Examination of Shape Complementarity in Docking of *Unbound* Proteins

Raquel Norel,[1] Donald Petrey,[2] Haim J. Wolfson,[1] and Ruth Nussinov[3,4]*
[1]*Computer Science Department, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel*
[2]*Department of Biochemistry & Molecular Biophysics, Columbia University, New York, New York*
[3]*Laboratory of Experimental and Computational Biology, SAIC, NCI-FCRDC, Frederick, Maryland*
[4]*Sackler Institute of Molecular Medicine, Faculty of Medicine, Tel Aviv University, Tel Aviv Israel*

ABSTRACT **Here we carry out an examination of shape complementarity as a criterion in protein-protein docking and binding. Specifically, we examine the quality of shape complementarity as a critical determinant not only in the docking of 26 protein-protein "bound" complexed cases, but in particular, of 19 "unbound" protein-protein cases, where the structures have been determined separately. In all cases, entire molecular surfaces are utilized in the docking, with no consideration of the location of the active site, or of particular residues/atoms in either the receptor or the ligand that participate in the binding. To evaluate the goodness of the strictly geometry-based shape complementarity in the docking process as compared to the main favorable and unfavorable energy components, we study systematically a potential correlation between each of these components and the root mean square deviation (RMSD) of the "unbound" protein-protein cases. Specifically, we examine the non-polar buried surface area, polar buried surface area, buried surface area relating to groups bearing unsatisfied buried charges, and the number of hydrogen bonds in all docked protein-protein interfaces. For these cases, where the two proteins have been crystallized separately, and where entire molecular surfaces are considered without a predefinition of the binding site, no correlation is observed. None of these parameters appears to consistently improve on shape complementarity in the docking of unbound molecules. These findings argue that simplicity in the docking process, utilizing geometrical shape criteria may capture many of the essential features in protein-protein docking. In particular, they further reinforce the long held notion of the importance of molecular surface shape complementarity in the binding, and hence in docking. This is particularly interesting in light of the fact that the structures of the docked pairs have been determined separately, allowing side chains on the surface of the proteins to move relatively freely.**

This study has been enabled by our efficient, computer vision-based docking algorithms. The fast CPU matching times, on the order of minutes on a PC, allow such large-scale docking experiments of large molecules, which may not be feasible by other techniques. Proteins 1999;36:307–317.

Key words: docking; protein-protein recognition; molecular surface complementarity; shape complementarity; rigid-body matching; surface normals

INTRODUCTION

It is well known that molecular shape plays a critical role in the binding of two molecules. Over the years, the notion of shape complementarity has been confirmed by inspection of a large number of complexed structures in the protein data bank (PDB, Bernstein et al.[1]). Consequently, shape complementarity has been used as a prime consideration in docking approaches that take into account entire molecular surfaces rather than strictly active site regions. Nevertheless, the more realistic question is how well does shape complementarity perform when the structures of the two molecules to be docked have been determined separately, and consequently display surface variability? Surface side chains move, and inevitably, the extent of molecular surface complementarity can be expected to be affected.

Under such circumstances, can shape complementarity still be used in searches for well-docked molecules? Shape complementarity at the interface of complexes that have been crystallized together is expected, and it is therefore logical that it would work satisfactorily in re-docking of

Grant sponsor: National Cancer Institute, Department of Health & Human Services; Grant sponsor: BSF, Israel; Grant number: 95-00208; Grant sponsor: Israel Science Foundation; Grant sponsor: Magnet Fund; Grant sponsor: Israel Ministry of Industry; Grant sponsor: Tel Aviv University; Grant sponsor: Adams Brain Center; Grant sponsor: Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University; Grant sponsor: National Institutes of Health; Grant numbers: 5-T32-GM08281 and 2ROI-GM41371; Grant sponsor: National Science Foundation; Grant number: DBI-9601463.

Raquel Norel's present address is Department of Biochemistry & Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032.

The contents of this publication do not necessarily reflect the views or policies of the Department of Health & Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

*Correspondence to: R. Nussinov at NCI-FCRF, Building 469, Room 151, Frederick, MD 21702. E-mail: ruthn@ncifcrf.gov

Received 15 September 1998; Accepted 5 April 1999

bound cases. However, the question of how it would perform in unbound cases has not been explored. Yet, this question is important for a number of reasons. First, on the technical side, docking procedures initiating from complete molecular surfaces where the binding sites are unknown cannot perform detailed time-consuming energy calculations. Only empirical, fast energy evaluation approaches are feasible at this stage. Hence, it is important to examine the goodness of the obvious alternative i.e., shape, at least in the first step in the docking. Second, shape complementarity is a geometry-based criterion. In general, criteria based on geometry are relatively fast to calculate. Third, by carrying out such docking experiments, the role of shape complementarity in binding of two, "free" molecules may be inferred.

Here we analyze shape complementarity as a major criterion, not only in the docking of bound protein-protein cases, but in particular in the docking of the unbound molecules. The obtained results are compared with those ranked by major favorable and unfavorable components in energy evaluation, i.e., non-polar buried surface area, polar buried surface area, buried surface area relating to buried unsatisfied charges, and hydrogen bonds. The latter are of particular interest in this regard, as they can serve not only with respect to their stabilizing contribution to the energetics, but in addition, they can straightforwardly provide geometric criteria as well. Hydrogen bonds provide both directionality and preferred distance considerations. Therefore, here we pay particular attention to the question of how well they perform in the docking of unbound protein-protein molecules.

Geometry Versus Energetics Docking Approaches

Traditionally, there have been two major approaches to the docking problem. The first is geometry-based, the second focuses on the energetics. While it has been recognized that both components need to be satisfied for correct positioning of a ligand onto a receptor surface, the question remained as to which way was most beneficial to approach the problem at the start. While energy-based docking schemes have been based on knowledge of the approximate positioning of the ligand in the receptor active site,[2,3] with a subsequent optimization of the interactions, in the geometry-based schemes the active site has not always been assumed as known a priori. To a large extent, the geometry-based methods have been based on the assumption that the molecular surfaces of the receptor and the ligand need to match if the molecules are to bind to each other with high affinity. This approach has been adopted early on by Kuntz and his colleagues,[4] and further elaborated and rationalized by Connolly.[5] Other geometry-based docking approaches have adopted and followed this rationale.[6–18]

Initiating the docking process from the geometrical rather than the energetics standpoint has several attractive advantages. First, one does not need to know a priori the positioning and approximate orientation of the ligand within the receptor-binding site. In addition, there is always the possibility that the ligand under consideration would bind at an alternate location. Furthermore, starting from geometrical considerations, the top scoring, highest ranking binding modes can subsequently be submitted to energetic optimization routines.

The problem of docking a ligand onto a receptor surface with the sole input being the atomic coordinates is extremely difficult. In general, one starts by describing the molecular surface. In the absence of any further biochemical data, in principle one would need to proceed by matching every piece of surface of one molecule with every piece of surface of the other, in every rotation and translation. Given the difficulty involved, and in particular, in order to limit the number of obtained binding modes and the ranking of the so-called "correct" docked configurations, one seeks additional constraints which could potentially be used. In principle, therefore, it would be advantageous if one could, from the beginning, incorporate into the docking procedure some chemical properties of the atoms, or groups of atoms involved. Here, the most natural components are hydrophobicity, electrostatics, and hydrogen bonds. To some extent, hydrophobicity is already implicitly taken into account in the matching of the geometrical shapes, however, it can be added as an additional filter as well. On the other hand, shape complementarity also considers the geometrical features of the matching. Electrostatics is long range, and is not easily implemented in a straightforward manner into the docking of interpenetrating molecules. In addition, there is a large variability in the actual, observed contributions of electrostatics to protein-protein interactions. On the other hand, implementing hydrogen bonds directly into the docking procedure appears attractive on several counts. First, it is relatively straightforward to implement, given a definition of H-bond donors and acceptors. Second, it is middle-range, with well-defined distances. Third, hydrogen bonds provide directionality as well because they form at preferred angular orientations. Hence, apart from their energetic contribution, which can be straightforwardly used as well, they provide important geometrical constraints directly in the matching of molecular-pairs.

Hydrogen bonds have been shown to work well and yield remarkably low RMSD solutions in the docking of a large set of protein-protein complexes (Meyer et al.[19]) where the two molecules have been co-crystallized. Furthermore, in particular, it has been assumed that hydrogen bonds may be more robust in the docking. This argument has been advanced in particular with respect to flexible molecules, and in consideration of surface variability/flexibility in solution. The rationale has been that in the latter, flexible cases, shape complementarity may be weak, particularly for the smaller (drug, cofactor) ligands. Hence, it has been argued that utilizing hydrogen bonds, with their preferred distances and directionalities would be advantageous.

Shape Complementarity and H-Bonds

The problem we are then faced with is how well does shape complementarity perform in the docking of the protein-protein bound, i.e., complexed cases, and in particular in the unbound cases, where the structures of the molecules have been determined separately. Furthermore, how general is the shape complementarity notion? Can shape complementarity, implicitly taking into account packing and total buried surface area considerations, predict well-docked configurations? These questions are particularly pertinent, given that extensive analyses of protein-protein interfaces have shown that even for crystal oligomers the variability of hydrogen bonds geometry is quite large, both with respect to angular and to distance distributions.[20] Superimposing an unbound receptor on a bound one, illustrates the conformational changes in charged residues upon binding. For example, Lys-97 in lysozyme shifts its side-chain from its conformation in the unbound state (in 5lymA) to form a salt bridge with Asp-32(H) in the bound state (3hfmY). Most large conformational changes of charged/polar groups are associated with charge complementarity, or with hydrogen bond formation upon binding. The conformational change shown by Lys-97 is appreciably higher than that of the adjacent Lys-96, where there is no negatively charged residue nearby in the complex. Furthermore, when a given receptor binds to different ligands, as in the case of the trypsin, there is an extensive variability in the geometry of the hydrogen bonds. The question of shape complementarity is also relevant for drugs. Drugs are small, appear to be more charged, and to form more hydrogen bonds with respect to their interface area than the larger protein-protein complexes, and to be more flexible.

Here we examine this question of shape complementarity extensively. We examine how well considerations of shape complementarity perform for 26 complexed protein-protein cases, which are large ligands, and in particular, for 19 unbound protein-protein cases. We evaluate the “goodness” of the shape complementarity as the sole criterion in the docking as compared to utilization of non-polar buried surface area, polar buried surface area, buried surface area corresponding to groups with unsatisfied buried charges, and with respect to the number of hydrogen bonds at the protein-protein interface. Kasinos et al.,[21] and more recently the extensive study of Meyer et al.,[19] have shown that hydrogen bonds can be used in the docking of protein-protein bound, complexed cases. The results obtained by Meyer et al.[19] are particularly impressive in this regard. Yet, in these studies protein complexes have been employed. In these cases, the shape complementarity and hydrogen bond geometry can be expected to be quite good. Results for the more realistic cases where the structures have been determined separately, and hence demonstrate surface variability, have not been reported.

A major difficulty in docking is the scoring of the docked configurations, and their ranking. Obtaining a correct docked configuration with a very high ranking is a serious problem. To date, for the unbound case, techniques improving the scoring and the ranking of the near native configurations have been shown to work only if the docking is constrained to the active site, and in particular, if one specifies at least one atom of the receptor which is in contact with the ligand.[22,23] While this specification may not necessarily be in the docking a priori, it is used subsequently in the filtering of the solutions, hence the same additional constraints apply. So far, automated fast scoring approaches have not been successful. If the active site is known in advance, and if residues playing a role in the binding are known, recent advances have enabled improving the ranking of the near native complexed configuration.[22,23]. However, in the absence of such knowledge, very little progress has been made. Hence, the ingredients used explicitly in the geometric docking procedure are critical for a successful docking and ranking.

We study the major determinants contributing to prediction of successful docking and ranking when only the atomic coordinates are known. We focus on shape complementarity particularly versus hydrogen bond geometrical constraints. We are able to carry out an extensive analysis owing to our extremely fast docking routines. For the largest protein-protein cases the CPU matching times do not exceed 25 min on a PC. We handle plasticity of molecular surfaces by allowing “soft” docking of interpenetrating molecular pairs. In all cases, we have obtained reasonable RMSDs. For the 26 complexed bound protein-protein cases, our RMSDs range from 0.5Å to 2.5Å, and for the 19 protein-protein cases in which the structures were determined separately, i.e., the unbound cases, the RMSDs range between 1.2Å to 5.2Å.

Our results illustrate that shape complementarity is still a major determinant in binding even when the structures have been obtained while in the unbound state. Given the flexibility of surface atoms, the positioning and directionality of hydrogen bonds may be too sensitive. On the other hand, it is well known that protein-protein binding does not necessarily take place where either the total buried surface area is largest, or when the non-polar buried surface area is maximal.

METHODS

Critical Points

Critical points that describe the molecular surface are selected from the Molecular Surface dots generated by the MS program.[24,25] The selection of critical points is similar to the way already presented by us as described in Norel et al.[13,16] The critical points describe local knobs and holes. A knob is a local minimum of the shape function. A hole is a local maximum. Both knobs and holes are calculated in each molecule. We seek matched knob-hole pairs from the two molecules, making no reference as to which molecule contributes the hole, and which molecule contributes the knob.

Docking

For the matching step we use the same routine with the same parameters as described previously.[16] Briefly, since three non-collinear points from each molecule are needed to compute a rigid transformation to superimpose one body onto the other, and because there may not be three independent matching critical points pairs between the receptor and the ligand, we have adopted a variant strategy. This strategy not only enables finding the matched surfaces, but also reduces the complexity of the program. We simply pick only two critical points. The additional points are provided by the tips of their respective surface normals. This is unlike the original work of Connolly,[5] which has inspired our approach. However, Connolly used four independent critical points, which were not always present in the binding interface.

For each pair of points from each molecule (any two critical points) a "signature" is computed. The information of the signature includes: the distance between the dots *(d)*; the two angles formed by the line between the critical points and their respective normals (α_1, α_2); and the torsion angle between the two normals (ω). The best rigid transformation[26] between the two pairs of points is computed only if the signatures are compatible. The signatures are defined as compatible if: a knob in one molecule matches a hole in the other; the difference between the two distances *(d)* is less than 2Å; the difference between the two corresponding α's is up to 0.6 radians and the difference between the ω's is up to 0.7 radians. To impose a stronger constraint on the alignment of the normals, two additional criteria are applied: the sum of the differences of the two α's should be less than 0.75 radians, and the sum of the differences of the three angles should be less than 1.2 radians. Under these conditions, if one of the normals is not well aligned, the other must be reasonably well aligned in order for the pair of dots to be considered a viable solution. There are several advantages in using the surface normals in the signature, and not just dots. First, we seek only two critical points in the interface area. Second, the combinatorics of finding the correct match is lower (we use pairs of points, rather than triplets or quartets). And third, the normal orientation is used for fast pruning of many wrong solutions.

Geometric Scoring and Overlap Test

Because the matching is computed locally, it is imperative to see if the solution is acceptable for the entire protein, i.e., if there is no overlap between the two molecules when bringing the ligand onto the receptor. We compute a scoring function based on geometric features, rewarding surface contact, penalizing overlaps, and rejecting serious overlaps. By allowing some intermolecular penetrations we implicitly take into account a certain extent of conformational flexibility. The only solutions that are discarded are those in which ligand atoms fall into the "core" of the receptor protein. "Core" atoms are those atoms that do not generate MS dots. Solutions with ligand atom centers that invade the outer shell of the molecular representation are retained. The scoring and overlap test is as described previously.[15]

In addition, we utilize a very simple hydrophobicity filter. The atoms are divided into polar and hydrophobic. Each MS dot (from each molecule) is labeled as polar or hydrophobic depending on its closest atom. When mapping the receptor molecule onto the 3D grid to compute the score, at each surface voxel two counters are kept, for polar MS dots and for hydrophobic MS dots that fall into that voxel. The ligand molecule is transformed and mapped onto the same grid. Three counters are then updated for each MS dot, one for polar-polar interactions *(pp)*, one for hydrophobic-hydrophobic interactions *(hh)* and one for hydrophobic-polar interactions *(hp)*. The total number of interactions is then $total_{\text{interactions}} = hh + pp + hp$. We compute the hydrophobicity factor as $hf = hh/total_{\text{interactions}}$.

We use two thresholds for the hydrophobicity factor: $hf > 0.12$ for immunoglobulins; otherwise $hf > 0.17$.

We have also implemented a Connectivity Filter for the solutions that passed the overlap test (for details see Norel et al.[16]). The connectivity filter gives preference to matchings of larger patches of surfaces. If there is no overlap between the ligand and the receptor, the MS dots from the ligand that are in contact with the receptor ('C' dots) are grouped into connected regions. The size of a connected component is simply the number of C dots that belong to that component. We seek the largest component and the second largest, if the second largest is "large enough," i.e., if its size is at least 10% of the size of the largest. The docked conformations whose connected components' (CC) size is at least 5% of MS_{ligand} are reported as potential solutions. MS_{ligand} is the number of MS dots in the ligand (computed at a density of 1 dot Å ^{2}).

Inspection of the obtained complexed conformations immediately reveals, however, that many molecular associations do not vary appreciably, representing virtually the same docked solution.[15] Clustering similar solutions both reduces their number and allows focusing on different and possibly alternate configurations. One of the problems with clustering schemes is the definition of the thresholds, which is rather subjective. A good clustering scheme should group similar solutions. However, at the same time it should properly distinguish between alternate binding sites. Here we cluster solutions with a relative rotation $< 60°$ angular distance and a relative translation < 5Å. The relative rotation between two conformations can be computed from their individual rotations against the initial (zero-transformation) conformation. If the rotations are R_a and R_b for conformations A and B, then B is rotated with respect to A by $R = R_b R_a^{-1}$. When the rotation matrix R is known between two conformers, the rotational angle between them is *acos* [tr[R] − 1/2]. The relative translation between two conformations is calculated as the distance between the mass centers of their interfaces.

RESULTS AND DISCUSSION

Our method is completely automated. There is no manual modification of the files. For the complexed bound cases, the receptor and the ligand are separated and put back together by the docking scheme. We measure the quality of the docked solutions by computing the RMSD between the native orientation of the ligand in the crystal complex and the ligand as oriented by the program. The only information required for these procedures is the atomic coordinates of each molecule (that is, the standard PDB file). Only heavy atoms are used. No hydrogens are included. Only one set of docking parameters has been utilized in all examples. The matching parameter set is the same as that presented in Norel et al.[16] Table I lists the protein-protein cases we have examined. Table Ia lists the 26 bound cases, PDB file names, resolution, and size, in terms of the number of atoms in the receptor and the ligand. Table Ib lists the same information for the 19 unbound cases we

TABLE Ia. The Protein-Protein Complexes Used in this Study

	pdb	Receptor name	#atoms	Ligand name	#atoms	Res. in Å
1	1cho	alpha-chymotrypsin 1–146 (E)	1047	alpha-chymotrypsin 149–245 (E)	701	1.8
2	1fdl	IG*G1 fab fragment (LH)	3306	2-lysozyme (Y)	1000	2.5
3	1tec	thermitase eglin-c (E)	2003	leech (I)	826	2.2
4	1tgs	trypsinogen (Z)	1645	pancreatic secretory trypsin inhibitor (I)	496	1.8
5	2hfl	IG*G1 fab fragment (LH)	3227	lysozyme (Y)	1000	2.5
6	2kai	kallikrein a (AB)	1798	bovine pancreatic trypsin inhibitor (I)	438	2.5
7	2mhb	hemoglobin α chain (A)	1068	β chain (B)	1133	2.0
8	2ptc	beta-trypsin (E)	1628	pancreatic trypsin inhibitor (I)	453	1.9
9	2sec	subtilisin carlsberg (E)	1919	genetically-engineered n-acetyl eglin-c (I)	529	1.8
10	2sni	subtilisin novo (E)	1937	chymotrypsin inhibitor (I)	512	2.1
11	2tgp	trypsinogen (Z)	1628	Pancreatic trypsin inhibitor (I)	453	1.9
12	3hfm	IG*G1 fab fragment (LH)	3293	lysozyme (Y)	1000	3.0
13	4cpa	carboxypeptidase	1536	potato carboxypeptidase a inhibitor (I)	275	2.5
14	4hvp	HIV-1 protease chain A	745	chain B	745	2.3
15	4sgb	serine proteinase (E)	1309	potato inhibitor pci-1 (I)	379	2.1
16	4tpi	trypsinogen (Z)	1628	pancreatic trypsin inhibitor (I)	455	2.2
17	1abi	hydrolase alpha thrombin (H)	2039	chain L	265	2.3
18	1acb	hydrolase alpha-chymotrypsin (E)	1769	eglin C (I)	522	2.0
19	1cse	subtilisin carlsberg (E)	1914	eglin C (I)	522	1.2
20	1tpa	anhydro-trypsin (E)	1628	trypsin inhibitor (I)	454	1.9
21	2sic	subtilisin (E)	1938	subtilisin inhibitor (I)	764	1.8
22	5hmg	influenza virus hemagglutinin (E)	2532	chain F	1417	3.2
23	6tim	triosephosphate isomerase chain A	1883	chain B	1883	2.2
24	8fab	fab fragment from IGG1 chain A	1544	chain B	1635	1.8
25	9ldt	lactate dehydrogenase chain A	2565	chain B	2565	2.0
26	9rsa	ribonuclease chain A	951	chain B	951	1.8

TABLE Ib. The Unbound Cases Used in This Study

	pdb	Receptor name	#atoms	Res. in Å	Ligand name	#atoms	Res. in Å
1	1hfm-1lym(A)	IG*G1 fv fragment	1714	model	lysozyme (A)	1001	2.5
2	1hfm-1lym(B)	IG*G1 fv fragment	1714	model	lysozyme (B)	1001	2.5
3	1tgn-4pti	trypsinogen	1621	1.6	trypsin inhibitor	453	1.5
4	1tgn-5pti	trypsinogen	1621	1.6	trypsin inhibitor	464	1.0
5	1tgn-6pti	trypsinogen	1621	1.6	trypsin inhibitor	458	1.7
6	1tld-4pti	beta-trypsin	1629	1.5	trypsin inhibitor	453	1.5
7	1tld-5pti	beta-trypsin	1629	1.5	trypsin inhibitor	464	1.0
8	1tld-6pti	beta-trypsin	1629	1.5	trypsin inhibitor	458	1.7
9	2hfl-1lyz	IG*G1 fab fragment	3220	2.5	lysozyme	1001	2.0
10	2hfl-6lyz	IG*G1 fab fragment	3220	2.5	lysozyme	1001	2.0
11	2pka-4pti	kallikrein a	1799	2.0	trypsin inhibitor	453	1.5
12	2pka-5pti	kallikrein a	1799	2.0	trypsin inhibitor	464	1.0
13	2pka-6pti	kallikrein a	1799	2.0	trypsin inhibitor	458	1.7
14	2ptn-4pti	trypsin	1629	1.5	trypsin inhibitor	453	1.5
15	2ptn-5pti	trypsin	1629	1.5	trypsin inhibitor	464	1.0
16	2ptn-6pti	trypsin	1629	1.5	trypsin inhibitor	458	1.7
17	2sbt-2ci2	subtilisin novo	1934	2.8	chymotrypsin inhibitor	521	2.0
18	5cha(A)-2ovo	alpha-chymotrypsin (A)	1735	1.7	ovomucoid third domain	418	1.5
19	5cha(B)-2ovo	alpha-chymotrypsin (B)	1736	1.7	ovomucoid third domain	418	1.5

The protein-protein cases used in this study. (a) The 26 bound cases; (b) the 19 unbound ones. We list the PDB file names, the names of the proteins, and their resolution and sizes, in terms of non-hydrogen atoms.

have docked. Table IIa displays the results we have obtained in the docking of the 26 cases enumerated in Table Ia. The table lists the PDB file name and the CPU in minutes on a PC workstation (586 PC clone, running at 133MHz) for the docking (matching) step. The CPU for the scoring step depends on whether a "connectivity component" is employed. The connectivity component essentially requires that there be patches of nearby MS dots that are in contact between the receptor and the ligand, rather than isolated ones. Checking for such connected, nearby

TABLE IIa. The Results Obtained for the Protein-Protein Bound Cases

	pdb	CPU (min) docking	# of clustered solutions	RMSD (Å)	Ranking
1	1cho	1.7	471	0.54	1
2	1fdl	8.6	2181	1.50	20
3	1tec	2.2	1042	1.18	1
4	1tgs	2.8	831	1.14	1
5	2hfl	10.4	2166	1.51	1
6	2kai	2.2	1227	1.17	11
7	2mhb	7.2	663	0.70	1
8	2ptc	2.6	1027	0.59	1
9	2sec	1.8	1114	2.08	1
10	2sni	2.2	1367	1.07	1
11	2tgp	1.6	828	0.59	1
12	3hfm	10.7	2274	0.76	1
13	4cpa	2.1	1310	1.02	3
14	4hvp	1.4	411	2.06	1
15	4sgb	0.9	591	1.88	5
16	4tpi	2.1	889	0.52	1
17	1abi	6.2	773	0.56	1
18	1acb	3.8	1121	0.94	1
19	1cse	1.7	1024	1.32	2
20	1tpa	2.6	950	0.23	1
21	2sic	3.2	1229	1.11	1
22	5hmg	17.7	329	1.09	1
23	6tim	11.0	351	0.50	1
24	8fab	2.3	93	1.97	1
25	9ldt	24.1	67	2.52	1
26	9rsa	2.9	511	1.30	21

TABLE IIb. The Results Obtained for the Unbound Cases

	pdb	CPU (min) docking	# of clustered solutions	RMSD (Å)	Ranking
1	1hfm-1lym(A)	11.8	11475	2.97	537
2	1hfm-1lym(B)	4.0	10685	2.80	281
3	1tgn-4pti	3.3	2619	1.85	53
4	1tgn-5pti	5.3	3453	1.22	1
5	1tgn-6pti	3.2	1455	1.75	2
6	1tld-4pti	2.5	2659	5.22	16
7	1tld-5pti	3.6	3471	4.71	619
8	1tld-6pti	2.5	1512	2.18	40
9	2hfl-1lyz	10.1	10989	1.79	110
10	2hfl-6lyz	12.6	10733	1.08	65
11	2pka-4pti	1.9	3184	3.29	29
12	2pka-5pti	3.1	4222	1.21	9
13	2pka-6pti	1.9	1756	1.82	27
14	2ptn-4pti	2.8	2156	3.53	9
15	2ptn-5pti	4.0	2880	3.11	34
16	2ptn-6pti	5.5	1200	1.28	56
17	2sbt-2ci2	2.8	3582	2.62	92
18	5cha(A)-2ovo	1.7	2194	1.49	11
19	5cha(B)-2ovo	3.1	2289	1.64	2

The results obtained in the docking of the protein-protein cases. (a) The results of the 26 bound cases; (b) the results for the 19 unbound protein-protein cases. We list, in this order, the PDB file names; the CPU of the docking (matching) step, in minutes, on a PC; the number of (clustered) solutions, the best RMSD, in Å which has been obtained, and the highest ranking solution with an RMSD under 5 Å. The ranking is given following the hydrophobicity and the connectivity filters.

points forming patches is a time-consuming operation. The scoring CPU time may reach hours when applying the connectivity filter. The scoring procedure without the connectivity filter is about three times faster, taking from minutes to a couple of hours on a PC. The scoring function itself is very simple, checking for the interpenetration of ligand atoms into the mapped receptor ones, as described previously.[16] In addition, a simple hydrophobicity score has also been added by taking the ratio of the hydrophobic-hydrophobic atom-*voxel* contacts divided by the total number of receptor-ligand *voxel* contacts. Table IIa also gives the number of solutions. Unlike previously,[16] here solutions having similar transformations have been clustered. The clustering thresholds are given in the Methods section. The highest ranking near-native conformation is noted next. Inspection of Table IIa illustrates that for all cases low RMSD solutions have been obtained in short docking (matching) times. The number of clustered solutions for these large protein-protein cases is tractable, and the ranking of the near native docked configuration is very high in most of the cases.

Table IIb presents the results obtained for the 19 unbound protein-protein examples. The table lists the PDB file names of both the receptors and the ligands used in the docking experiments. It gives the CPU of the docking (matching), again in minutes on a PC. The RMSDs of the best solutions are noted. The number of clustered solutions is displayed, as well as the ranking of the near native docked configurations. Again, the docking times are short, the RMSDs of the obtained configurations (with respect to their unbound placed on their crystal complexed counterpart) are acceptable, although not as good as for the bound cases. As expected, the real problem emerges with the ranking of the near native conformations. In some cases the ranking is particularly problematic (e.g., for the 1hfm-1lym(A), the highest ranking near native solution is at 537; for 1tld-5pti the best rank is at 619; for 1hfm-1lym(B) the best rank is 281), indicating a need for an improvement in the ranking procedure. Figures 1a and 1b illustrate two of the unbound protein-protein cases we have docked. Figure 1a depicts the docking of 2pka-4pti, which obtained an RMSD of 3.29Å, and Figure 1b shows 2sbt-2ci2, with an RMSD of 2.62Å. In both cases the entire docked molecules are displayed, with the predicted solution (in red) superimposed on the mock complex (with the mock complex being the unbound PDB molecules superimposed on the PDB complex).

The docking and scoring presented in Tables IIa and IIb are based solely on shape complementarity. Even a quick glance at these tables suffices to illustrate that the shape complementarity generally performs well in the docking of the bound examples. For the unbound cases, the results are substantially different. Even when based only on shape complementarity between the two molecules, solutions with reasonable (though, as expected, higher) RMSDs are still obtained, and despite the variable surfaces, the ranking is considerably worse. We have next proceeded to examine the performance of other, empirical energy-based considerations. We have explored non-polar buried surface

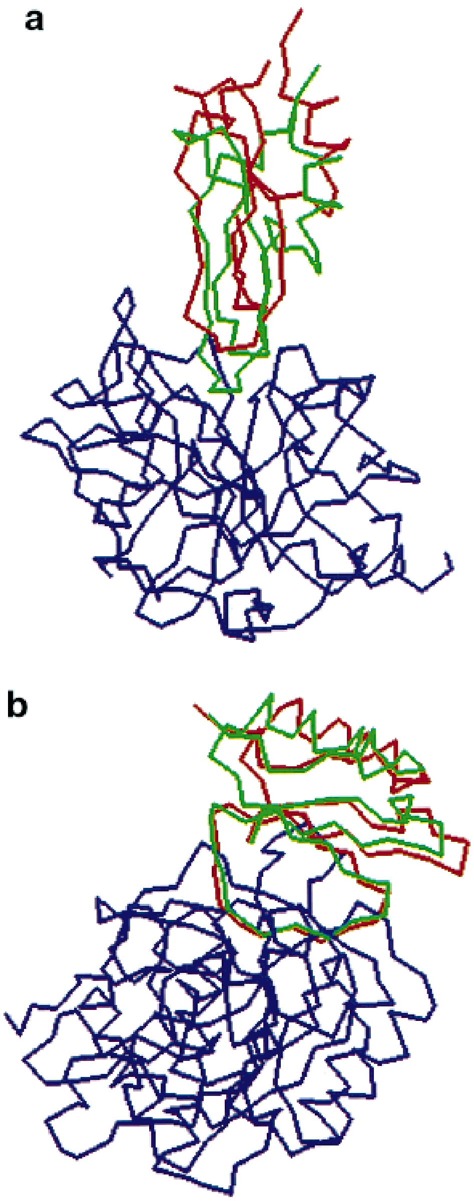

Fig. 1. Two examples of the docked configurations obtained for unbound protein-protein docking. (**a**) 2pka-4pti. The RMSD of the obtained solution with respect to the unbound-crystal superimposed on the complexed crystal is 3.29Å. The unbound receptor is depicted in light blue. It is superimposed on the bound receptor. The unbound ligand, superimposed on the bound is in green. The orientation of our solution is in red. The graphics program used is RasMol V 2.5, by Roger Sayle, Biocomputing Research Unit, University of Edinburgh. (**b**) 2sbt-2ci2. The RMSD is 2.62Å. The color code is as in (a).

area, polar buried surface area, buried surface area corresponding to groups bearing unsatisfied charges, and the number of hydrogen bonds. We were particularly interested in the latter, as hydrogen bonds can serve in the dual roles of providing both a means for energy scoring, and a fast geometric criterion, which can be employed both in the docking and in the scoring. The results we have obtained are depicted in Figures 2a through 2d for four arbitrarily chosen unbound protein-protein docked cases.

To calculate the energy parameters we have employed the following procedures: Hydrogen bond identification and surface area calculations in each conformation were performed with a program written specifically for this purpose using the Troll software library of macromolecular analysis tools (D. Petrey and B. Honig, unpublished results). Hydrogen bonds were identified based on the geometry of the donor-acceptor pair, using angle constraints and donor/acceptor definitions as specified by Stickle et al.[27] However, a distance constraint of 3.8Å between heavy atoms was employed. This distance is larger than the one used by Stickle et al. and hence may result in the overcounting of hydrogen bonds. Nevertheless, it has been chosen so that a description of an interface based on such electrostatic properties would be broad, and the decision on rejection of docked conformations would not be based on parameters which might be too strict.

Accessible surface area calculations were performed using an algorithm developed by Sridharan.[28] The algorithm spreads dots on each atom at a distance equal to the radius of the atom plus a probe radius of 1.4Å. The dots are spread in an icosahedral pattern, which results in a dot density of approximately one point per Å^2. An area is assigned to each of the dots based on their density and the radius of the atom. Each dot is examined sequentially to determine if it is occluded by another atom. Speed and efficiency are obtained by placing the dots in an order which makes examination of later dots unnecessary, if earlier dots have already been observed to be occluded by some single atom. Hence, if a single face of the icosahedron is occluded by another atom, no further examination of dots on that face is carried out.

Accessible surface areas are determined by summing the areas represented by dots not occluded by other atoms. Buried total, polar, and non-polar accessible areas are determined by flagging each atom as either polar or non-polar, and counting the accessible dots lost upon formation of a complex. Backbone N, C, and O atoms are defined as polar and C_α is defined as non-polar. Side-chain polar and non-polar atoms are listed in Table III. Buried polar area associated with unsatisfied polar groups is defined as any polar area which is lost upon formation of the complex and, additionally, is not associated with an atom that participates in a hydrogen bond. Groups bearing such an unsatisfied charge are mostly composed of atoms that have an absolute charge of greater than 0.3 as defined in the charmm22[29] parameter set.

Figures 2a through 2d show wide distributions of the solutions for all of these empirical energy parameters. Examination of all the examples reveals that for none of the parameters we obtain consistent, low RMSD solutions using either the non-polar or the polar buried surface areas. Similarly, no low RMSD solutions are observed with consistently large numbers of hydrogen bonds. Also similarly, for the buried surface areas associated with groups bearing unsatisfied charges, no low RMSD solutions are observed with the smallest such areas. In all cases, a scatter is observed. And, while for all parameters low RMSD solutions are obtained with a relatively large polar or non-polar buried surface areas, hydrogen bonds, or low

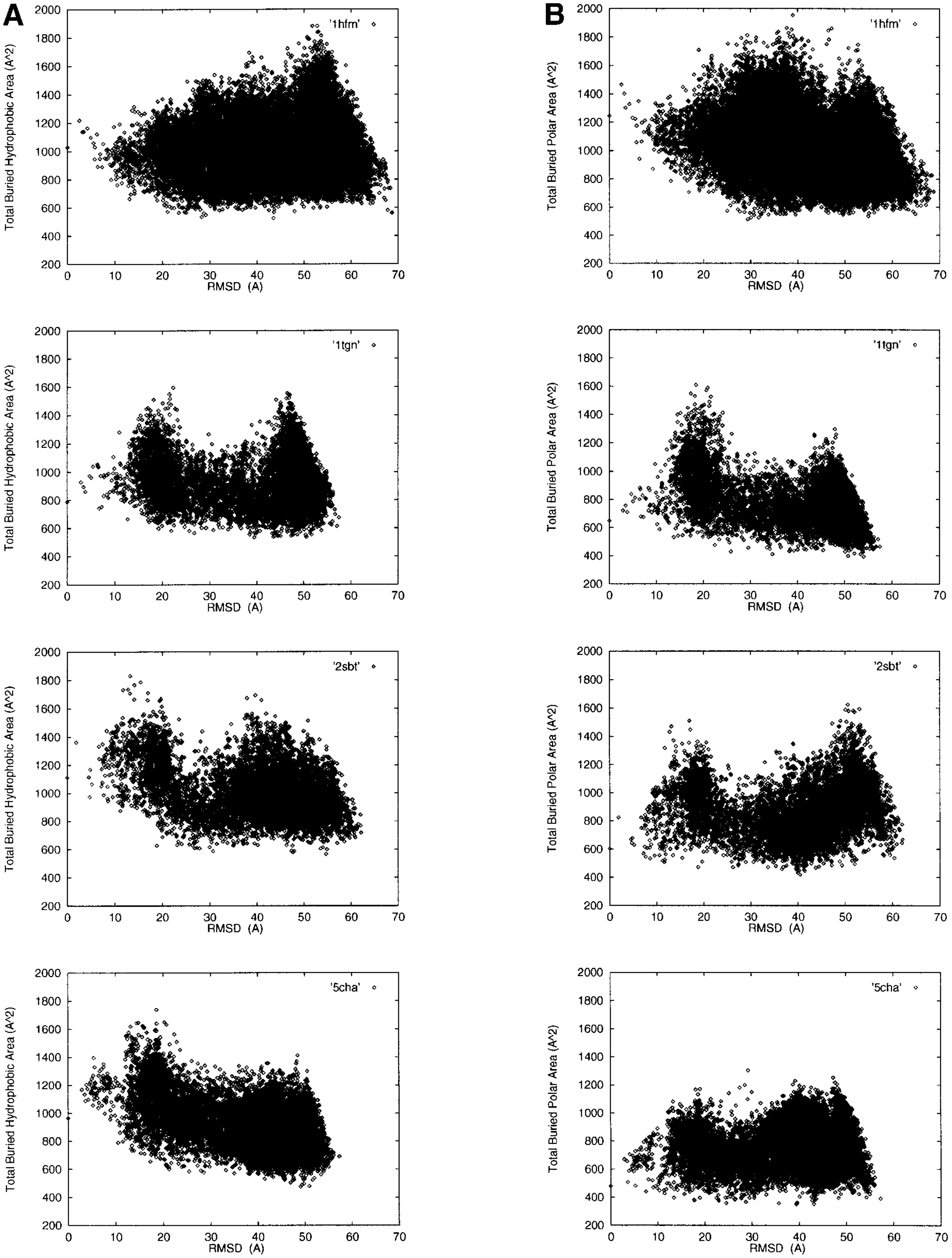

Fig. 2. Four examples of docking of unbound molecules are illustrated here. For each case, the obtained solutions of their associations with respect to the separately available crystal complex are shown. The solutions are plotted for four empirical energy calculations: (**a**) non-polar buried surface area; (**b**) polar buried surface area; (**c**) the number of hydrogen bonds; and (**d**) buried surface area relating to groups bearing buried unsatisfied charges. The mode of calculation is outlined in the text. The solutions with a 0 RMSD are those of the unbound crystal structures placed on the corresponding crystal complex. These are the reference points in the calculation of the RMSDs of all solutions. The file names are noted next to the plots. The full names of the proteins are correspondingly given in Table Ib.

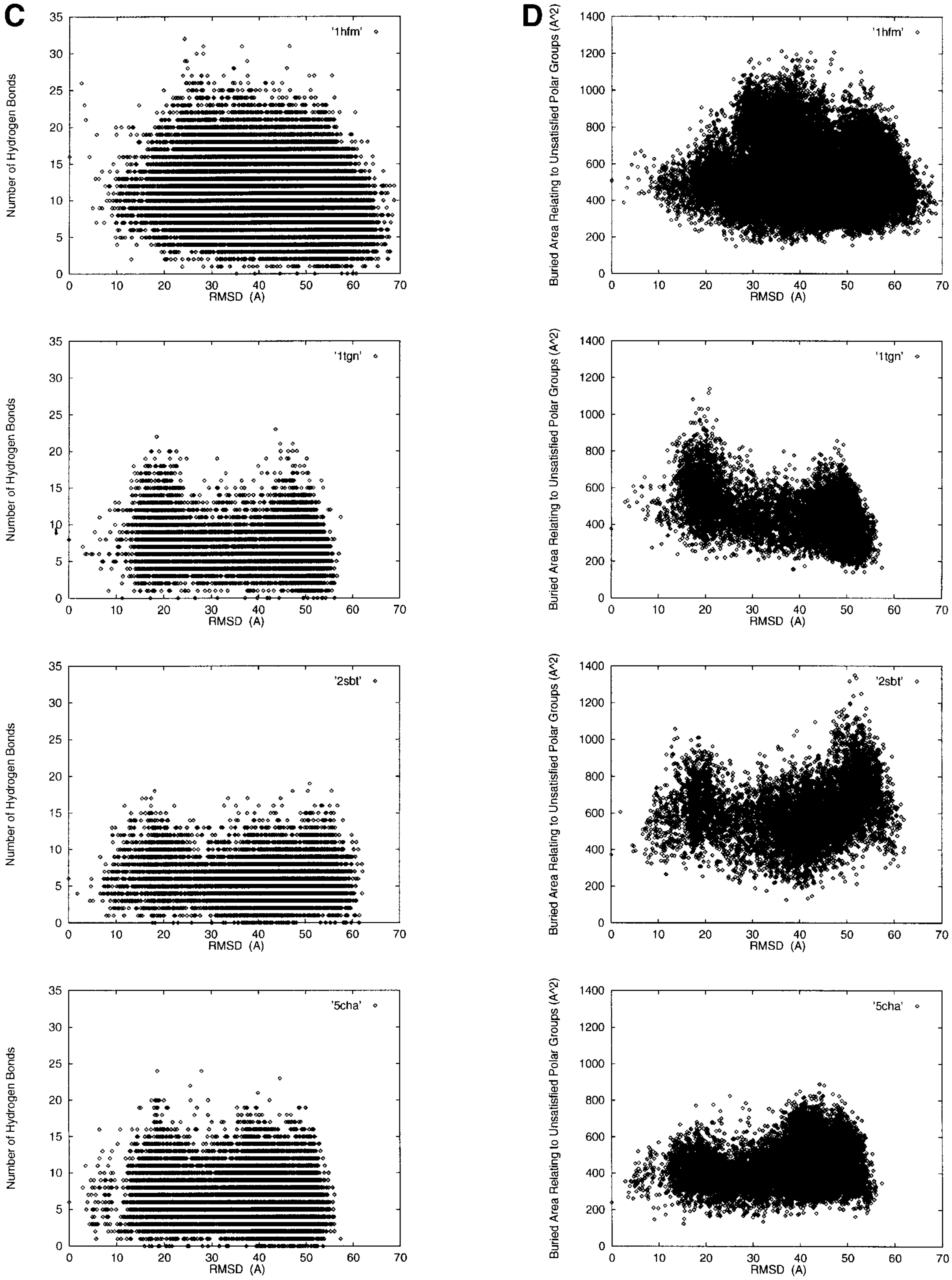

Figure 2.

buried surface areas associated with buried unsatisfied charges, there are solutions with larger RMSDs having more attractive values. We have also plotted (not shown) the total buried surface areas, and these show similar scatter. Hence, while it is still possible that a certain combination and weighting of these parameters would work, it appears that such an approach cannot replace the simple, straightforward shape complementarity, at least

TABLE III. Polar and Non-Polar Atoms

Amino Acid	Polar Atoms	Non-Polar Atoms
Ala	—	CB
Val	—	CB, CG1, CG2
Leu	—	CB, CG, CD1, CD2
Ile	—	CB, CG1, CG2, CD
Pro	—	CB, CG, CD
Met	—	CB, CG, SD, CE
Phe	—	CB, CG, CD1, CD2, CE1, CE2, CZ
Ser	OG	CB
Thr	OG1	CB, CG2
Asn	OD1, ND2, CG	CB
Gln	OE1, OE2, CD	CB, CG
Trp	NE1	CB, CG, CD1, CD2, CE1, CE2, CE3, CZ2, CZ3, CH2
Cys	SG	CB
Tyr	OH	CB, CG, CD1, CD2, CE1, CE2, CZ
Asp	OD1, OD2, CG	CB
Glu	OE1, OE2, CD	CB, CG
Lys	NZ	CB, CG, CD, CE
Arg	NE, NH1, NH2, CZ	CB, CG, CD
His	ND1, NE2	CB, CG, CD1, CD2, CE1

A listing of polar and non-polar atoms as classified for the buried surface area calculations.

in the first stage of the docking. In particular, the fact that no correlation between low RMSD solutions and the number of hydrogen bonds is observed, is in contrast to the situation observed in the bound cases, where extremely good results have been obtained.[19] This is understandable in light of the sensitivity of hydrogen bond criteria. In solution, the directionality, i.e., the angles, may be quite different than those observed for the already associated molecules. The latter have optimized their complementarity in the complex.

Here we treat the more general problem, namely, assuming no knowledge of the binding site region. Since this necessitates consideration of entire molecular surfaces, the procedure needs to be both efficient, and to handle a larger number of potential solutions. In principle, this points to the possibility of using simple geometric criteria. Since shape complementarity has long been observed in bound molecules, the question arises as to the validity of such considerations in the binding of molecules which exist separately in solution, and hence manifest surface variability. Here we show that even under such circumstances, shape complementarity may still be utilized. Nevertheless, two conditions should be considered. First, the extent of shape complementarity which can be expected between the molecular pair is a function of molecular flexibility. The more flexible the molecules, and the further away their unbound states are from their corresponding bound conformations, a smaller shape complementarity could be expected. The study carried out here addresses molecules manifesting what is often termed surface plasticity, rather than entire flexibility. The second consideration concerns water molecules. To achieve optimal surface complementarity, bound crystal waters would need to be taken into account. Currently, despite the existence of several algorithms to predict such water molecules, there is still a considerable uncertainty.

CONCLUSIONS

Here we address a practical problem: To what extent does complementarity of molecular shapes at the binding interfaces play a role in protein associations? This problem is particularly realistic for cases where the molecular structures have been determined separately, and thus owing to surface side-chain variability and plasticity, the shapes might be expected to be deformed. While a large distribution in the extent of surface variability can be expected, here we have examined a reasonable sample size, containing 19 large protein-protein examples. In all cases we studied, the structures have been determined both for each of the molecules separately, and in addition, while associated in a complex. Hence, a reference state is available for a comparison.

Prediction of docked complexes when entire molecular surfaces are utilized is an extremely difficult problem. Yet, for protein-small drug binding, even if the active site is unknown it can be predicted with a certain degree of confidence,[30,31] while this is not the case in prediction of protein-protein associations. Given the magnitude of the problem on one hand, and the need for approaches to tackle it on the other, it is important to see which criteria can be utilized. Here we have shown that despite the need for further improvement in the ranking, shape complementarity can be used for "real life" unbound cases.

ACKNOWLEDGMENTS

We thank Dr. J.V. Maizel for discussions and for encouragement. We thank Dr. B. Honig for encouragement and for making programs available to us prior to their publication. The research of R. Nussinov has been sponsored by the National Cancer Institute, DHHS, under Contract No. 1-CO-74102 with SAIC. The research of R. Nussinov and H. Wolfson in Israel has been supported in part by: grant number 95–00208 from the BSF; a grant from the Israel Science Foundation, Israel, administered by the Israel Academy of Sciences; the Magnet fund, administered by the Israel Ministry of Industry; and a basic science grant administered by Tel Aviv University and the Adams Brain Center grant. The research of H. Wolfson is partially supported by the Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University. D.P. acknowledges support from grants numbered 5-T32-GM08281 and 2ROI-GM41371 from NIH and grant number DBI-9601463 from NSF.

REFERENCES

1. Bernstein F, Koetzle T, Williams G, et al. The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
2. Goodsell D, Olson A. Automated docking of substrates to proteins by simulated annealing. Proteins 1990;8:195–202.
3. Bacon D, Moult J. Docking by least-square fitting of molecular surface patterns. J Mol Biol 1992;225:849–858.
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. Cherfils J, Duquerroy S, Janin J. Protein-protein recognition analyzed by docking simulations. Proteins 1991;11:271–280.
7. 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.
8. Jiang F, Kim S. Soft docking: matching of molecular surface cubes. J Mol Biol 1991;219:79–102.
9. Shoichet B, Kuntz I. Protein docking and complementarity. J Mol Biol 1991;221:327–346.
10. Wang H. Grid-search molecular accessible algorithm for solving the protein docking problem. J Comp Chem 1991;12:746–750.
11. Cherfils J, Janin J. Protein docking algorithms: simulating molecular recognition. Curr Opin Struct Biol 1993;3:265–269.
12. Norel R, Fischer D, Wolfson H, Nussinov R. Molecular surface recognition by a computer vision based technique. Protein Engineering 1994;7:39–46.
13. Norel R, Lin SL, Wolfson H, Nussinov R. Shape complementarity at protein-protein interfaces. Biopolymers 1994;34:933–940.
14. Helmer-Citterich M, Tramonato A. Puzzle: a new method for automated protein docking based on surface shape complementarity. J Mol Biol 1994;235:1021–1031.
15. Fischer D, Lin SL, Wolfson H, Nussinov R. A geometry-based suite of molecular docking processes. J Mol Biol 248:459–477.
16. Norel R, Lin SL, Wolfson H, Nussinov R. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse points in docking. J Mol Biol 1995;252:263–273.
17. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol 1996;6:402–406.
18. Walqvist A, Covell D. Docking enzyme-inhibitor complexes using a preference-based free-energy surface. Proteins 1996;25:403–419.
19. Meyer M, Wilson P, Schomburg D. Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. J Mol Biol 1996;264:199–210.
20. Xu D, Lin S, Nussinov R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. J Mol Biol 1997;265:68–84.
21. Kasinos N, Lilley G, Subbarao N, Haneef I. A robust and efficient automated docking algorithm for molecular recognition. Protein Engineering 1992;5(1):69–75.
22. Gabb H-A, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
23. Jackson R, Gabb H, Sternberg M. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
24. Connolly M. Analytical molecular surface calculation. J Appl Cryst 1983;16:548–558.
25. Connolly M. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
26. Besl PJ, McKay ND. A method for registration of 3-D shapes. IEEE Trans on Pattern Analysis and Machine Intelligence 1992; 14(2):239–256.
27. Stickle DF, Presta LG, Dill KA, Rose GD. Hydrogen bonding in globular proteins. J Mol Biol 1992;226:1143–1159.
28. Sridhiran S, Nicholls A, Honig B. A new vertex algorithm to calculate solvent-accessible surface areas. Biophys J 1992;61: A174.
29. MacKerell AD Jr, Bashford D, Bellot M, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
30. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
31. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Prot Sci 1996;5:2438–2452.