Protein–protein docking with multiple residue conformations and residue substitutions

DAVID M. LORBER,[1] MARIA K. UDO,[1,2] AND BRIAN K. SHOICHET[1]

[1]Northwestern University, Department of Molecular Pharmacology and Biological Chemistry,
Chicago, Illinois 60611, USA
[2]Loyola University Chicago, Department of Physics, Chicago, Illinois 60626, USA

(RECEIVED July 18, 2001; FINAL REVISION March 4, 2002; ACCEPTED MARCH 5, 2002)

Abstract

The protein docking problem has two major aspects: sampling conformations and orientations, and scoring them for fit. To investigate the extent to which the protein docking problem may be attributed to the sampling of ligand side-chain conformations, multiple conformations of multiple residues were calculated for the uncomplexed (unbound) structures of protein ligands. These ligand conformations were docked into both the complexed (bound) and unbound conformations of the cognate receptors, and their energies were evaluated using an atomistic potential function. The following questions were considered: (1) does the ensemble of precalculated ligand conformations contain a structure similar to the bound form of the ligand? (2) Can the large number of conformations that are calculated be efficiently docked into the receptors? (3) Can near-native complexes be distinguished from non-native complexes? Results from seven test systems suggest that the precalculated ensembles do include side-chain conformations similar to those adopted in the experimental complexes. By assuming additivity among the side chains, the ensemble can be docked in less than 12 h on a desktop computer. These multiconformer dockings produce near-native complexes and also non-native complexes. When docked against the bound conformations of the receptors, the near-native complexes of the unbound ligand were always distinguishable from the non-native complexes. When docked against the unbound conformations of the receptors, the near-native dockings could usually, but not always, be distinguished from the non-native complexes. In every case, docking the unbound ligands with flexible side chains led to better energies and a better distinction between near-native and non-native fits. An extension of this algorithm allowed for docking multiple residue substitutions (mutants) in addition to multiple conformations. The rankings of the docked mutant proteins correlated with experimental binding affinities. These results suggest that sampling multiple residue conformations and residue substitutions of the unbound ligand contributes to, but does not fully provide, a solution to the protein docking problem. Conformational sampling allows a classical atomistic scoring function to be used; such a function may contribute to better selectivity between near-native and non-native complexes. Allowing for receptor flexibility may further extend these results.

Keywords: Protein–protein docking; mutant; combinatorial; flexibility

Supplemental material: See www.proteinscience.org.

Interactions between proteins are critical in biology, and have been widely studied. With the advent of genome and proteome projects, there is much interest in predicting the structures of protein–protein complexes. This, however, turns out to be difficult, and is often referred to as the "Protein Docking Problem" (Richmond 1984; Connolly 1986). This problem has two aspects: enumeration of possible states, and evaluation of their complementarity.

Since the early 1990s, docking programs have been able to regenerate near-native structures of protein–protein complexes using the complexed (bound) conformations of the two proteins (Cherfils et al. 1991; Shoichet and Kuntz 1991;

Reprint requests to: Brian K. Shoichet, Northwestern University, Department of Molecular Pharmacology and Biological Chemistry, 303 E. Chicago Ave., Chicago, Illinois 60611, USA; e-mail: b-shoichet@northwestern.edu; fax: (312) 503-5349.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.2830102.

Hart and Read 1992; Vakser 1995). The protein docking problem only becomes acute when docking the uncomplexed (unbound) conformations of the two proteins; these are the relevant states for true prediction (Totrov and Abagyan 1994; Vakser 1995; Jackson et al. 1998; Norel et al. 1999; Vakser et al. 1999; Camacho et al. 2000; Kimura et al. 2001). Although unbound proteins often adopt main-chain conformations similar to their bound counterparts, their solvent-exposed side chains commonly adopt conformations that are not complementary to their binding partner (Conte et al. 1999). Thus, when docking algorithms generate near-native complexes from the unbound conformations of the partners, atoms in the interface clash. Such near-native fits score poorly in classical, atomistic energy potentials because of these clashing atoms. Even a single noncomplementary atom can lead to very unfavorable energies because of the steepness of the steric repulsion term of the van der Waals energy (Weiner et al. 1984).

The problem of docking unbound proteins can be addressed either by explicitly sampling many conformations or by modifying the scoring function to accommodate clashes. Several methods have been published that use modified scoring functions. Soft docking (Jiang and Kim 1991; Palma et al. 2000), Fourier correlation (Gabb et al. 1997; Ritchie and Kemp 2000), and shape fitting methods (Shoichet and Kuntz 1991; Norel et al. 1994, 1995, 1999), smooth the details of protein–protein interactions, thereby allowing clashes to occur, at least before minimization. Empirically derived, or trained, scoring functions (Weng et al. 1996; Moont et al. 1999; Palma et al. 2000) have also been used to address the protein docking problem. Although these methods allow near-native structures to be identified, they cannot reliably distinguish the near-native complexes from non-native complexes when docking unbound proteins. Vajda and coworkers have explored applying extensive minimization to a series of predocked complexes (Camacho et al. 2000). Although this has shown some promise in discriminating near-native complexes from non-native complexes, the discrimination is not always as convincing as one might like, and the high computational expense of the procedure makes it prohibitive for on the fly docking.

An alternative to avoiding the van der Waals violation problem with a trained or attenuated scoring function is to use a full atomistic scoring function and to consider multiple conformations of the docking proteins. In principle, a very large number of coordinated motions would have to be considered. In fact, the backbones of most proteins remain largely unchanged upon complex formation (Conte et al. 1999), and so one might be able to limit flexibility to side chains. Even assuming a rigid backbone, explicitly sampling the possible ligand side-chain conformations in protein docking might be difficult. For surface-exposed residues where the effect of excluded volume is small, the number of possible conformations increases as the power of the number of rotatable bonds. However, most side chains on the convex surface of a protein, especially a protein ligand, may be considered as independent, uncoupled rotors. This would reduce an exponential problem to one that is additive in the number of flexible side chains.

Here we consider the questions: "Can we discretely sample enough states to approximate the native complex?" and then "Are these near-native docked complexes distinguishable from non-native complexes in an atomistic energy potential?" We make two simplifying assumptions. First, only ligand side chains are flexible; we do not consider conformational changes in the backbone. Second, there is complete additivity among residues; the conformations for each residue are calculated assuming that they are independent of all other flexible residues. These assumptions, along with a hierarchical organization of side-chain conformations, allow us to implicitly consider at least 10^{40} ligand conformations while only explicitly representing hundreds. A simple extension of these ideas allows us to consider residue substitutions (mutant proteins) as well as residue conformations.

Overview of the method

Starting with the unbound ligand (Table 1), we select residues to be made flexible based on their exposed surface area. We then generate conformations for all of the exposed side chains, keeping the rest of the ligand rigid and in the

Table 1. *Docking systems*

	Complex	Resolution (Å)	Receptor	Resolution (Å)	RMSD (Å)[b]	Inhibitor	Resolution (Å)	RMSD (Å)[b]
Trypsin/BPTI	2PTC	1.90	2PTN	1.55	0.34	4PTI	1.50	0.42
α-Chymotrypsin/ovomucoid	1CHO	1.80	5CHA	1.67	0.33	2OVO	1.50	0.78
Protease B/ovomucoid	1SGP	1.40	—	—	—	2OVO	1.50	0.42
TEM-1/BLIP	—[a]	1.70	—[a]	1.80	0.33	—[a]	2.10	0.60
FAB/lysozyme	1VFB	1.80	IVFA	1.80	0.93	132L	1.80	0.87
Barnase/barstar	1B27	2.10	1A2P	1.50	0.63	1A19	2.76	0.52
Subtilisin/Cl-2	2SNI	2.10	2ST1	1.80	0.30	2CI2	2.00	0.46

[a] Natalie Stynadka, personal communcation.
[b] Cα-RMSD between unbound and bound structure.

same orientation. In the docking calculations, the rigid part of the ligand (the backbone, Cβ atoms, and all buried residues) is oriented in the site first. Because all conformations were calculated in the same frame of reference prior to docking, the same rotation matrix can be used to move all the conformations into the frame of reference of the binding site during docking. This approach is similar to a method for multiconformer docking that we have previously described (Lorber and Shoichet 1998). The novelty of the present method comes from the organization of side-chain conformers into a hierarchical data structure. This data structure eliminates the redundancy of the internally rigid parts of the conformations, because only one copy of the rigid part of the protein need be represented for any number of side-chain conformations. More importantly, it organizes the atoms of each side chain so conformers that clash with the receptor can be efficiently pruned off at docking time, and it encodes connectivity information so side chains can be recombined between conformations. There is one key assumption and simplification of this method: complete additivity of side-chain conformations.

A modest extension of this method allows one to not only substitute a particular side-chain conformation with an ensemble of calculated alternate conformations but to substitute one amino acid residue with another. Each of these substituted residues may itself have multiple conformations. The additivity assumption, which allows the recombination of conformations at multiple sites, also allows recombination of residue substitutions at multiple sites with, again, only a small computational cost. In this way, many mutant proteins with many conformations can be docked.

As is typical for algorithms built off of the DOCK suite of programs (Kuntz et al. 1982; Ewing et al. 2001), the protein ligands are first oriented and then scored in the binding site. For each orientation of the ligand in the binding site the internally rigid portion of the ligand is docked, each conformation of each flexible residue of the ligand is fit into the site according to the rotation matrix found for the rigid fragment, and the best conformers of each residue are recombined to create a best-scoring ligand conformation for that orientation. Ligands are evaluated in an atomistic potential function composed of a Poisson-Boltzmann electrostatic term precalculated for the receptor using DelPhi (Gilson and Honig 1987) and a van der Waals term based on the AMBER potential (Weiner et al. 1984; Meng et al. 1992). In all of the docking calculations described here, only the ligand is made flexible; the receptor is held rigid. Although the receptor could also be made flexible, our current scoring method, based on a precalculated potential grid, makes this impracticable. Other approaches to receptor flexibility and scoring have been described (Schnecke et al. 1998; Claussen et al. 2001).

Throughout the paper, the inhibitors, ligands, and their associated mutants will be collectively referred to as the "ligand" and the pregenerated, hierarchically organized ensemble of ligand conformations will be referred to as the "flexible ligand." Additionally, all ligands are in their unbound conformations; no complexed ligands were used in docking.

Results

Conformations and docking statistics

The number of side chains treated as flexible ranged from seven for bovine pancreatic trypsin inhibitor (BPTI) to 40 for β-lactamase inhibitory protein (BLIP). On average, each of these side chains had about 10 conformations, leading to between 130 and 617 side-chain conformations to be evaluated. Assuming complete additivity (i.e., independence) of these side chains, the conformations were recombined on the fly during docking to create 10^9 to 10^{40} conformations of each ligand (Table 2). For three ligands, we explored not only different residue conformations but also different residue substitutions. Following available experimental data, all 20 amino acids were substituted at the P1 residue of BPTI and ovomucoid third domain (ovomucoid). Specific substitutions at multiple sites on BLIP were also made. Using the same additivity assumptions that we used for conformations, we explicitly evaluated 20 variants (mutants) at each of 10 amino acid positions in the loop region of ovomucoid. These 200 mutations were recombined to produce 20^{10}, or 10^{13} mutants, from which we selected the best scoring molecules. During docking, between one million and 20 million orientations were evaluated for each conformation/mutation (Table 2). Docking calculations on Pentium III computers (up to 800 MHz) took up to 12 h of CPU time (Table 2).

All docking calculations were performed on the unbound conformation of the ligands. To explore the role of ligand flexibility, the unbound ligands were docked to their cognate receptors in four separate calculations: (a) the unbound ligand was docked without conformational sampling to the unbound conformation of the receptor; (b) the unbound ligand was docked without conformational sampling to the bound conformation of the receptor; (c) multiple conformations of the unbound ligand were docked to the unbound conformation of the receptor; (d) multiple conformations of the ligand were docked to the bound conformation of the receptor. The results of each of these calculations are shown in four panels for BPTI docking to trypsin (Fig. 1). For clarity, these four panels contain only the data points with the best energies at each RMSD value. To guide the eye, a line is drawn delimiting the lowest energy dock scores at each RMSD value. For each of the six other systems, only the lines representing the lowest energy dock score at each RMSD value are shown; for reasons of space, a single graph is presented for each system (Fig. 3) (the complete data set for all systems may be found in the supplementary materi-

Table 2. *Docking flexible and rigid ligands*

	Flexible ligand						Rigid Ligand			
System	Time (h)[a]	Orientations[a]	Energy (kcal/mol)[b]	RMSD (Å)[b]	No. flex residues	No. confs.[c]	Time (h)[a]	Orientations[a]	Energy (kcal/mol)[b]	RMSD (Å)[b]
Unbound trypsin/BPT1	11.56	18,430,899	−93.60	2.90	7	2.04E + 09	2.92	20,337,728	−64.47	5.75
Bound trypsin/BPTI	10.55	18,874,514	−96.38	1.75	7	2.04E + 09	3.02	16,523,192	−60.85	6.23
Unbound α-chymotrypsin/ovomucoid	3.35	9,420,436	−59.65	1.60	11	2.82E + 11	3.20	10,141,904	−51.95	1.60
Bound α-chymotrypsin/ovomucoid	4.69	13,413,826	−67.51	1.91	11	2.82E + 11	2.44	14,248,499	−56.73	2.05
Bound protease B/ovomucoid	5.44	9,222,117	−58.90	1.22	11	2.82E + 11	2.21	13,846,254	−39.06	2.50
Unbound TEM-1/BLIP	7.02	10,703,014	−74.21	2.39	40	2.96E + 40	3.66	12,088,477	−36.37	4.13
Bound TEM-1/BLIP	8.64	12,266,500	−97.61	1.89	40	2.96E + 40	5.73	15,837,698	−29.13	5.21
Unbound FAB/lysozyme	3.41	8,332,413	−45.98	23.44	11	1.12E + 11	0.71	2,594,998	−33.70	8.79
Bound FAB/lysozyme	3.80	5,809,043	−52.90	1.69	11	1.12E + 11	1.86	4,134,314	−31.16	7.86
Unbound barnase/barstar	1.84	2,430,168	−104.68	18.35	19	2.39E + 21	2.46	6,541,046	−70.05	17.69
Bound barnase/barstar	2.47	2,940,918	−90.26	2.41	19	2.39E + 21	1.21	3,123,504	−66.07	8.39
Unbound subtitisin/C1-2	1.98	8,733,777	−47.20	2.67	13	2.35E + 17	1.78	19,341,624	−42.20	19.16
Bound subtilisin/C1-2	3.51	16,340,507	−67.66	1.81	13	2.35E + 17	2.86	35,944,313	−33.62	19.26

[a] Summed over all docking interfaces for the system.
[b] For the complex with the most favorable interaction energy.
[c] Number of possible conformations through recombination at the time of docking.

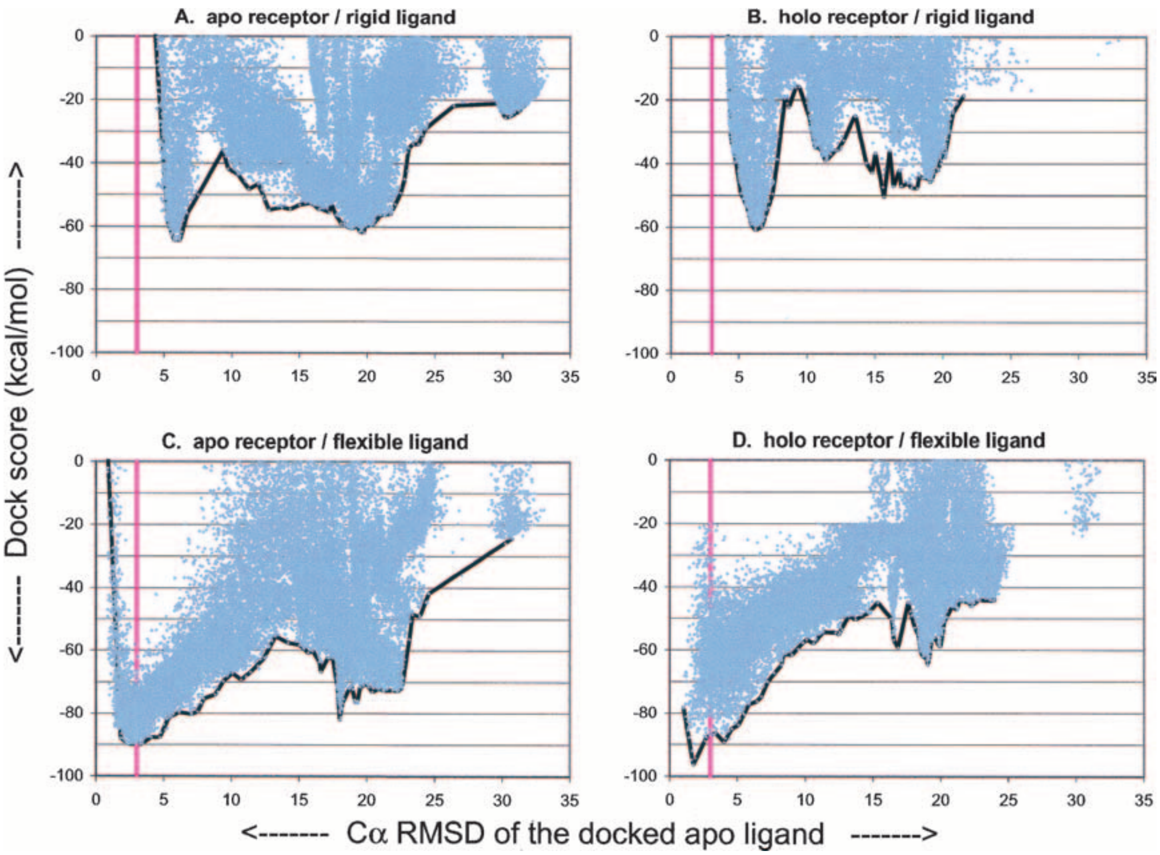

Fig. 1. Comparing rigid (*A*, *B*) and flexible (*C*, *D*) docking of an unbound structure of BPTI (4PTI) to unbound (2PTN) and bound (2PTC) trypsin. Data points for the best docking energy scores (y-axis) for all RMSD values (x-axis) are shown. RMSD values are calculated between all Cα atoms of each docked ligand and the Cα atoms of the unbound ligand superpositioned onto the bound ligand. As a convenience to the reader, a black line is drawn to define an envelope of the best energy values. This line includes the lowest RMS data point and 50 additional data points representing the lowest energies in each of 50 divisions of the data. Points for the line are distributed on the x-axis based on the density of the data points (higher resolution where more data points exist). The vertical line at 3 Å indicates an upper RMSD bound for a near-native conformation.

als). To calculate the discrimination between the near-native and non-native complexes, the difference between the best scoring near-native complex was subtracted from the best scoring non-native complex for both the rigid and flexible docking (Fig. 4).

Trypsin/BPTI

As expected, docking generated no low energy near-native orientations of the unbound conformation of BPTI when docked as a rigid body (Figs. 1A,B, 4). This was because key interface residues, such as Lys15, Arg17, and Arg39, are in the “wrong” conformations in the unbound BPTI crystal structure (Fig. 2). In the best-scoring complex of the unbound conformation of BPTI, Lys15 does not fit into the deep S1 specificity pocket of trypsin, as it is observed to do in the trypsin/BPTI complex. Instead, a lysine on the opposite surface of BPTI, Lys26, is docked into the S1 specificity pocket, leading to more favorable energies for non-native dockings against both the bound and unbound conformations of trypsin (not shown).

Docking the unbound conformation of BPTI as a flexible molecule alters the binding preference in favor of near-native configurations in both the unbound and bound conformations of trypsin (Figs. 1C,D, 4). BPTI had seven residues that were more than 60% exposed, resulting in 228 independent side-chain conformations to evaluate. These were recombined during the docking calculation to produce as many as 10^9 conformations for each orientation. In the near-native complexes, Lys15 of BPTI adopts a conformation that allows it to fit into the S1 pocket of trypsin which, with the other conformational adaptations, leads to better docking energy scores for near-native configurations than for non-native configurations. The RMSD between the Cα of the flexibly docked unbound form of the ligand and the bound form of the ligand was 2.9 Å, corresponding to a 70% conservation of native contacts observed in the complex.

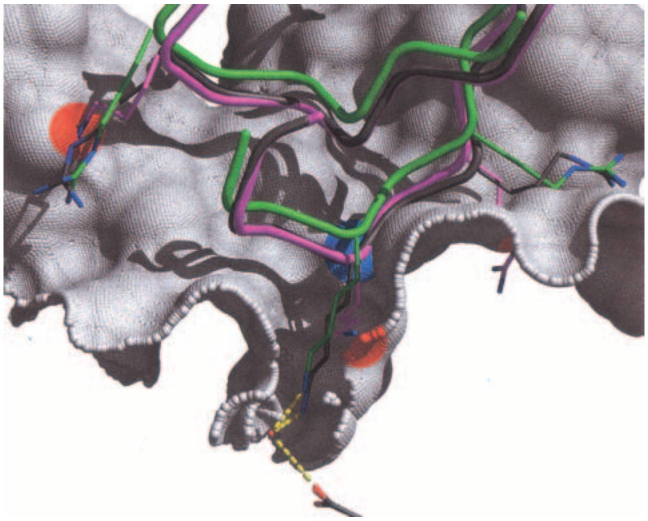

Fig. 2. The high scoring conformation of unbound BPTI (backbone in green) docked in multiple conformations into the unbound conformation of trypsin (surface in gray). The rigid unbound BPTI (magenta) has been superpostioned onto the crystallographic bound BPTI (gray). The molecular surface is colored red where the unbound, rigid BPTI would clash into the surface. The blue surface indicates the position of the catalytic Oγ of Ser195. Asp189 of trypsin and a bridging water molecule (2.6 Å from the docked ligand) are shown for context. No water molecules were present in the docking calculation.

TEM-1/BLIP

Docking the unbound conformation of BLIP as a rigid body, led to non-native complexes of TEM-1/BLIP having better energy scores than near-native complexes (Figs. 3A, 4). This was because two key interface residues on the unbound ligand, Asp49 and Phe142, are in the “wrong” conformations for optimal fit to TEM-1 (Fig. 5). Consequently, the near-native complexes cannot be distinguished from the non-native docked complexes when docking to either the unbound or bound conformations of the receptor.

Docking the unbound conformation of BLIP as a flexible molecule alters the binding preference in favor of near-native configurations in both the unbound and bound conformations of TEM-1 (Figs. 3A, 4). From the unbound structure of BLIP, only two ligand residues, Asp49 (93% exposed) and Phe142 (90% exposed), needed to change conformation to shift the preferred binding mode toward a near-native complex (Petrosino et al. 1999). Taking advantage of the relatively large size of BLIP, we made a large number of residues flexible to test the scalability of the method. For this protein we allowed all residues that were more than 40% exposed to adopt multiple conformations instead of the 60% used in the other systems. This resulted in 40 flexible residues with a total of 617 independent side-chain conformations. These were recombined during the docking calculation to produce as many as 10^{40} conformations per orientation. Docking with the standard residues (60% exposed) made flexible produced the same results, but decreased the run time to slightly over 2 h. Allowing ligand flexibility produced dock scores favoring near-native complexes over non-native complexes by more than 40 kcal/mole (Fig. 4). The RMSD between the Cα of the flexibly docked unbound form of the ligand and the bound form of the ligand was 2.4Å, corresponding to a 59% conservation of native contacts observed in the complex.

Subtilisin/chymotrypsin inhibitor 2

Subtilisin Novo (SN) from the complex with chymotrypsin inhibitor 2 (CI-2) was used as the bound receptor and subtilisin from *Bacillus amyloliquefaciens* (BAS) was used as the unbound receptor. The RMSD value for active site Cα

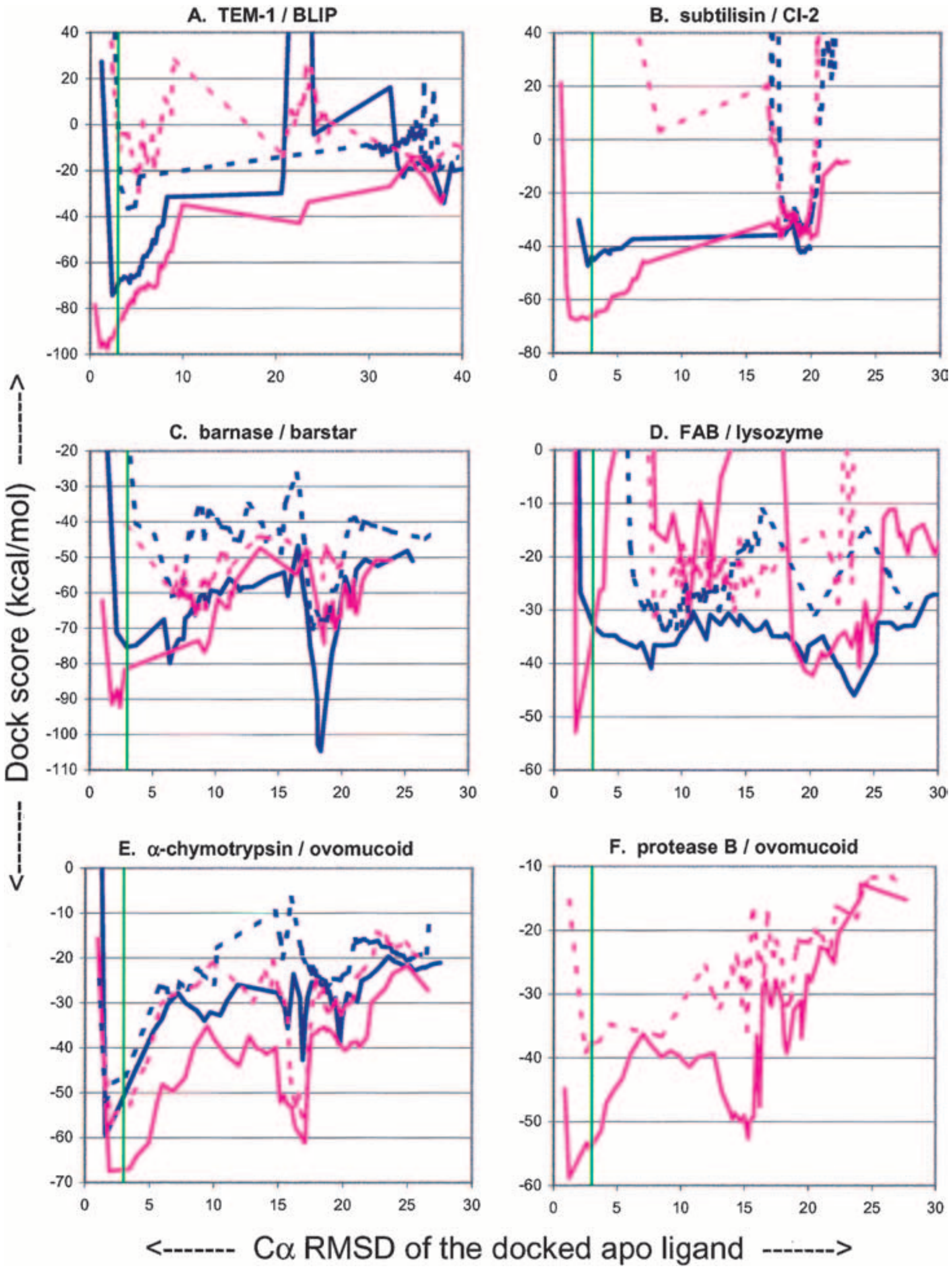

Fig. 3. Lines representing the best docking energy scores (y-axis) versus RMSD from the experimental complex (x-axis) for: (*A*) TEM-1/BLIP; (*B*) subtilisin/CI-2; (*C*) barnase/barstar; (*D*) FAB/lysozyme; (*E*) α-chymotrypsin/ovomucoid; (*F*) protease B/ovomucoid. The dashed lines represent docking the rigid unbound ligand, and the solid lines represent docking the flexible unbound ligand. The blue lines represent docking to the unbound conformations of the receptors, and the pink lines represent docking to the bound conformations of the receptors. A vertical line is drawn at 3 Å to indicate an upper RMSD bound for a near-native conformation.

atoms between SN and BAS is 0.30 Å. The residues critical for binding are identical in both proteins.

Docking the rigid conformation of unbound CI-2 to both the unbound and bound receptor structures favored non-native binding modes (Figs. 3B, 4). This is because the P2 and P1 residues of CI-2, Thr58, and Met59, are in the "wrong" conformation; upon superposition of the unbound ligand onto the bound ligand, they clash with residues His64 and Ala152 of subtilisin (Fig. 6). Multiple conformations were generated for the 12 residues of unbound CI-2 that had 60% or more of their surface area exposed. The P1' residue, Glu60 (48% exposed), was also made flexible. These 13

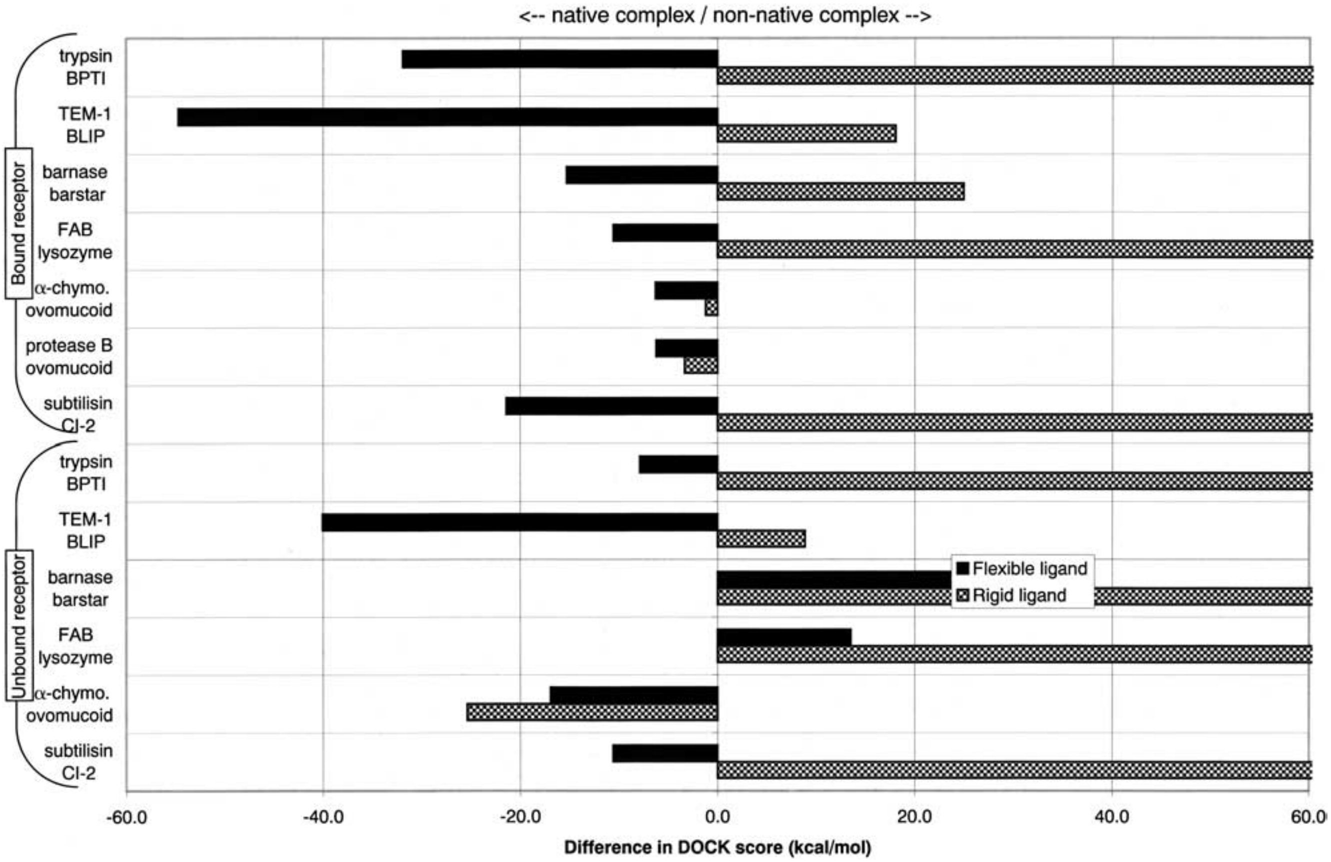

Fig. 4. Subtraction of the best near-native (RMSD values <3 Å from the bound complex) energy score from the best non-native (>3 Å) energy score for each system. Bars to the left indicate a near-native complex was preferred, and bars to the right indicate that a non-native complex scored best. The magnitude of the bar indicates how well the preferred docked complex is distinguished from other complexes.

flexible residues resulted in 642 independent side chain conformations that were recombined during docking to produce up to 10^{17} conformations in each orientation. Docking multiple conformations of the unbound ligand favored near-native binding modes in both the unbound and bound receptor structures (Figs. 3B, 4). The RMSD between the Cα of the flexibly docked unbound form of the ligand and the bound form of the ligand was 2.7 Å, corresponding to a 60% conservation of native contacts observed in the complex.

Barnase/barstar

The rigid unbound barstar preferentially docks in a non-native mode to both the bound and unbound barnase (Figs. 3C, 4). Several small differences between the bound and unbound ligand structures contribute to the preference for the non-native binding mode. Salient among them is Asp39 of barstar (61% exposed), which assumes a different conformation in the bound and unbound structures, rotating by 95° around the Cα–Cβ. In the bound complex, this residue makes a hydrogen bond to His102 of barnase, a residue that contributes to the electrostatically driven binding of the two proteins (Buckle et al. 1994).

To investigate the role of ligand flexibility, conformations were calculated for the 19 residues on barstar that were more than 60% exposed, generating a total of 389 side chain conformations. These conformations were recombined during docking to produce up to 10^{21} conformations of the ligand in each orientation. Docking these multiple barstar conformations to the unbound barnase improved the scores of the near-native complexes relative to the non-native complexes (Fig. 3C), but the non-native complexes still scored better than the near-native complexes (Fig. 4). It was only when the unbound, multiconformer barstar was docked to the bound conformation of barnase that the near-native dockings could be distinguished from the non-native dockings (Fig. 3C).

The conformational differences between the unbound and bound receptors were of greater importance in barnase than for other receptors. Different conformations of two key residues in the unbound and bound barnase cause a non-native complex to be favored. His102 of barnase hydrogen bonds

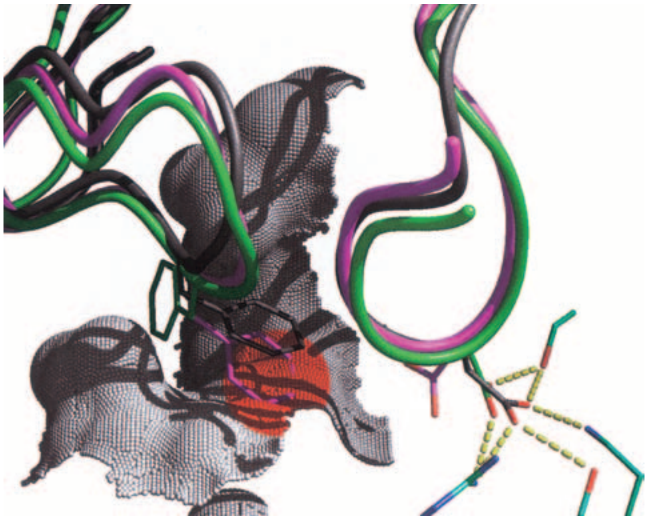

Fig. 5. The high-scoring unbound BLIP structure (green), generated from multiconformer docking to the unbound conformation of TEM-1 (cyan), is shown. The rigid unbound BLIP (magenta) has been superimposed onto the bound ligand (gray). A partial molecular surface for the complexed receptor is shown to illustrate hydrophobic interactions. Important intermolecular hydrogen bonds are shown in yellow. The conformations of two key interface residues from BLIP, Asp49 and Phe142, are shown for the best scoring docked structure (green), for the original unbound structure (magenta), and for the bound complex (gray). The molecular surface of TEM-1 is colored red where the superpositioned rigid unbound ligand clashes into the receptor.

with Asp39 and Gly31 of barstar, and a second residue, Arg59, packs tightly against residues Glu76 and Trp38 of barstar. The conformations of His102 and Arg59 in the unbound barnase, although not precluding interactions with barstar, do not allow these favorable interactions to occur. As a result, α-helix 2 of barstar still binds to the binding site of barnase but is flipped 180° (Fig. 7). Intriguingly, a number of common interactions are observed between the crystallographically determined complex and the flipped ligand orientation generated from docking. Glu46 from the docked unbound barstar mimics interactions observed crystallographically by Asp36 from barstar (Fig. 7); Trp38 from the docked structure occupies the same space that Trp45 from the complex does, and, in general, hydrogen bond donor positions in the complexed ligand structure are mimicked by donors in the flipped structure. The conformational problems in barnase are not present in the bound enzyme, allowing the multiconformer barstar to dock in high scoring, near-native configurations. The RMSD between the Cα of the flexibly docked unbound form of the ligand and the bound form of the ligand was 18.3 Å, corresponding to an 8% conservation of native contacts observed in the complex.

FAB/lysozyme

When docking the single conformation represented by the unbound crystal structure (PDB code 132L), lysozyme preferentially docks to FAB D44.1 in a non-native complex. This is because Arg45 from the unbound structure adopts different conformations in the unbound and bound structures (Fig. 8). No low-energy, near-native, complexes are produced from docking the rigid ligand to either the unbound or bound receptor (Figs. 3D, 4).

Conformations were calculated for the 11 residues on lysozyme that were more than 60% exposed, generating a total of 130 side-chain conformations. These conformations were recombined during docking to produce up to 10^{11} conformations of the ligand in each orientation. As with barnase/barstar, docking these multiple lysozyme conformations to the unbound FAB improved the ranking of the near-native complexes relative to the non-native complexes; however, the non-native complexes still scored better than the near-native complexes (Fig. 4). It was only when the unbound, multiconformer lysozyme was docked to the bound conformation of FAB that the near-native complexes could be distinguished from the non-native complexes (Fig. 3D). The binding site for lysozyme on FAB D44.1 is at the interface of the A and B monomer. Backbone movements of two complement determining loops in the B monomer of

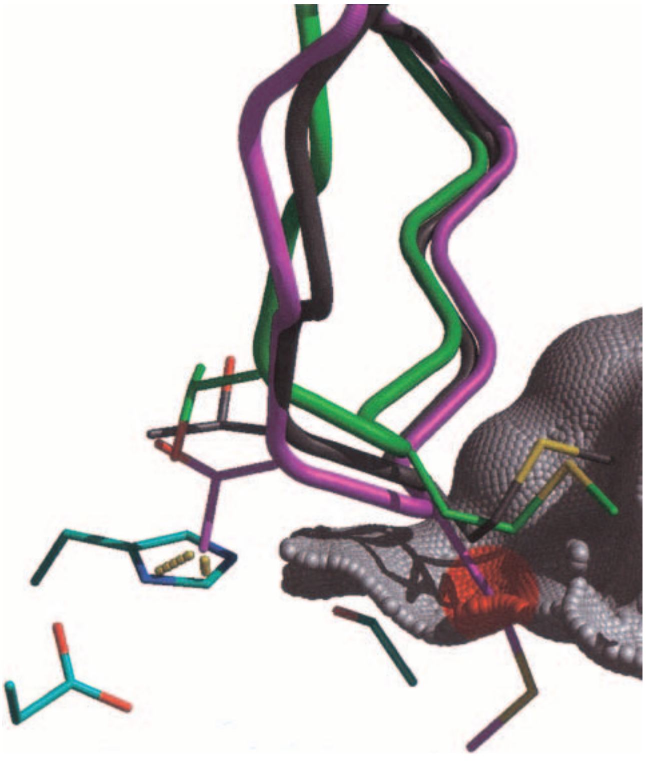

Fig. 6. The best-scoring CI-2 ligand (green) generated from multiconformer docking is shown with the unbound conformation of subtilisin (catalytic triad shown in cyan). The rigid unbound CI-2 (magenta) has been superpositioned onto the complexed CI-2 (gray). The dashed red lines indicate clashes between Thr58 of the rigid unbound ligand and the receptor. The red surface indicates the region where the P1 residue from the rigid unbound CI-2 clashes into the receptor surface.

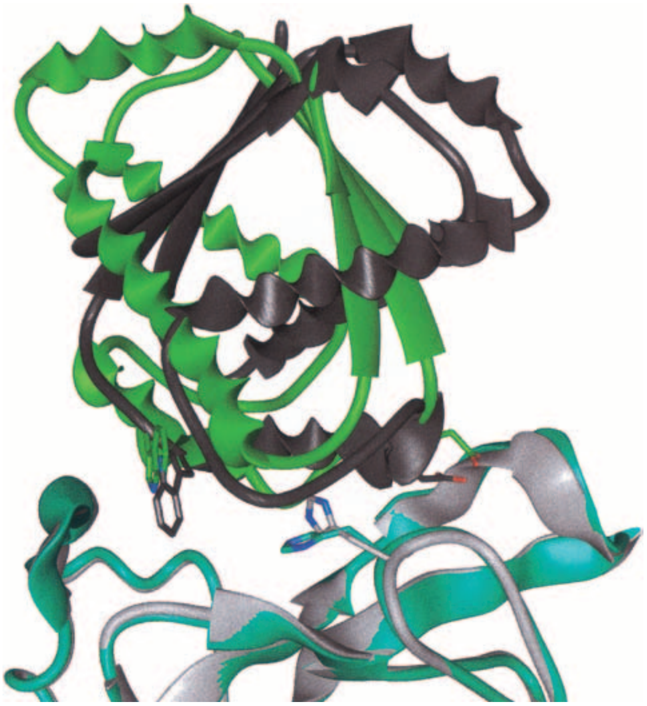

Fig. 7. The best-scoring structure generated from docking multiple conformations of unbound barstar (green) into the unbound conformation of barnase (cyan) is shown. The experimental complex of barstar (gray) and barnase (light gray) has been superimposed on the unbound receptor. The different conformations adopted by His102 of barnase are shown. Glu46 from the unbound structure (green) and Asp36 from the bound structure (gray) are proximal to each other, as are Trp38 (green) and Trp45 (gray).

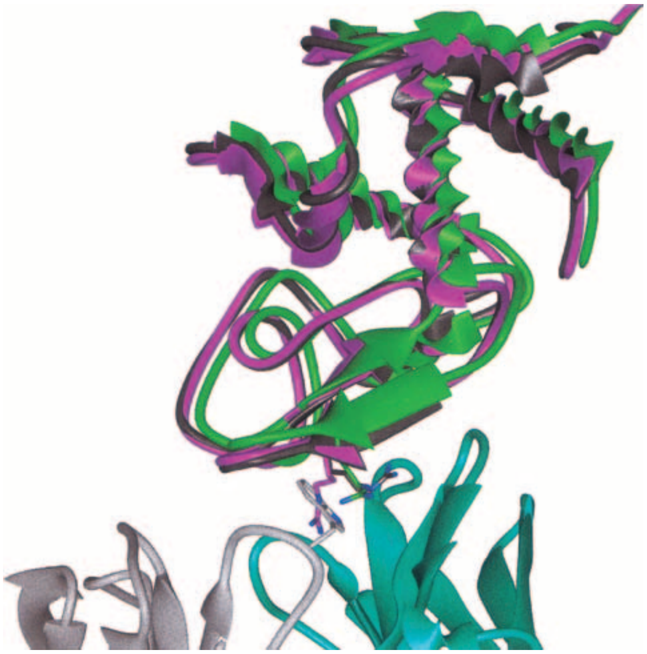

Fig. 8. The high-scoring conformation of lysozyme from multiconformer docking (green) to the bound structure of FAB D44.1 is shown. The bound form of lysozyme (gray) and the unbound structure of FAB (*A*, monomer in light gray, *B*, in cyan) are shown. Arg45 from the unbound ligand structure (magenta), when superpositioned onto the complexed structure of lysozyme, clashes with Trp94 of FAB.

FAB, residues 28–33 and 53–57, and their corresponding side-chain displacements, represent the largest conformational changes of the receptor in the ligand binding site. The RMSD between the Cα of the flexibly docked unbound form of the ligand and the bound form of the ligand was 23.4 Å, corresponding to a 63% conservation of native contacts observed in the complex.

α-Chymotrypsin/ovomucoid

The preferred binding mode of the rigid unbound ovomucoid to both the unbound and bound α-chymotrypsin recep-

Table 3. *Docking mutant proteins*

	Mutant flexible ligand						Rigid unbound ligand	
Receptor/inhibitor	Number of flexible residues	Site of substitution	Number of conformations	Number of substitutions	Number of orientations	Time (h)	Number of orientations	Time (h)
Bound trypsin/BPT1	7	K15[a]	2.04E + 09	25[e]	3,961,947	3.61	4,892,841	1.44
Unbound trypsin/BPT1	7	K15[a]	2.04E + 09	25	4,250,425	6.04	4,169,402	2.29
Bound TEM-1/BLIP	40	Sch/Pal[b]	2.96E + 40	15	6,723,016	4.63	8,537,456	5.93
Unbound TEM-1/BLIP	40	Sch/Pal[b]	2.96E + 40	15	5,874,706	4.39	5,874,706	3.26
Bound protease B/ovomucoid	11	Loop[c,d]	2.82E + 11	1.02E + 13	6,938,997	58.31	3,951,216	4.26
Bound protease B/ovomucoid	11	M18[d]	2.82E + 11	25	12,244,048	10.43	3,951,216	4.26
Bound α-chymotrypsin/ovomucoid	11	M18[d]	2.82E + 11	25	1,937,340	3.11	6,653,072	1.85
Unbound α-chymotrypsin/ovomucoid	11	M18[d]	2.82E + 11	25	1,584,450	1.63	2,404,468	1.01

[a] Twenty amino acids substituted at position 15 of BPTI (Krowarsch et al. 1999).
[b] Mutants with experimental binding affinities reported in recent papers (Huang et al. 2000; Selzer et al. 2000).
[c] Twenty amino acids substituted at each of 10 loop positions (13–15, 17–21, 32, 36) of ovomucoid.
[d] Twenty amino acids substituted at position 18 of ovomucoid (Lu et al. 1997).
[e] Twenty-five substitutions include charged and neutral forms of Asp, Glu, His, Arg, and Lys.

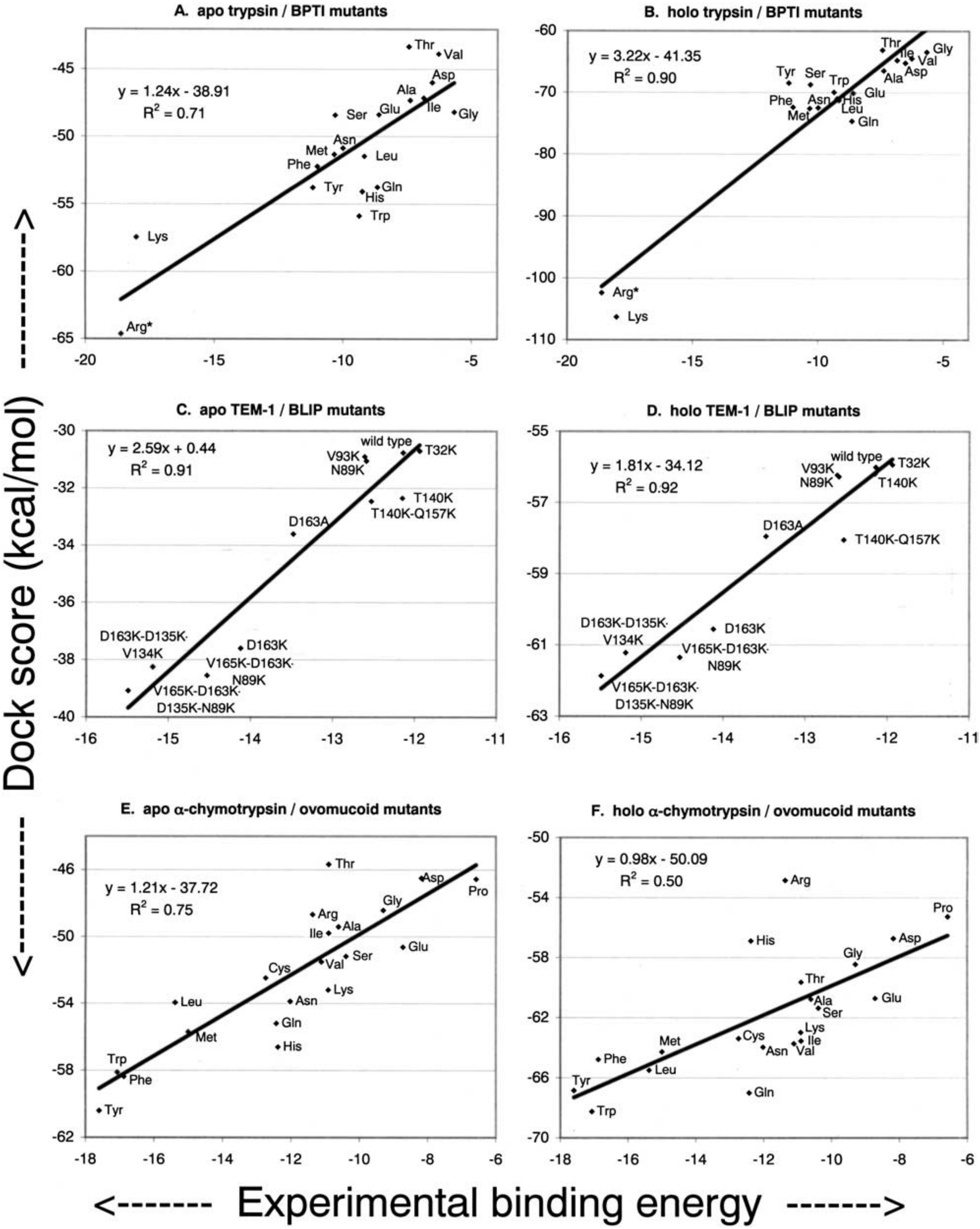

Fig. 9. *Continued on facing page.*

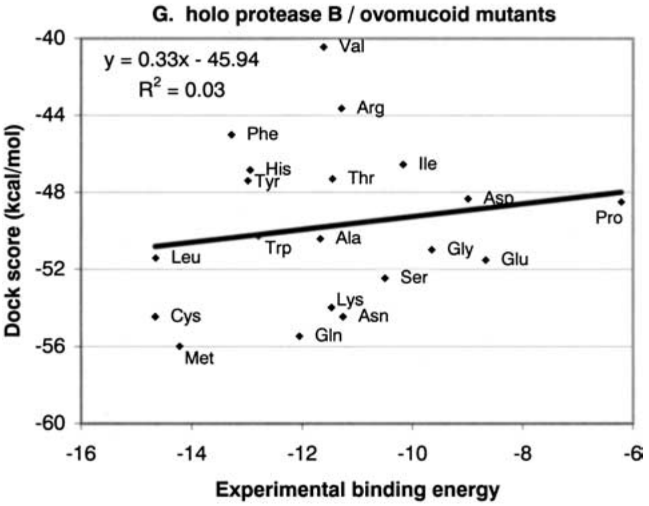

Fig. 9. Comparisons of experimentally determined and docking-predicted binding affinities for mutant protein inhibitors docked into their cognate enzymes. BPTI variants docked to unbound (*A*) and bound (*B*) trypsin; BLIP mutants docked to unbound (*C*) and bound (*D*) TEM-1; ovomucoid variants docked to unbound (*E*) and bound (*F*) α-chymotrypsin, and ovomucoid variants docked to bound protease B (*G*).

tor is a near-native pose. Surprisingly, even though a near-native complex is favored over non-native complexes when docking to the bound receptor, it is only favored by about one kcal/mole—a smaller discrimination than observed from docking to the unbound receptor (Figs. 3E, 4). This small preference for near-native complexes may partially owe to the bound structure used for docking. The bound structure was the complex between α-chymotrypsin and an ovomucoid that had a leucine as the P1 residue (Fujinaga et al. 1987) instead of the methionine found the in the unbound structure.

To investigate the role of ligand side chain flexibility, 147 conformations were generated from 11 residues (60% exposed) on ovomucoid. These were recombined during the docking calculation to produce up to 10^{11} conformations of the ligand in each orientation. The introduction of flexibility increased the energy separation of the native-like and non-native-like docked complexes from one to six kcal/mole when docking to the bound receptor (Figs. 3E, 4). When docking to the unbound form of the receptor, ligand flexibility decreased our ability to distinguish native from non-native structures from 25 to 17 kcal/mole. The addition of flexibility improved the docking score of non-native poses more than it did the docking score of near-native poses—perhaps reflecting the unusually favorable scores of the near-native, rigid ligand dockings. Despite this improvement, near-native complexes are still clearly distinguished from the non-native complexes. The RMSD between the Cα of the flexibly docked unbound form of the ligand and the bound form of the ligand was 1.6 Å, corresponding to an 83% conservation of native contacts observed in the complex.

Protease B/ovomucoid

All docking calculations with protease B were conducted using the bound form of the receptor because no unbound structure was found in the PDB. This structure of protease B was crystallized with a mutant ovomucoid that had an alanine instead of a methionine at the P1 position (Read et al. 1983). The unbound form of ovomucoid used for docking was the same as used in docking to α-chymotrypsin (Met18 at P1). As with α-chymotrypsin, the rigid ovomucoid was preferentially docked in a near-native pose. Introducing flexibility into this system, with 11 residues in up to 10^{11} conformations per orientation, increased our ability to distinguish native from non-native poses (Figs. 3F, 4).

Docking mutant ligands

A natural extension to combinatorially evaluating ligand residue conformations is to combinatorially evaluate ligand residue substitutions. We created point mutations in three ligands (Table 3) and docked them to four different receptors. Docking all 20 amino acid substitutions at the P1 residue, residue 15 for BPTI (Fig. 9A, B), and residue 18 for ovomucoid (Fig. 9E–G), generally took 2.5 to six times as long per orientation as docking a single molecule. Evaluating 15 selected substitutions at specific residues in BLIP (Fig. 9C, D) took about the same amount of time per orientation as did docking a single molecule. Additionally, assuming complete additivity of the side chain conformations and substitutions, 20 substitutions of 10 loop residues (10^{13} mutant ligands) were docked for ovomucoid. Each of these 10^{13} ligands had about 10^{11} conformations (11 flexible residues distributed over the surface of the ligand each with about 10 conformations) were docked for ovomucoid. This calculation took about 15 times longer than docking a single protein (i.e, no substitutions) in a single conformation.

Our initial attempt to predict relative mutant binding affinities led to charged mutants ranking best for all systems. We also noted that some acidic and basic mutants might be neutral upon binding. To address these issues, apartic acid, glutamic acid, lysine, arginine, and histidine, were represented in both their charged and neutral forms. We then precalculated a desolvation term for all 25 amino acid forms. For each mutant residue in contact with the receptor for each complex, that residue's desolvation value was subtracted from the dock score (correcting for the different protonation states according to their pKa values and the pH at which the experiment was conducted).

In all but one case, there was a significant positive correlation between the experimental result and the docking predictions. The slopes of the lines for both the unbound and bound α-chymotrypsin with ovomucoid (Fig. 9E, F) and unbound tryspin with BPTI (Fig. 9A) were close to 1. Slopes for the TEM-1 with BLIP (Fig. 9C, D) and bound trypsin with BPTI (Fig. 9B) were significantly greater than one (Fig. 9G). There was no correlation ($R^2 = 0.03$) between the binding affinities reported by the docking program and the experimental values for ovomucoid mutants binding to protease B. For all other systems, docking to both the unbound and bound receptors, R^2 values ranged from 0.50 to 0.91.

With either Arg or Lys as the P1 residue, BPTI has been shown to bind to trypsin with an affinity that is five orders of magnitude greater than other amino acids at this position (Krowarsch et al. 1999). Consistent with this finding, our results from simultaneously docking all 20 amino acid variants of BPTI to both the bound and unbound structures of the receptor indicate a clear preference for the basic amino acids (Fig. 9A, B). A trend for the remaining amino acids is present with Thr, Trp, His, and Gln being the farthest outliers when docking to the unbound receptor.

Two different research groups (Huang et al. 2000; Selzer et al. 2000) experimentally determined binding affinities for a total of 15 mutants of BLIP binding to TEM-1. For both sets of mutants, the relative dock energy scores correlate well with experimental results, indicating a binding preference for polar or charged mutants. The 15 mutations were partitioned into two sets of data because of differences in experimental baseline binding affinities determined by the two groups. Only the 11 data points corresponding to the results of Schreiber and coworkers have been plotted for ease of viewing (Fig. 9C, D).

Binding affinity trends from docking the 20 P1 variants of ovomucoid to the unbound and bound structures of α-chymotrypsin (Fig. 9E, F) indicate a distinct correlation with experiment (Lu et al. 1997), with the larger, nonpolar amino acids being favored at this position in both the docking predictions and the experimental results. On the other hand, no correlation was observed between the experimental binding affinities and the predicted bindings for ovomucoid with protease B (Fig. 9G). In the docked structures, the different P1 residues fold back upon the surface of the inhibitor, rather than "down" into an S1 pocket, possibly diminishing specificity interactions with the enzyme.

Discussion

The protein docking problem has two components: sampling conformations and orientations for the docking molecules, and implementing a scoring function that can distinguish near-native complexes from non-native complexes. Here, we investigate how far we can progress by sampling conformations of ligand residues that are solvent exposed, leaving the receptor rigid, and using a full atomistic potential function to evaluate the docked complexes.

For all seven systems, the addition of ligand residue flexibility was sufficient to produce low energy, near-native complexes when docking the unbound ligands to their cognate bound receptors (Fig. 4). When the unbound ligand structures were docked as rigid bodies, without flexible side chains, this was not the case; the non-native docked complexes scored better than the near-native complexes in all cases except the ovomucoid dockings. In every system, adding flexibility to the ligand improved the scores of the near-native complexes relative to the non-native complexes with the exception of ovomucoid docked to the unbound structure of α-chymotrypsin. For this system, although the near-native complexes were clearly distinguished from the non-native, the introduction of flexibility slightly decreased our ability to discriminate between the two. This was probably due to our failure to account for different intramolecular energies in the different ligand conformations. These results suggest that adding ligand flexibility is important, and in some systems sufficient, to distinguish near-native from non-native complexes in protein–protein docking. We find that many conformations need to be sampled; a hierarchical representation of conformational possibilities provides one method to do so.

The challenge of explicitly sampling the enormous number of conformations accessible to protein ligands has led to the introduction of modified scoring schemes. In particular, the van der Waals component of the interaction energy between two proteins is exquisitely sensitive to conformation. A single atom positioned a fraction of an Angstrom too close to the receptor can lead to very large repulsive terms. Given the coarseness with which we sample conformations and the fact that we do not allow backbone flexibility, it was entirely possible that our conformations might not complement the receptor. Our results indicate that, at least for conformational changes that are largely restricted to ligand side chains, coarse conformational sampling, combined with a large amount of orientation sampling, is sufficient to reproduce near-native complexes. Additionally, these complexes are similar enough to the native complex that they score well and are distinguishable from non-native complexes when using a Lennard-Jones 6–12 potential term combined with Poisson-Boltzmann electrostatics for our scoring function. The Lennard-Jones term, though unforgiving, appears to help discriminate the near-native complexes from others.

Although adding ligand conformational flexibility seems to be important, it is clearly not sufficient in all systems. Of the six unbound receptors docked (no unbound structure for protease B was found), two favored non-native docked complexes. Because our docking to the bound conformations of these receptors produced "correct" answers, we attribute the preference for a non-native binding mode to a conformational change between the bound and unbound receptors.

An interesting question in protein–protein interactions is how individual side chains contribute to overall binding affinity. To evaluate a large number of possibilities, investigators have turned to combinatorial methods of exploring side-chain diversity, such as phage display. It occurred to us that the same hierarchical method that allowed for efficient recombination of side-chain conformations would also allow us to recombine side-chain substitutions, in effect recapitulating the combinatorial experiments in the docking calculations (Table 3).

For three of the four inhibitor/enzyme systems, docking to both unbound and bound receptors, there was a significant correlation between the predicted and the experimental binding affinities. In docking mutants of BLIP to the unbound TEM-1, the R^2 value was 0.91; for docking ovomucoid to unbound α-chymotrypsin, the correlation was 0.75. A high correlation was also observed between predicted and experimental energies for docking BPTI mutants into unbound trypsin. Considering the failure to fully treat desolvation and lack of receptor accommodations, these correlations are surprisingly good, and should not be expected to generally hold. Indeed, for docking ovomucoid to protease B, there is essentially no correlation. This may owe partly to the bound structure of protease B used for docking. The protease B/ovomucoid complex with an alanine mutant at P1 (rather than the native leucine or methionine) may have resulted in a more constrained S1 pocket, preventing other residues from binding. The relatively shallow slope (0.33) supports this. A more complete treatment of mutant desolvation (Lamb et al. 2001) and the possible addition of receptor flexibility may make this method more useful for considering residue substitutions to complement a receptor of known structure.

Several caveats deserve mention. The method assumes additivity in recombining conformations from different residues that are independently generated and evaluated in the receptor site. Without this assumption, the method would suffer a combinatorial explosion and the problem would be intractable, at least as we have represented it. For highly exposed residues on convex surfaces, additivity is a reasonable assumption. Although it will never be strictly true, additivity has been observed in several ligand protein interfaces (Wells 1990; Weiss et al. 2000; Lu et al. 2001). Violations to additivity will occur when the residues being sampled are close enough to one another that they can significantly influence each other's internal energies. When this happens, this method will lead to unreasonable results. Indeed, even when residues may be assumed to be independent of each other, we found that it was still important to consider the internal energies of the different conformations being calculated. For instance, mobilizing residues that were more than 50% buried often led to spurious results. A better integration of internal energies with interaction energies would make this approach more robust. Similarly, it is clear that the absolute energies of the docked complexes were often over estimated in magnitude. As in small-molecule docking (Shoichet and Kuntz 1996) predicting absolute binding energies in protein–protein docking remains problematic. Key components of the interaction energy, such as the cost of desolvating the receptor, have been ignored in our calculations. In favorable circumstances, we might hope for a monotonic relationship between docking energy and experimental binding affinity. It is not clear to us whether the sometimes high correlations that were observed between docking and experimental energies for the mutant ligands will be extensible to other systems.

The method of generating and recombining conformations is well suited to side-chain rotamers, and one might imagine, extending it to small loop movements as well. Other methods, including the use of rotamer libraries, might work as well or better than using SYBYL to calculate low-energy side-chain conformations. More generally, it seems clear that the current method will not address global changes in the conformation of the ligand or the receptor, where the additivity assumption will break down. Finally, it should be clear that this method is best suited for the independent movement of side chains, and may break down when substantial backbone movements are involved in the docking event.

These limitations notwithstanding, these studies suggest that it is feasible to sample a large number of protein ligand side-chain conformations, and that doing so significantly improves our ability to distinguish near-native from non-native configurations in protein–protein docking. Although there is clearly room for investigation of alternative scoring schemes, the use of a standard Lennard-Jones term may have improved the signal to noise in these calculations by discarding many nonphysical configurations early in the docking calculation. Considering receptor side chain flexibility may further extend these results. For protein–protein complexes that form without significant main-chain accommodation, sampling side-chain conformations and scoring with a classical energy function may be sufficient to address the protein docking problem.

Materials and methods

Hierarchical representation of conformations

The method presented here extends the standard DOCK protocol where rigid molecules are fit in multiple orientations in the binding site and then evaluated for fit. Before docking, multiple low-energy conformations for each flexible residue are calculated offline. These conformations are stored in a database, and are not recalculated subsequently. The ensemble of pregenerated ligand conformations is processed into a hierarchical data structure such that atom connectivity is implicitly represented across all members (conformations and substitutions) of the ensemble. The method will be sketched here and a detailed description will be published elsewhere (D. M. Lorber and B.K. Shoichet, in prep.). The hierarchical data structure for representing protein ligands is optimized to perform three features:

1. Identify atoms that are common among different conformations to eliminate redundant calculations. For globular proteins, most of the atoms can be treated as internally rigid. Most conformations of proteins containing hundreds of atoms differ by only a few atoms.
2. Order the atoms within side chains so that side chains clashing into the receptor can be pruned off as early as possible. In most side chains, terminal atoms have more positions than atoms closer to the main chain. Hierarchical ordering of atoms avoids evaluation of many terminal atom positions by identifying clashes close to the main chain.
3. Recognize that side chains can be moved independently allowing for combinatorial evaluation of conformations and mutations. Assuming complete additivity, an exponential number of conformations can be evaluated while computation time increases linearly.

Selection of docking systems

All structures were taken from the PDB except for the BLIP and TEM-1 structures that were obtained as part of an earlier docking prediction test (Strynadka et al. 1996). Except for protease B/ovomucoid, we chose systems that had crystal structures for the complex as well as unbound structures for both partners (Table 1). Where multiple records were available for the same structure, we looked for the most recent, highest resolution, and most complete structure. The exception to this rule was that of the lysozyme structure, which was selected for its backbone similarity to the complexed lysozyme structure we chose. Compared to some of the unbound ligand structures, the lysozyme in the FAB complex undergoes significant main-chain accommodations. We excluded these ligand structures because our method has not been used to sample main-chain conformational changes. This is a limitation of the method as currently implemented. Selection of residues for mutation was based on the availability of experimental data (Table 3).

Preparation of the receptors and energy grids

The systems were prepared for docking using a script to ensure uniform treatment. All water molecules were removed from the structure. Where alternate conformations for side chains existed, the first conformation was selected. Only the known interface region of the receptor proteins was targeted for docking. The His57 residues of the serine proteases were protonated on the Nδ atoms, consistent with their role in the catalytic triad. Each receptor was then protonated using the Biopolymers package of SYBYL 6.6 (Tripos Software), and the atoms were renamed using the AMBER naming convention. No other modifications were made to either the unbound or bound receptors. The program CHEMGRID (Meng et al. 1992) was used to generate a van der Waals potential grid based on a standard Lennard-Jones 6–12 potential around the area of the binding site. For the receptor, a molecular surface was generated for all residues within 15 Å of a key active-site residue, and this surface was used by SPHGEN (Kuntz et al. 1982) to generate spheres representing potential ligand atom locations. The locations of these spheres were used to define a region of low dielectric in the active site of the receptor—consistent with the ultimate occupancy of this site by the ligand (Shoichet and Kuntz 1993; Lorber and Shoichet 1998). DelPhi (Gilson and Honig 1987) produced an electrostatic potential grid based on the receptor structure and all of the spheres generated by SPHGEN. DISTMAP (Shoichet et al. 1992) was run on each receptor to generate a shape complementary grid. A subset of the SPHGEN spheres was analytically selected (Shoichet et al. 1992) for use in orienting the ligand. A second, larger subset of spheres was selected for orientation refinement using focusing (Shoichet et al. 1992).

Preparation of the ligands

All ligand surfaces were docked into the active sites of the receptors, unless otherwise noted. Ligand backbone atoms with B-factors greater than 80 and their corresponding side chains (residues 60–65 of barstar and 101–103 and 128–129 of lysozyme) were removed. The program ACCESS (Lee and Richards 1971) was used to determine the solvent accessible area of ligand residues. Residues over 60% exposed were treated as flexible; all others were unchanged from the experimental structure unless otherwise noted. The program SPHGEN was used to generate spheres filling the volume of the ligand. Subsets of these spheres were used for matching at the ligand orientation stage of docking.

Multiple conformations were calculated for each exposed side chain in SYBYL using a systematic search. The key to this conformer generation is that all conformations are generated in the same reference frame. This extends an earlier method for docking multiple ligand conformations (Lorber and Shoichet 1998), significant differences being that previously neither recombination nor a full atomistic potential function was used. By requiring all conformations to be in the same reference frame, a single rotation matrix can be used to bring the entire ensemble of conformations into the binding site. When generating conformations of a given side chain, all other side chains were present in their unbound crystallographic conformation. Rotatable bonds for aromatic amino acids were rotated in 30° increments, all other Cα-Cβ bonds were rotated in 60° increments, and all remaining rotatable bonds were rotated in 120° increments. Internal energies were calculated for each conformation, taking into account electrostatic and van der Waals terms; the sum of these terms was constrained to within 10 kcal/mole of the minimum energy conformation. All conformations meeting the energy requirements were used in the docking calculations. Atoms with only one conformation (main-chain and Cβ atoms, and buried residues) were defined as the rigid fragment of the molecule. The remaining flexible atoms were converted to hierarchy format. This typically resulted in seven to 40 residues with multiple conformations.

Treatment of the mutants

The biopolymers package in SYBYL was used to substitute selected residues, and a systematic search generated multiple conformations. Fifteen mutants were explicitly created for BLIP (Huang et al. 2000; Selzer et al. 2000), and 25 were explicitly created for ovomucoid and BPTI. The 25 mutants included charged and neutral representations for Asp, Glu, Arg, Lys, and His. Desolvation values for each residue were calculated using HYDREN (Rashin and Namboodiri 1987). The value for glycine was then subtracted from the desolvation of each residue to determine a relative desolvation value for the side chain. If, upon docking, the residue was in contact with the receptor, the full desolvation energy was subtracted from the substituted residue's energy score.

Generating sphere sets and focusing

Ligand orientations were determined using the sphere matching method (Shoichet et al. 1992). An initial sphere set and at least one larger sphere set were used for each receptor. Because each ligand was relatively large, the number of resulting spheres used to produce orientations was also large. To limit the exponential growth of the number of orientations, the ligand surface and spheres were spatially divided in up to four smaller subclusters to represent different parts of the ligand. These subclusters were each docked sequentially, and their results summed. Alternate surfaces of the ligand barstar were not evaluated because the other convex regions of the ligand had B-factors greater than 80 on main-chain atoms, and were hence removed from the calculation. Similarly, alternate surfaces were not docked for lysozyme. The alternate convex regions were localized near residues 128–129 and 101–103. The 132L structure has both backbone and side-chain atoms in these regions with B-factors of 100. These residues were omitted from the docking calculations, and these interfaces were not explicitly docked to FAB. Like the receptor, another larger set of spheres was generated in the region of each subcluster. These larger sets were used in orientation refinement through focusing (Shoichet et al. 1992) to allow additional orientations, in the neighborhood of an early, potentially favorable orientation, to be sampled. The number of orientations and times reported for docking reflects the sum of the multiple independent docking runs using different surfaces of the ligands. When docking the mutants, the orientation search space was limited to the sphere cluster nearest the mutation.

Docking scheme

For each orientation of the ligand in the binding site, DOCK places and scores the rigid fragment and then attempts to build each side chain. At this stage each side chain is explored only until one conformation meets the docking requirements. After the algorithm has confirmed that at least one side chain can be built at each position, the remaining side chain conformations are explored. All conformations of each side chain are explored, pruning only when clashes occur with the binding site. The best scoring conformation of each side chain is saved. After all conformations of all side chains have been evaluated, the best side chain conformations are recombined to produce the best ligand conformation for each orientation. The recombination process is repeated for each residue substitution. The entire build-up process is repeated for each orientation of the ligand.

Once the docking calculations were completed, the best scoring near-native and non-native docked complexes were displayed graphically and evaluated to ensure there were no clashing groups due to violations of the additivity assumptions or to molecules extending outside of the energy grids, and hence, not being fully scored. In one case, docking CI-2 to the unbound conformation of subtilisin, two flexible side chains, Lys72 and Leu73, were found in conformations that clashed with each other; such conformations were removed manually from the docking results.

Acknowledgments

This work was supported by GM59957 from the NIH (to B.K.S.). We thank Binqing Wei, Susan McGovern, and Beth Beadle for reading this manuscript.

The publication costs of this article were defrayed in part by payment of page charges.This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

References

Buckle, A.M., Schreiber, G., and Fersht, A.R. 1994. Protein–protein recognition: Crystal structural analysis of a barnase–barstar complex at 2.0-A resolution. *Biochemistry* **33:** 8878–8889.

Camacho, C.J., Gatchell, D.W., Kimura, S.R., and Vajda, S. 2000. Scoring docked conformations generated by rigid-body protein–protein docking. *Proteins* **40:** 525–537.

Cherfils, J., Duquerroy, S., and Janin, J. 1991. Protein–protein recognition analyzed by docking simulation. *Proteins* **11:** 271–280.

Claussen, H., Buning, C., Rarey, M., and Lengauer, T. 2001. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **308:** 377–395.

Connolly, M.L. 1986. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers* **25:** 1229–1247.

Conte, L.L., Chothia, C., and Janin, J. 1999. The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285:** 2177–2198.

Ewing, T.J., Makino, S., Skillman, A.G., and Kuntz, I.D. 2001. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **15:** 411–428.

Fujinaga, M., Sielecki, A.R., Read, R.J., Ardelt, W., Laskowski, M., Jr., and James, M.N. 1987. Crystal and molecular structures of the complex of alpha-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 A resolution. *J. Mol. Biol.* **195:** 397–418.

Gabb, H.A., Jackson, R.M., and Sternberg, M.J. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272:** 106–120.

Gilson, M.K. and Honig, B.H. 1987. Calculation of electrostatic potentials in an enzyme active site. *Nature* **330:** 84–86.

Hart, T.N. and Read, R.J. 1992. A multiple-start Monte Carlo docking method. *Proteins* **13:** 206–222.

Huang, W., Zhang, Z., and Palzkill, T. 2000. Design of potent beta-lactamase inhibitors by phage display of beta-lactamase inhibitory protein. *J. Biol. Chem.* **275:** 14964–14968.

Jackson, R.M., Gabb, H.A., and Sternberg, M.J. 1998. Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. *J. Mol. Biol.* **276:** 265–285.

Jiang, F. and Kim, S.H. 1991. “Soft docking”: matching of molecular surface cubes. *J. Mol. Biol.* **219:** 79–102.

Kimura, S.R., Brower, R.C., Vajda, S., and Camacho, C.J. 2001. Dynamical view of the positions of key side chains in protein–protein recognition. *Biophys. J.* **80:** 635–642.

Krowarsch, D., Dadlez, M., Buczek, O., Krokoszynska, I., Smalas, A.O., and Otlewski, J. 1999. Interscaffolding additivity: Binding of P1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. *J. Mol. Biol.* **289:** 175–186.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. 1982. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161:** 269–288.

Lamb, M.L., Burdick, K.W., Toba, S., Young, M.M., Skillman, A.G., Zou, X., Arnold, J.R., and Kuntz, I.D. 2001. Design, docking, and evaluation of multiple libraries against multiple targets. *Proteins* **42:** 296–318.

Lee, B. and Richards, F.M. 1971. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55:** 379–400.

Lorber, D.M. and Shoichet, B.K. 1998. Flexible ligand docking using conformational ensembles. *Protein Sci.* **7:** 938–950.

Lu, S.M., Lu, W., Qasim, M.A., Anderson, S., Apostol, I., Ardelt, W., Bigler, T., Chiang, Y.W., Cook, J., James, M.N., Kato, I., Kelly, C., Kohr, W., Komiyama, T., Lin, T.-Y., Ogawa, M., Otlewski, J., Park, S.-J., Qasim, S., Ranjbar, M., Tashiro, M., Warne, N., Whatley, H., Wieczovek, M., Wilusz, T., Wynn, R., Zhang, W., and Laskowski, Jr., M. 2001. Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proc. Natl. Acad. Sci.* **98:** 1410–1415.

Lu, W., Apostol, I., Qasim, M.A., Warne, N., Wynn, R., Zhang, W.L., Anderson, S., Chiang, Y.W., Ogin, E., Rothberg, I., Ryan, K., and Laskowski, Jr., M. 1997. Binding of amino acid side-chains to S1 cavities of serine proteinases. *J. Mol. Biol.* **266:** 441–461.

Meng, E.C., Shoichet, B.K., and Kuntz, I.D. 1992. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13:** 505–524.

Moont, G., Gabb, H.A., and Sternberg, M.J. 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35:** 364–373.

Norel, R., Lin, S.L., Wolfson, H.J., and Nussinov, R. 1994. Shape complementarity at protein–protein interfaces. *Biopolymers* **34:** 933–940.

———. 1995. Molecular surface complementarity at protein–protein interfaces: The critical role played by surface normals at well placed, sparse, points in docking. *J. Mol. Biol.* **252:** 263–273.

Norel, R., Petrey, D., Wolfson, H.J., and Nussinov, R. 1999. Examination of shape complementarity in docking of unbound proteins. *Proteins* **36:** 307–317.

Palma, P.N., Krippahl, L., Wampler, J.E., and Moura, J.J. 2000. BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins* **39:** 372–384.

Petrosino, J., Rudgers, G., Gilbert, H., and Palzkill, T. 1999. Contributions of aspartate 49 and phenylalanine 142 residues of a tight binding inhibitory protein of beta-lactamases. *J. Biol. Chem.* **274:** 2394–2400.

Rashin, A.A. and Namboodiri, K.A. 1987. A simple method for the calculation of hydration enthalpies of polar molecules with arbitrary shapes. *J. Phys. Chem. US* **91:** 6003–6012.

Read, R.J., Fujinaga, M., Sielecki, A.R., and James, M.N. 1983. Structure of the complex of *Streptomyces griseus* protease B and the third domain of the turkey ovomucoid inhibitor at 1.8-A resolution. *Biochemistry* **22:** 4420–4433.

Richmond, T.J. 1984. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.* **178:** 63–89.

Ritchie, D.W. and Kemp, G.J. 2000. Protein docking using spherical polar Fourier correlations. *Proteins* **39:** 178–194.

Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A., and Kuhn, L.A. 1998. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins* **33:** 74–87.

Selzer, T., Albeck, S., and Schreiber, G. 2000. Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.* **7:** 537–541.

Shoichet, B.K. and Kuntz, I.D. 1991. Protein docking and complementarity. *J. Mol. Biol.* **221:** 327–346.

———. 1993. Matching chemistry and shape in molecular docking. *Protein Eng.* **6:** 723–732.

———. 1996. Predicting the structure of protein complexes: A step in the right direction. *Chem. Biol.* **3:** 151–156.

Shoichet, B.K., Bodian, D.L., and Kuntz, I.D. 1992. Molecular docking using shape descriptors. *J. Comput. Chem.* **13:** 380–397.

Strynadka, N.C., Jensen, S.E., Alzari, P.M., and James, M.N. 1996. A potent new mode of beta-lactamase inhibition revealed by the 1.7 A X-ray crystallographic structure of the TEM-1–BLIP complex. *Nat. Struct. Biol.* **3:** 290–297.

Totrov, M. and Abagyan, R. 1994. Detailed ab initio prediction of lysozyme–antibody complex with 1.6 A accuracy. *Nat. Struct. Biol.* **1:** 259–263.

Vakser, I.A. 1995. Protein docking for low-resolution structures. *Protein Eng.* **8:** 371–377.

Vakser, I.A., Matar, O.G., and Lam, C.F. 1999. A systematic study of low-resolution recognition in protein–protein complexes. *Proc. Natl. Acad. Sci.* **96:** 8477–8482.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106:** 765–784.

Weiss, G.A., Watanabe, C.K., Zhong, A., Goddard, A., and Sidhu, S.S. 2000. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci.* **97:** 8950–8954.

Wells, J.A. 1990. Additivity of mutational effects in proteins. *Biochemistry* **29:** 8509–8517.

Weng, Z., Vajda, S., and Delisi, C. 1996. Prediction of protein complexes using empirical free energy functions. *Protein Sci.* **5:** 614–626.