PROTEINS: Structure, Function, and Genetics 47:409–443 (2002)

INVITED REVIEW

Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions

Inbal Halperin,[1] Buyong Ma,[2] Haim Wolfson,[3] and Ruth Nussinov[1,4†]
[1]*Sackler Institute of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*
[2]*Laboratory of Experimental and Computational Biology, NCI-Frederick, Frederick, Maryland*
[3]*School of Computer Science, Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[4]*Laboratory of Experimental and Computational Biology, Intramural Research Support Program, SAIC, NCI-Frederick, Frederick, Maryland*

ABSTRACT **The docking field has come of age. The time is ripe to present the principles of docking, reviewing the current state of the field. Two reasons are largely responsible for the maturity of the computational docking area. First, the early optimism that the very presence of the "correct" native conformation within the list of predicted docked conformations signals a near solution to the docking problem, has been replaced by the stark realization of the extreme difficulty of the next scoring/ranking step. Second, in the last couple of years more realistic approaches to handling molecular flexibility in docking schemes have emerged. As in folding, these derive from concepts abstracted from statistical mechanics, namely, populations. Docking and folding are interrelated. From the purely physical standpoint, binding and folding are analogous processes, with similar underlying principles. Computationally, the tools developed for docking will be tremendously useful for folding. For large, multidomain proteins, domain docking is probably the only rational way, mimicking the hierarchical nature of protein folding. The complexity of the problem is huge. Here we divide the computational docking problem into its two separate components. As in folding, solving the docking problem involves efficient search (and matching) algorithms, which cover the relevant conformational space, and selective scoring functions, which are both efficient and effectively discriminate between native and non-native solutions. It is universally recognized that docking of drugs is immensely important. However, protein–protein docking is equally so, relating to recognition, cellular pathways, and macromolecular assemblies. Proteins function when they are bound to other molecules. Consequently, we present the review from both the computational and the biological points of view. Although large, it covers only partially the extensive body of literature, relating to small (drug) and to large protein–protein molecule docking, to rigid and to flexible. Unfortunately, when reviewing these, a major difficulty in assessing the results is the non-uniformity in the formats in which they are presented in the literature. Consequently, we further propose a way to rectify it here. Proteins 2002;47:409–443.** © 2002 Wiley-Liss, Inc.*

Key words: docking; matching; scoring; protein-ligand recognition; flexible docking

INTRODUCTION

In this post-genomic era, research increasingly focuses on proteomics. Experimental and computational efforts are devoted to large-scale generation and analysis of information derived from 3D structures and dynamics of proteins, with the goal of scientific and commercial breakthrough in drug discovery.[1,2] Computational generation of protein structures via modeling by homology and threading, and by ab initio prediction, and docking of a protein structure with potential interacting partners are two related steps in computational proteomics. While folding is largely an academic practice (at least until this millennium), docking has been heavily used in industry in rational drug design. The principles of docking, and the progress that has been made during the last decade have been described.[2–8] Here we attempt to look back on what has been achieved and to suggest what might be tried in the next steps. Due to the enormous size of the literature, unfortunately, we are unable to cover all important contributions in the field.

Grant sponsor: Ministry of Science, Israel; Grant sponsor. "Center of Excellence in Geometric Computing and Its Applications"; Grant sponsor: Hermann Minkowski-Minerva Center for Geometry at Tel Aviv University; Grant sponsor: National Cancer Institute, NIH, Grant number: NO1-CO-12400.

†Correspondence to: R. Nussinov, NCI-Frederick Bldg 469, rm 151, Frederick, MD 21702. E-mail: ruthn@ncifcrf.gov.

Received 6 November 2001; Accepted 18 January 2002

Docking is a term used for computational schemes that attempt to find the "best" matching between two molecules: a receptor and a ligand. The *molecular docking problem* can be defined as follows: Given the atomic coordinates of two molecules, predict their "correct" bound association. In its most general form, no additional data are provided. In practice, however, additional biochemical information may be given, in particular knowledge of the binding sites. Clearly, this considerably facilitates the docking problem. Nevertheless, it should be borne in mind that there are additional potential binding sites on the protein surface. While it is assumed that the primary (known) site would be the one to participate in the bound conformation, there is no guarantee that this will be the case.

The simpler problem in docking is referred to as "bound" docking. It relates to computational schemes that attempt to reconstruct a complex using the bound structures of the receptor and the ligand. A "bound" structure is extracted from a structure of more than one molecule, typically a co-crystal of the receptor and the ligand. The goal is, however, the more difficult predictive docking, also referred to as the "unbound" docking. The unbound problem relates to computational schemes that attempt to reconstruct a complex using the unbound structures of the receptor and the ligand. An unbound structure may be a native structure, a pseudo-native structure, or a modeled structure. In this terminology, a native structure is the structure of a molecule when it is free in solution, in its uncomplexed state. A pseudo-native structure is the structure of a molecule when complexed with a molecule different from the one used for the docking. For example, a native structure exists for receptor 1 but not for ligand 1. Ligand 1 was co-crystallized with receptor 2. The structure of ligand 1 extracted from the complex with receptor 2 is a pseudo-native structure. The use of modeled structures is an even more challenging task.[9]

One of the first practical suggestions for docking came from Crick who suggested that complementarity in helical coiled coils could be modeled as knobs fitting into holes.[10] Nevertheless, only in the mid-1980s did the docking field begin to flourish. The first computational program developed for surface representation described it as a set of dots spread on the van der Waals surface.[11] While this method is currently being used for a variety of purposes, it is not the method of choice for docking. In this method, the surface is described away from the actual van der Waals surface, since the radius of a water sphere is added to the atomic van der Waals radii. Consequently, the molecular surface that is obtained smooths crevices into which ligand/receptor atoms can intrude. A method for analytically calculating a smooth three-dimensional contour about a molecule that is more suitable for such a purpose was developed by Connolly.[12,13] This program was crucial to the development of docking algorithms. The deposition of a large number of proteins in the Protein Data Bank (PDB[14]) was equally important. The early algorithms were based mainly on geometric criteria[15–19] although a few energy-based algorithms were also developed.[20–23] The first pioneering and widely used docking program was DOCK, conceived by Kuntz and his colleagues.[15] This program, and its attractive binding site description by intersecting spheres, has inspired the computational docking field. Goodford's GRID has also been integrated into many algorithms.[22]

There are three key ingredients in the docking: (1) representation of the system, (2) conformational space search, and (3), ranking of potential solutions. Docking essentially simulates the interaction of the protein surface. Therefore, the first question is how to define a protein surface. The surface can be described by mathematical models, such as for example by geometrical shape descriptors or by a grid; Alternatively, it can involve treatment (static or dynamic) of the protein frame, such as, for example, rigid vs. flexible.

Docking involves two separate molecules. It initiates from folded protein chains and ligand conformations. In contrast, protein folding initiates from some non-native protein conformations. Hence, docking is often viewed as distinct from folding. Yet, while currently computational prediction of protein structures largely addresses relatively small, single domain proteins, for large multidomain proteins, one faces the problem of domain docking. Such an approach is consistent with experiment. Experimentally, complementary fragments provide a system for studying protein folding,[24–27] consistent with intermolecular binding resembling intramolecular folding events.[28–30] Intramolecular domain docking appears simpler owing to chain linkage. However, here one needs to confront multipart docking, with consequently a large increase in the number of possible arrangements. This problem is reminiscent of docking multicomponent, supramolecular assemblies. Additionally, protein "domains" are not necessarily stable, and they may have low population times. Some domains fold on a second domain template. This suggests that, on average, they may be more flexible than entire protein chains.

Just like in protein folding, solving the docking problem also involves two components: an efficient search procedure and a good scoring function. The two critical elements in a search procedure are speed and effectiveness in covering the relevant conformational space. On the other hand, the scoring function should be fast enough to allow its application to a large number of potential solutions and, in principle, effectively discriminate between native and non-native docked conformations. The scoring function should include and appropriately weigh all the energetic ingredients. Hence, as in folding, the performance of a particular docking program should not be viewed as representing one complete piece. To solve the docking problem, ideally, the best matching algorithms and scoring schemes should be combined. Similar considerations and division have recently been discussed.[31–33]

The three aspects of the docking are mutually interrelated: The choice of the system (surface) representation decides the types of conformational search algorithms, and the ways to rank potential solutions. Below, we review the principles of the representation, available search algo-

rithms, and scoring schemes. Based on these, we highlight some potential promising approaches.

REPRESENTATION OF THE SYSTEM

Mathematical Models of Surface Representation

The basic description of the protein (or ligand) surface is the atomic representation of exposed residues. However, such a representation is usually used only when the ranking is based on real potential energy functions (for example, CHARMM's). One example is the docking program DARWIN, which communicates with CHARMM to calculate the energy.[34] In the MM/Grid docking approach, the atomic details of the ligand and the receptor binding site are simulated explicitly, while the other bulk portions of system are represented as grids.[35]

More often, the surface is represented by its geometric features. Connolly laid the foundation for protein surface analysis. The Connolly surface consists of the part of the van der Waals surface of the atoms that is accessible to the probe sphere (contact surface) connected by a network of convex, concave, and saddle shape surfaces that smooths over the crevices and pits between the atoms. Based on the Connolly analysis,[12,13] the surface may be described by sparse critical points,[36,37] defined as the projection of the gravity center of a Connolly face. In the docking program ESCHER, the solvent-accessible surface from the Connolly analysis is cut into parallel slices 1.5 Å thick, with each slice transformed into a polygon to be used in a rigid surface matching.[38]

To align the surfaces of two molecules in a complementary manner, we need to compute a rigid transformation that superimposes the surfaces without allowing one molecule to deeply penetrate, or overlap the other. To obtain hypotheses for such transformations, it suffices to align a triplet of ordered non-collinear points (congruent triangles) from both molecules. However, it may happen that there are no three independent matching point-pairs between the receptor and the ligand. For docking, the points are those describing the molecular surfaces. These are computed to accurately represent the maxima (*holes*) and minima (*knobs*) of the shape function.[39–41] Sparseness is critical, since the complexity of the algorithm depends on the number of points. A surface normal, associated with each point, is also computed. Below, these points are dubbed *critical points*. The docking strategy reduces to matching only pairs of critical points with the additional geometric information of their surface normals. In order to compute a candidate rigid transformation, we need to detect a pair of critical points in both molecules that share the same internal distance, and, if superimposed, have opposing surface normals. This reduces the number of potential docked configurations, and concomitantly reduces the run-time complexity of the program.

In practice, Connolly's MS-DOT program[12,13] yields discrete points along with three types of surface faces representing the molecular shape. For each face an interest point and a normal are computed. The interest point is a cap, belt, or pit for convex, toroidal, and concave faces, respectively.[36,37] Figure 1 illustrates the interface of an artificial complex and the surface match.[37] The caps (red) of the top moiety (lemon) and the pits (dark blue) of the bottom moiety (brown) are coupled in quality pairing. Alternatively, "critical points" and their associated surface normals, based on solid angles, can be computed.[42,43] Given the interest points, a 3D rotation and translation of the ligand is sought, such that a large portion of its surface "cap" interest points are in the vicinity of corresponding receptor surface "pit" interest points, with normals that are (almost) antiparallel. Since the tips of the surface normals provide additional *critical point-associated* points, the complexity of the docking using the Geometric Hashing algorithm reduces from $O(n^4)$ to $O(n^3)$, where n is the number of interest points. In practice it is still better, since only critical points pairs within a certain distance are taken.[42–44] This yields very fast execution times, on the order of minutes for even large protein–protein docking. We further note that, in principle, the Geometric Hashing can handle any type of molecular surface representation. Table I lists the matching times (in minutes, on a 586 PC clone, running at 133 MHz) for the Geometric Hashing-based rigid-body programs, for the bound, complexed (Table Ia) and unbound (Table Ib) protein–protein docking. The values are taken from Norel et al.[40]

As mentioned ealier, surface representation relates to conformational search and ranking of potential solutions. Physicochemical features of the protein surface are added into the purely geometrical description. For example, colored negative images[45] have polar, nonpolar, and a changeable portion. The negative image is a mold of a putative ligand and is generated in two steps: (1) all possible spheres within the binding interface are constructed, and (2) the spheres are reduced to a relatively small, representative set.[45] Protein surface may also be fitted with spherical harmonic functions to include electrostatics.[46] GRASP uses the Poisson-Boltzman equation to map the receptor electrostatic potential.[47]

Rigid to Flexible: Types of Conformational Changes Observed Between the "Bound" and "Unbound"

Predictive docking is far more complex than bound docking. The additional complexity derives from conformational changes that take place between the bound and unbound structures. There are three types of conformational changes. The first involves small-scale, fast motions as observed, for example, in ensembles of NMR conformers. The second derives from large-scale, slow domain motions. The third is the outcome of protein "disorder." In such a case, owing to larger protein flexibility, no coordinates are obtained in either the X-ray or NMR structures. Here the population time of the native conformer is too low to enable experimental detection.[48] Typically, in such cases the unbound molecule is either locally or globally disordered. However, binding stabilizes the bound conformer, shifting the equilbrium in its direction. In such cases, the native state has a small hydrophobic core, or the molecule (or its disordered domain) contains uncompensated buried charges.

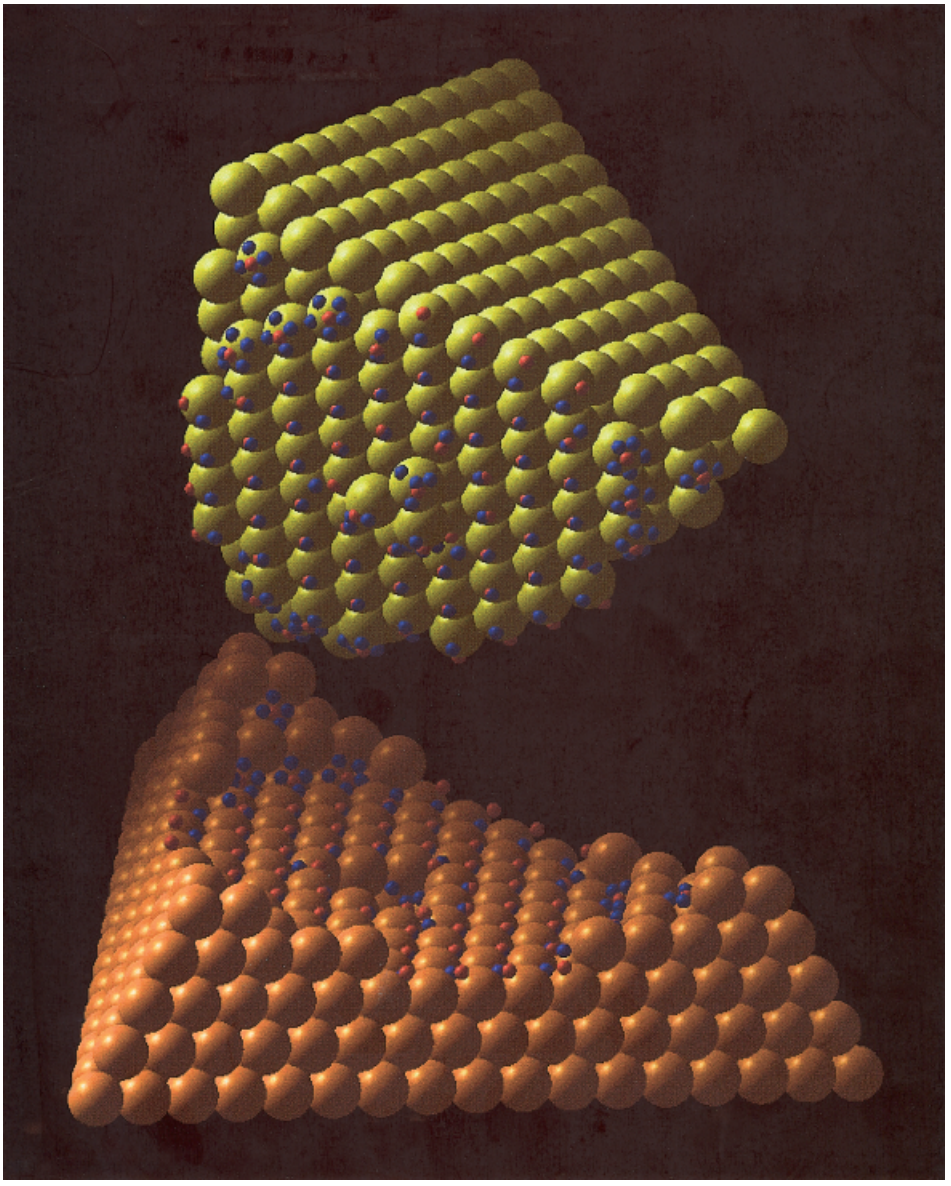

Fig. 1. The interface of an artificial complex and the surface match.[37] The caps (red) of the top moiety (lemon) and the pits (dark blue) of the bottom moiety (brown) are coupled in quality pairing. (Reprinted from J Mol Graphics, Vol 14, S.L. Lin, R. Nussinov, Molecular recognition via face center representation of a molecular surface, pp 78–90, © 1996, with permission from Elsevier Science)

With respect to small-scale motions, Betts and Sternberg examined 12 cases of multiple X-ray structures.[49] The rmsd (root mean squared deviation) of the C_α-atoms ranged between 0.1 and 0.5 Å, whereas side-chain atoms ranged between 0.5 and 1.6 Å. These data are probably an underestimation of solution flexibility movements, since in the analysis residues with poor electron density or B factor greater than or equal to 50 A^2 were excluded from the calculations.

Movements of the interface may be greater than in other exposed parts of the structure. This is consistent with the finding that proteins frequently display regions of instability around the binding sites. Freire and his colleagues[50–53] have shown that binding sites are typically part rigid and part flexible. Further, conformational differences between two bound structures are likely to be less pronounced than conformational differences between bound and unbound structures. Betts and Sternberg confirmed the first prediction but neither confirmed nor refuted the second.[49] Examination of C_α and side-chain atom movements during complex formation in 39 complexes has shown that half have values no larger than the differences between different unbound structures of the same protein. However, the extent of the movements was greater in the interface than in other exposed parts of the structure. Further, there are cases where the unbound structure could not be crystallized. Interestingly, by examining residue perturbation upon binding, Baysal and Atilgan[54] have recently shown

TABLE I. Matching Times (in Minutes, on a 586 PC Clone, Running at 133 MHz) for the Geometric Hashing-Based Rigid-Body Program†

	pdb	Receptor name	No. of atoms	Ligand name	No. of atoms	CPU (min) Docking	CPU (min) Rank/cluster	RMSD (Å)
a: Protein-protein complexes								
1	1cho	Alpha-chymotrypsin 1-146 (E)	1,047	Alpha-chymotrypsin 149-245 (E)	701	1.7	1 out of 471	0.54
2	1fdl	IG*G1 fab fragment (LH)	3,306	2-lysozyme (Y)	1,000	8.6	20 out of 2,181	1.50
3	1tec	Thermitase eglin-c (E)	2,003	Leech (I)	826	2.2	1 out of 1042	1.18
4	1tgs	Trypsinogen (Z)	1,645	Pancreatic secretory trypsin inhibitor (I)	496	2.8	1 out of 831	1.14
5	2hfl	IG*G1 fab fragment (LH)	3,227	Lysozyme (Y)	1,000	10.4	1 out of 2,166	1.51
6	2kai	Kallikrein a (AB)	1,798	Bovine pancreatic trypsin inhibitor (I)	438	2.2	11 out of 1,227	1.17
7	2mhb	Hemoglobin α chain (A)	1,068	β chain (B)	1,133	7.2	1 out of 663	0.70
8	2ptc	Beta-trypsin (E)	1,628	Pancreatic trypsin inhibitor (I)	453	2.6	1 out of 1,027	0.59
9	2sec	Subtilisin carlsberg (E)	1,919	Genetically-engineered n-acetyl eglin-c (I)	529	1.8	1 out of 1,114	2.08
10	2sni	Subtilisin novo (E)	1,937	Chymotrypsin inhibitor (I)	512	2.2	1 out of 1,367	1.07
11	2tgp	Trypsinogen (Z)	1,628	Pancreatic trypsin inhibitor (I)	453	1.6	1 out of 828	0.59
12	3hfm	IG*G1 fab fragment (LH)	3,293	Lysozyme (Y)	1,000	10.7	1 out of 2,274	0.76
13	4cpa	Carboxypeptidase	1,536	Potato carboxypeptidase a inhibitor (I)	275	2.1	3 out of 1,310	1.02
14	4hvp	HIV-1 protease chain A	745	Chain B	745	1.4	1 out of 411	2.06
15	4sgb	Serine proteinase (E)	1,309	Potato inhibitor pci-1 (I)	379	.09	5 out of 591	1.88
16	4tpi	Trypsinogen (Z)	1,628	Pancreatic trypsin inhibitor (I)	455	2.1	1 out of 889	0.52
17	1abi	Hydrolase alpha thrombin (H)	2,039	Chain L	265	6.2	1 out of 773	0.56
18	1acb	Hydrolase alpha-chymotrypsin (E)	1,769	Eglin C (I)	522	3.8	1 out of 1,121	0.94
19	1cse	Subtilisin carisberg (E)	1,914	Eglin C (I)	522	1.7	2 out of 1,024	1.32
20	1tpa	Anhydro-trypsin (E)	1,628	Trypsin inhibitor (I)	454	2.6	1 out of 950	0.23
21	2sic	Subtilisin (E)	1,938	Subtilisin inhibitor (I)	764	3.2	1 out of 1,229	1.11
22	5hmg	Influenza virus hemagglutinin (E)	2,532	Chain F	1,417	17.7	1 out of 329	1.09
23	6tim	Triosephosphate isomerase chain A	1,883	Chain B	1,883	11.0	1 out of 351	0.50
24	8fav	Fab fragment from IGG1 chain A	1,544	Chain B	1,635	2.3	1 out of 93	1.97
25	9ldt	Lactate dehydrogenase chain A	2,565	Chain B	2,565	24.1	1 out of 67	2.52
26	9rsa	Ribonuclease chain A	951	Chain B	951	2.9	21 out of 511	1.30
b: Unbound cases								
1	1hfm-1lym(A)	IG*G1 fv fragment	1,714	Lysozyme (A)	1,001	11.8	537 out of 11,475	2.97
2	1hfm-1lym(B)	IG*G1 fv fragment	1,714	Lysozyme (B)	1,001	4.0	281 out of 10,685	2.80
3	1tgn-4pti	Trypsinogen	1,621	Trypsin inhibitor	453	3.3	53 out of 2,619	1.85
4	1tgn-5pti	Trypsinogen	1,621	Trypsin inhibitor	464	5.3	1 out of 3,453	1.22
5	1tgn-6pti	Trypsinogentrypsinogen	1,621	Trypsin inhibitor	458	3.2	2 out of 1,455	1.75
6	1tld-4pti	Beta-trypsin	1,629	Trypsin inhibitor	453	2.5	16 out of 2,659	5.22
7	1tld-5pti	Beta-trypsin	1,629	Trypsin inhibitor	464	3.6	619 out of 3,471	4.71
8	1tld-6pti	Beta-trypsin	1,629	Trypsin inhibitor	458	2.5	40 out of 1,512	2.18
9	2hfl-1lyz	IG*G1 fab fragment	3,220	Lysozyme	1,001	10.1	110 out of 10,989	1.79
10	2hfl-6lyz	IG*G1 fab fragment	3,220	Lysozyme	1,001	12.6	65 out of 10,733	1.08
11	2pka-4pti	Kallikrein a	1,799	Trypsin inhibitor	453	1.9	29 out of 3,184	3.29
12	2pka-5pti	Kallikrein a	1,799	Trypsin inhibitor	464	3.1	9 out of 4,222	1.21
13	2pka-6pti	Kallikrein a	1,799	Trypsin inhibitor	458	1.9	27 out of 1,756	1.82
14	2ptn-4pti	Trypsin	1,629	Trypsin inhibitor	453	2.8	9 out of 2,156	3.53
15	2ptn-5pti	Trypsin	1,629	Trypsin inhibitor	464	4.0	34 out of 2,880	3.11
16	2ptn-6pti	Trypsin	1,629	Trypsin inhibitor	458	5.5	56 out of 1,200	1.28
17	2sbt-2ci2	Subtilisin novo	1,934	Chymotrypsin inhibitor	521	2.8	92 out of 3,582	2.62
18	5cha(A)-2ovo	Alpha-chymotrrrypsin (A)	1,735	Ovomucoid third domain	418	1.7	11 out of 2,194	1.49
19	5cha(B)-2ovo	Alpha-chymotrrrypsin (B)	1,736	Ovomucoid third domain	418	3.1	2 out of 2,289	1.64

†**a:** The bound, complexed cases, and **b:** unbound protein–protein docking. The values are taken from Norel et al. (1999).[40] In either case, no additional biochemical information has been incorporated. The binding sites are assumed to be unknown. Entire molecular surfaces are considered. The only input information are the atomic coordinates, taken from the PDB (Bernstein et al., 1997). In the scoring/ranking procedure, carried out following the matching stage, the binding sites are still assumed to be unknown. Hence, this is the most general approach. This is a rigid-body approach. Molecular flexibility is implicitly taken into account through a certain allowance (though penalized) intermolecular penetration. The first column gives the PDB file name; the second gives the receptor name, followed by the number of atoms; the fourth column lists the ligand name followed by the number of atoms. The sixth column lists the CPU times for the matching stage. As can be seen, the longest is 25 min, with most complexes under 10 mins. The next column gives the rank of the lowest (under 5 Å) RMSD solution, and the number of clusters (i.e., equivalent to the number of solutions, as all similar solutions are grouped into a cluster). The last column gives the RMSD of the top ranking solution.

that the stabilization of the binding region is accomplished at the expense of loss of stability in other parts of the structure. While the rmsd deviations between the bound and unbound structures (of chymotrypsin inhibitor 2) are small, residue fluctuations and stability differ significantly in their response to the binding. The packing density changes only at the binding loop. However, residue fluctations changed in the rest of the protein.

With respect to the second type of conformational change, domain movements in proteins have been classified into shear and hinge bending motions.[55] "Induced fit" involves such hinge-type motion. It was originally suggested in the late 1950s.[56] Since then it was illustrated in a large number of cases, e.g. in enzyme-inhibitor,[57] DNA-protein and antigen-antibody binding. Hinge-bending motions are low-energy transitions, with hinge-bent conformers populating the solution around the native state. In the presence of the ligand, the conformer that binds is the one whose conformation is most favorable. Binding leads to a population shift, propagating the binding reaction. Hinge motion might be critical for predictive docking. In antibody-antigen, the initial binding to solvent-exposed residues may promote local side-chain displacements and thereby allow the participation of other, previously buried residues. The crystal structure of β4Gal-T1 with bound UDP shows a conformational change with a large 20-Å loop movement with concomitant changes in residue burial, related to substrate binding.[58]

The third type of conformational change observed upon binding is the most drastic "disordered" to "ordered." Shoemaker et al. have recently proposed a 'fly-casting' binding mechanism.[59] In vivo, a large fraction of the proteins are in an unfolded, "disordered" state,[60] illustrating that folding and function are coupled. Shoemaker et al. have suggested that a relatively unstructured protein molecule can have a greater capture radius for a specific binding site than the folded state with its restricted conformational freedom.[59] In their calculations, the increased capture radius operates by exploiting the available folding free energy. Such a mechanism should help in avoiding metastable non-specific bound complexes, which may arise from the ruggedness of ligand–protein landscape. This binding case resembles protein folding (on a template).

The computational procedures inherent to docking are a function of the extent of flexibility that they attempt to address. These can be classified into three levels by their degree of approximation[61]: (1) Rigid body docking. Rigid body is a highly simplistic model that regards the two proteins as two rigid solid bodies. (2) Semi-flexible docking. The semi-flexible model is asymmetric; one of the molecules, usually the smaller ligand, is considered flexible, while the receptor is regarded as rigid. (3) Flexible docking. Both molecules are considered flexible, although clearly the extent of flexibility of either (or of both) is necessarily limited, or simplified. Mangoni et al. have carried out docking of a flexible ligand to a flexible receptor via molecular dynamics simulation,[62] using an explicit water model. This feat has been enabled by an enhanced method: While most of the simulation was in the standard way, the center of mass of the protein was heated to a higher temperature, giving it a higher velocity. This yielded rapid additional sampling, however, with largely relevant conformations.

In a classical sense, docking schemes are divided into rigid body and flexible algorithms. Owing to this division, docking papers (and algorithms) are frequently separated into protein–protein and protein–small molecule (or protein–drug docking). The underlying notion in such a classification is that since drugs are smaller molecules, they are likely to undergo larger conformational fluctuation. Furthermore, given their smaller size, this is computationally affordable, as compared to the large protein–protein docking. Nevertheless, for both types, docking algorithms display a spectrum of flexibility. Rigid docking handles a certain extent of surface variability by allowing some inter-molecular penetration. At the other extreme, for small molecules, depending on their size, one can allow motion on (almost) every ligand bond. Until very recently, even when allowing a large extent of flexibility for drug docking, typically the receptor has been held rigid, or with only a limited extent of flexibility, largely at the binding site. However, the last few years have seen a revolution in concepts and approaches addressed at solving the molecular docking problem. As in folding, these largely derive from considerations of populations and ensembles. Hence, by resorting to concepts derived from energy landscapes and statistical mechanics, and employing pre-generated and recombined ensembles rather than single bond stretching and bending, receptor flexibility is beginning to be handled in small molecule-docking.[63] As outlined below, similar integrated approaches can also be applied to flexible protein–protein docking. Similar concepts have recently been presented by Carlson and McCammon for drug design.[7]

The relevance of using ensembles is also indicated in recent studies showing that even apparently specific receptors bind a range of ligands of different sizes, shapes, and composition, often with higher affinities than the presumably specific ligand.[64–67] Current data suggest that they preferentially bind at the same binding site. This is not surprising. Proteins exist in a range of conformational substates, with low-energy barriers separating them. Different ligands will associate with the most favorable conformers.[30,68,69] This suggests the important role of hinge-bending, large-scale domain (or loop) motion.[58] It further suggests that binding sites may be distinguished from other sites on the protein surface by their enhanced flexibility.[52,54] Hence, analysis of the interactions between biological molecules cannot be reduced to the description of (static) molecular structures. Integrated functional approaches need to consider the binding partner and the time component of the interaction.[69,70] The function of a protein and its properties are decided not only by the static folded three-dimensional structure, but by the distribution and *redistributions* of the populations of its conformational substates under different (physical or binding) environments.[30,68,71] Such a mechanism provides multiple path-

ways and allows a single molecular surface to interact with numerous structurally distinct binding partners, accommodates mutations through shifts in the dynamic energy landscape, and as such is evolutionarily advantageous. A ligand will similarly change the environment, affecting the (preexisting) most populated state of the protein.[72] Combined, these point to the clear advantages of using ensembles in docking, and to the need to consider drug (ligand, inhibitor) diversity.[73] Furthermore, the existence of multiple binding modes has to be accounted for in predictions of binding affinity.[74]

Flexibility can be simulated in many different ways. One of the most rigorous methods is by Lamb and Jorgensen,[75] who calculate the free energy of the system. While highly reliable, producing accurate results (consistent with experiment, up to 1 kcal/mol), they are too slow for extensive ligand docking. Luty et al.[35] have simplified the simulation by employing an implicit solvent model. Most of the protein was held rigid, with only the binding site and the ligand sampled. They handled the border zone between the rigid and flexible protein layers using a small buffer region next to the rigid (crystal coordinate) part with a small harmonic potential. This allowed a relatively rapid simulation of the docking, however still with limited sampling. Rigid docking has the advantage of speed, being able to explore the entire receptor and ligand surfaces, and performing database docking. In such approaches, flexibility is handled through a "soft" belt into which atoms from the second molecule can penetrate[40–44,76,77] reducing drastically the complexity. In Vakser et al.'s study of low-resolution recognition of protein-protein complexes,[78] the rigid docking of low-resolution structure essentially attempts to simulate the average effects of the conformational flexibility.

Computational Approaches to Presentation of Protein Flexibility

Proteins can be docked as rigid bodies. Molecules, just like any other rigid body, have six degrees of freedom: three rotational and three translational. For some cases, the rigid body approximation is justified by comparing the X-ray structures of complexes with those of their unbound free components.[79] Najmanovitch et al.[80] have shown that in most cases only a few side-chains change their conformations in the active site. However, for many other known cases this assumption is not valid. For example, hirudin undergoes a large structural change upon complexation with thrombin.[81] Large localized rearrangements at the protein surface are frequent, especially for large flexible amino acid chains.[82]

Implementing full conformational flexibility into a search stage, separately docking a large number of conformers, is infeasible. A reasonable approach is to take account of ensembles of populations, generated prior to the docking, and dock the ensemble rather than single conformers. Depending on the strategy, docking an ensemble highlights the more conserved regions by, for example, assigning these larger weights, whereas lower weights maybe given to regions of space visited more rarely. Experimentally, ensembles can be assembled by collecting all crystal structures (unbound, or bound to different ligands), or using NMR conformers. There are two systems where a large number of structures have been determined. The first is the HIV-1 protease, and the second is DHFR (dihydrofolate reductase). Over 86 crystal structures have been determined for HIV-1 in PDB[14] and over 50 structures are available in the NCI HIV protease database.[83,84] Additionally, an NMR conformer ensemble[85] is present in the PDB. A large body of inhibitors exist and combinatorial libraries are also available.[86,87] For DHFR, there are three sets of NMR ensembles[88] in the PDB in addition to near 80 crystal structures with good (2.5 Å or better) resolution.

Recently, Philippopoulos and Lim[89] have compared an ensemble of *Escherichia coli* ribonuclease H1 (RNase H1) conformers derived from NMR experiments both with an ensemble obtained from molecular dynamics simulations, and with two X-ray structures. They have shown that the 15 conformers of the NMR ensemble sample more conformational space of the RNase H1 than the 1.7-ns molecular dynamics simulations. Further, multiple crystal structures may cover more space than sampled in a 1-ns MD trajectory.[90] The collection of crystal (or NMR) structures can be used in the simulations to enhance the conformational sampling. For those cases where there are no experimentally determined (NMR) ensembles, an extensive coverage of conformational space can be achieved through computational sampling. Protein ensembles can be generated by molecular dynamics by random thermodynamic sampling[91] or, for example, by genetic algorithms[92–94] or through multiple MD trajectories.[90]

Several groups have superimposed protein conformations, with the goal of structure-based drug discovery. Among the first are Knegtel et al.[95] who have used a composite grid incorporating multiple crystal and/or NMR structures of protein–ligand complexes. Protein conformations were superimposed using residues at/near the binding sites. A grid was calculated for each conformer of the set, with subsequent averaging of all grids to obtain a picture of the binding site. Using either simple or weight-averaged grids improved the accuracy of the identification of known inhibitors. Docking has been carried out using DOCK on several systems, HIV protease, ras p21, uteroglobin, and bovine retinol binding protein. However, while several protein conformations are taken into account, the combinatorial nature of explicit conformations has not been considered.[63] In a second approach, Sudbeck et al.[96] have used nine crystal structures of inhibitors complexed with the HIV reverse transcriptase, superimposing by the most stable region. The subsequent superpositioning of the inhibitors allowed them to create a composite map of the binding site. Small molecules were docked into a single receptor binding site, using conjugate gradient minimization of ligands, and of atoms of the receptor within 5 Å of the ligand. A subsequent experimental test confirmed two highly potent new inhibitors out of their solutions. In a third approach, Broughton[97] used MD simulations to generate conformations of protein–inhibitor complexes. As in the first approach, the structures were superimposed,

with the inhibitor serving as the reference. Weight-averaged single grids yielded the average grid used for the docking. The FLOG code[98] was used for the docking purpose. Most recently, in a fourth approach, Claussen et al.[63] have introduced FlexE. Here, the authors have averaged conserved coordinates, rather than using grids. Flexible side-chain orientations were retained as a set, as in a rotamer library. FlexE is able to mix rotamers taken from the superimposed structures, creating new combinations. This fast and attractive method has been tested on aldose reductase, α-monorcharin, carboanhydrase II, carboxypeptidase, DHFR, isocitrate dehydrogenase, mandelate racemase, ricin, seryl-tRNA synthase, and trypsin. Excellent results were obtained with this method. It also allows for ligand flexibility.[99] However, the ligands docked by FlexE were in the complexes used to extract the protein structures. While some loop motions are captured by this method, domain motions cannot be handled. Bouzida et al.[100] have actually acted on the premise that a protein exists in multiple conformations, with the ligand binding to the most favorable of these.[72] Their procedure involves docking of a ligand to every protein conformer, scoring each. Two different ligands were docked to ten rigid HIV-1 protease crystal structures, using MC simulated annealing minimization. A soft scoring function further implicitly addressed the enzyme surface flexibility. Their findings were interesting: Whereas the first ligand was observed to bind most favorably to one particular conformer, the second bound practically equally well four enzyme conformations. This study shows the importance of using ensembles, and the preferential ligand binding. Unfortunately, such an approach is too computationally expensive to allow large-scale docking experiments, particularly if the number of conformers is large.

Movements of domains with respect to each other are also essential in simulating protein flexibility. An efficient way of docking, allowing domain motions, has been presented by Sandak et al.,[101–105] who have docked a ligand onto a receptor surface, allowing hinge-bending movements of domains, subdomains, or substructural parts. All angular rotations are allowed, while still avoiding a conformational space search. By picking a hinge point, the molecule is divided into two parts. However, rather than dock each of the molecular parts separately, with subsequent reconstruction of the consistently docked molecules, all parts are docked simultaneously. Furthermore, the position of the hinge is utilized from the start. Like pliers closing on a screw, the receptor automatically closes on its ligand. Movements are allowed either in the ligand or in the larger receptor, hence mimicking the so-called "induced" molecular fit. In principle, more than one hinge can be allowed in the docking. Hinge-bending movements are frequently observed when molecules associate. The movements can involve entire domains, subdomains, loops, helices, or occur between any groups of atoms connected by flexible joints. Sandak et al. have implemented the hinges at points and at bonds. By allowing full 3D rotations around a point, rather than around a bond, several rotations about (consecutive or nearby) bonds are implicitly taken into account. Nevertheless, if required, the complete rotation about a point can be restricted to bond rotation. Several simultaneous hinges may also be allowed. By allowing several hinge motions to occur at the same time, we simulate the cumulative effect of larger conformational flexibility or of multiple mutations, each introducing a limited motion. Additionally, several hinge points are examined in nearby (non-adjacent) residues, still obtaining similar docked configurations. In practice, so far the algorithm has been implemented to enable two simultaneous hinges. The algorithm was applied to a number of *bound* and *unbound* molecular configurations, achieving fast matching (recognition) times of their surfaces, for both rigid and flexible docking. As in the rigid-body case, the atomic coordinates are taken from the PDB. The location of the hinge has been determined by a comparison of similar structures in different, i.e., "open" and "closed" conformations. Nevertheless, in this efficient robotics-based method, while the domains are allowed to move with respect to each other, the domain (or part) itself is held rigid.

A more modest simulation of protein flexibility involves side-chain flexibility.[106,107] Determination of side-chain conformations can be modeled according to a rotamer library.[108] Rotamers are usually defined as low-energy side-chain conformations. Certain rotameric states will be higher in energy than others because of steric interactions that force the side-chain to twist out of the way of neighboring atoms, inflicting a high dihedral energy on the residue. These interactions can be "backbone-independent," that is, not depending on the conformation of the local backbone of the residue. Alternatively, they can be "backbone-dependent," depending on the local backbone conformation as determined by the backbone dihedrals ϕ and ψ.[109] Leach[106] used single conformations for Gly, Ala, and Pro. Fourteen side-chains were given 3–10 rotamers, and Met, Lys, and Arg sampled 21, 51, and 55 rotamers. The advantage of utilization of rotamers is the relative speed in the sampling, and avoiding minimization barriers. Leach and Lemon[107] have recently used a dead-end elimination and the A* algorithm to explore the side-chain conformational space. The advantage of such an algorithm is that it allows diverse sampling of conformations not present in the protein dataset that is used. Schaffer and Verkhiver[110] use large rotamer libraries. However, their procedure of generating likely side-chain conformations and minimizing docked structures using the receptor binding site side-chains while allowing also local backbone atoms optimization is fast. Apostolakis et al.[111] have docked ligands in a flexible binding site using conjugate gradient minimization, with the non-bonded interactions gradually switched on during the process. These were followed by short MC minimization runs on the more promising candidates. The nice feature in such a procedure is that initially there may be some receptor–ligand overlaps. These are handled as the van der Waals terms and are gradually switched on.

Small-scale simulation of protein flexibility can be performed by a surface belt of nonpenalized penetration

area.[38,77,112,113] Zhao and colleagues[114] have analyzed a dataset of paired protein structures, which have multiple high quality uncomplexed structures available in the PDB. They have quantified side-chain flexibility, obtaining a set of residue- and environment-specific confidence levels, which describe a range of motions. These can be used to evaluate models, and as a metric in docking. When protein flexibility is modeled, inclusion of intra-molecular overlaps in a scoring function might be beneficial in a search of self-consistent solutions.[101–105,115]

Computational Approaches to Presentation of Ligand Flexibility

Simulating ligand flexibility is also a computationally expensive process. Lomber and Shoichet[116] note that the number of possible ligand conformations rises in proportion to the power of the number of rotatable bonds. They calculate that for an organic molecule with ten rotatable bonds, the number of possible conformations is 59,049, if only three minima are considered per bond. However, allowing six minima per bond yields 3.48×10^9 conformations. To address the problem, one approach involves Monte Carlo simulation and simulated annealing to sample the ligand flexibility.[23,117–120] In Goodsell et al.,[23] the binding site is rigid, whereas in Stoddard and Koshland[119] some binding site flexibility is allowed.

To reduce the computational time, alternate methods have been devised, which divide the ligand into fragments, and incrementally build it in the receptor binding site. In other approaches, the fragments are docked separately, and consistent fragment-docked solutions are joined.[19,121–128] In general, the drawback of incremental growth is that one frequently needs to resort to an exhaustive grid-search of each added fragment. On the other hand, docking each part separately and searching for consistent solutions overlooks a piece of information that is available a priori, namely, the fact that the two fragments are linked, and that their linking point is known. A more efficient incremental method is FlexX.[99] Here the authors model conformational flexibility of ligands by using a set of minimized geometries[129] derived for the Cambridge Database. Up to 12 low-energy torsion angles were assigned to each bond. FlexX then automatically forms a set of alternative fragments by selecting single components or their combinations. To dock each fragment, triples or, if needed, pairs of interaction centers are used. Owing to geometric ambiguity, multiple placements are generated by rotations around the axis defined by the interaction points and centers. Domain movements can be performed by hinge bending.[101–105] This robotics-based approach is very fast, and uses all the information simultaneously in the docking process. There is no need to rotate bonds. However, there are two drawbacks in this method: The hinges need to be pre-picked, and each part should be informative enough to be recognized.

Genetic algorithms[130–132] have been used to generate conformers. Genetic algorithms work by representing the ligand conformations in a modular way, using operations similar to mutations and crosses. The quality of the results is a function of the starting genes, the number of evolutionary events, i.e., the mutations and crosses, and the scoring function to pick the more favorable conformers. Using GOLD, Jones et al. have successfully tested their method on over 100 complexes extracted from the PDB.[132] This list was selected on the basis of pharmacological interest. However, genetic algorithms are too slow for extensive flexible drug docking. Morris et al.[133] have developed a new, Lamarckian-based genetic algorithm, and incorporated it into AutoDock. This algorithm predicts the docked conformations of flexible ligands in rigid macromolecular targets. In the Lamarckian model of genetics, environmental adaptations of an individual phenotype are reverse transcribed into its genotype, becoming a heritable trait. They have further incorporated into it a new scoring function that estimates the free energy change upon binding. The energy function was calibrated on a set of 30 known complexes, with experimentally determined binding constants. Comparison of the Lamarckian-based genetic algorithm with a traditional genetic algorithm and a Monte Carlo simulated annealing in seven protein–ligand systems, revealed that both genetic algorithms handled the ligands with more degrees of freedom than the simulated annealing as used in earlier versions of AutoDock. However, the Lamarckian genetic algorithm was observed to be the most efficient and reliable of the three methods.

Lomber and Shoichet[116] have used SYBYL (Tripos Associates, Inc., St. Louis, MO) to generate a library of drug conformations. In their method, the conformers were docked together as an ensemble into a receptor binding site. The ligand was broken into constituent fragments. Fragments that are conformationally similar are considered rigid, and docked only once into the receptor. Docking of the flexible fragments follow. DOCK was used for this docking purpose.

Schnecke et al.[134,135] have developed Spectitope and Slide. Their approach is also based on matching fragments, using a fast, multilevel hashing algorithm. Anchor fragments, derived from ligand conformations in their dataset, are matched to template points describing the binding site. The ligand (and receptor) flexibility is introduced following the matching, using mean field theory to select rotations and resolve collisions.

OVERVIEW OF SEARCH PROCEDURES AND MATCHING ALGORITHMS

Computational Approaches to the Search Stage

Docking is computationally difficult because there are many ways of putting two molecules together (three translational and three rotational degrees of freedom). The number of possibilities grows exponentially with the size of the components.[136] Combining all patches of the surface of one protein molecule with all patches of a second molecule takes on the order of 10^7 trials.[82] The computational problem is even more profound when considering protein flexibility and the increasing demand to screen large databases (of protein structures and of potential drugs).

As in protein folding, the binding energy landscape has a rugged funnel shape.[72,137,138] Hence, docking and folding programs employ similar searching algorithms to locate the most stable state (global minimum) in the energy landscape. The search for candidate solutions in a docking problem is addressed in two essentially different approaches: (1) a full solution space search in contrast to (2) a gradual guided progression through solution space. The first scans the entire solution space in a predefined systematic manner. In contrast, the second either scans only part of the solution space in a partially random and partially criteria-guided manner, or generates fitting solutions. The second approach consists mainly of Monte Carlo (MC), simulated annealing, molecular dynamics (MD), and evolutionary algorithms such as genetic algorithms (GA) and Tabu search.

The docking program DOT[139] provides a complete search of all orientations between two rigid molecules by systematically rotating and translating one molecule about the other. DOT predicted successfully the electron transfer complex between bovine cytochrome c oxidase and horse cytochrome c. Energies for over 36 billion configurations were calculated, providing a free-energy landscape showing the guidance of the positively charged cytochrome c to the negative region on the cytochrome c oxidase surface formed by subunit II.[140]

Vieth et al.[31] have carried out a study addressed at assessing the search strategies of three methods, molecular dynamic, Monte Carlo, and genetic algorithms, on five representative complexes. The three algorithms use a modified CHARMM-based energy function. In all cases, the receptors were held rigid, while the ligands were flexible. Two types of search space were used: an 11-Å radius sphere and a 2.5-Å, both centered on the active site. Vieth et al.[31] observed MD to be the most efficient with the larger sphere and the genetic algorithm with the small sphere. They also found that on average molecular dynamics provided structures lower in energy and closer to the crystallographic complexes. They further demonstrated that genetic algorithms require the longest time for a single energy calculation (the result of the nonbonded interaction calculations), and hence are the least efficient.[31]

Unlike the traditional optimization methods, the Geometric Hashing–based matching algorithm is one of the unique fundamental methods used in docking that can deal with such a complexity efficiently. The algorithm was originally suggested for object recognition in computer vision.[141] Combined with an adequate molecular surface representation, it yields a state-of-the-art tool-kit for docking.[36,37,39–44] Fourier correlation techniques are also widely used. Usually, three-dimentional grid-based fast-Fourier transform (FFT) docking correlation methods are slow; however, more efficient searching may be obtained using spherical polar Fourier correlations.[46] Consider, for example, the so-called *rigid body* docking (which, despite its name, still allows surface variability). Currently, there is a vast variability in the search performance of the available algorithms. Some matching programs, such as those represented by the FFT,[142] take on the order of days of CPU.[78,143–148] This is owing to the exhaustive conformational space search procedure inherent to the FFT procedure. Others, such as the Geometric Hashing,[39–44] focus only on the *relevant* conformational space, and reach similar results in minutes. This is done by considering only actual reference frames derived from the model (say, the receptor) protein, and avoiding searching space where in any case an adequate solution would not be found.

A motion-planning approach to flexible ligand docking also avoids a full solution space search.[149] This method uses several predicted intermediate configurations of the ligand. It obtains a distribution of energetically favorable paths to the binding site. For each path, a "difficulty weight" represents the energy barriers that the ligand encounters along the path.

The search stage of molecular docking of ligands to proteins can be divided into two independent procedures, depending on whether the binding site is known.[150] In many cases, the binding site can in principle be predefined using experimental data, e.g., site-directed mutagenesis, chemical cross-linking, protein family comparisons, or through computational predictions of the binding sites.[19,121,151–156] Some docking efforts concentrate on only one of these: In the first, the site is assumed known, typically in methods allowing flexibility, but in some rigid body–based methods too.[112,115,142] In the second procedure, it is unknown.[36,43,44,78,113,142,144–146]

In all methods, the search part creates a population of solutions, each assessed by some energy function, whether coarse (e.g., as in rigid-body, or hinge-bending algorithms; Sandak et al.[104]) or more rigorous (as in MC, MD, simulated annealing, or in steepest descent optimization). The algorithms differ from each other in the computational methods used (genetic algorithms, graph theory, molecular dynamics, Monte Carlo, etc.), as well as in the physicochemical criteria composing the scoring function (geometric complementarity, nonpolar buried surface area, electrostatic interactions, hydrogen bonds, unsatisfied buried charges, pairwise amino acid contacts, solvation energy, similarity to a known ligand, etc.).

We further note that, in principle, any algorithm that is applicable to structural comparisons, i.e., with the input being points in space, without a predefined order to these, can be applied to rigid or hinge-bending domain motion docking.[44,157] The converse also holds: any geometry-based docking algorithm can, in principle, be applied to amino acid sequence order-independent structural comparison.

Docking is one of the most creative research areas in computational chemistry/biology. It is hard to enumerate all the algorithms that have been used. Appart from efforts to speed up the search stage and include protein flexibility, one of the trends in the area is to combine the computational tools or to include experimental (e.g., NMR) information in the docking calculation. Combining docking with NMR benefits both docking and the NMR assignment, particularly for large protein–protein complexes in the current structural genomic initiatives.[158,159] Instead of a

simple search and ranking protocol, a two-stage docking first carries out a rapid search to generate a large number of plausible candidates. Candidate conformations are next subjected to an MD simulation/minimization using classical force field and are ranked also using the classical force field.[160,161] This approach offers a compromise between search speed and accuracy compared with the more rigorous free energy evaluation.

Docking of Small Molecules

Although the physical principles that govern protein–protein association are similar to those responsible for other ligands, docking algorithms designed for protein–protein association differ somewhat from those for small ligands used in drug design.[82] It has been well documented that a single conformation drug docking is likely to fail.[162] On the other hand, a single conformation protein–protein docking is widely carried out and has a markedly higher chance of success.[40,41,113] Second, while electrostatics plays an important role in the energetics of protein–protein associations, it is nevertheless more important for the smaller drugs. In protein–small ligand docking, the complementary contact surfaces between the ligand and the receptor are substantially smaller and less discriminating than in the case of protein–protein docking. Small ligands are often highly flexible, with a broad range of populations, adapting their surface to optimally complement the receptor pocket. Therefore, one of the main features of drug docking schemes is enhanced ligand flexibility.[163–165] Small ligand-oriented docking schemes are either semiflexible[122,166–169] or flexible.[106,170] Single water molecules in the interface may be particularly important in small ligand docking, mediating hydrogen bonds.[171]

The goals of protein–drug (small molecule) docking are twofold: The most common goal relates to drug design. However, protein–small molecule docking includes also cofactors. In such cases, the goal may relate to prediction of the binding site, or the binding orientation, pointing to residues crucial in such interactions. Approaches to drug docking fall into two main categories. Typically, in both, the binding sites are assumed known. In the first approach, one starts from a library of fragments, combinatorially selecting a fragment at a time. These are docked into the binding site, growing the molecule while testing all permissible degrees of freedom, minimizing the energies, and searching for the most favorable combinations. Such algorithms are usually breadth-search-first approaches (reviewed in Bohacek et al.[172]). In the second type, one docks entire molecules. Here, rather than step-by-step ligating groups of atoms, one carries out a (frequently large-scale) database matching and scoring. To carry out such a task, a readily available database is employed, such as ACD (Available Chemical Directory), the NCI database of drugs, or the Comprehensive Medicinal Chemistry (CMC) database. One may then either dock these directly, one by one, or, alternatively, dock compounds containing given *pharmacophores*.

Pharmacophore generation and docking

To identify a pharmacophore, the starting point is a collection of small molecule ligands that were experimentally observed to interact with the given receptor. The underlying assumption is that such an interaction is obtained either via a set of geometric features common to the data set of ligands, or alternatively, they may be chemical attributes, translated into geometrical features (e.g., hydrogen bonds, coordinates of hydrophobic atoms, points representing charged groups, etc.). These features compose the pharmacophore, which is recognized by the receptor. Once the pharmacophore is identified, other ligands with a potential for similar functionality can be found by screening for molecules containing a similar constellation. A multiple structural alignment algorithm is the natural method for identifying pharmacophores. Additionally, through a multiple structural alignment of receptor proteins that interact with a given drug molecule (or any ligand), one can detect the functional site of these proteins. Figure 2 illustrates one of the algorithms for multiple structure allignment. Further details are given in Leibowitz et al.[173] Below, we sketch representative algorithms developed for this purpose.

A graph representation of the molecule enables applying graph theory methods to solve the pair-wise alignment. A molecule is described as a graph. The nodes and edges have labels, corresponding to atom types and interatom distances, respectively. Given the graph representation of two molecules, an alignment between them corresponds to a common subgraph, appearing in both graphs. One way to find common subgraphs is by creating their correspondence graph. In the new graph, a node is formed from each pair of nodes, each contributed by each of the original two graphs. An edge is formed between two nodes in the correspondence graph only if the edges in the original graphs that correspond to the nodes have similar labels. A clique in the correspondence graph is equivalent to a common subgraph of the size of the clique.[174,175]

The earliest yet most rigorous *multiple* approaches have been suggested by Brint and Willet,[174] based on an algorithm by Crandell and Smith.[176] The approach is iterative. At each stage, all common substructures of size i are found and stored. Next, single atoms from the molecule are repeatedly added to each of these. If an enlarged substructure is not found in all the other molecules' lists, it is not considered further. Surviving substructures form the common substructures of size $i + 1$. To verify that they are common, an efficient comparison of the substructures is carried out by generating a canonical name for each substructure. The name consists of the sorted list of interdistances of the substructure. The comparison of substructures is then achieved by simply comparing their canonical names.

This truly *multiple* comparison approach guarantees the optimal solution. At all stages, it retains substructures that are common to all molecules rather than rely on common pairwise substructures. Interestingly, this algorithm is only lightly dependent on M, the number of compared molecules, but is exponentially dependent on

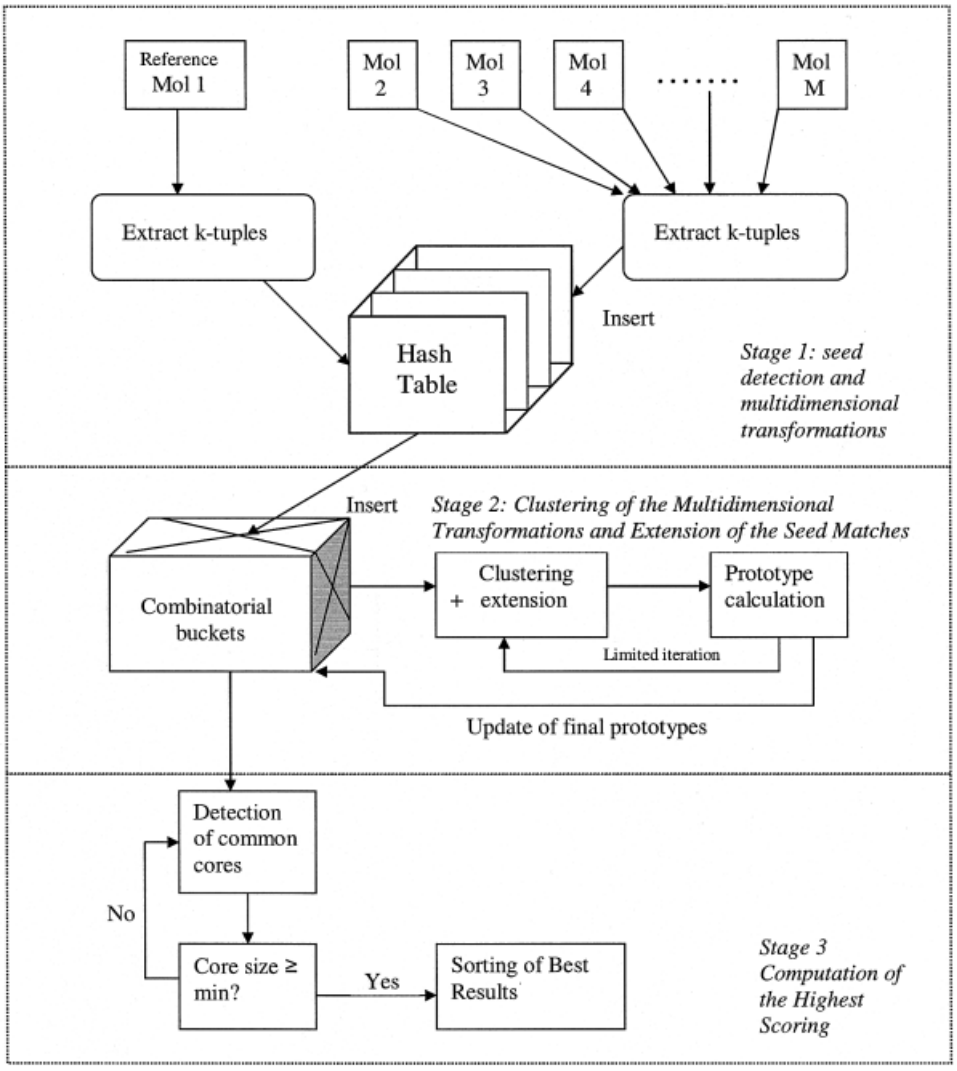

Fig. 2. A flowchart of the algorithm for multiple structure allignment (MUSTA).[173]

the size of the largest common substructure. This is a significant handicap, since it implies a run time of this order, and worse, a storage space of this order. Indeed, these results are limited to molecules of size 20, and the common pharmacophore of up to 9 atoms.

Brint and Willet[174] have further suggested an extension of the clique detection algorithm for finding maximal common subgraphs. This method is applicable to the multiple problem, by forming the correspondence graph. The graph is obtained by taking all *M-tuples* of nodes, a node from each molecule. A clique in such a graph corresponds to a substructure common to all the molecules, of the size of the clique. As in the previous approach, this is a truly multiple approach, as it both guarantees the optimal solution, and does not rely on the substructures that are common only pairwise. However, it is obviously exponential in M, as the correspondence graph has the complexity of the number of *M-tuples* generated by the M molecules. This is on top of the exponential complexity of clique detection. The authors propose that the clique algorithm be applied to find the maximal subgraphs of the pairwise problems, and the multiple solution is computed by performing all the intersections of these graphs.

Holliday and Willet[177] take into account the possibility that the core does not necessarily appear in all the molecules. They search for the smallest set of points that "covers" each of the input molecules by at least P points. The idea behind their method is to utilize an initial genetic algorithm stage, which produces a collection of alternative sets, each consisting of points that are "highly common" in many of the molecules. The second stage of the algorithm decides which of these proposed sets is a solution, and attempts to improve it. It applies a pairwise clique algorithm between each molecule and a proposed set to verify the solution and to further reduce the size of the set. The Holliday and Willet's[177] running times are several minutes, but all their experiments were carried out on small molecules of around 10 atoms each.

Finn et al.[178] developed an algorithm for pharmacophore identification, Randomized Pharmacophore Identifica-

tion for Drug Design (RAPID). The algorithm is designed to find the structural alignment between a pair of molecules. A parameter α is defined, which is a fraction of the smallest molecule, and the goal is to detect a solution of a size larger than this fraction. The basic operation is to randomly select a triplet from one molecule and find a congruent triplet in the other. The transformation Finn et al.[178] induce is computed, and is considered a solution if the match list it yields is larger than required.

To extend the algorithm to multiple structures, Finn et al.[178] iteratively take all solutions of a certain pairwise problem. For each of these, they generate a new molecule composed of the core found by that solution. RAPID compares the next molecule from the original ensemble, against each of these new molecules. The authors recognize that the multiple solution is not necessarily composed from the pairwise solutions. Therefore, at each stage, all the solutions propagate onwards, and the next molecule must be compared with each of them. This strategy is exponential with M. At each iteration, the problem must be solved again between the current solution and the next molecule.

However, again, in these approaches, the difficulty resides in drug flexibility.

Rigoutsos et al.[179] have developed an efficient algorithm for flexible 3-D structure matching against massive data bases of small molecules. For a given database of three-dimensional structures and a query molecule as an input, the method determines those molecules from the data base, which contain substructures in common with substructures in the query molecule, allowing torsional flexibility around rotatable bonds. The molecules in the database are represented as a set of rigid atom groups, with rotatable bond connectors. For each such molecule, there might exist a large number of different conformers. Nevertheless, for the method it is enough to store only one arbitrary conformer. As a result, the algorithm produces a novel conformer for the query molecule and a three-dimensional transformation for each rigid part.

The method benefits from the *geometric hashing* and the *pose clustering* techniques. Stockman[180] introduced the pose clustering method, also termed *generalized Hough transform* (or *transformation clustering*). Given two images, a model and a scene, the method discovers a transformation between them, which maps sufficient features of one, onto the other. Rather than begin with transformations and compute their match lists, this approach works the other way around: Small pairwise matching fragments are inspected, collecting evidence of the transformations that they infer. Each pairwise matching fragment that is inspected, increments a vote for the transformation that generates it. Transformations accumulating the largest number of votes at the end of the process are the transformations that generate the best alignment. Since in practice no two fragments are associated by the exact same transformation, a quantization of the transformation space is performed. At the first stage of Rigoutsos et al.,[179] for each molecule in the database, every rigid part is represented in a translation and rotation invariant manner and stored in the look-up table. This stage is done only once, off-line. The look-up table is updated when new molecules become available. The second matching stage identifies the rigid groups of the query molecule. For every rigid part, after finding a number of 3-D invariant representations, the look-up table is accessed and similar molecule parts are retrieved and stored. Consider some database molecule that gave a high number of similar substructures. Not all rigid parts have to be aligned with the rigid parts of the query molecule. However, some rigid parts may receive a number of different hypothesized matches. The flexible alignment could be achieved by selecting one such hypothesis for every rigid part, but all aligned subparts should be consistent with the rotatable bonds between the rigid groups. Since in practice the molecules are small and the number of rigid parts is also bounded by a small constant (around 6), it is possible to explore all relevant combinations of hypothesized matched rigid substructures and to choose the best scoring ones.

Based on similar principles, we have also developed flexible structural comparisons.[181,182]

Pharmacophoric-pattern searches of three-dimentional databases have been routinely used in the search of novel active compounds for more than a decade.[183,184] Miller et al.[185] have recently developed SQ, an atom-based clique-matching, followed by an alignment scoring function that recognizes pharmacologically relevant atomic properties. Pharmacophoric-pattern searches have also been used in similarity-driven flexible ligand docking,[186]combined with DOCK. Carlson et al.[187] have developed a dynamic pharmacophore construction algorithm, and tested it on HIV-1 integrase. The method accounts for inherent flexibility of the active site, and attempts to reduce the entropic penalty that is associated with binding a ligand. Yet, while constructing pharmacophores and docking compounds containing them has obvious advantages, such an approach also has limitations. The major drawback is that it limits drug diversity. This is particularly important as it has been shown that the volume and shape of the binding site can change, and drugs of different shape/size/composition can bind at the same site.

Drug diversity in docking

That the volume and shape of the binding site can change has been very attractively shown in retroviral proteases, specifically with regard to structural implications for drug design. Rose et al.[188] have found that rigid body rotation of five domains and movements within their interfacial joints provide a rational context for understanding why HIV protease mutations that arise in drug-resistant strains are often spatially removed from the drug or substrate-binding sites. They have identified and characterized domain motions associated with substrate binding in the retroviral HIV-1 and SIV proteases. These motions are in addition to closure of the flaps, and result from rotations of 6–7° at primarily hydrophobic interfaces. The crystal structure of the unliganded SIV protease is in the most "open" conformation of any retroviral protease determined to date. Comparisons of this structure and of

unliganded HIV structures with their corresponding liganded complexes, have illustrated that five domains of the protease dimer move as rigid bodies with respect to one another. These five domains include a terminal domain of the dimer (containing the *N* and *C* terminal β-sheet of the dimer), two core domains, which contain the catalytic aspartic residues, and two flap domains. Rose et al.[188] have shown that the two core domains rotate toward each other, reshaping the binding pocket. Further, they have shown that mutations at the interdomain interfaces that favor the unliganded form increase the off-rate of the inhibitor, allowing the substrate greater access for catalysis. This indicates a potential mechanism of resistance to competitive inhibitors, especially when the forward enzymatic reaction rate exceeds the rate of substrate dissociation. It has been noted for many cases, ranging from HIV-1 infection to diseases such as cancer, that often naturally occurring mutations selected to combat one drug will confer resistance to some others. Drugs are generally engineered to have a tightly fitting interface with the protein active site, with specific favorable interactions. Hence, this cross-drug resistance may well imply that it is not necessarily the mutations at the receptor-ligand binding interface that are solely responsible for the cross resistance.[188] Straightforward reasonable alternatives are changes in the binding site size, shape, and epitope, hampering the highly favorable drug-receptor binding.

Further, even presumably specific enzymes or receptors may bind ligands of different shape, size, and composition, with (sometimes) higher affinities.[67] Currently, synthetic inhibitors largely mimic natural substrates, and are frequently transition-state analogs. However, if the binding range is substantially broader, better-fitting inhibitors with higher affinities can potentially be designed.[67] Yet, by systematically docking databases of drugs, it is difficult to achieve diversity. Su et al.[73] have pointed out that when one compound fits the binding site well, close analogues typically do the same. Thus, in ranking docked drugs in the receptor-binding site, similar drugs are likely to appear next to each other in the list. Consequently, somewhat less well-fitting drugs might be down the list, and ignored. In an attempt to increase the diversity, Su et al.[73] have grouped the Available Chemical Directory into familes of related structures. Using DOCK, they docked all members of each family, picking only the highest scoring member of a well-docked family into the list. They have then compared the obtained list with a molecule-by-molecule docking for DHFR, thymidylate synthase, and a T4 mutant of lysozyme. In all cases, the family-based strategy yielded higher diversity. These were subsequently tested experimentally, and found to bind satisfactorily.

Nucleic Acid Docking

For DNA, as for proteins, the deterministic structure can be replaced with the conformation-population concept. Comparative analysis of DNA–protein complexes with protein–protein complexes revealed the similar features of the two, i.e., binding specificity and induced fit upon binding.[189–191] DNA–protein interfaces appear to be more polar, with many more intermolecular hydrogen bonds and buried water molecules than protein–protein interfaces. Binding specificity is dictated by specific DNA-binding motifs. Although many distinct families of DNA-binding proteins were defined, no simple "code" describing the side-chain/base interactions between proteins and DNA was found.[192]

Inspection of the literature reveals that very few docking algorithms have been applied to predict protein–DNA interactions. There are two main reasons for these rare applications. First, despite the well-recognized sequence-dependence of DNA conformation, for the most part the variability in DNA structure, most of which exists as variants of the canonical B-DNA, is on a relatively smaller scale, as compared to proteins or drugs. Hence, from the purely geometric consideration, the conformers might be too similar for a standard docking procedure. Second, while there is a considerable number of protein–DNA complexes[191] in the nucleic acid database (NDB[193]), no structure for a Watson-Crick base-paired B-DNA longer than a dodcamer[194] has been reported. Additionally, a larger fraction of the proteins that bind to DNA are believed to be natively disordered.[60,195] The natively disordered state may be an outcome of a small hydrophobic core, and uncompensated buried charges, a likely consequence of large binding interfaces and of their binding to a negatively charged polymer backbone. A larger diversity of conformations exists for single-stranded RNA. However, to date, the number of experimentally determined RNA structures is still relatively small. Additionally, as for DNA, it appears that an appreciable number of proteins that bind to RNA are natively disordered.[60] Nevertheless, owing to conformational variability and to its practical recognition and drug potential, RNA docking may be expected to grow in magnitude with database growth. One of the few algorithms dedicated to DNA-protein docking is MONTY, a Monte Carlo simulation program.[196] This algorithm includes phosphate ethylation interference and mutagenesis data.

Experimental studies have shown that in solution the p53 tumor suppresser protein cooperatively binds to the DNA response elements as a tetramer and results in a bent conformation of the complex.[197] This is in contrast to the "straight" conformation observed in the crystal complex, where only one p53 DNA binding domain (DBD) is specifically bound. To investigate the restraints imposed by the protein tetramer, Durell et al.[198] have developed a systematic procedure to exhaustively search the relative orientations accessible to the bound p53 DBD subunits. The model building and energy calculations predicted a remarkable amount of conformational variability, including correlated changes of the bend, twist, and slide degrees of freedom.

Zacharias and Sklenar[199] used harmonic modes to describe protein flexibility. They have used relaxation of harmonic modes to improve a steric fit between the ligand and the minor groove of the DNA. This is an example of application of large-scale mobility of domains to the dock-

ing problem. We shall come back to such potential applications below.

Protein–Protein Docking

Protein–protein docking simulates molecular recognition. Owing to the sizes of the molecules, this is the most challenging task. The number of degrees of freedom is huge, making it impractical to perform entire conformational search. This is the main reason why protein–protein docking algorithms handle the molecules as relatively rigid bodies. Despite the fact that such methods may miss correct prediction of the native complexes, nevertheless for a range of complexes such a drastic simplification has worked reasonably well. The problem is, however, having a necessarily fast scoring function that would be able to evaluate a huge number of solutions. Despite the tremendous effort invested in developing such functions (an outline of which is presented below), this goal has not yet been achieved satisfactorily.

The problem of protein–protein docking has immense implications: First, with respect to inhibitor design. However, second, it is of particular importance with regard to predicting cellular pathways, macromolecular interactions, and macromolecular assemblies. Given the difficulties in experimentally determining the structures of macromolecular assemblies, being able to computationally predict potential binding modes is a major aim of protein–protein docking algorithms. Furthermore, to be truly useful, such algorithms should be able to model the associations of computationally modeled protein structures, derived from amino acid sequences. As in any approach, be it experimental or computational, the first practical step is to demonstrate that the method works on known cases. Protein–protein docking algorithms have followed this convention.[15,36,39–44,76,78,101–105,112,113,115,139,142,144,200–211]

Within the framework of the rigid body treatment, flexibility is typically handled by surface variability, with a soft belt of allowed (though sometimes penalized) intermolecular surface atom penetration. Following prediction of binding associations, some routines carry out optimization of the interactions. However, given the difficulty in the ranking of the unbound solutions, such a procedure is often impractical. The majority of the rigid-body docking studies assume knowledge of the binding site. Only a few are able to handle the entire molecular surfaces (e.g., the Fast Fourier Transform-based matching, FFT[142]; the Geometric Hashing[39–44,204] and BiGGER.[113] And even among these, while sometimes they initially search the full conformational space (FTDOCK), subsequently the solutions are pruned by a binding site filter (see references[112,115,143,208] and to a lesser extent reference[148]). The CPU times required for the matching part of the docking vary widely: in the minutes range for the Geometric Hashing (Table I), in hours for BiGGER, and in days for the FFT.[78,143,148] BiGGER[113] is faster than the FFT (or FTDOCK) Fast Fourier Transform based-algorithm. While executing a real space search, BiGGER is guided by effective heuristic rules, which reduce the search space and the computational times.

The results obtained from protein–protein docking algorithms are in general satisfactory when reconstructing known complexes. However, when applying these to the so-called unbound docking problem, the results typically depend on the extent of the rearrangement that has taken place between the input data coordinates and the native complexed structures. In some examples, docking algorithms do very well in unbound cases (Table Ib, see also Tables II and III, obtaining relatively low rmsd between the predicted and the actual crystal-complexed structure. However, the quality of the predictions deteriorates with the extent of the rearrangement that has taken place. This is expected. The surfaces of the molecules are in constant motion. Movements of side-chains and surface atoms implicitly force docking programs to take account of intermolecular penetrations of the docked molecule-pair. Yet, in solution, such surface penetrations are alleviated by the movements of the groups of atoms on the molecular surface.

Several schemes have been adopted beyond soft-layer scoring. If the binding site is assumed to be unknown, two major approaches have been undertaken. In the first, "rough"[78,138,145] (or, low resolution) docking is carried out. Here, only the C_α backbone atoms are used rather than the full side-chain atom description. In this implementation, such a low-resolution docking implicitly takes into account surface atom movements. While the idea is attractive, and yields enhanced performance (in minutes, like the Geometric Hashing full atom runs[40]), on the down side the quality of the solutions deteriorates, although this can change with improved scoring schemes. The second drawback of this approach is that it handles only side-chain motions, disregarding backbone movements. The second approach is application of a method such as that devised by Abagyan and his colleagues,[212–214] which is based on the method of Li and Scheraga.[215] This Internal Coordinate Mechanics (ICM) strategy and the ICM psedo-Brownian algorithm followed by optimization[216] have proven quite successful. Such an approach can in principle be applied following rigid docking by other algorithms. However, the drawback of this approach is that it too largely handles side-chains, rather than large backbone movements.

The third approach is docking allowing hinge-bending motions.[101–105] The ligands are allowed to undergo translations and rotations of their parts in order to optimally dock to the surface of the receptor. The ligand information is stored in a look-up table, generated in the preprocessing phase, which is invariant to this type of transformation. The location of the hinge is predetermined. Since this robotics-based algorithm is fast, many other hinge positions may be tried. The structure of the receptor is presented to the system in the next, recognition phase of the algorithm. If a ligand has an "interest point" (surface descriptor point) configuration similar to the receptor "interest point" configuration, the algorithm casts a vote for the computed location of the hinge. This hinge location is computed from the transformation between the corresponding receptor and ligand interest point surface configurations. Highest scoring (voted for) hinge locations are

TABLE II. Classification of Docking Algorithms According to Function Parameters[†]

Algorithm name (reference)	Scoring stage in algorithm flow	Reference for the solution	Geometric complementarity	Hydrogen bonds	Contact area	Intramolecular overlap	Intermolecular overlap	Pairwise amino acid contacts	Electrostatic interactions	Solvation energy	Active site residues	Free energy
Sobolev et al.[241]	Integrated	Self	+	+	+	−	+	−	+	−	−	−
SP-DOCK Fradera et al.[186]	Edge	Known structure	+	+	−	+	+	−	+	−	−	−
SG-DOCK Fradera et al.[186]	Integrated	Known structure	+	+	−	+	+	−	+	−	−	−
Norel et al.[230]	Edge	Self	+	−	−	−	−	−	+	−	−	+
FTDOCK, Katchalski-Katzir et al.[142]	Edge	Self	+	−	−	−	−	−	−	−	−	−
Fischer 1995	Integrated	Self	+	−	−	−	+	−	−	−	−	−
DARWIN, Burnett and Taylor[34]	Integrated	Self	−	+	−	−	+	−	+	+	−	+
PUZZLE, Helmer-Citterich et al.[205]	Edge	Self	−	−	+	−	+	−	−	−	−	−
Hybrid algorithm, Hou et al.[242]	Integrated	Self	+	+	−	−	+	−	+	−	−	−
Gardiner et al.[77]	Integrated	Self	+	+	−	−	+	−	−	−	−	−
Jackson et al.[115]	Edge	Self	+	−	−	−	+	+	−	−	+	−
Norel et al.[40]	Edge	Self	+	−	−	−	+	−	+	−	−	−
ESCHER, Ausiello et al.[38]	Integrated	Self	+	+	−	−	+	−	+	−	−	−
Camacho et al.[148]	Integrative	Self	−	−	−	−	−	−	+	+	−	+
BiGGER, Palma et al.[113]	Integrative	Self	+	−	−	−	+	+	+	+	−	−

[†]Classification of some common algorithms according to the scoring function parameters they use, and the stage in the algorithm flow.

TABLE IIIa. Comparison of Some Rigid-Body Algorithms for Bound Cases[†]

Complex PDB	Res (Å)	Norel et al.[a] Rank solutions	Norel et al.[a] RMSD	FTDOCK[b] Rank solutions	FTDOCK[b] RMSD	BiGGER[c] Rank solutions	BiGGER[c] RMSD
Protease-inhibitor							
1acb	2.00	1 out of 1121	0.9	—	—	18 out of 1,000	0.6
1cho	1.80	1 out of 471	0.5	40 out of 218	0.8	—	—
1cgi	2.30	—	—	3 out of 161	1.0	—	—
2kai	2.50	11 out of 1,227	1.2	38 out of 502	0.4	—	—
2sni	2.10	1 out of 1,367	1.1	8 out of 54	0.6	—	—
2sic	1.80	1 out of 1,229	1.1	22 out of 30	0.8	2 out of 1,000	3.8
1cse	1.20	2 out of 1,024	1.3	—	—	—	—
2tec	1.98	1 out of 1,042	1.2	—	—	77 out of 1,000	3.6
2ptc	1.90	1 out of 1,027	0.06	91 out of 513	0.7	—	—
Antibody-antigen							
1mlc	2.10	—	—	2 out of 507	0.8	—	—
1vfb	1.80	20/2181 (1.5)	1.5	240 out of 631	0.7	—	—

TABLE IIIb. Comparison of Some Rigid-Body Algorithms for Some Inbound Cases[†]

Complex PDB	Receptor PDB	Receptor Res (Å)	Ligand PDB	Ligand Res (Å)	Norel et al.[a] Rank solutions	Norel et al.[a] RMSD	FTDOCK[b] Rank solutions	FTDOCK[b] RMSD	FTDOCK[c] Rank solutions	FTDOCK[c] RMSD	BiGGER[d] Rank solutions	BiGGER[d] RMSD
Protease-inhibitor												
1cho	5cha	1.67	2ovo	1.50	2 out of 2,289	1.60	11 out of 86	1.20	1 out of 86	1.30	6 out of 1,000	2.90
1cgi	1chg	2.50	1hpt	2.30	—	—	3 out of 94	1.80	3 out of 94	2.20	9 out of 1,000	3.70
2kai	2pka	2.05	6pti	1.70	9 out of 4,222	1.20	130 out of 364	1.50	2 out of 364	1.30	Not found	
2sni	2sbt	2.80	2ci2	2.00	92 out of 3,582	2.60	8 out of 26	1.80	4 out of 26	2.60	16 out of 1,000	1.30
2sic	2stl	1.80	3ssi	2.30	—	—	Not found		—	—	15 out of 1,000	3.30
2ptc	1tgn	1.60	5pti	1.50	1 out of 3,453	1.20	16 out of 229	1.50	11 out of 229	1.60	52 out of 1,000	2.70
Antibody-antigen												
1mlc	1mlb	2.10	1lza	1.60	—	—	41 out of 590	1.20	—	—	Not found	
1vfb	1vfa	1.80	1lza	1.60	—	—	176 out of 707	2.10	—	—	Not found	
1hfl	3hfl	2.50	1lza	1.60	65 out of 10,733	1.08	228 out of 519	1.80	—	—	Not found	
3hfm	3fhm		1lza	1.60	281 out of 10,685	2.80	65 out of 762	1.10	—	—	Not found	

[†]Ranking, number of solutions and the RMSD for the highest ranking docked prediction, for several rigid-body docking algorithms. **a:** Complexed, bound cases; **b:** unbound cases. Table IIIa: [a]The Norel et al. cases are taken from Norel et al., 1999[40]; [b]The FTDOCK cases are taken from Gabb et al., 1997[112]; [c]The BiGGER results are taken from Palma et al., 2000[113]. In all cases, rigid-body docking is performed. In all cases, initially the entire surfaces of the two molecules are used. However, at the second stage, Gabb et al.[112] use active site residues to prune the solutions. **Table IIIb:** The [a]Norel et al.[40] and the [b]FTDOCK and [d]BiGGER results are from Gabb et al.[112] and Palma et al.[113] The FTDOCK are taken from Jackson et al.[115] As described in the text, the scoring is somewhat different. However, the basic docking procedures are the same (entire molecular surfaces and rigid docking in the first stage, followed by active site pruning). The major difference in scoring between Gabb et al. and Jackson et al. is in the solavation treatment.

sought. No knowledge of the binding site, or of the hinge locations relative to the receptor is assumed.

In the preprocessing step, the ligand molecule (model) is described as a set of interest points. The (predetermined) hinge location is positioned at the origin of a 3-D Cartesian coordinate frame. This frame is the *ligand frame*. The orientation of this frame is set arbitrarily. For each noncollinear triplet of interest points, in each of the ligand parts, a unique triplet-based Cartesian frame is defined (the *triplet frame*). The shape signature of each triplet of points is the ordered triangle side lengths. This geometric shape signature of the triplet constitutes an address to a look-up hash table. The information that is stored in this entry at this address is the ligand identification, the part number, and the transformations between the *triplet frame* and the *ligand frame*.

In the recognition step, the molecular surface of the receptor is similarly described by its set of interest points. All noncollinear triplets of the interest points of the receptor are considered in the docking stage. For each of these triplets, the triplet-based Cartesian frames are computed. Each is the *receptor triplet frame*, calculated as above. This calculation is invariant under rotation and translation, with congruent ligand triangles having similar values. The look-up table calculated in the preprocessing phase is entered using as an address the currently computed ordered triplet of triangle side lengths of the receptor. For each ligand-record present at that entry in the table, a *candidate ligand frame* is computed by applying the pre-recorded ligand transformation at that (hash table) entry,[141,217] to the current *receptor triplet frame*. The origin of the *candidate ligand frame* is the candidate hinge location. We vote for the location and orientation of the candidate, hinge-centered, ligand frame. Finally, we seek hinge locations with high scores. A high-scoring hinge location defines the 3-D translation that the ligand would

need to undergo in this candidate docking. The appropriate rotations are calculated separately for each part only at a later scoring and filtering stage. In the current step, hinges that have received a large number of votes are selected. Further details are given in Sandak et al.[101–105] This hinge-bending algorithm exploits the fact that both parts of the molecule share the same hinge. The essential point here is the way it is taken into account by locating the origin of the reference frame of the ligand at the hinge. In this way, both parts contribute votes to a reference frame at the same location, although the orientations of both parts with respect to each other may differ. Most importantly, by picking up votes from both molecular parts, a ligand, which might otherwise have only a small portion of surface complementary to the receptor surface in each of the parts, may still score high. Thus, while each of the individual parts of the ligand can obtain an insignificant score, the sum of the votes obtained from both of the ligand's parts, may yield an overall acceptable match, which can be automatically detected. This algorithm can be applied to multiple hinges and a database search.

This approach is fast, also on the order of minutes of CPU. It allows backbone movements. However, it has two major drawbacks: the first is the necessity to prepick likely hinges; and the second is that each part is handled as a rigid body. However, combined with an algorithm that predicts the location of the hinges and the motions[218,219] or through structural comparisons of the protein family allowing hinge-bending movements,[181–182] the likely locations of the hinges can be predicted. To address the rigid parts, a similar approach as, for example, taken by Abagyan and his colleagues, can be applied. Alternatively, the surface flexibility may be derived from ensembles. However, it remains to be seen how well such a combined method would actually work.

Docking in Protein Folding

Docking and protein folding are all too often considered to be distinct fields. One resorts to using docking algorithms either in efforts for drug design or in protein–protein docking when considering enzyme–inhibitor, antibody-antigen associations or receptor–ligand. Yet, today it is increasingly realized that protein folding is a hierarchical process.[210–220] Hierarchy implies a process of assembly of smaller folding entities, be it conformationally fluctuating building blocks, or the more stable hydrophobic folding units or domains. Currently, predictive protein folding schemes focus on single domain proteins, as these constitute simpler systems. However, the next stage would involve putting the domains together. Multidomain association resembles the formation of multimolecular cellular assembly. To carry out such a task, docking is likely to be a tool of choice.

Current docking schemes do not attempt to recreate multimolecular assemblies, unless they possess restrictive symmetries, such as in viral coat proteins. The reason is the huge computational complexity in the number of ways several molecules can recombine. This problem is aggravated by the size of each of the molecules. Nevertheless, in applying docking tools to folding, there are a number of considerations that, at least in principle, may reduce the heavy computational load. These include the size and the backbone connectivity.

On the practical side, cutting crystal structures of monomers to their domains and redocking these may not address the real problem. Reconstructing a protein from its dismembered domains would present the combinatorial assembly nature of the multiple domains. However, as the domains would be cut from the three-dimensional structure, this problem resembles the "bound" docking, rather than mimics the real-life "unbound" domains. One way to overcome this difficulty is again by generating ensembles of conformations of the separate domains through multiple trajectory runs.

Figure 3 illustrates the concept of docking of domains of a large, multidomain protein, within the folding scheme. This kind of relationship between docking and folding is best illustrated by the domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments.[220,221] Dihydrofolate reductase may be divided into three fragments[223–225] (Fig. 3). Neither of the fragments is stable by itself in solution,[225] even though some population of the near native conformation for the individual fragments may exist[226] (Fig. 3). Fragment 37–159 has been observed to be stable in nonnative conformation. However, when fragments are linked with GCN4 leucine zipper domain (fragment 1 and 2 in one part and fragment 3 in another), the dimerization of the leucine zipper domains assembles all fragments, and the dihydrofolate reductase is active.[223,224] Here, the folding of dihydrofolate reductase may be viewed as a step-wise docking of near native conformations from the fragments, as illustrated in Figure 3.

APPROACHES TO SCORING SCHEMES

A search algorithm may produce an immense number of solutions,[113] unmanageable for any practical need: 10^9. Theoretically, free-energy simulation can be a reliable discrimination to check the solutions.[227] However, it is not practical to use such an approach in docking searches. The purpose of the scoring function is to discriminate between "correct" native solutions with low rmsd from the crystal complex and others within a reasonable computation time. Vieth et al.[32] have assessed the energy functions for *flexible* docking in term of efficiency and selectivity. They assess these two components in a broad range of energy functions, derived from systematic modification of the CHARMM param19/toph19 energy function. In particular, they examine the effects of the dielectric constant, the solvation model, the scaling of surface charges, reduction of van der Waals repulsion and nonbonded cutoffs. Overall, Brooks and colleagues[32] favor an efficient function for docking, since only an efficient function can dock the ligand in the active site. In the early docking stages, the soft core vdW is critical. Once docked, a hard-core potential can take over to optimize it.

Although some algorithms are able to rank correct solutions within the top hundred or even within the top ten

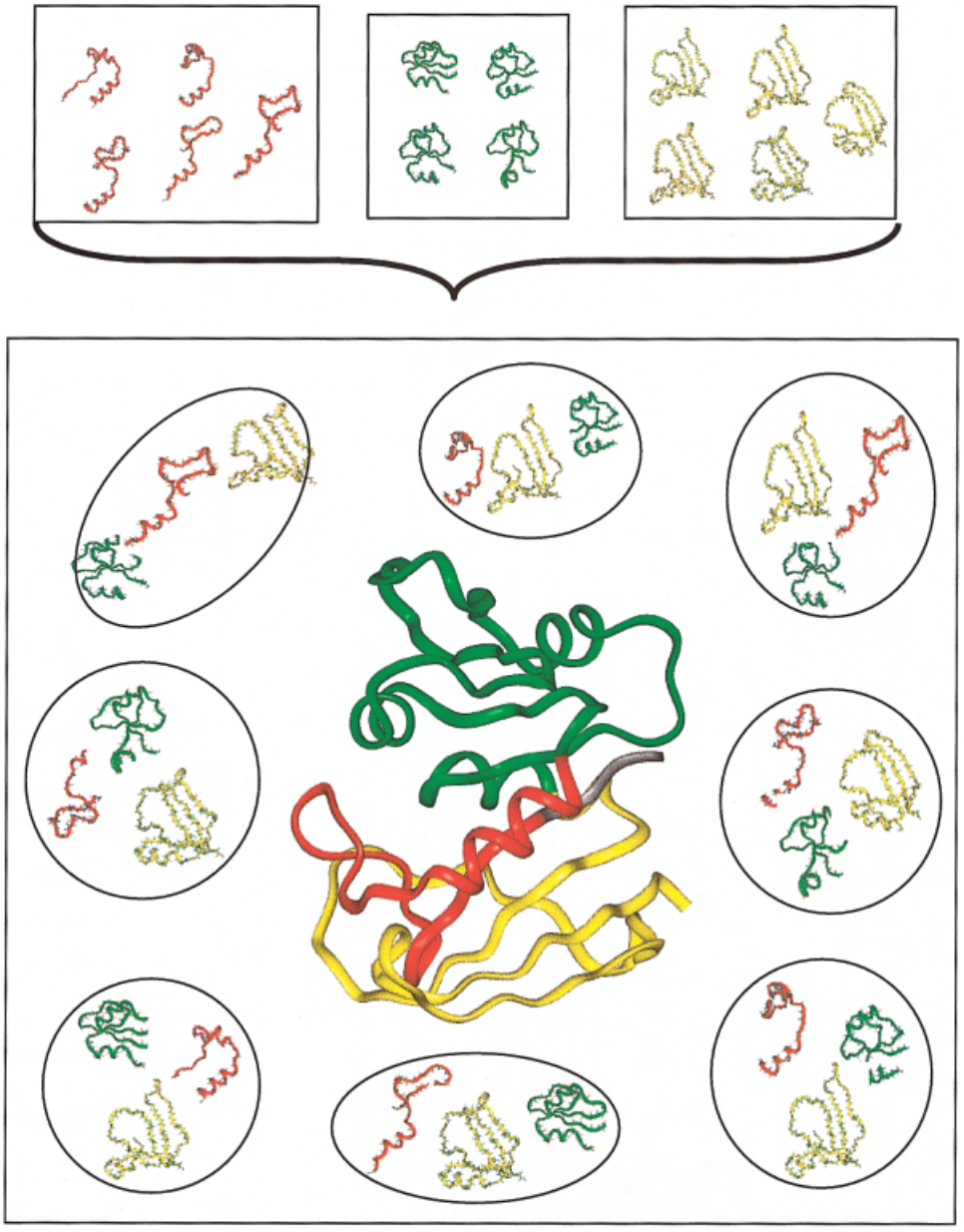

Fig. 3. The concept of docking of domains of multidomain proteins within the folding scheme. Dihydrofolate reductase may be divided into three fragments: F1 (1–36) red; F2(37–88) green; and F3(89–159) yellow. Some population of near native conformations for the individual fragments may exist (top box).[226] The collision complexes from these transient fragments is a docking–folding problem. There are many collision complexes (small boxes in the large box) and the native-like folded conformation (central one) is one of them. When the fragments are linked via dimerization domains, the native conformation is stabilized and shifts to a folded active dihydrofolate reductase.

places for some predictive docking cases,[40,113,225,115,208,228] for most complexes the highest ranked structures are still false positives, i.e., solutions with a high rmsd from the complex, a high score, and a low rank[148] (Tables I–III). Thus, in spite of the advanced sophisticated schemes applied in some scoring functions, no efficient method for reliable discrimination between correct solutions and false positives generated by predictive docking algorithms is currently available.[40] A lack of a reliable method for quickly locating correct solutions, in particular if the binding site is unknown, is the major obstacle in using predictive docking for practical applications.

Analyses of protein–protein interfaces[40] have illustrated that the binding interfaces do not necessarily have the largest extent of buried surface areas. Furthermore, native-like bound conformations do not manifest the largest nonpolar buried surface areas as compared to other potentially feasible docked solutions. They do not contain the largest number of hydrogen bonds, or the smallest number of unsatisfied buried polar groups. In solution,

these are likely to be handled by surface motions, eliminating such unfavorable energy contributions. Hence, the problem is how, despite these hurdles, to still conceivably be able to detect candidate lead compounds/protein inhibitors or native protein–protein interactions.

Currently, the solution appears to be a two-stage ranking, i.e., using traditional scoring to rapidly scan possible solutions and obtain initial "good" candidates, followed by more advanced methods to further discriminate the limited conformations.[160] Below, we first describe scoring functions (for fast scanning) and briefly mention recent progress in electrostatic solvation and free-energy calculations.[140,148,161,227,229,230]

Additionally, we outline a potentially complementary way of handling this problem, through utilization of a *library of functional epitopes*. Such a library can be generated using efficient computational techniques.[173,181,182,231] In deriving these, we should bear in mind the recently observed determinants of binding sites: flexibility[50–52,54,64,69,70] and energy hot spots.[232–238]

Stage of Scoring in the Algorithm Flow and Reference for the Solution

Docking algorithms can be classified by the stage of scoring in the algorithm flow into two groups: *integrated* and *edge* functions. In integrated algorithms, scoring is integrated into the search stage and filter emerging solutions. In edge algorithms, scoring is applied at the end of the search stage. The major difference, therefore, is that the scoring function forms part of the design of the solutions in integrated algorithms but not in edge algorithms. Integrated algorithms are required in some of the computational schemes used in docking, for example, in genetic algorithms and in anchoring algorithms. Genetic algorithms require a fitness score that is applied after each generation and used for the selection pressure operation.[77,133,227] An anchoring algorithm divides the docked protein into segments. The segments are docked to the target protein layer by layer. Since the solution number grows exponentially with the addition of layers, a scoring function is needed to limit the number of growing seeds that are kept for further calculations. It is a matter of course that the scoring function will include intramolecular scores.

The score can be determined with regard to (1) other solutions, (2) a known structure, or (3) the solution itself. The parameters that can be used for self-reference are listed below. If a correct solution (a solution with a low rmsd from the complex) is found, one can expect to find other solutions that differ only slightly from it. Therefore, a comparison of each solution with other solutions, by a direct comparison or by clustering, is generally beneficial. In principle, the cluster size may also be used as a parameter in a scoring function. However, it may also depend on the shape of the binding site. A larger number of similar solutions is expected when the binding site is flat, as compared to, say, deep holes.

An example for a scoring function that refers to a known structure is the similarity driven algorithm by Fradera et al.[186] This is an anchoring algorithm to assess the improvement of the scoring function by a similarity parameter used at different stages of the algorithm flow. SPDOCK, similarity-penalized docking, and SGDOCK, similarity-guided docking, differ from each other only by the stage in which the similarity to a reference ligand is used to correct the score of the solutions. In SPDOCK, the docking scores are corrected according to the similarity to a reference structure only at the end of the docking procedure, whereas in SGDOCK, the docking scores are corrected according to the similarity to a reference structure every time a DOCK energy calculation is performed. The similarity in performance between SPDOCK and DOCK appears greater than between SGDOCK and DOCK. Consistent with this finding, integrated algorithms are expected to have more impact on the docking results than edge algorithms.

Modes of parameter consideration: positive contribution vs. penalty and exclusion vs. relative contribution

The parameters of a scoring function can either contribute to or subtract from the score of a given solution or be used as an exclusion filter. Most of the current scoring functions combine a few parameters. The mode of parameter consideration is semantic, since it might have an impact on the stringency of other scoring criteria as well as on the algorithm flow and complexity. The addition of exclusion filters might enable the use of less stringent criteria for other scoring parameters. Therefore, an exclusion filter might contribute to dealing with flexibility by "softening" the scoring function. An exclusion parameter also enables a module flow. In ESCHER[38] for example, there are three modules: SHAPES, which creates rough solutions based on geometric complementarity; BUMPS, which identifies molecular collisions; and CHARGES, which evaluates the electrostatic complementarity. Since the SHAPES module excludes solutions, the other two modules are operated only on the remaining solutions, reducing the complexity of these stages. This principle of applying a rapid calculated parameter for reducing the solution population before using a high-cost calculated parameter, is widely used. In most cases, geometric criteria are used as the primary exclusion parameter owing to their speed, and energy criteria are applied on the significantly reduced solution population.

Tables II classifies the docking algorithms according to their function parameters. Appendix A suggests a unified format for presentation of the results by the different algorithms, so they can be straightforwardly assessed. Appendix B lists some available structures for docking trials.

Parameters Used for Scoring

The parameters' choice depends on the breadth of sampling: entire surface vs. binding sites. Scoring parameters can be roughly divided into two groups: *collective* parameters and *individual* parameters. A *collective* parameter refers to a property, which characterizes the entire molecule, whereas an individual parameter refers to a specific

atom or residue. An example of a collective parameter is the diameter parameter used in the graph theoretic technique for macromolecular docking.[239] The algorithm finds the maximal complementary sets of donor/acceptor H-bond pairs between two proteins. Many sets of hydrogen bonds are sparsely distributed over the surface of the proteins, whereas the area of interaction between two proteins is usually reasonably compact. A diameter parameter is used to constrain the distance between any two members of a set of hydrogen bond donors/acceptors. An example for an *individual* parameter is the pairwise amino acid contacts parameter (P_{AB}, the probability of amino acid A being in contact with amino acid B) used in the BiGGER algorithm.[113]

Docking programs use essentially similar computational elements. These include energy minimization of rigid-body docking, conformational minimization including side-chains, solvation, electrostatics, a van der Waals term in the free-energy expression, or a quasi-global Monte Carlo minimization.[214] These are computationally demanding. On the other hand, simpler and much faster scoring schemes include elements such as geometric complementarity, contact and overlap checks; counts of hydrogen bonds; counts of unsatisfied buried charges; extent of buried nonpolar surface area, or total buried surface area. However, the programs differ in the implementation of these, and in the way they are put together. Below, we outline the major computational parameters. It is, however, important to distinguish between those that are effective only for discrimination between near-native conformations, all in the relative *vicinity of binding sites*[112,115,148] and those applied to *globally different conformations*.[40,78,113] These latter need faster scoring schemes. Hence, in particular, which element a user should employ depends on the situation at hand. Further, the shape of the binding site, concave (as in enzymes) vs. flat (as in subunit–subunit), charged vs. hydrophobic, large vs. small, may also dictate some parameter preference.

The following parameters have been used in scoring functions.

Geometric complementarity. From the early days of docking, it has been postulated and repeatedly reaffirmed that geometric matching plays an important role in determining the structure of a complex.[15–17,39,240] The three-dimensional (3D) structures of most protein complexes reveal a close geometric match between interfaces of a receptor and a ligand.[29–44,142] The scoring functions of early docking algorithms used practically exclusively geometric complementarity criteria.[17,39–44,136,142,154,211] These functions are understandably more successful in "bound" docking as compared to "unbound" docking. As a result, the consideration of conservation of geometric similarity between the unbound and bound, and consequently using the bound structure in rigid docking of unbound structures were supplemented by efforts to characterize protein flexibility and its effect on geometric complementarity.[49,62,114] Although usually bound complex complementarity is better than in computationally proposed solutions,[241] there are cases where false-positive solutions display a better shape complementarity than correct solutions.[40] Consequently, current scoring functions frequently use additional criteria in combination with geometric complementarity.[40,77,113,228,242] Since geometric complementarity calculations are highly efficient, they usually serve as a primary filter before costly evaluation criteria, reducing the number of solutions at the end of the search stage.

Geometric complementarity has been assessed by a number of methods. Gardiner et al.[239] sought a similar shape. The maximal area of complementarity between two proteins is detected using a clique of a graph. A clique of a graph G, is a subgraph of G in which every vertex is connected to every other vertex and that is not contained in any larger subgraph with this property. In a later work, Gardiner et al.[77] replaced this definition of geometric complementarity by the relative directions of the surface normals and additionally the type of surface. Geometric complementarity was redefined as nearly opposite surface normals (the angle between the normals is close to 180°) and the Connolly shape is different, unless it is type 2 (i.e., sadddle). This definition of geometric complementarity has been suggested earlier as part of the search stage.[36,37,39–43,157] The implementation of this definition in the scoring stage was performed not only by Gardiner et al.[77] but also previously by others.[36,37,39–44,206,242] Although the definition of geometric complementarity appears to be similar, in the score based on this definition it is calculated differently. Lin et al.[36] and Hou et al.[242] use the area shared between two matching (i.e., complementary) dots, whereas Norel et al.[40] and Gardiner et al.[77] use the number of matching dots. Therefore, there will be a variance in the contributions of some matching point pairs sharing different areas to these scoring funtions.

How to score geometric complemenarity is strongly coupled with how the surface is represented. Palma et al.[113] used a simplified approach. The surface of a protein is represented by a collection of a 1 Å grid cubes. Geometric complementarity is defined as an overlap between surface cells from different proteins. In ESCHER,[38] the complementarity score is defined as the number of corresponding consecutive polygon vertices whose distance is under 1.6 Å. Hidden in the ESCHER algorithm is the assumption that a geometric complementarity between the top of the ligand and the bottom of the receptor is indicative of areas showing the best geometric fit. No validation of this assumption was presented, although a verification that the results of the algorithm are independent of the orientation of the target receptor in the reference system has been done. The optimization of solutions with highest complementarity using a fine step of translation and rotation is a common method of geometric complementarity optimization.[241]

Intermolecular overlap. The quest for geometric complementarity is often balanced by consideration of intermolecular overlap. The general approach to intermolecular overlaps is tolerance to slight interface clashes and penalty for protein interior clashes. The tolerance is usually implemented by a surface belt of nonpenalized penetration area.[38,77,112,113] Gardiner et al.[77] defined inter-

molecular overlap as a clash between the ligand surface dots and the interior dots of the receptor. Interior points were defined as grid points that were near a receptor atom and were further than 2 Å from any surface dot. This is not an even-handed approach. A penetration of a ligand side-chain into the receptor is penalized, whereas a receptor side-chain penetration into the ligand is not. On the other hand, Palma et al.[113] use an even-handed approach. The core of a protein is defined as a 1 Å grid cube, the center of which, and all of its neighbors[1] centers, lies within 1 Å of a van der Waals sphere of any protein atom. An overlap between cores of the two proteins is used to exclude solutions, with the exception of five amino acids: Arg, Lys, Asp, Glu, Met. These amino acids possess the highest frequencies and amplitudes of movements between the structures of free and cocrystallized proteins. Norel et al.[40–43] compute a scoring function, also based on geometric features. They award surface contact, penalize overlaps, and reject serious overlaps. Allowing some intermolecular penetrations implicitly takes 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. Ligand atom centers that invade the outer shell of the molecular representation are retained.

For the protein–protein docking, in the scoring and overlap test Norel et al.[43] map the receptor onto a 3D grid, with a 1 Å resolution. Interior and exterior atoms are mapped as balls with radii equal to their van der Waals radii plus the radius of the probe size, and their voxels are marked *i* and *e*, respectively. MS dots (generated using the Connolly surface description, at a density of 5 dots/1Å^2 are also mapped onto the grid as balls of 1Å radii. These are *s* voxels. With such mapping, the receptor is represented as a "core" of interior voxels *i*, a wide layer of intermediate voxels *e*, and a thin layer of surface *s* voxels. After the receptor is mapped, the ligand atoms are transformed using the best rigid transformation matrix obtained in the matching step. If any of the ligand atoms fall into a voxel designated *i* or *e*, the solution is discarded. If the solution passes the overlap test, the MS dots of the ligand are also transformed and used to compute a scoring function. Three counters are kept: one for ligand dots that fall in *i* voxels (I), one for MS dots that fall into *e* voxels (E), and one for MS ligand dots that fall into *s* voxels (S). For the protein–protein and protein–DNA cases, the score is computed as: S − 4E-10I. For the protein–drugs and DNA–drugs, the score is 5S + E − I. Surface contact score increases the score, and overlap of the ligand surface dots into the receptor's interior reduces the score. This scoring scheme is used for ranking the solutions.

ESCHER[38] takes into consideration the collisions sum, rather than each surface dot separately. In Sandak et al.[101–105] in the verification stage, the respective transformations of each of the parts are applied to the atoms in each of the parts of the ligand (or the receptor, depending on the location of the hinge). Transformations that result in the penetration of a ligand part into the receptor (collision check) or yield collisions between the parts of the ligand (self-collision check, see below), are discarded. The receptor and the ligand molecules are assumed to collide, if the distance between a ligand atom and a receptor atom is smaller than the sum of their respective van der Waals radii minus a proximity threshold.

An original approach to intermolecular overlaps is in the scoring function of the least-squares algorithm.[243] Instead of using the widespread belt tolerance, the B factor was taken into account. The advantage of this method is the uneven weight coupled with penetration of different atoms. High B factor atoms indicate uncertainty in an atoms' position. A high mobility atom is more likely to be associated with a false penetration than a low mobility atom. Therefore, the weight of the collision has been taken to be inversely proportional to the B factor. The B factor can be used not only in the scoring stage but also in the search stage. Docking using a stable subset of atoms, such as backbone atoms, has proven to be a feasible approach.[63,96] The B factor can be used to define an even more robust subset to be used for docking.

Intra-molecular overlap. The geometric complementarity and the inter-molecular overlap criteria are used to evaluate shape complementarity with respect to entire receptor/ligand molecules. When ligand[101,121,186] or receptor backbone flexibility[104,105] are taken into account, an additional criterion of shape complementarity is used. One of the possibilities enabling domain flexibility is division of the ligand to sets of rigid fragments that are docked separately. An anchor rigid fragment is selected, and docked into the receptor. Other fragments are docked layer by layer. The flexibility is implemented in a range of positions at the joints of the rigid fragments. This method bears some resemblance to protein folding algorithms that identify building blocks on the primary sequence, which are then docked with respect to each other to reconstruct the folded protein structure. However, there a combinatorial assembly of the domains is considered.[244] In order to limit the number of seeds (i.e., the number of different trees of fragment orientations) and to allow a reasonable calculation time, intramolecular overlap is used to score the growing solutions. In both algorithms,[121,186] a threshold was used to eliminate solutions. Kuntz et al. discarded any orientation in which a fragment atom is within 2.5 Å of a receptor atom.[15]

Sandak et al.[101–105] first check for intermolecular penetration, and calculate the "contact percentage" between the receptor and the ligand. Only binding modes receiving a contact percentage that is higher than the contact threshold are considered for the self-penetration check. Depending on the location of the hinge (in the ligand or in the receptor), the self-collision would be carried out. The self-collision check employs the same criterion for rejecting self-penetration causing transformations, as being done by the intermolecular collision check, described above. Development of additional intramolecular collision (penetration) checks is expected, given that docking of domains is likely to be increasingly applied to the folding of large proteins.

Hydrogen bonds. One of the most widely used criteria minor only to shape complementarity is hydrogen bonding. There are 1.13 ± 0.47 hydrogen bonds per 100^2 buried surface area.[49] Information on hydrogen atom positions is missing for the majority of X-ray structures of proteins. Several methods exist for constructing these hydrogen atom positions. Methods such as Barlow and Thornton's[245] or WHAT IF, place hydrogen atoms according to standard geometries for every amino acid type. More sophisticated methods, such as the standard coordinate method in Insight 97, take the local environment into account and optionally perform an energy minimization.[246] There is a diverse classification of atoms with respect to hydrogen bond formation potential.[77,239] For example, four types of classes: H donor, H acceptor, H donors/acceptors, and non-H bonding may be defined.[77] Satisfying the hydrogen bonding potential is defined as formation of the following matches: H donor matches H acceptor or H donor/acceptor, H acceptor, or matches H donor or H donor/acceptor, H donor/acceptor matches H donor, H acceptor, or H donor/acceptor, and non-H bonding matches non-H bonding. The same group classification was already suggested previously.[38,76] The division of the atoms into these groups is practically identical (except N of His, which is classified as a hydrogen bond donor by Gardiner et al.[77] and by Jiang and Kim[76] but as a hydrogen non-bonding by Avsiello et al.[38]). The major difference between these scoring functions regarding the hydrogen bond parameter, is that Gardiner et al.[77] and Jiang and kim[76] do not discriminate between the bond types, whereas Avsiello et al.[38] attach different weights to different bond types. For example, a donor-donor interaction is weighted as +3 and a donor-acceptor interaction is weighted as −3. A low sum value implies multiplicity of hydrogen bonds. Another difference is the distance between the atoms forming the hydrogen bond. According to Gardiner et al.[77] the hydrogen satisfaction potential is considered only if atoms have the same grid value (i.e., the distance between the atoms is not larger than 2 Å), while Ausiello et al.[38] consider atoms up to 3.4 Å apart. On the other hand, Jiang and Kim[76] adopt a different approach. One predefined distance is replaced by a series of cube sizes (1 to 6 Å with a 1 Å step). These are used to represent the receptor and the ligand surfaces. A solution is selected only if it consistently has positive total interactions. A different classification of atoms with respect to hydrogen bonds generated 4 types. These replace the H donors/acceptors and non-H bonding with concepts of Neutral donor and Neutral acceptor.[241]

Contact area. Consideration of contact area is frequently employed by a variety of schemes, in particular in algorithms that are rigid-body, geometry-based.

Janin and Chothia[79] determined the interface area of a number of protein–protein complexes to be in the range 1,200–1,600 $Å^2$. These data were implemented into the Gardiner et al.[239] docking algorithm. Assuming an approximately circular interface, an interface area of 1,200–1,600 $Å^2$ is equal to 20–30 Å diameter in each protein. Gardiner et al.[239] used this diameter parameter to restrict the area of hydrogen bonds as was detailed previously. Shoichet et al.[136] proposed the parameter of volume inside a receptor pocket as an alternative to the contact surface area parameter.

Kuntz et al.[247] found that binding energies initially increase with contact area. However, they quickly reach maxima and do not correlate with the contact area beyond this point. This is understandable since binding energy cannot increase indefinitly. In practice, contact area largely translates to taking account of hydrophobicity at the intermolecular interfaces. Hydrophobicity has been well known to play important, albeit variable, role at the interfaces. Hydrophobic patches are present, although it is not the largest hydrophobic patches that determine the interaction sites.[40,41,248,249]

For the protein–protein cases, Norel et al.[40,41] have further utilized a very simple "hydrophobicity filter." The receptor and ligand atoms are divided into polar and hydrophobic. Each MS dot (Connolly's Molecular Surface dots) is labeled, 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 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 computed as

$$total_{interactions} = hh + pp + hp.$$

The hydrophobicity factor is

$$hf = \frac{hh}{total_{interactions}}.$$

A Connectivity Filter for solutions that pass the overlap test is also computed.[43] The connectivity filter awards matches of larger patches of surfaces. 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 defined as the number of "C" dots that belong to that component. Largest components are sought, with the threshold set such that the size of a "large" component should be 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 $dot/Å^2$). The conectivity filter is employed only for the protein–protein docking.

Wallqvist and Covell[206] use an energetic criterion based on surface burial. The free-energy approximation is derived from their analysis of surface burial of atom pairs in crystal complexes. They parameterize the occurrence of specific atom-atom surface burial to mimic the free energy of binding.

Pairwise amino acid and atom–atom contacts. Pairwise amino acid contacts is a purely empirical term derived from observed statistical frequency of amino acid contacts in a database of well-resolved X-ray protein

structures. There are two methods for calculating the expected number of pairs between residues i and j: mole fraction and contact fraction. Mole fraction is proportional to the product of the fractional abundance of the two residues in the pair. Contact fraction is proportional to the propensities of the two residues to be paired with any residue.[228] The scoring function of both BiGGER[113] and Pair potentials[228] uses the contact fraction method, however, in a slightly different way. The Pair potentials score is the log fraction of the actual count divided by the expected count, whereas the BiGGER score is the probability of contact. The Moont et al.[228] score is calculated as

$$score_{i,j} = score_{j,i} = \log(c_{i,j}/e_{i,j}) \quad (1)$$

where $c_{i,j}$ is the residue level potential, defined as occurring between residues i and j if the C_β atoms in the two residues are under a given cutoff. The value of the score for each pair is considered as a statistical measure of the likelihood of that pair occurring. Since this is a log fraction, the total likelihood for the conformation is a summation of all individual scores. The major difference between these computational schemes is the type of pairwise potential that were used and the dataset for the calculation. Only one residue level potential is used (based on side-chain atoms) in BiGGER, based on the atlas of protein side-chain interactions.[250] In contrast, Pair potentials use four types of pairwise potentials: a residue level potential based on C_β atoms, a residue level potential based on all atoms, a residue level potential based on all side-chain atoms and an atom level potential. The difference between the residue level potentials is in the atoms used for the definition of pairing residues. The atom level potential is the occurrence of contacts between atom types i and j within a given distance cutoff. Forty atom types were assigned according to Melo and Feytmans.[251]

Wallqvist and Covell[206] apply knowledge-based potential energy functions to solutions passing the first geometry-based matching and overlap checks. Their pairwise-atom type based potential functions have been separately derived from a dataset of enzyme-inhibitors.[252] In their enzyme-inhibitors docking applications, Wallqvist and Covell[206] apply the corresponding enzyme-inhibitor pair potential function set. The docking is carried out on "bound" crystal complexes. Pairwise atom–atom potential functions have also been derived by Weng et al.[210] and used by Camacho et al.[148] as discussed above.

The derivation and utilization of knowledge-based pairwise potential functions for docking has its origin in studies of protein folding.[253] Subsequently, they were applied to predict protein structures, particularly in fold recognition.

Electrostatic interactions and solvation energy in quick scan. Electrostatic interactions play a key role in many aspects of proteins including binding.[254] Well-known cases in which electrostatics are important are superoxide dismutase,[255] trypsin-BPTI complex,[254] and the barstar-barnase system.[256] Electrostatic potentials can be calculated by a variety of programs: Delphi,[257] GRASP,[47] and UHBD (University of Houston Brownian dynamics[258]). These programs solve the Poisson-Boltzman equation for a protein–protein solvent system.[246]

Estimates of binding affinities based on simplified potential functions have been used in scoring.[112,214,228] The BiGGER algorithm[113] uses atom point charges from the Amber molecular mechanics force field. The electrostatic interactions are calculated using point to point Coulombic potential: $V_{elec} = k * Q_i Q_j (r_{ij} + c)^2$ where c is a dampening constant added to the distance separating both nuclei. The c constant is needed due to the allowance of limited interpenetration of grid positions of both molecules. Some atoms become unrealistically close to each other giving rise to high interaction energies. The c constant is set to the minimum distance allowed between two interacting atoms. A value of 1.5 Å was used in BiGGER. A different representation of electrostatic interactions was performed using the FlexX algorithm.[169] FlexX handles hydrophobic ligand docking to proteins by placing ligand fragments. In FlexX, interaction types and geometries describe the protein–ligand interactions. The interactions are divided into 3 levels according to the geometrical restriction of the interaction. Level 3 interactions are salt bridges and hydrogen bonds. Level 2 interactions are specific hydrophobic interactions between an aromatic ring center and aromatic ring atoms, amides, or methyl groups. Level 1 interactions are non-specific hydrophobic contacts between aliphatic and aromatic carbon atoms. Level 1 interactions are spherical with a radius of about 4 Å. As long as a fragment contains interacting groups belonging to a geometrically restrictive interaction type, these are prefered. If the fragment has a very limited number of interactions of this type, the algorithm starts using less geometrically restrictive interaction types for the fragment placement. In a data set of 200 hydrophobic ligands there were, on average, 444 level 3, 167 level 2, and 325 level 1 interactions per complex.

Solvation energy in a scoring function was intensively applied to the Fast Fourier Transform-based algorithm by two groups.[115,148] In the first Jackson et al.[115] Implementation, the water solvent is described by a soft sphere Langevin dipole model, with discrete dipoles that interact with the electric field of the protein but subject to random thermal fluctuations that reduce the effective electric field at the dipole itself. van der Waals and field-dependent hydrophobic terms are also included. If the energy for the interaction (the solvation enthalpy) is greater than a 0 kcal/mol cutoff, the solution is excluded from further consideration. This cutoff was chosen since the unfavorable vdW repulsion term quickly overcomes a favorable electrostatic interaction. The authors found the soft sphere treatment and fine grid spacing to be critical in reproducing observed solvation energies. Their energy refinement consists of two steps: (1) Determination of side-chain conformations. Side-chains are modeled according to a rotamer library.[108] A side-chain interacts with the protein backbone and with probability-weighted average of the surrounding protein side-chains and solvent molecules. The presence of a particular water molecule at a particular site around a side-chain rotamer is modeled probabilisti-

cally and depends on the occupancy of that site by a side-chain atom of any other amino acid side-chain. (2) A rigid body energy minimization to relax the protein interface. The resulting solutions are self-consistent. The energy refinement was applied to five cases of enzyme–inhibitor and four cases of antibody–antigen docking. The enzyme–inhibitor were more encouraging than the antibody–antigen results. We note that enzyme–inhibitor cases typically obtain better results, in many docking schemes, possibly owing to a high sequence (and structure) conservation and the concave shape of the binding site.

Camacho et al.[229] have developed electrostatic and desolvation free-energy maps. These maps essentially show two different types of behavior: If there is a weak electrostatic complementarity, the desolvation free-energy map shows a well-defined minimum with a broad area of attraction. However, in the case of oppositely charged interfaces, there is a region of low electrostatic energy surrounding the binding site. Here desolvation provides further adhesion. If the electrostatics are strong, the binding site can then be predicted by a minimum of desolvation free energy at the low-electrostatic energy region. Based on these, the authors suggest that either desolvation free energy (in the case of neutral surfaces) or the electrostatic energy (in the strongly charged surfaces case) provide information on the relative orientation of the two molecules, even prior to *tight* complex formation. Nevertheless, it is important to note that here, as in the Jackson et al.[115] work above, *near-native* conformations were used.

Consensus scoring and two-stage ranking

It is extremely difficult to reliably discriminate between binding modes by a single algorithm. A combination of several available ranking packages may, however, lead to binding modes that top the scoring lists in most docking/scoring combinations.[33,259,260]

Charifson et al.[259] conducted an extensive computational study in which they show that combining scoring functions in an intersection-based consensus approach results in an enhancement in the ability to discriminate between active and inactive enzyme inhibitors. An analysis of two different docking methods and 13 scoring functions provides insights into which functions perform well, both singly and in combination. The consensus scoring further provides a dramatic reduction in the number of false positives identified by individual scoring functions, leading to a significant enhancement in hit-rates.

Similar approaches have also been taken by Bissantz et al.[33] and Terp et al.[260] In Terp et al.'s study, eight different scoring functions have been combined with the aim of improving the prediction of protein-ligand binding conformations and affinities. The obtained scores were analyzed using multivariate statistical methods to generate expressions, with the ability (1) to select the best candidate between different docked conformations of an inhibitor (MultiSelect) and (2) to quantify the protein-ligand binding affinity (MultiScore). In Bissantz et al.'s case, three different database docking programs (Dock, FlexX, Gold) have been used in combination with seven scoring functions (Chemscore, Dock, FlexX, Fresno, Gold, Pmf, Score) to assess the accuracy of virtual screening methods against two protein targets (thymidine kinase, estrogen receptor) with known three-dimensional structures. Even though the consensus scoring improves the ranking hits, however, as found by Bissantz et al.,[33] no clear relationships could be found between docking and ranking accuracies. Moreover, predicting the absolute binding free energy of true hits was not possible regardless of the docking accuracy that was achieved and the scoring function that was used.[23]

Knowledge-based free-energy scoring functions were used in docking.[261,262] However, a promising ranking scheme may involve simulation of the free energy using either thermodynamic integration or continuum electrostatics following the filtering stage.

Camacho et al.[148] employ a two-step scoring algorithm. The first step includes two rigid-body filters that use the desolvation free energy and the electrostatic energy to limit the number of docked conformations. The desolvation free energy ΔG_{ACE} is calculated for all docked conformations where $\Delta G_{ACE}^{min} < G_{ACE}^{\min} + C_{ACE}$. ΔG_{ACE}^{min} is the lowest ΔG_{ACE} found and C_{ACE} is an empirical threshold (in that work chosen as 5 kcal/mol). ΔG_{ACE} is the atomic contact energy (*ACE*), an extension of the quasi-chemical residue pair-wise potential (Miyazawa and Jernigan[253]). In *ACE*, they sum $\Sigma_i \Sigma_j e_{ij}$ for all i,j that are up to 6 Å apart. The authors note that their preference for ΔG_{ACE} rather than $\Delta G_{ACE} + \Delta E_{elec}$ is owing to the lower sensitivity to small changes in the atomic positions, which the ΔE_{elec} term is expected to show. In the next step, ΔE_{elec} is calculated. The conformations selected have $\Delta E_{elec} < \Delta E_{elec}^{min} + C_{elect}$, where ΔE_{elec}^{min} is the lowest ΔE_{elec} and C_{elect} is a threshold (C_{elect} = 10 kcal/mol is the current value). The second filter is designed to reject false positives. It minimizes the molecular mechanics energy of the structures and reranks them. It employs a combined free-energy function that includes electrostatics, solvation and van der Waals energy terms. The CHARMM energy of the conformations that pass the above filters are minimized using a large number (500 to 1,000) of steps of the adopted basis Newton-Raphson minimization algorithm. The free energy is calculated as $\Delta G = \Delta G_{ACE} + \Delta E_{elec} + \Delta E_{vdw}$. These filters have been tested on a set of docked decoys, generated by the Fast Fourier correlation method for several protein complexes, including protease-inhibitors and antibody-antigen. DOT[139] uses similar parameterization as used by Camacho et al.[148]

In one of the last elegant contributions[161] from Peter Kollman, MD simulations are combined with MM-PBSA (molecular mechanics Poisson-Boltzman/surface area) to rank binding modes suggested by DOCK 4.0. One of the exciting results from this simulation is that they predicted a conformation with 1.1 Å RMSD of an HIV-1 RT inhibitor before the cystal structure was published.

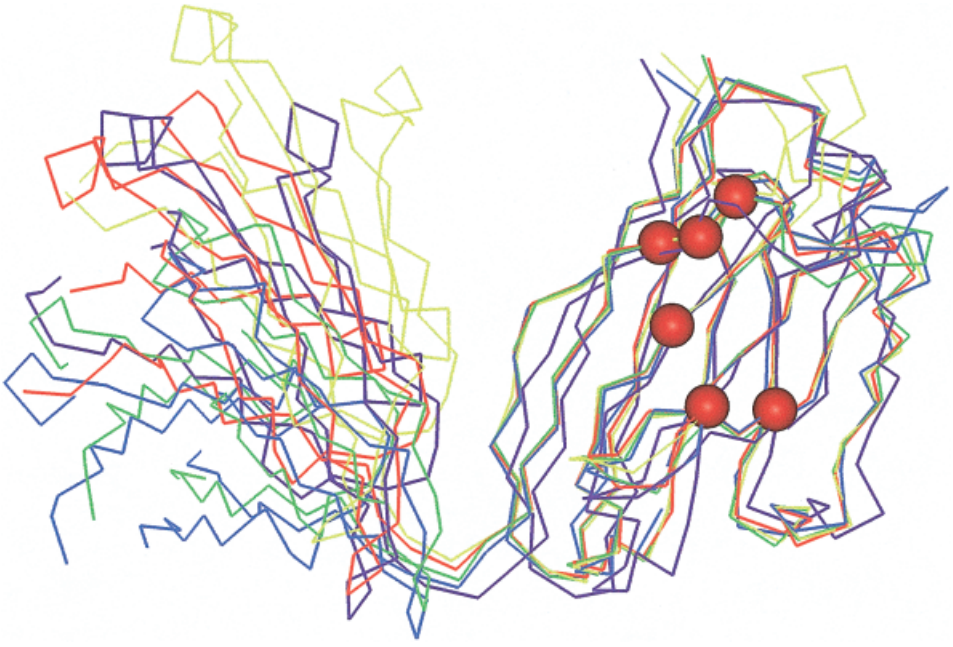

Fig. 4. Hotspots on the interfaces of 1hil chainH, using MUSTA.

Binding site information may be included in scoring

Knowledge of the location of the binding site on one or both proteins drastically reduces the number of possible solutions. Knowing specific binding site residues reduces the search space even further. Information about active site residues is sometimes available from site-directed mutagenesis, chemical cross-linking, and phylogenetic data.[112,155] In the absence of experimental data, it is sometimes possible to predict the correct binding site.[263] Potential hydrogen bonding groups, enzyme clefts and charged sites on a protein surface were all used for binding site prediction.[10,151–156,264] Structural comparisons with molecules with known binding sites may also lead to binding site identification. Since binding sites are at least partially flexible, searches for part-flexible part-rigid sites have also produced encouraging results.[50–53] Algorithms that predict the location of hinges and modes of motions,[215,216] or those that carry out structural comparisons of the protein family, in particular if they allow hinge bending movements,[181,182,265] should be useful as well.

Serine proteases and immunoglobulins represent systems where at least the major binding sites are known in advance. The catalytic triad of the serine proteases (His, Asp, Ser) and the CDR of immunoglobulins are both well characterized.[112] Three scoring functions for antibody-antigen docking with varied filter stringency have been suggested.[112] These functions are based on analysis of structural principles of antibody-antigen contacts.[266] The loose filter requires at least one contact between any part of the antigen with the L3 or H3 CDRs of the antibody. The medium filter requires at least one contact between any part of the antigen with both L3 and H3 CDRs of the antibody. The tight filter requires at least one contact between an epitope residue of the antigen with both L3 and H3 CDRs of the antibody. Equivalent functions were defined for serine protease inhibitor docking. This scoring function was tested on 8 unbound cases. A loose biochemical filter successfully narrowed the solution number from around 4,000 to a few hundred, within which was a prediction with better than 2.5 Å rmsd for C_α atoms of the interface. Medium filtering reduced the total number of predictions by an order of magnitude with respect to the loose filter. The tight filter, on the other hand, does not significantly further alter the number of false positives.[112]

Hu et al.[238] have shown that in enzyme-inhibitors residues are more conserved at the interfaces than at other locations. This can explain why pairwise potential function derived from enzyme–inhibitors are more successful when applied to this same class. On the other hand, overall, antibody-antigen interfaces have similar surface conservation as compared to their corresponding linear sequence alignment, consistent with the suggestion that evolution has optimized protein interfaces for function. This explains why lesser success is achieved in such an application for the antibody-antigen class. However, Norel et al.[230] have used the fact that antibody binding sites are enriched in tyrosines and tryptophans. Applying a filter based on this enrichment has resulted in considerable improvement in antibody-antigen complex prediction.

For rigid-body docking, knowledge of the location of the binding site and of specific residues is extremely helpful.[267,268] However, they can be run in the absence of such information too.[40,43,78,113,144,269] For algorithms employing detailed molecular dynamics calculations, knowledge of the location of the binding site is absolutely essential. Monte Carlo, or MD simulations, initiate by placing the ligand in the active site,[270,271] and often restrain its movement away from the site. Flexible drug docking by

TABLE IVa. Some of the Available Pseudo Unbound Cases for Predictive Heteroprotein Docking

Number	Complex	Receptor unbound	Ligand unbound
Enzyme/inhibitor			
1	1acb	4cha	1acb
2	1brc	1bra	1brc
3	1cho	4cha	1cho
4	1cse	1scd	1cse
5	1ppe	3ptn	1ppe
6	1sbn	1sup	1sbn
7	1sft	1ppn	1sft
8	1tab	3ptn	1tab
9	1tgs	1tgt	1tgs
10	2tec	1thm	2tec
11	4htc	2hnt	4htc
12	1udi	1udh	1udi
Antibody/antigen			
1	1nca	1nca	7nn9
2	1nmb	1nmb	7nn9
3	1igc	1igc	1igd
4	1jel	1jel	1poh
5	3hfl	3hfl	1lza
6	3hfm	3hfm	1lza
Others			
1	1atn	1atn	3dni
2	1gla	1gla	1f3g
3	1spb	1sup	1spb
4	2btf	2btf	1pne
5	3hhr	3hhr	1hgu

TABLE IVb. Some of the Available Unbound Cases for Predictive Heteroprotein Docking

Number	Complex	Receptor chain	Ligand chain	Receptor unbound	Ligand unbound
Enzyme/inhibitor					
1	1brb	E	I	1bra	1bpi,4pti,5pti,6pti
2	1cgi	E	I	1chg,2chg(AB)	1hpt
3	2kai	AB	I	2pka	1bpi,4pti,5pti,6pti
4	2ptc	E	I	3ptn,1tld,1bty	1bpi,4pti,5pti,6pti
5	2sic	E	I	1sup,2stl	3ssi
6	2sni	E	I	2sbt,isup	2ci2,1ypc
7	1brs	A	D	1a2p,1bao	1al9,1bta
8	1bvn	P	T	1pif	2ait
9	1cao	ABC FGH	DI	5cha(AB)	1app(AB)
10	2tgp,4pti	Z	I	1tgn	1bpi,4pti,5pti,6pti
11	1cbw	ABC FGH	DI	5cha(AB)	1bpi,4pti,5pti,6pti
12	1cho	E	I	5cha(AB)	2ovo
13	1hia	AB XY	IJ	1ao5	1bx8
14	1ugh	E	I	1akz	1ugi(A)
15	1brc	E	I	1bra	1aap(AB)
16	1dfj	E	I	2bnh	7rsa
Antibody/antigen					
1	1mlc	ABCD	EF	1mlb	1lza,1lyz,6lyz,3lzt
2	1vfb	AB	C	1vfa	1lza,1lyz,6lyz,3lzt
3	1fdl	LH	Y	3hfl(LH)	1lza,1lyz,6lyz,3lzt
4	3hfm	LH	Y	1hfm	5lym(A) 5lym(B)
5	1ahw	ABDE	CF	1fgn(LH)	1boy
6	1bvk	AB	C	1bvl(AB)	1lza,1lyz,6lyz,3lzt
7	1dqi	AB	C	1dqq(AB)	1lza,1lyz,6lyz,3lzt
Others					
1	1mda	LH	A	2bbk	1aan
2	4hvp	A	B	3hvp	3hvp
3	1fss	A	B	2ace	1fsc
4	2pcb	AC	B	1ccp	1hrc
5	2pcc	AC	B	1ccp	1ycc
6	1wej	LH	F	1qbl(LH)	1hrc
7	1avz	A	C	1avv	1shf(A)
8	1wql	Q	R	1wer	5p21
9	1bdj	A	B	3chy	2a0b

other algorithms (e.g., by DOCK) whether fragment-wise, using ensembles, or flexible drugs picked from the drug database,[45,67,95,116] also invariably assume knowledge of the binding site on the protein surface (see the discussion on small molecule docking for details).

Binding hot spots may also be incorporated in the scoring process. Hot spots can be defined as a small subset of residues that contribute to the binding energy more than can be expected from an even distribution across the interface.[232] The energetic contribution of individual side-chains was experimentally examined via alanine scanning mutagenesis.[272] By combining the alanine scanning with kinetic and thermodynamic measurements, it has been shown that, despite the large size of the binding interfaces, single residues can contribute a large fraction of the binding free energy in an interface.[233,234] In a computational study on families of protein–protein interfaces, Hu et al.[238] have confirmed and generalized the alanine scanning data analysis, which was of a limited size. These hot-spot conserved residues have been detected consistently in all interface families. Figure 4 illustrates the hot spots in a family of protein interfaces. Here, the MUSTA (MUltiple STructure alignment algorithm)[173,231] has been used for the alignment. A cluster of hot spots appears to be a good indication of the presence of a binding epitope on the protein surface (Elkayam et al, unpublished results).

CONCLUSIONS

The so-called *computational molecular docking problem* is far from being solved.[2–8,273] Nevertheless, despite the drawbacks in each docking strategy, significant progress has been made. First, rigid-body algorithms have been remarkably successful,[15,40,113,274] especially in addressing the protein–protein docking problem, even in the absence of knowledge of the binding site, *if* the conformational change is limited to surface side-chain atoms. These are handled via a "soft" surface belt. Second, it is apparent that "docking in steps" is a promising strategy. A computational design where initial rigid-body, entire-surface matching algorithm is applied followed by a dynamic method that can overcome the energy barriers in a reasonable time, such as the pseudo-Brownian algorithm,[214] is one such possibility. Nevertheless, here too, since the sampling initiates from rigid-body docking, backbone movements are limited. Third, an algorithm that can carry out docking allowing *hinge-bending* motions has been developed.[101–105] However, there the hinges need prepicking, and the domains are held rigid. Application of such an algorithm, followed by, e.g., the pseudo-Brownian algorithm,[214] might be successful, sampling the conformational space more extensively[275] *if* likely hinges can be automatically picked. Fourth, docking of ensembles, especially also allowing combination of conformers to increase the sampling,[63] is an approach founded on the physical behavior of molecules in solution. These multistep approaches initiate by docking the conserved parts, followed by the more variable ones.[63,96,97,196] Currently these can handle only a limited extent of backbone flexibility. Nevertheless, combined with hinge-bending motion algorithms, possibly such an extension is feasible. Likely hinges can be picked through *multiple* structural comparison algorithms, allowing hinges. Since topologically related proteins show similar large- and small-scale motions, protein families can be used for this purpose. Alternatively, the ensembles may derive from simulations.

The second major bottle-neck is the availability of selective and efficient scoring functions. To address this critical hurdle, additional large-scale combined studies[31–33,259,260] are likely to help in making a good choice. Yet, it should be borne in mind that two different types of scoring schemes are needed. The first are for global searches. These should be fast and empirical,[276–278] course-graining and filtering the solutions. The second are the more detailed free-energy simulations, initiating from given likely conformations. For the first, deriving likely binding sites on the protein surface should be immensely useful. In identifying *binding epitopes*, not only the "classical" parameters of geometry and conservation in families should be considered. Additionally, the presence of *hot spots*[238] and flexibility[7,52,54,69,279,280] can prove to be tremendously useful. With such initial information, additional free-energy calculations might be used as a reliable index for identifying correct binding modes.[140,148,160,161,227,230]

High throughput docking has been used extensively in library design. The practice has been extended from docking a single protein to a ligand library to evaluation of multiple receptor libraries against multiple targets.[281–283] Computational generation of protein structures and the docking of modeled protein structures with potential interacting partners will have great impact on the life sciences.

ACKNOWLEDGMENTS

We thank our former graduate student Raquel Norel, currently a postdoctoral fellow at Columbia University for many discussions, and for critical reading of the manuscript. Her work contributed to a very large extent to the progress we have made. We thank our docking group at Tel Aviv University, and in particular Dina Duhuvny, Vladimir Polak, and Hadar Benyaminy. We thank our group in Frederick, in particular C.-J. Tsai, and Dr. J. V. Maizel for discussions, and for encouragement. The research of R. Nussinov and H. J. Wolfson in Israel has been supported in part by the Ministry of Science grant, and 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.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 organization imply endorsement by the U.S. Government.

REFERENCES

1. Maggio ET, Ramnarayan K. Recent developments in computational proteomics. Trends Biotechnol 2001;19:266–272.
2. Abagyan R, Totrov M. High-throughput docking for lead generation Curr Opin Chem Biol 2001;5:375–382.
3. Kuntz I. Structure-based strategies for drug design and discovery. Science 1992;257:1078–1082.
4. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol 1996;6:402–406.
5. Joseph-McCarthy D. Computational approaches to structure-based ligan design J Mol Biol 1997;267:727–748.
6. Gane PJ, Dean PM. Recent advances in structure-based rational drug design. Curr Opin Struct Biol 2000;10:401–404.
7. Carlson HA, McCammon JA. Accomodating protein flexibility in computational drug design. Mol Pharmacol 2000;57:213–218.
8. Sotriffer CA, Flader W, Winger RH, Rode BM, Liedl KR, Varga JM. Automated docking of ligand to antibodies: methods and applications. Methods 2000;20:280–291.
9. Schafferhans A, Klebe G. Docking ligands onto binding site representations derived from proteins built by homology modelling. J Mol Biol 2001;307:407–427.
10. Crick FHC. The packing of α-helices: simple coiled-coils. Acta Cryst 1953;6:689–697.
11. Lee BK, Richards FM. The interpretation of protein structures. Estimation of static accessibility. J Mol Biol 1971;55:379–400.
12. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:709–713.
13. Connolly M. Analytical molecular surface calculation. J Appl Cryst 1983;16:548–558.
14. Bernstein FC, Koetzle TF, Williams GJ, Meyer E, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
15. Kuntz I, Blaney J, Oatley S, Langridge R, Ferrin T. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982; 161:269–288.
16. Zielenkiewicz P, Rabaczenko A. Protein–protein recognition: Method for finding complementary surfaces of interacting proteins. J Theor Biol 1984;111:17–30.
17. Connolly M. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. Biopolymers 1986;25:1229–1247.
18. Lee RH, Rose GD. Molecular recognition. I. Automatic identification of topographic surface features. Biopolymers 1985;24:1613–1627.
19. DesJarlais RL, Sheridan RP, Dixon JS, Kuntz ID, Venkataraghavan R. Docking flexible ligands to macromolecular receptors by molecular shape. J Med Chem 1986;29:2149–2153.
20. Wodak SJ, Janin J. Computer analysis of protein–protein interaction. J Mol Biol 1978;124:323–342.
21. Wodak SJ, De Crombrugghe M, Janin J. Computer studies of interactions between macro-molecules. Prog Biophys Mol Biol 1987;49:29–63.
22. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28:849–857.
23. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. J Mol Recog 1996;9: 1–5.
24. Taniuchi H, Parr GR, Juillerat MA. Complementation in folding and fragment exchange. Methods Enzymol 1986;131:185–217.
25. Fisher A, Taniuchi H. A study of core domains, and the core domain-domain interaction of cytochrome c fragment complex. Arch Biochem Biophys 1992;296:1–16.
26. Yang XM, Yu WF, Li JH, Fuchs J, Rizo J, Tasayco ML. NMR evidence for the reassembly of an α/β domain after cleavage of an α-helix: implications for protein design. J Am Chem Soc 1998;120: 7985–7986.
27. Spolaore B, Bermejo R, Zambonin M, Fontana A. Protein interations leading to conformational changes monitored by limited proteolysis: apo form and fragments of horse cytochrome c. Biochemistry 2001;40:9460–9468.
28. Xu D, Tsai CJ, Nussinov R. Mechanism and evolution of protein dimerization. Protein Sci. 1998;7:533–544.
29. Tsai CJ, Xu D, Nussinov R. Protein folding via binding, and vice versa. Fold Design 1998;3:R71–R80.
30. Tsai CJ, Ma B, Nussinov R. Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci USA 1999;96:9970–9972.
31. Vieth M, Hirst JD, Dominy BN, Dailer H, Brooks CL III. Assessing search strategies for flexible docking. J Comp Chem 1998;19:1623–1631.
32. Vieth M, Hirst JD, Kolinski A, Brooks CL III. Assessing energy functions for flexible docking. J Comp Chem 1998;19:1612–1622.
33. Bissantz C, Folkers G, Rognan D Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations J Med Chem 2000;43:4759–4767.
34. Burnett RM, Taylor JS. DARWIN: A program for docking flexible molecules. Proteins 2000;41:173–191.
35. Luty BA, Wasserman ZR, Stouten PFW, Hodges CN, Zacharias M, McCammon JA. A molecular mechanics/grid method evaluation of ligand-receptor interactions. J Comp Chem 1995;16:454–464.
36. Lin SL, Nussinov R, Fischer D, Wolfson HJ. Molecular surface representation by sparse critical points. Proteins 1994;18:94–101.
37. Lin SL, Nussinov R. Molecular recognition via face center representation of a molecular surface. J Mol Graph 1996;14:78–90.
38. Ausiello G, Cesareni G, Helmer-Citterich, M ESCHER: A new docking procedure applied to the reconstruction of protein tertiary structure. Proteins 1997;28:556–567.
39. Norel R, Lin SL, Wolfson H, Nussinov R. Shape complementarity at protein-protein inter-faces. Biopolymers 1994;34:933–940.
40. Norel R, Petrey D, Wolfson H, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999; 35:403–419.
41. Norel R, Wolfson H, Nussinov R. Small molecule recognition: solid angles surface representation and shape complementarity. Comb Chem High Throughput Screen 1999;2:177–191.
42. Norel R, Fischer D, Wolfson H, Nussinov R. Molecular surface recognition by a computer vision based technique. Prot Eng 1994;7:39–46.
43. 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.
44. Fischer D, Lin SL, Wolfson HL, Nussinov R. A geometry based suite of molecular docking processes. J Mol Biol 1995;248:459–477.
45. Oshiro CM, Kuntz ID. Characterization of receptors with a new negative image: use in molecular docking and lead optimization. Proteins 1998;30:321–336.
46. Ritchie DW, Kemp GJL. Protein docking using spherical polar fourier correlations. Proteins 2000;39:178–194.
47. Nicholls A. GRASP: graphical representation and analysis of surface properties. New York: Columbia University Press; 1992.
48. Tsai CJ, Ma B, Sham Y, Kumar S, Nussinov R. Structured disorder and conformational selection. Proteins 2001;44:418–427.
49. Betts MJ, Sternberg MJE. An analysis of protein conformational changes on protein-protein association: implications for predictive docking. Prot Eng 1999;12:271–283.
50. Todd MJ, Semo N, Freire E. The structural stability of the HIV-1 protease. J Mol Biol 1998;283:475–488.
51. Todd MJ, Freire E. The effect of inhibitor binding on the structural stability and cooperativity of the HIV-1 protease. Proteins 1999;36:147–156.
52. Freire E. The propagation of binding interactions to remote sites in proteins: Analysis of the binding of the monoclonal antibody D1.3 to lysozyme. Proc Natl Acad Sci 1999;96:10118–10122.
53. Luque I, Friere E. Structural stability of binding sites: consequences for binding affinity and allosteric effects. Proteins 2000; (Suppl) 4:63–71.
54. Baysal C, Atilgan AR. Coordination topology and stability for the native and binding conformers of chymotrypsin inhibitor 2. Proteins 2001;45:62–70.
55. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. Biochemistry 1994;33:6739–6749.
56. Koshland DE Jr. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 1958;44:98–123.
57. Creighton TE. Proteins: structure and molecular properties, 2nd ed. New York: WH Freeman and Company; 1993.

58. Ramakrishnan B, Qasba PK. Crystal structure of lactose synthase reveals a large conformational change in its catalytic component, the β1,4-galactosyltransferase-1. J Mol Biol 2001;310: 205–218.
59. Shoemaker BA, Portman JJ, Wolynes PG. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc Natl Acad Sci USA 2000;97:8868–8873.
60. Wright PE, Dyson HJ. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.
61. Fraga S, Parker JM, Pocok JM. Computer simulations of protein structures and interactions. New York: Springer Verlag, 1995; 2081 p.
62. Mangoni M, Roccatano D, Di Nola A. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. Proteins 1999;35:153–162.
63. Claussen H, Buning C, Rarey M, Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. J Mol Biol 2001;308:377–395.
64. DeLano WL, Ultsch MH, de Vos AM, Wells JA. Convergent solutions to binding at a protein-protein interface. Science 2000; 287:1279–1283.
65. Vazquez-Laslop N, Zheleznoca EE, Markham PN, Brennan RG, Neyfakh AA. Recognition of multiple drugs by a single protein: a trivial solution of a old paradox. Biochem Soc Transact 2000;28: 517–520.
66. Zwahlen C, Li SC, Kay LE, Pawson T, Forman-Kay JD. Multiple modes of peptide recognition by the PTB domain of the cell fate determinant Numb. EMBO J 2000;19:1505–1515.
67. Tondi D, Slomczynska U, Costi MP, Watterton DM, Ghelli S, Shoichet BK. Structure-based discovery and in-parallel optimization of novel competitive inhibitors of thymidylate synthase. Chem Biol 1999;6:319–331.
68. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Folding and binding cascades: dynamic landscapes and population shifts. Prot Sci 2000;9:10–19.
69. Ma B, Wolfson H, Nussinov R. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. Curr Opin Struct Biol 2001;11:364–369.
70. van Regenmortel MHV. Molecular recognition in the post-reductionist era. J Mol Recog 1999;12:1–2.
71. Tsai CJ, Kumar S, Ma, B, Nussinov R. Folding funnels, binding funnels, and protein function Prot Sci 1999;8:1181–1190.
72. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. Prot Eng 1999;12:713–720.
73. Su AI, Lorber DM, Weston GS, Baase WA, Matthews BW, Shoichet BK. Docking molecules by families to increase the diversity of hits in database screens: computational strategy and experimental evaluation. Proteins 2001;42:279–293.
74. Brem R, Dill KA. The effect of multiple binding modes on empirical modeling of ligand docking to proteins. Prot Sci 1999;8: 1134–1143.
75. Lamb ML, Jorgensen WL. Computational approaches to molecular recognition. Curr Opin Chem Biol 1997;1:449–457.
76. Jiang J, Kim SH. 'Soft Docking' matching of molecular surface cubes. J Mol Biol 1991;219:79–102.
77. Gardiner EJ, Willett P, Artymiuk PJ. Protein docking using a genetic algorithm. Proteins 2001;44:44–56.
78. Vakser IA, Matar OG, Lam CF. A systematic study of low resolution recognition in protein-protein complexes. Proc Natl Acad Sci 1999;96:8477–8482.
79. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990;265:16027–16030.
80. Najmanovitch R, Kuttner J, Sobolev V, Edelman M. Side-chain flexibility in proteins upon ligand binding. Proteins 2000;39:261–268.
81. Rydel TJ, Tulinsky R, Bode W, Huber R. Refined structure of the Hirudin-Thrombin complex. J Mol Biol 1991;221:583–601.
82. Janin J, Cherfils J. 1993 Protein docking algorithms: simulating molecular recognition. Curr Opin Struct Biol 1993;3:265–269.
83. Vondrasek J, Wlodawer A. New database. Science 1996;272:337–338.
84. Vondrasek J, van Buskirk C, Wlodawer A. Database of three dimensional structures of HIV proteinases. Nat Struct Biol 1997;4:8.
85. Yamazaki T, Hinck AP, Wang YX, Nicholson LK, Torchia DA, Wingfield P, Stahl SJ, Kaufman JD, Chang CH, Domaille PJ, Lam PY. Three-dimensional solution structure of the HIV-1 protease complexed with DMP323, a novel cyclic urea-type inhibitor, determined by nuclear magnetic resonance spectroscopy. Prot Sci 1996;5:495–506.
86. Wittaker M. Discovery of protease inhibitors using targeted libraries. Curr Opin Chem Biol 1998;2:386–396.
87. Dolle RE. Comprehensive survey of combinatorial library synthesis. J Comb Chem 1999;2:383–433.
88. Feeney J. NMR studies of ligands binding to dihydrofolate reductase. Angew Chem Int Ed 2000;39:290–312.
89. Philippopoulos, M. & Lim, C. (1999). Exploring the dynamic information content of a protein NMR structure: Comparison of a molecular dynamics simulation with the NMR and X-ray structures of *Escherichia coli* ribonuclease H1. Proteins 36:87–110.
90. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips, Jr GN. A sampling problem in molecular dynamics simulations of macromolecules. Proc Natl Acad Sci USA 1995;92:3288–3292.
91. Totrov M, Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins 1997; 1(Suppl):215–220.
92. Bloomers MJ, Lucasius CB, Kateman G, Kaptein R. Conformational analysis of a dinu-cleotide photodimer with the aid of genetic algorithm. Biopolymers 1992;32:45–52.
93. LeGrand S, Merz K. The application of the genetic algorithm to conformational search. FASEB J 1992;6:A132
94. Payne AWR, Glen RC. Molecular recognition using a binary genetic search algorithm. J Mol Graph 1993;11:74–91.
95. Knegtel RMA, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures. J Mol Biol 1997;266:424–440.
96. Sudbeck EA, Mao C, Venkatachalam TK, Tuel-Ahlgren L, Uckun FM. Structure-based design of novel dihydroalkoxybenzyloxopyrimidine derivatives as potent nonnuclease inhibitors of the human immunodeficiency virus reverse transcriptase. Antimicrob Agents Chemother 1998;42:3225–3233.
97. Broughton HB. A method for including protein flexibility in protein-ligand docking: Improving tools for database mining and virtual screening. J Mol Graph 2000;18:247–257.
98. Miller MD, Kearsy SK, Underwood DJ, Sheridan RP. FLOG: A system to select 'quasi-flexible' ligands complementary to a receptor of known three dimensional structure. J Comput-Aided Mol Des 1994;8:153–174.
99. Kramer B, Metz G, Rarey M, Lengauer T. Ligand docking and screening with FlexX. Med Chem Res 1999;7/8:463–478.
100. Bouzida D, Rejto PA, Arthurs S, Colson AB, Freer ST, Gehlhaar DK, Larson V, Luty BA, Rose PW, Verkhiver GM. Computer simulations of ligand-protein binding with ensembles of protein conformations: a Monte Carlo study of HIV-1 protease binding energy landscape. Int J Quant Chem 1999;72:73–84.
101. Sandak B, Nussinov R, Wolfson H J. An automated computer-vision and robotics based technique for 3-D flexible biomolecular docking and matching. Comp Appl BioSci 1995;11:87–99.
102. Sandak B, Wolfson HJ, Nussinov R. Hinge-bending at molecular interfaces: Automated docking of a dihydroxyethylene-containing inhibitor of the HIV-1 protease. J Biomol Struct Dyn Proceedings of the Ninth Conversation, Sarma RH, Sarma MH, editors. New York: Adenine Press, 1996;1:233–252.
103. Sandak B, Nussinov R, Wolfson HJ. Docking of conformationally flexible proteins. Seventh Symposium on Combinatorial Pattern Matching, Laguna Beach, California. Lecture Notes in Computer Science. New York: Springer Verlag 1996;1075:271–287.
104. Sandak B, Wolfson HJ, Nussinov R. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers. Proteins 1998;32:159–174.
105. Sandak B, Nussinov R, Wolfson HJ. A Method for biomolecular structural recognition and docking allowing conformational flexibility. J Comput Biol 1999;5:631–654.
106. Leach AR. Ligand docking to proteins with discrete side-chain flexibility. J Mol Biol 1994;235:345–356.
107. Leach AR, Lemon A. Exploring the conformational space of protein side-chains using dead-end elimination and the A* algorithm. Proteins 1998;33:227–239.
108. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267–1289.
109. Dunbrack RL Jr, Karplus M. Backbone-dependent Rotamer Library for Proteins: application to side-chain prediction. J Mol Biol 1993;230:543–574.

110. Schaffer L, Verkhiver GM. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain conformations. Proteins 1998;33:295–310.
111. Apostolakis J, Plucktun A, Caflisch A. Docking small ligands in flexible binding sites. J Comp Chem 1998;19:21–37.
112. Gabb HA, RM Jackson, Sternberg MJE. Modeling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
113. Palma PN, Krippahl L, Wampler JE, Moura JJG. BIGGER: a new soft docking algorithm for predicting protein interactions. Proteins 2000;39:178–194.
114. Zhao S, Goodsell DS, Olson AJ. Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. Proteins 2001;43:271–279.
115. Jackson RM, Gabb HA, Sternberg MJE. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
116. Lomber DM, Shoichet BK. Flexible ligand docking using configurational ensembles. Prot Sci 1998;7:938–950.
117. Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. Proteins 1990;8:195–202.
118. Caflisch A, Niederer P, Anliker M. Monte Carlo docking of oligopeptides to proteins. Proteins 1992;13:223–230.
119. Stoddard BL, Koshland DE. Prediction of the structure of a receptor–protein complex using binary docking method. Nature 1992;358:774–776.
120. Wasserman ZR, Hodge CN. Fitting an inhibitor into the active site of thermolysin: a molecular dynamics study. Proteins 1996; 24:227–237.
121. DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID, Venkataraghavan R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three dimentional structure. J Med Chem 1988;31:722–729.
122. Leach AR, Kuntz ID. Conformational analysis of flexible ligands in macromolecular receptor sites. J Comput Chem 1992;13:730–748.
123. Mirankar A, Karplus M. Functionality maps of binding sites: a multicopy simultaneous search method. Proteins 1991;11:29–34.
124. Mirankar A, Karplus M. An automated method for dynamic ligand design. Proteins 1995;23:472–490.
125. Moon JB, Howe WJ. Computer design of bioactive molecules: a method for receptor-based de novo ligand design. Proteins 1991; 11:314–328.
126. Bohm HJ. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput-Aided Mol Des 1992;6:61–78.
127. Lewis RA, Roe DC, Huang C, Ferrin TE, Langridge R, Kuntz ID. Automated site-directed drug design using molecular lattices. J Mol Graph 1992;1:66–78.
128. Ho CMW, Marshall GR. Foundation: A program to retrieve all possible structures containing a user-defined minimum number of matching elements from three-dimensional databases. J Comput-Aided Mol Des 1993;7:3–22.
129. Klebe G, Mietzner T. A fast and efficient method to generate biologically relevant conformations. J Comput-Aided Mol Des 1994;8:583–606.
130. Oshiro CM, Kuntz ID, Dixon JS. Flexible ligand docking using a genetic algorithm. J Comput-Aided Mol Des 1995;9:113–130.
131. Verkhiver GM, Rejto PA, Gehlhaar DK, Freer ST. Exploring the energy landscapes of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes. Proteins 1996;25:342–353.
132. Jones G, Willet P, Glen R, Leach A, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
133. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem 1998;19:1639–1662.
134. Schnecke V, Swanson CA, Getzoff ED, Kuhn LA. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. Proteins 1998;33:74–87.
135. Schnecke V, Kuhn LA. Virtual screening with solvation and ligand-induced complementarity. Perspect Drug Discov Des 2000; 20:171–190.
136. Shoichet BK, Bodian DL, Kuntz ID. Molecular docking using shape descriptors. J Comp Chem 1992;13:380–397.
137. Miller DW, Dill KA Ligan binding to proteins: the binding landscape model Prot Sci 1997;6:2166–2179.
138. Tovchigrechko A, Vakser IA. How common is is the funnel-like energy landscape in protein–protein interactions? Prot Sci 2001; 10:1572–1583.
139. Ten Eyck LF, Mandell J, Roberts VA, Pique ME. Surveying molecular interactions with DOT. In: Hayes A, Simmons M, editors. Proceedings of the 1995 ACM/IEEE Supercomputing Conference. New York: ACM Press; 1995.
140. Roberts VA, Pique ME. Definition of the interaction domain for cytochrome c on cytochrome c oxidase. J Biol Chem 1999;274: 38051–38060.
141. Wolfson HJ, Lamdan Y. Geometric hashing: A general and efficient model-based recognition scheme. In: Proceedings of the IEEE Int Conf on Computer Vision Tampa, FL. 1988;238–249.
142. 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.
143. Walls PH, Sternberg MHJ. New algorithm to model protein-protein recognition based on surface complementarity. J Mol Biol 1992;228:277–297.
144. Vakser IA. Protein docking for low resolution structures. Prot Eng 1995;8:371–377.
145. Vakser IA. Low resolution docking: prediction of complexes for underdetermined structures. Biopolymers 1996;39:455–464.
146. Vakser IA. Main-chain complementarity in protein recognition. Prot Eng 1996;9:741–744.
147. Vakser IA, Afialo C. Hydrophobic docking a proposed enhancment to molecular recognition techniques. Proteins 1994;20:320–329.
148. Camacho JC, Gatchell DW, Kimura SR, Vajda S. Scoring dockied conformations generated by rigid body protein protein docking. Proteins 2000;40:525–537.
149. Singh AP, Latombe JC, Brutlag DL. 1999. A motion planning approach to flexible ligand binding. Proceedings of the 7th Conference on Intelligent Systems in Molecular Biology (ISMB). Menlo Park, CA: AAAI Press, 1999; p 252–261.
150. Balbes LM, Mascarella SW, Boyd DB. A perspective of modern methods in computer-aided drug design. In: Lipkowitz KB, Boyd DB, editors. Reviews in computational chemistry 1994; p 337–378.
151. Gilson MK, Honig B. Calculation of electrostatic potentials in an enzyme active site. Nature 1987;330:84–86.
152. Laskowski RA. SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. J Mol Graph 1995;13:323–330.
153. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Prot Sci 1996;5:2438–2452.
154. Frommel C, Peters KP, Fauck J. The automatic search for ligand binding sites in proteins of known three dimentional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
155. Bliznyuk AA, Gready JE. Simple method for locating possible ligand binding sites on protein surfaces. J Comp Chem 1999;20: 983–988.
156. Pettit FK, Bowie JU. Protein surface roughness and small molecular binding sites. J Mol Biol 1999;285:1377–1382.
157. Nussinov R, Wolfson H. Efficient detection of three-dimentional structural structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88: 10495–10499.
158. Morelli XJ, Palma PN, Guerlesquin F, Rigby AC. A novel approach for assessing macro-molecular complexes combining soft-docking caculations with NMR data Prot Sci 2001;10:2131–2137.
159. Kohlbacker O, Burchardt A, Moll A, Hildebrandt A, Bayer P, Lenhof HP. Structural prediction of protein complexes by an NMR-based protein docking algorithm. J Biomol NMR 2001;20: 15–21.
160. Hoffmann D, Kramer B, Washio T, Steinmetzer T, Rarey M, Lengauer T. Two-stage method for protein-ligand docking. J Med Chem 1999;42:4422–4433.
161. Wang J, Morin P, Wang W, Kollman PA. Use of MM-PBSA in reproducing the binding Freee Energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of

efavirenz by docking and MM-PBSA. J Am Chem Soc 2001;123; 5521–5230.
162. Haraki KS, Sheridan RP, Venkataraghavan R, Dunn DA. 1990 Looking for pharmacophores in 3-D databases: does conformational searching improve the yield of actives? Tetrahedron Comput Methodol 1990;3:365–573.
163. Murrall NW, Davies EK. Conformational freedom in 3-D databases Techniques. J Chem Inf Comput Sci 1990;30:312–316.
164. Guner OF, Henry DR, Pearlman RS. Use of flexible molecules in databases of three-dimensional structures. J Chem Inf Comput Sci 1992;32:101–109.
165. Willett P, Clark DE, Jones G. Pharmacophoric pattern matching in files of three-dimensional chemical structures: comparison of conformational-searching algorithms for flexible searching. J Chem Inf Comput Sci 1994;34:197–206.
166. Mizutani MY, Tomioka N, Itai A. Rational automatic search method for stable docking models of protein and ligand. J Mol Biol 1994;243:310–326.
167. Rarey M, Kramer B, Lengauer T. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261: 470–489.
168. Rarey M, Wefing S, Lengauer T. Placement of medium-sized molecular fragments into active sites of proteins. J Comp-Aided Mol Des 1996;10:41–54.
169. Rarey M, Kramer B, Lengauer T. Docking of hydrophobic ligands with interaction-based matching algorithms. Bioinformatics 1999; 15:243–250.
170. Di Nola A, Roccatano D, Berendsen HFC. Molecular dynamics simulations of the docking of substrates to proteins. Proteins 1994;19:174–182.
171. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol 1996;6:402–406.
172. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 1996;16:3–50.
173. Leibowitz N, Fligelman Z, Nussinov R, Wolfson HJ. An automated multiple structure alignment and detection of a common substructural motif. Proteins 2001;43:235–245.
174. Brint AT, Willet P. Pharmacophoric pattern matching in files of three-dimensional chemical structures: comparison of geometric searching algorithms. J Mol Graph 1987;5:49–56.
175. Smellie A, Crippen G, Richards WG. Fast drug receptor mapping by site-directed distances: A novel method for predicting new pharmacological leads. J Chem Inf Comput Sci 1991;31:386–394.
176. Crandell CW, Smith DH. Computer assisted examination of compounds for common three-dimensional substructures. Chem Inf Comput Sci 1983;23:186–197.
177. Holliday JD, Willet P. Using a genetic algorithm to identify common structural features in sets of ligands. J Mol Graph Model 1997;15:221–232.
178. Finn PW, Kavarki LE, Latombe LC, Motwani R, Shelton C, Venkatasubramanian S, Yao A. RAPID: randomized pharmacophore identification for drug design. Comput Geom ACM 1997;97: 324–333.
179. Rigoutsos I, Platt D, Califano A. Flexible 3D-substructure matching and novel conformer derivation in very large databases of 3D-molecular information. IBM Research Division. Yorktown Heights, NY: T.J. Watson Research Center, 1996.
180. Stockman G. Object recognition and localization via pose clustering. J Comput Vis, Graph Image Process 1987;40:361–387.
181. Shatsky M, Fligelman Z, Nussinov R, Wolfson H. Alignment of flexible protein structures. In: Altman et al., editors. Proceedings of the 8th Conference on Intelligent Systems in Molecular Biology (ISMB). Menlo Park, CA: AAAI Press, 2000; p 329–343.
182. Shatsky M. Alignment of flexible protein structures. M.Sc. Thesis, Tel Aviv University, 2001.
183. Bures MG. The Discovery of novel Auxin transport inhibitors by molecular modeling and three dimentional pattern analysis. J Comput-Aided Mol Des 1991;5:323–334.
184. Martin YC. 3D database searching in drug design. J Med Chem 1992;35:2145–2154.
185. Miller MD, Sheridan RP, Kearsy SK. SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. J Med Chem 1999;42:1505–1514.
186. Fradera X, Knegtel RMA, Mestres J. Similarity driven flexible ligand docking. Proteins 2000;40:623–636.
187. Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL, Lins RD, Briggs JM, McCammon JA. Developing a dynamic pharmacophore model for HIV-1 integrase. J Med Chem 2000;43: 2100–2114.
188. Rose RB, Craik CS, Stroud RM. Domain flexibility in retroviral proteases: structural implications for drug resistant mutations. Biochemistry 1998;37:2607–2621.
189. Harison SC. A structural taxonomy of DNA-binding domains. Nature 1991;353:715–719.
190. Larson CL, Verdine GL. 1996. The chemistry of protein-DNA interactions bioorganic chemistry: nucleic acids (Hecht SM ed.) New York: Oxford University Press, 1996; p 324–346.
191. Jones S, Heyningen PV, Berman HM, Thorton JM. 1999 Protein-DNA interactions: a structural analysis. J Mol Biol 1999;287:877–896.
192. Pabo CO, Nekludova L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol 2000;303:597–624.
193. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, Schneider B. The Nucleic Acid Database A comprehensive relational database of three-dimentional structures of nucleic acids. Biophys J 1992;63:751–759.
194. Drew HR, Wing RM, Takano T, Broka C, Tanaka S, Itakura K, Dickerson RE. Structure of a b-DNA dodcamer-conformation and dynamics. Proc Natl Sci USA 1981;78:2179–2183.
195. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427.
196. Knegtel RM, Boelens R, Kaptein R. Monte Carlo docking of protein-DNA complexes: incorporation of DNA flexibility and experimental data. Prot Eng 1994;7:761–767.
197. Nagaich AK, Zhurkin VB, Durell SR, Jernigan RL, Appella E, Harrington RE. p53-induced DNA bending and twisting: p53 tetramer binds on the outer side of a DNA loop and increases DNA twisting. Proc Natl Acad Sci USA 1999;96:1875–1880.
198. Durell SR, Jernigan RL, Appella E, Nagaich AK, Harrington RE, Zhurkin VB. DNA Bending induced by tetrameric binding of the p53 tumor suppressor protein: steric constraints on conformation. In: Sarma RH, Sarma MH, editors. Structure, motion, interaction and expression of biological macromolecules. Proceedings of the Tenth Conversation, 1997. New York: Adenine Press, 1998, Vol. 2, p 277–296.
199. Zacharias M, Sklenar H. Harmonic modes as variables to approximately account for receptor flexibility in ligand-receptor docking simulations: application to DNA minor groove ligand complex. J Comp Chem 1999;20:287–300.
200. Cherfils J, Duquerroy S, Janin J. Protein-protein recognition analyzed by docking simulations. Proteins 1991;11:271–280.
201. Cherfils J, Janin J. Protein docking algorithms: simulating molecular recognition. Curr Opin Struct Biol 1993;3;265–269.
202. Shoichet B, Kuntz I. Protein docking and complementarity. J Mol Biol 1991;221;327–346.
203. Wang H. Grid-search molecular accessible algorithm for solving the protein docking problem. J Comp Chem 1991;12;746–750.
204. Norel R, Lin SL, Xu D, Wolfson H, Nussinov R. Molecular surface variability and induced conformational changes upon protein-protein association. In: Sarma RH, Sarma MH, editors. Structure, motion, interaction and expression of biological macromolecules. Adenine Press, Albany: Adenine Press, 1998; p 33–51.
205. 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.
206. Wallqvist A, Covell DG. Docking enzyme-inhibitor complexes using a preference based free surface. Proteins 1996;25:403–419.
207. Kasinos N, Lilley G, Subbarao N, Haneef I. A Robust and efficient automated docking algorithm for molecular recognition. Pharmacol Theraapeut 1999;84;179–191.
208. Jackson RM. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. Prot Sci 1999;8:603–613.
209. Meyer M, Wilson P, Schomburg D. Hydrogen bonding and molecular surface complementarity as a basis for protein docking. J Mol Biol 1996;264:199–210.
210. Weng Z, Vajda S, DeLisi C. Prediction of complexes using empirical free energy functions. Prot Sci 1996;5:614–626.
211. Yue SY. Distance constrained molecular docking by simulated annealing. Prot Eng 1990;4:177–184.

212. Abagyan R, Argos P. Optimal protocol and trajectory visualization for conformational searches of peptide and proteins. J Mol Biol 1992;225:519–532.
213. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235:983–1002.
214. Totrov M, Abagyan R. Detailed ab initio prediction of lysozyme-antibody complex with a 1.6A accuracy. Nat Struct Biol 1994;4: 259–263.
215. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615.
216. Abagyan R, Totrov M, Kuznetsov D. ICM: a new method for structure modeling and design: application to docking and structure prediction from distorted native conformation. J Comp Chem 1994;15:488–506.
217. Wolfson HJ. Generalizing the generalized Hough transform. Pattern Recog Lett 1991;12:565–573.
218. Hayward S, Kitao A, Berendsen HJC. Model-free methods of analyzing domain motions in proteins from simulations of lysozyme. Proteins 1997;27:425–437.
219. Bahar I, Erman B, Jernigan RL, Atilgan AR, Covell DG. Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. J Mol Biol 1999;285:1023–1037.
220. Baldwin RL, Rose GD. Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 1999;24:26–33.
221. Baldwin RL, Rose GD. Is protein folding hierarchic? II. Folding intermediates and transition states. Trends Biochem Sci 1999;24: 77–84.
222. Tsai CJ, Maizel JV, Nussinov R. Anatomy of protein structures: visualizing how a 1D protein chain folds into a 3D shape. Proc Natl Acad Sci USA 2000;97:12038–12043.
223. Pelletier JE, Campbell-Valois FX, Michnick SW. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. Proc Natl Acad Sci USA 1998;95:12141–12146.
224. Ostermeier M, Benkovic SJ. Finding Cinderella's slipper-protein that fit. Nature Biotechnol 1999;17:639–640.
225. Gegg CV, Bower KE, Mattews CR. Probing minimal independent folding units in dihydrofolate reductase. Prot Sci 1997;6:1885–1892.
226. Sham YY, Ma B, Tsai CJ, Nussinov R. Molecular dynamics simulation of Escherichia coli dihydrofolate reductase and its protein fragments: relative stabilities in experiment and simulations. Prot Sci 2001;10:135–148.
227. Pearlman DA, Charifson, PS. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system J Med Chem 2001;44:3417–3423.
228. Moont G, Gabb HA, Sternberg MJE. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
229. Camacho JC, Weng Z, Vajda S, DeLisi C. Free energy landscapes of encounter in protein-protein association. Biophys J 1999;76: 1166–1178.
230. Norel R, Sheinerman F, Petrey D, Honig B. Electrostatic contributions to protein-protein interactions: fast energetic filters for docking and their physical basis. Prot Sci 2001;10:2147–2161.
231. Leibowitz N, Nussinov R, Wolfson HJ. MUSTA: a general, efficient automated method for multiple structure alignment and detection of a common motif. J Comp Biol 2001;8:93–121.
232. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
233. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–386.
234. Clackson T, Ultsch MH, Wells JA, de Vos AM. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 1998;277:1111–1128.
235. Wallis R, Leung KY, Osborne M, Moore GR, James R, Kleanthous C. Specificity in protein-protein recognition: conserved Im9 residues are the major determinants of stability in the colicin E9 DNase-Im9 complex. Biochemistry 1998;37:476–485.
236. Li W, Hamill SJ, Hemmings AM, Moore GR, James R, Kleanthous C. Dual recognition and the role of specificity-determining residues in colicin E9 DNase-immunity protein interactions. Biochemistry 1998;37:11771–11779.
237. Kuhlmann UC, Pommer AJ, Moore GR, James R, Kleanthous C. Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. J Mol Biol 2000;301, 1163–1178.
238. Hu Z, Ma B, Wolfson W, Nussinov R. Conservation of polar residues as hot spots at protein-protein interfaces. Proteins 2000;39:331–342.
239. Gardiner EJ, Willett P, Artymiuk PJ. Graph-theoretic techniques for macromolecular docking. J Chem Inform Comput Sci 2000;40:273–279.
240. Greer J, Bush BL. Macromolecular shape and surface maps by solvent exclusion. Proc Natl Acad Sci USA 1978;75:303–307.
241. Soblev V, Wade RC, Vriend G, Edelman M. Molecular docking using surface complementarity. Proteins 1996;25:120–129.
242. Hou T, Wang J, Chen L, Xu X. Automated docking of peptides and proteins by using genetic algorithm combined with a tabu search. Prot Eng 1999;12:639–647.
243. Bacon DJ, Moult J. Docking by least squares fitting of molecular surface patterns. J Mol Biol 1992;225:849–858.
244. Tsai CJ, Ma B, Sham Y, Kumar S, Wolfson H, Nussinov R. A hierarchical, building-blocks based computational scheme for protein structure prediction. IBM J Res Dev Life Sci 2001;45:513–523.
245. Barlow BJ, Thornton JM. Ion-pairs in proteins. J Mol Biol 1983;168:867–885.
246. Nielsen JE, Andersen KV, Honig B, Hooft RWW, Klebe G, Vriend G, Wade RC. Improving macromolecular electrostatics calculations. Prot Eng 1999;12:657–662.
247. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. Proc Natl Acad Sci USA 1999;96:9997–10002.
248. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Prot Sci 1994;3:717–729.
249. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Prot Sci 1997;6:53–64.
250. Singh J, Thornton JM. Atlas of protein side chain interactions. Oxford: IRL Press, 1992.
251. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol 1997;267:207–222.
252. Wallqvist A, Covell DG, Jernigan RL. A preference based free energy parameterization of enzyme inhibitor binding. Applications to HIV-1 protease inhibitor design. Prot Sci 1995;4:1881–1903.
253. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
254. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
255. Sharp K, Fine R, Honig B. Computer simulations of the diffusion of a substarte to an active site of an enzyme. Science 1987;236: 1460–1463.
256. Schreiber G, Fersht AR. Rapid, electrostatically assisted association of proteins. Nature Struct Biol 1996;3:427–431.
257. Nicholls A, Honig B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. J Comput Chem 1991;12:435–445.
258. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA, Lin A, Antosiewicz J, Gilson MK, Bagheri B, Scott LR, McCammon JA. Electrostatics and diffusion of molecules in solution: simulations with the university of Houston Brownian dynamics program. Comput Phys Commun 1995;91:57–95.
259. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42:5100–5109.
260. Terp GE, Johansen BN, Christensen IT, Jorgensen FS. A new concept for multidimensional selection of ligand conformations (multiselect) and multidimensional scoring (multi-score) of protein-ligand binding affinities. J Med Chem 2001;44:2333–2343.
261. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein–ligand interactions. J Mol Biol. 2000;295:337–356.
262. Logean A, Sette A, Rognan D. Customized versus universal scoring functions: application to class I MHC-peptide binding

free energy predictions. Bioorg Med Chem Lett 2001;11:675–679.
263. Aloy P, Querol E, Aviles FX, Sternberg MJE. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
264. Gilson MK, Honig B. Calculation of the total electrostatic energy of a macro-molecular system: solvation energies, binding energies and conformational analysis. Proteins 1998;4:7–18.
265. Verbitsky G, Nussinov R, Wolfson HJ. Structural comparison allowing hinge bending, swiveling motions. Proteins 1999;33:232–254.
266. MacCullum RM, Martin ACR, Thornton JM. Protein–protein recognition. J Mol Biol 1996;262:732–745.
267. Bohm HJ, Klebe G. What can we learn from molecular recognition in protein–ligand complexes for the design of new drugs? Angew Chem Int 1996;35:2588–2614.
268. Padlan EA. On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands. Proteins 1990;7:112–124.
269. Peters KP, Fauck J, Frommel C. The automatic search for ligand binding sites in proteins in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 1996;256:201–213.
270. Hart TN, Read RJ. A multiple start Monte Carlo docking method. Proteins 1992;13:206–222.
271. Leach AR, Klein TE. A molecular dynamics study of the inhibitors of dihydrofolate reductase by a phynyl triazine. J Comput Chem 1995;16:1378–1393.
272. Wells JA. Systemic mutational analysis of protein–protein interfaces. Methods Enzymol 1991;202;390–411.
273. Dunbrack RL, Gerloff DL, Bower M, Chen X, Lichtarge O, Cohen FE. Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (CASP2) Asilomar, California, December 13–16, 1996. Fold Des 1997;2: R27–R42.
274. Strynadka NC, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Tortov M, Janin J, Cherfils J, Jackson R, Sternberg MJE, James MNG. Molecular docking programs successfully predict the binding of a b-lactamese. Nature Struct Biol 1996;3:233–239.
275. Maiorov V, Abagyan R. A new method for modeling large-scale rearrangements of protein domains. Proteins 1997;27:410–424.
276. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
277. Cummings MD, Hart TN, Reader RJ. Atomic solvation parameters in the analysis of protein protein docking results. Prot Sci 1995;4:2087–2099.
278. Lawrence MC, Colman PM. Shape complementarity at protein-protein interfaces. J Mol Biol 1993;234:946–950.
279. Freire E. Statistical thermodynamic linkage between conformational and binding equilibria. Adv Prot Chem 1998;51:255–279.
280. Sundberg EJ, Mariuzza RA. Luxury accommodations: the expanding role of structural plasticity in protein-protein interaction. Structure 2000;8:R137–R142.
281. Diller DJ, Merz KM. High throughput docking for library design and library prioritization. Proteins 2001;43:113–124.
282. Lamb ML., Burdick KW., Toba S, Young MM, Skillman AG, Zou X, Arnold JR, Kuntz ID. Design, docking, and evaluation of multiple libraries against multiple targets. Proteins 2001;42:296–318.
283. Chen YZ, Zhi DG. Ligand-protein inverse docking and its potntial use in the computer search of protein targets of a small molecule. Proteins 2001;43:217–226.

APPENDIX A
DRUF: Docking Results Unified Format

The presentation of results by the different methods is quite variable. For example, there are at least 32 different theoretical methods for calculation of a solution rmsd with regard to the complex. The superposition can refer to both the receptor and the ligand or the ligand alone, to all atoms or only C_α atoms. After the transformation of the predicted relative to the bound complex is computed, the rmsd can be calculated for the ligand alone or for both receptor and the ligand, the entire molecule or only the interface, for all atoms or only C_α. Overall, there are 5 parameters with 2 options for each of them. Not all of these methods make sense, and accordingly not all of them are used.

In order to enable comparison between docking methods and eliminate the need for repeating detailed protocols, a Docking Results Unified Format (DRUF) is suggested below. We recommend that the rmsd calculation should be performed by superimposing the C_α atoms of the unbound receptor on the bound receptor. C_α atoms of the entire ligand molecule will be considered for distance measurements. DRUF includes different parameters. The parameters are divided into 3 groups: simple rmsd data, relative rmsd data, residue contacts, and binding residue prediction data.

Simple rmsd data

1. Highest rank rmsd: The rmsd of the highest-ranking solution with rmsd less than 5Å from the complex.
2. Highest rank: The rank of the highest-ranking solution with rmsd less than 5Å. The highest rank parameter together with the highest rank rmsd is an indication of the optimal performance that can be expected from the docking scheme.
3. Best rmsd: The rmsd of the lowest rmsd solution. The best rmsd parameter is an indication of the optimal performance that can be expected from the search stage.
4. Rank of best rmsd: The rank of the lowest rmsd solution. This parameter, like the complex rank parameter, is an indication of the scoring function quality
5. N10: The number of solutions with rmsd less than 3A among the 10 highest ranked solutions. N10, like N50 and N100, is the most accurate parameter for evaluation of the entire docking scheme, both search and scoring stages. It is more stable than the highest and the best parameters since it refers to a fixed number of solutions as opposed to a single solution out of an uneven number of solutions. The total solution number of different algorithms can differ significantly. The number of solutions can vary up to an order of magnitude for different docking cases using the same docking scheme. For example, the number of solutions after clustering varied between 1,200 to 11,475 using the same Shape Complementarity scheme (Norel et al.[40]). This typically depends on the molecular size.
6. N50: The number of solutions with rmsd less than 4A among the 50 highest ranked solutions.
7. N100: The number of solutions with rmsd less than 5A among the 100 highest ranked solutions. Here, and in the N10 and N50, the number of solutions can be replaced by the number of clusters.

Relative rmsd data

1. Complex rank: The rank of the known complex. This parameter indicates more than any other parameter

the scoring function quality. Unlike N10, N50, and N100, this parameter does not depend on the quality of the search stage, but on the goodness of the emerging solutions.

2. Unbound rmsd: The rmsd of the unbound ligand after superimposition on the bound ligand. This parameter is indicative of the difficulty degree of the studied case. A high unbound rmsd indicates significant changes between the bound and the unbound states.
3. Hyper highest rank rmsd: The difference between the highest rank rmsd and the unbound rmsd. The hyper-highest rank rmsd parameter is more informative than the highest rank rmsd. Since each unbound case differs from the bound case to a different extent (reflected by the unbound rmsd parameter), one can expect to find differences in the highest rmsd parameter as a function of the unbound rmsd. High unbound rmsd is expected to accompany high values of the highest rank rmsd. The hyper highest rank rmsd is the normalized parameter.
4. Hyper best rmsd: The difference between the best rmsd and the unbound rmsd. The hyper-best rmsd is the normalized version of the best rmsd parameter just as the hyper-highest rank rmsd is the normalized version of the highest rank rmsd parameter.

Binding residue prediction data and C_α *contacts* at a threshold distance are also measures to be considered.

The number of solutions and the ranking achieved in some common methods, where data are available on the same cases, are given in Table IIIa (for the bound cases) and in Table IIIb for the unbound cases.

APPENDIX B
Available Data Base for Protein-Protein Predictive Docking

One of the major problems regarding predictive docking is the limited number of complex structures. Currently there are about 100 protein complexes in the PDB. Only 39 of them have an unbound structure (either native or pseudo native) of at least one of the components in the complex, and only 8 of them have an unbound structure of both complex components (Janin and Chothia, 1990[79]). The available cases for predictive heteroprotein docking are summarized in Table IVa,b. The available complexes can be divided into four groups: (1) Enzyme-Inhibitor (EI); (2) Antibody-Antigen (AA); (3) Subunit–Subunit (SS); (4) Receptor-Ligand (RL). The division into these groups is essential since complexes in different groups exhibit intrinsically different interaction characteristics. It has been suggested that EI and AA complexes represent two different classes of binding (Lawrence and Colman[278]). According to Jackson,[208] EI complexes and AA complexes differ in the interaction mechanism, the residue types that contribute to the interaction, and the binding affinity. EI complexes interact through a main-chain–main-chain mechanism, with six types of residues contributing 70% of the interaction energy, and their binding affinity is of nanomolar order. AA complexes interact through a side-chain–main-chain mechanism. Diversified types of residues contribute to the interaction energy, and their binding affinity is of fentomolar order. Another major difference between EI and AA complexes is the role of shape correlation. Enzymes and their inhibitors have coevolved to form an interface with a high degree of surface complementarity. On the other hand, the immune system produces many different antibodies in response to an antigen, some of which bind quite poorly. So a particular AA complex does not necessarily possess the best possible binding interface. Shape correlation may not be as important in AA complexes (Jackson et al.[115]). This might provide an explanation for the success of all research groups in the first docking challenge of b-lactamase/inhibitor complex (Strynadka et al.[279]), in contrast to the collective failure in the second docking challenge of antibody/haemagglutinin complex (Dunbrack et al.[273]). The complexes are not equally dispersed between these groups. Most of the complexes are EI complexes, mainly serine protease inhibitor complexes.