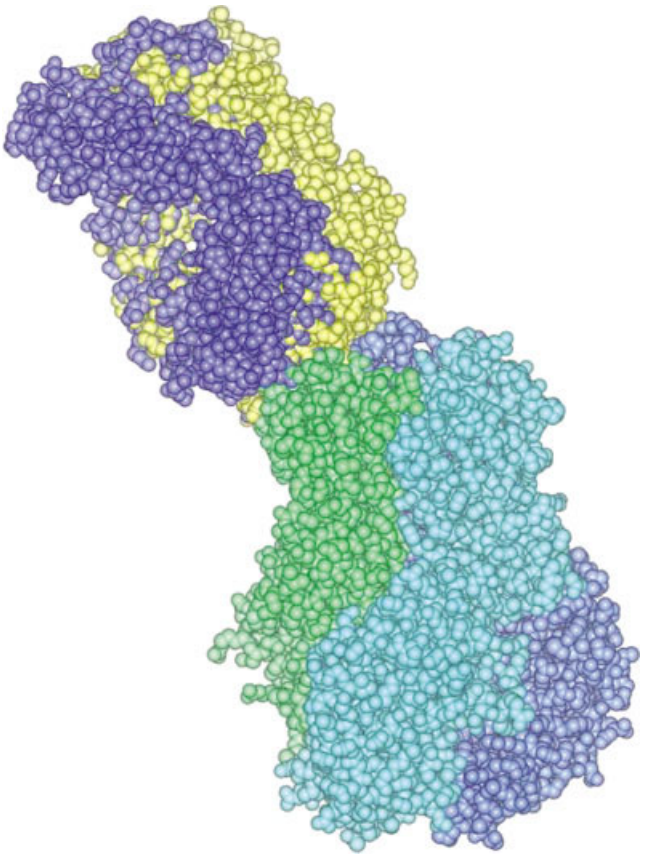

Fig. 2. Target 2. Highest ranked solution compared to the native: the interface area of the native (blue) complex is ~400 Å^2, whereas the interface area of our highest ranked solution (yellow) is ~600 Å^2. In this result, the light chain of the antibody is shifted toward the center of virus capsid, enlarging shape complementarity. The heavy chain is very close to its original location.

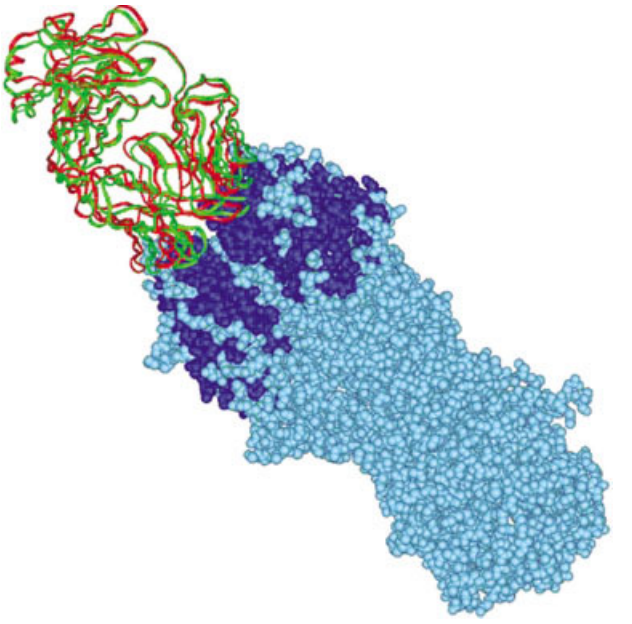

Fig. 3. Target 3. Of 320 residues 138 structurally conserved residues (in blue). Some of those residues exhibit significant sequence variability. Focusing on these regions, the 6th ranked result with 3.10 Å RMSD. The native complex antibody is in red and our solution is in green. Run time was 5 min.

tion and the main reason to failure in the experiment!). Further, clustering of the solutions obtained for the different chains of the trimer was performed.

Lessons learned

As in target 2, we restricted the potential antigen-binding site to the exposed domain of the virus capsid. We also discard results that cause steric clash of the three antibodies (symmetry constraint) or include only one chain of the virus capsid in the interface. In addition, we have focused on the structurally conserved regions of the influenza hemagglutinin. There are several reasons for this hypothesis:

- The virus constantly mutates the exposed domain surface of the antigen to avoid neutralization by the antibodies. On the other hand, the binding site of the antigen to the cellular receptors is located in a small cavity on the exposed top of the molecule, so the mutations in these regions should not significantly change the structure to preserve activity.[19]
- The antibody will be most effective in its simultaneous action against various mutants of the antigen, if it binds those regions of the antigen that are less likely to be changed.
- In particular, the antibody has to block the access to the antigen receptor-binding site, which is structurally conserved.

By applying the MultiProt algorithm[20] to multiply align all available 25 hemagglutinin structures from PDB, we have detected 138 structurally conserved residues out of 320 residues (Fig. 3), some of them exhibiting significant sequence variability. The "complementary pattern detection step" (see Materials and Methods) was further restricted to "critical points" belonging to the antigen structurally conserved residues. As a result of this binding site focusing, we received a solution with 3.10 Å RMSD from the native, ranked 6 (Fig. 3). The run time of the procedure was 5 min.

Targets 4–6: α-Amylase-Camelide Antibody-VH Domains 1,2,3

Submitted results

We decided to focus on the sequence variable regions of the mammalian amylase, which have been detected by multiple sequence alignment. We were given antibodies for pig α-amylase that were produced by the camel. The camel has its own α-amylase, so he would produce antibodies for pig molecules that will not bind to his amylase to avoid autoimmune response. From the structural viewpoint, the produced antibodies should include in their interfaces residues from the amylase molecule that differ between the two species.

The sequence of the camel α-amylase is not known, so we decided to focus on the variable regions of the amylase molecule. We preferred results with larger interface area of the CDR heavy loop H3, as well as larger interface area belonging to the sequence variable regions of the amylase. This latter constraint was the main reason for our failure to get reasonable predictions in all three targets. Actually, only 15%, 13%, and 20% of the interfaces in targets 4, 5, and 6, respectively, belonged to these nonconserved residues, whereas the submitted results included much higher percentage of variable regions.

Lessons learned

We disregarded the sequence variable region constraint and applied only one restriction, which required that at least 70% of the antibody interface in the candidate complexes belongs to the CDRs. As a result for target 6, the closest solution to the native ranked 4 with 1.90 Å RMSD. For targets 4 and 5, the closest solutions to the native ranked 169 and 156 with RMSD 2.67 Å and 1.82 Å respectively. The run times for these targets were about 25 min on average. It is important to note that although the interfaces of the native complexes are of size 405 $Å^2$, 435 $Å^2$, and 570 $Å^2$, respectively, the interfaces ranked highest by our geometric docking procedure were 765 $Å^2$, 700 $Å^2$, and 600 $Å^2$, respectively. This large gap between the best geometric fit and the native fit for targets 4 and 5 emphasize the limitations of purely geometric docking in some cases.

Target 7: T-Cell Receptor β-Chain-Streptococcal Pyrogenic Exotoxin

Submitted results

This a classical case that can be solved by structural homology. In a PDB search we found a complex of TCR with staphylococcal enterotoxin. The superimposition of the toxins with high structural similarity is our first solution, which proved to be correct. In addition, we run two docking experiments with binding site focusing. The first experiment focused on the binding sites of both the TCR and SAG, obtained by structural alignment. The best result ranked 3rd with 3.37 Å RMSD from the native. The run time was 1 min. When focusing only on the binding site of the TCR the best result with 3.37 Å RMSD ranked 36 with run time of 7 min.

CONCLUSIONS

We have presented results of fast rigid docking algorithms, which are based on geometric shape complementarity only. One of the algorithms has been easily extended to include flexibility (hinge bending) following the CAPRI rounds 1 and 2 experiments. Despite the heuristic nature of the algorithms, which are based on local shape complementarity and not on exhaustive search of the transformation space, correct solutions are not lost. A correct solution always appears among the first few hundred, yet the best geometric solution might exhibit significantly higher shape complementarity than the native one. We have presented both the results of our original submissions and the results of a posteriori experiments, which have mainly concentrated on improved binding site focusing procedures. We learned that such binding site-focusing procedures, based on biological knowledge, significantly improve the success of the geometric docking algorithms. Because no energy function was directly used in our experiments, it remains an open question whether reranking of the top few hundred geometric solutions by one of the available energy functions would significantly improve the results.

ACKNOWLEDGMENTS

We thank the organizers of CAPRI and the experimentalists, who contributed structures for this exciting experiment. We acknowledge the help and advice of Snait Tamir and Raquel Norel. The research of R. Nussinov and H.J. Wolfson 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). The research of H.J. Wolfson 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 organization imply endorsement by the U.S. government.

REFERENCES

1. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A, Aflalo C, Vakser I. Molecular surface recognition: determination of geometric fit between protein and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
2. Jiang J, Kim SH. "Soft Docking" matching of molecular surface cubes. J Mol Biol 1991;219:79–102.
3. Gardiner EJ, Willett P, Artymiuk PJ. Protein docking using a genetic algorithm. Proteins 2001;44:44–56.
4. Kuntz I, Blaney J, Oatley S, Langridge R, Ferrin T. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982; 161:269–288.
5. Connolly M. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. Biopolymers 1986;25:1229–1247.
6. Fischer D, Norel R, Nussinov R, Wolfson HJ. 3-D docking of protein molecules, 4th Symposium on Combinatorial Pattern Matching, June 1993, Padova, Italy, Lecture Notes in Computer Science 684, pp. 20–34, Springer-Verlag.
7. Norel R, Lin SL, Wolfson H, Nussinov R. Shape complementarity at protein-protein interfaces. Biopolymers 1994;34:933–940.
8. Norel R, Petrey D, Wolfson H, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999;35: 403–419.
9. Polak V. Backbone-based unbound docking. M.Sc. thesis. School of Computer Science, Tel Aviv University, 2002.
10. Duhovny D, Nussinov R, Wolfson HJ. Efficient unbound docking of rigid molecules. In: Guido R, Gusfield D, editors. Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI) Rome, Italy, Lecture Notes in Computer Science 2452, pp. 185–200, Springer Verlag, 2002.
11. Sandak B, Nussinov R, Wolfson HJ. An automated Computervision & Robotics based technique for 3-D flexible biomolecular docking and matching. Comp Appl BioSci 1995;11:87–99.
12. Camacho JC, Gatchell DW, Kimura SR, Vajda S. Scoring docked conformations generated by rigid body protein protein docking. Proteins 2000;40:525–537.

13. Connolly M. Analytical molecular surface calculation. J Appl Crystallogr 1983;16:548–558.
14. Lin SL, Nussinov R, Fischer D, Wolfson HJ. Molecular surface representation by sparse critical points. Proteins 1994;18:94–101.
15. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
16. Hu Z, Ma B, Wolfson W, Nussinov R. Conservation of polar residues as hot spots at protein–protein interfaces. Proteins 2000;39:331–342.
17. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities: relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. J Immunol 1991;147:1709–1719.
18. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 1987;196:901–917.
19. Skehel JJ, Wiley DC. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem 2000;69:531–569.
20. Shatsky M, Nussinov R, Wolfson HJ. MultiProt—a multiple protein structural alignment algorithm. In: Guido R, Gusfield D, editors. Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI) Rome, Italy, Lecture Notes in Computer Science 2452, pp. 235–250, Springer Verlag, 2002.