Protein docking along smooth association pathways

Carlos J. Camacho* and Sandor Vajda

Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215

Edited by Peter G. Wolynes, University of California at San Diego, La Jolla, CA, and approved July 6, 2001 (received for review March 26, 2001)

We propose a docking method that mimics the way proteins bind. The method accounts for the dominant driving forces at the different length scales of the protein binding process, allowing for an efficient selection of a downhill path on the evolving receptor-ligand-free energy landscape. Starting from encounter complexes with as much as 10 Å rms deviation from the native conformation, the method locally samples the six dimensional space of rigid-body receptor-ligand structures subject to a van der Waals constraint. The sampling is initially biased only by the desolvation and electrostatic components of the free energy, which capture the partial affinity of unbound structures that are more than 4 Å away from the native state. Below this threshold, improved discrimination is attained by adding an increasing fraction of the van der Waals energy to the force field. The method, with no free parameters, was tested in eight different sets of independently crystallized receptor-ligand structures consistently predicting bound conformations with the lowest free energies and appropriate stability gap around 2 Å from the native complex. This multistage approach is consistent with the underlying kinetics and internal structure of the free energy funnel to the bound state. Implications for the nature of the protein binding pathways are also discussed.

The goal of protein docking is to obtain a model for the bound complex from the coordinates of the unbound component molecules. Current docking methods evaluate a vast number of docked conformations by simple functions that measure surface complementarity. However, in addition to near-native states, these methods produce many false positives, i.e., structures with good surface complementarity but high root mean square deviations (rmsd). Substantial efforts have been devoted to the development of methods to eliminate the false positives. Approaches involve the reranking of the complex structures by using scoring functions that account for the chemical affinity between the individual molecules, and the refinement of interacting surfaces (1–7). Although these procedures improve the discrimination such that conformations with less than 5 Å rmsd are generally found within the top ten to hundred structures, for most complexes the highest ranked structures are still far from the native.

Based on the mapping of the interaction free energy between a receptor and its ligand (8), we have previously concluded that a reasonable approach to successfully predicting docked conformations was to divide the problem into two steps (4). The first step entails the identification of the binding region (within 10 Å rmsd), emulating the diffusional search of the ligand for its target on the receptor surface. Before establishing substantial surface contacts, receptor-ligand association is governed by electrostatic and desolvation interactions, and hence the approximate binding region can be found by mapping these *smooth* components of the free energy in the conformational space of encounter complexes (8). Another option is the use of the low resolution docking method (7), which removes the details of interacting proteins to match the resulting smooth geometric forms. Both methods predict the broadly defined binding mode of the two proteins, but are unable to describe the specific interactions at the atomic level. The second step consists on the refinement of this broad binding region to atomic scale. At this stage, surface complementarity (9), led by short-range van der Waals (vdW) forces, plays a crucial role on the stability and specificity of the high affinity complex.

In this paper, we present a docking algorithm that, inspired by the general principles governing protein binding, docks or refines complex structures with as much as 10 Å rmsd from the bound state to 2 Å rmsd. In particular, the method embodies the changes in protein–protein interactions as the process moves along the association pathways. Fig. 1 illustrates the stages in the free energy as the receptor and ligand approach the bound state along some association pathway. At separation between the proteins on the order of 10 Å, the interactions are purely electrostatic and partial desolvation effects, resulting in a free energy surface that is relatively smooth along some arbitrary configurational coordinate measuring the rmsd from the bound or native conformation (8). As the proteins get closer, the occurrence of vdW interactions yield favorable contributions in several states, including the native state, leading to mostly steric energy barriers (Fig. 1*B*). Finally, once the molecules are fully desolvated, the free energy surface becomes very rugged because of the high sensitivity of the vdW interactions to structural perturbations (Fig. 1*C*).

From the point of view of kinetics, Fig. 1 suggests that protein binding should entail distinct kinetic regimes where different driving forces govern the binding process at different times. This scenario is summarized in Fig. 2 (8, 10, 11), where a sketch of the free energy funnel corresponding to Fig. 1 is plotted against a reaction coordinate. The funnel distinguishes three kinetic regimes. First, nonspecific diffusion (regime I) brings the molecules to close proximity. Second, in the recognition stage (regime II), the chemical affinity steers the molecules into relatively well oriented encounter complexes ($\approx$5 Å), overcoming the mostly entropic barrier to binding. Brownian dynamics simulations of this regime (10) were also found to be consistent with a significant narrowing of the binding pathway to the final bound conformation, as suggested by Fig. 1*B*. Finally, regime III corresponds to the docking stage where short-range forces mold the high affinity interface of the complex structure.

As already mentioned, earlier attempts to address the problem of predictive protein docking have been based on the thermodynamic hypothesis, which reduces the search of the complex structure to the minimization of a potential approximating the free energy. However, as sketched in Fig. 1*C*, states that are separated by few angstroms in the configurational space may be separated by large steric barriers in energy space. Thus, the straightforward minimization on such a landscape results in the well known multiple minima problem. Significant steps toward the resolution of this problem have been achieved by novel methods proposed by Scheraga and collaborators (12), and others (13, 14). These new algorithms avoid the multiple minima problem by smoothing the landscape in Fig. 1*C*. The above notwithstanding, the large number of possibilities in which a protein can bind to a substrate, together with the ruggedness of

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: rmsd, rms deviation; vdW, van der Waals.

*To whom reprint requests should be addressed. E-mail: ccamacho@bu.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

 www.pnas.org/cgi/doi/10.1073/pnas.181147798

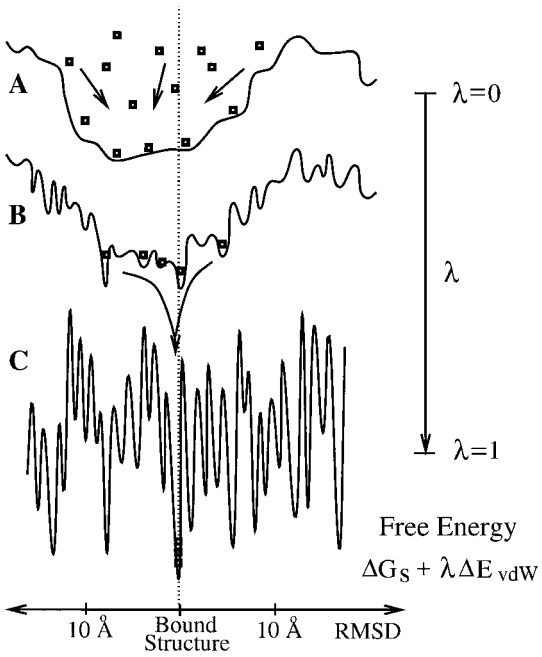

Fig. 1. Shapes of the binding free energy landscape as a function of some arbitrary coordinate measuring the rmsd from the native conformation. (*A*) Free energy corresponding to the smooth component ΔG_s, which dominates the interactions of partially desolvated encounter complexes (8) near the binding region ($\lambda = 0$). (*B*) Intermediate free energy mimicking the transition between the ''smooth'' and the ''rugged'' free energy ($\lambda \simeq 0.5$). (*C*) Free energy of a fully desolvated interface of a receptor-ligand system ($\lambda = 1$). Square symbols portray structures that, driven by the tunable landscape, are funneled into a single minimum of the free energy.

the free energy surface, has rendered it almost impossible for theoreticians to consistently use the thermodynamic principle to predict bound structures from separately crystallized proteins.

The tunable docking method proposed here is based on a kinetic approach that builds on the aforementioned multistage

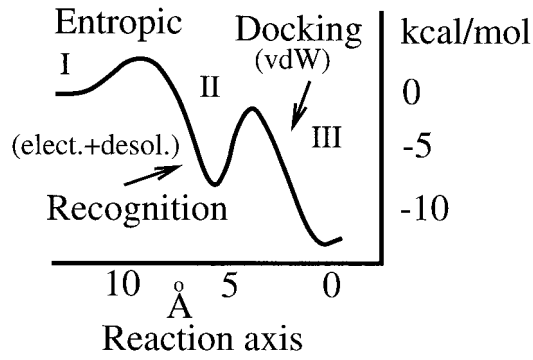

Fig. 2. Binding free energy funnel. The first barrier to binding is for the most part entropic, accounting for the loss in rotational and translational degrees of freedom. The height of this barrier depends on the interplay between the long-range electrostatic steering and the intermediate range desolvation forces. The former can, in some circumstances, eliminate this barrier (23, 24). As suggested by Fig. 1, the steric barrier associated with structural rearrangements is deeper in the binding pocket. For diffusion limited processes, the maximum height of this barrier must be lower than the barrier for the ligand to escape back to solution, implying that the recognition process is the rate-limiting step. Finally, kinetic estimates suggest that the steric activation barrier is significantly higher than the entropic one, suggesting a relatively slow transition, on the order of milliseconds, to the bound structure.

protein binding process, and the inherently smooth free energy governing the initial affinity between a receptor and its ligand. The method slowly tunes in the contribution of the vdW energy in the ranking free energy (see Fig. 1), from the smooth components of the free energy ($\lambda = 0$), to the full free energy dominated by vdW interactions ($\lambda = 1$). Starting from complexes within the basin of attraction, around 10 Å rmsd from the native complex structure, the method successfully predicts bound conformations within 2 Å rmsd. Interestingly, productive binding pathways also display some of the expected internal structure of the free energy funnel shown in Fig. 2.

Methods

Crystal Structures. We present docking applications to eight different sets of independently crystallized (unbound) protein pairs, with known complex structure (see Table 1). The protein structures are denoted by their four letter code in the Protein Data Bank (PDB). We study representative systems from four major classes of protein–protein complexes: the barnase-barstar (PDB code 1brs) system from the RNase-inhibitor family; five systems from the protease-inhibitor family, including the trypsin-bovine pancreatic trypsin inhibitor (2ptc) and subtilisin-streptomyces (2sic) and -chymotrypsin (2sni) inhibitor; the serine esterase complex acetylcholinesterase-fasciculin II (1fss); and hen egg white lysozyme bound to antibody Fab D44.1 (1mlc) from the antigen-antibody family. Some of these structures have been analyzed in numerous docking studies (4–7), and most of them are generally considered to be challenging systems when the separately determined protein structures are used for docking. To test the generality of the method, we also make sure that these systems include different binding specificities. For instance, binding of barnase and barstar (1brs) is driven by electrostatic forces, whereas, for the protease inhibitor complexes 1cho and 1ppf, desolvation is the dominant driving force. The complexes 2ptc and 1mlc have basic residues at key positions in the binding pocket, and their flexibility is a further challenge.

Initial Conformations. In this paper, we do not directly address the question of how the initial set of docked structures is generated. The only restriction or assumption is that these structures are within the basin of attraction of the binding region which, based on previous computations, has been established to be around 10 Å rmsd from the crystal complex structure (4, 8). The initial set of conformations, $\{A\}$, for 1ppg/2ovo, 5cha/2ovo, and 1mlb/1lza correspond to a cluster of encounter complexes around the lowest binding free energy pocket between receptor and ligand found in a previous discrimination analysis (4). The starting conformations for the other five protein pairs were extracted from a set of 10,000 structures generated by using the program DOT (15). The conformations were chosen such that their rmsd from the crystal structure were between 7 and 14 Å. It is interesting to note that in several cases the output from DOT resulted in few or no structures below 5 Å rmsd.

Both of these methods produce a relatively large number of conformations around the binding region. Because the docking method depends on the set of initial structures, for this study we selected clusters composed of structures that would differ among each other by 8–12 Å rmsd. We note that in general the average structure would be less that 10 Å rmsd from the complex—typically, around 7 Å rmsd. We also checked that the deviations from the native structure were both translational and rotational. Indeed, superimposing the centers of the initial structures and the target we find that the typical rmsd arising from the difference in rotational states was between 4 and 7 Å.

Table 1. Protein complexes studied

Complex PDB	Receptor	Ligand	Unbound PDB	Unbound PDB
1ppf	Human leukocyte elastase	OMTKY	1ppg	2ovo
1cho	α-chymotrypsin	OMTKY	5cha	2ovo
1fss	Acetylcholinesterase	Fasciculin-II	2ace	1fsc
1brs	Barnase	Barstar	1a2p	1a19
2sic	Subtilisin BPN′	Streptomyces inhibitor	2st1	3ssi
2ptc	Trypsin	BPTI	2ptn	6pti
2sni	Subtilisin novo	Chymotrypsin inh.2	2sbc	2ci2
1mlc	Fab D44.1	Hen egg lysozyme	1mlb	1lza

Free Energy Decomposition. The binding free energy is computed by the expression (4)

$$\Delta G = \Delta G_s + \Delta E_{vdW}, \quad [1]$$

where ΔG_s will be referred to as the smooth component of the free energy,

$$\Delta G_s = \Delta E_{coul} + \Delta G_{sol} - T\Delta S_{sc.} \quad [2]$$

The term ΔE_{coul} corresponds to the *direct* electrostatic energy; ΔG_{sol} is the desolvation free-energy change due to transferring the atoms, buried in the interface, from the solvent to a protein environment; ΔS_{sc} is the side-chain entropy loss; and T is the temperature. The van der Waals energy is denoted by ΔE_{vdW}. We note that the binding free energy expression generally also includes the change in internal energy and a further term, $T\Delta S$, associated with the loss of rotational, translational, and vibrational entropy. However, in the present analysis, the protein backbones are held rigid at all times, and the relatively small changes in the internal energy are neglected. The $T\Delta S$ term is omitted because, for a given pair of molecules, it depends weakly on the structure.

The electrostatic term ΔE_{coul} is calculated by using a distance-dependent dielectric (16) equal to $4r$, enforcing a minimum atom-to-atom distance separation equal to the sum of their corresponding vdW radii to avoid artificial overlaps. We estimate the full contribution of the desolvation forces by using Zhang *et al.* atomic contact potential (ACP) (17). This method is an adaptation of a method first introduced by Miyazawa and Jernigan (18). The ACP potential includes the self-energy change on desolvating charge or polar atom groups and side-chain entropy loss, i.e., $\Delta G_{ACP} = \Delta G_{sol} - T\Delta S_{sc}$. The van der Waals energy, ΔE_{vdW}, is computed by using the standard Lennard-Jones potential as implemented in CHARMM (19). All structural minimizations have been done by using ABNR (adopted basis Newton-Raphson) steps and the CHARMM-19 potential with polar hydrogens only, distance-dependent dielectrics $\varepsilon = 4r$, and *fixed backbone*.

Method: Tunable Docking. The algorithm refines a cluster $\{A\}$ of ten rigid body receptor-ligand structures, constrained to around 10 to 13 Å rmsd from the complex. To specify the geometry of a complex structure, the receptor is centered at the origin of the coordinate system, and the position of the ligand is described in terms of 12 variables. Three of the these, the center-to-center distance d_{cm} between the two molecules, and two Euler angles θ_{cm} and ϕ_{cm}, are used to specify the ligand's geometrical center. The relative orientation of the ligand is specified by the remaining 9 variables, corresponding to the Cartesian coordinates of the three unitary vectors, (x_i, y_i, z_i) with $i = 1,2,3$. The three vectors define a coordinate system fixed on the ligand's center. Because the vectors are orthonormal to each other, only three degrees of freedom are independent. However, the use of this redundant Cartesian system facilitates the sampling of the conformational space. We also define the vector $\vec{\sigma}$, where σ_k is the standard deviation of the kth variable in the set $\{A\}$, constrained to a minimum and maximum of 1° and 6° for θ_{cm} and ϕ_{cm}, 0.5 Å and 3 Å for d_{cm}, and 0.08 Å and 0.5 Å for (x_i, y_i, and z_i), respectively.

The docking method consists of the following six steps.

(*i*) *Preprocessing.* The overlaps in the initial set of conformations $\{A\}$ are removed by rigidly pulling apart the molecules along their center-to-center axis. A second set, $\{B\}$, is initialized with the adjusted set $\{A\}$ ranked according to ΔG_s, and contains the top ten best ranked complexes that are sampled during the docking procedure.

(*ii*) *Sampling.* The goal of this step is to generate new structures by a modified version of the nonlinear simplex algorithm (20) using the structures in $\{A\}$ as templates. We randomly select two structures α and β from $\{A\}$ to obtain a new point $\vec{x}$ with the coordinates d_{cm}^{α}, θ_{cm}^{α}, ϕ_{cm}^{α}, and $(x_i^{\beta}, y_i^{\beta}, z_i^{\beta})$, and define the centroid of the simplex $\vec{c}$ as the average of the coordinates in $\{A\}$ not including $\vec{x}$. The basic operations in the simplex method are reflection, expansion, and contraction (20). We perform randomized versions of these operations to sample along the vector $\vec{r}$ from $\vec{x}$ to $\vec{c}$. First, reflection of $\vec{x}$ is used to generate the structure $\gamma = \vec{c} + \vec{r} + \vec{\varepsilon}$, where $\vec{\varepsilon}$ is a vector of uniformly distributed random real numbers between $\pm 2\vec{\sigma}$. If γ does not improve the set $\{B\}$ (see step *v* of the algorithm), then we perform a contraction to generate a new $\gamma = \vec{c} + 0.5\vec{r} + \vec{\varepsilon}$; or, if γ improves the top ranked structure in $\{B\}$, then we perform an expansion and generate $\gamma = \vec{c} + 2\vec{r} + \vec{\varepsilon}$; otherwise, a new point $\vec{x}(\alpha,\beta)$ is selected.

(*iii*) *Constrained vdW optimization.* The energy of the receptor-ligand structure γ from *ii* is minimized by using 30 adopted basis Newton-Raphson (ABNR) steps in CHARMM, and the vdW energy of the resulting complex is evaluated. The compactness of the docked conformation is further improved by using the following steps: (*a*) the ligand is brought closer to the receptor by translating it by an amount equal to $-0.02 \times d_{cm}$ along the axis joining the centers of the molecules; (*b*) 30 ABNR steps are applied to minimize the CHARMM energy of the structure, and its vdW energy is evaluated; (*c*) if the new vdW energy is larger than its previous value by more than 10 kcal/mol, then the lower energy structure is minimized by 300 steps. Otherwise, we return to step *a*.

(*iv*) *Ranking by the free energy.* We compute the electrostatic and desolvation free energy of the minimized complex structure γ. Structures with large overlaps, i.e., with vdW energies larger than 60% of the lowest observed vdW energy, are discarded. If γ is the ith structure sampled during the search, then the target function is evaluated as

$$\Delta G_i^{\gamma} = \Delta G_s + \lambda_i \Delta E_{vdw}, \quad [3]$$

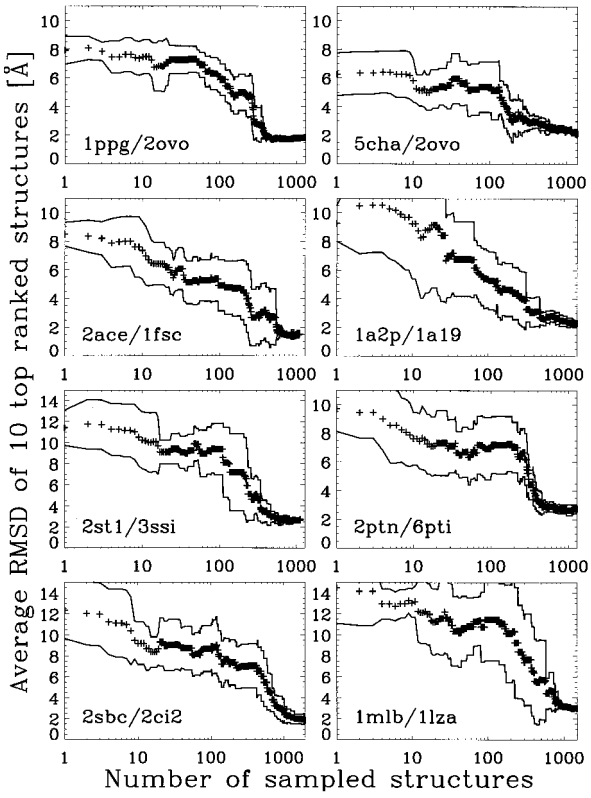

Fig. 3. rmsd as a function of the number of sampled structures. The rmsd is computed with respect to the optimal alignment of the unbound ligand in the bound native structure. The + symbols correspond to the average rmsd of the top 10 structures stored in set {*B*}. Solid lines indicate the standard deviation of the average rmsd.

where $\lambda_i = i/1{,}000$ for $i < 1{,}000$, and $\lambda_i = 1$ otherwise. If ΔG_i^{γ} is lower than ΔG_i for any of the structures in $\{B\}$, then γ is added to $\{B\}$, while the highest energy structure is dropped.

(*v*) *Updating* {*A*}. As described, set $\{B\}$ always contains the 10 lowest free energy structures sampled. These structures are used to periodically update set $\{A\}$, but with some delay to avoid shifting the entire cluster because of single outliers. The first update is at $i = 140$. For $i > 140$, $\{A\}$ is updated every time when the best ranked structure in $\{B\}$ improves, and also at periods of 140 sampled structures.

(*vi*) *Convergence.* Once all the σ_k are less than their corresponding minima, the method has converged to a solution where the docked structures in the set $\{B\}$ are within 0.25 Å of each other. Otherwise, the search resumes in step *ii*.

Results

Our results are summarized in Fig. 3 where the average rmsd of the top 10 best ranked structures in $\{B\}$ are plotted as a function of the number of structures sampled. The rmsd is computed with respect to the unbound "native-like" ligand—i.e., aligned as in the native complex structure. As indicated by the standard deviation of the average rmsds, the method converges toward a unique binding site (within 0.5 Å rmsd) after ≈500 sampled structures, or about a day of CPU time on a RISC 10000 Silicon Graphics computer. For two cases, 2sbc/2ci2 and 1mlb/1lza, we observed a slower rate of convergence.

To appreciate how the refinement proceeds from the initial structures to the final prediction, Fig. 4 shows the *locus* of the initial set of conformations, the unbound "native-like" ligand, and the predicted docked structure for four complexes, as well

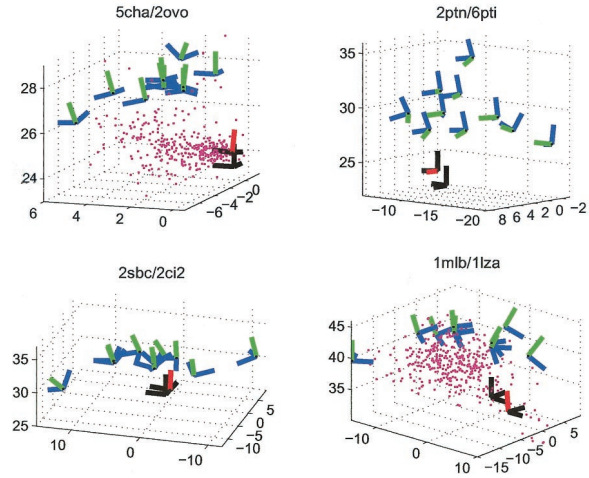

Fig. 4. *Loci* of the initial set of structures and of the predicted structures with respect to the native state. The 10 initial structures have their *z* axis drawn in green. The predicted structure has its *z* axis drawn in red, and the axes of the "native-like" structure are in black. For clarity, we show only the position of the center of mass of the first 400 structures sampled in two systems (magenta dots). All distance units are in Å. The center of the receptors are at (0,0,0).

as the centers of the 400 initial structures sampled for two systems. In all cases, we find that the method flows nicely to the correct region of the phase space. The reason is that, as shown in Fig. 5, the total free energy of the sampled structures as a function of rmsd for all complexes is a good discriminator, with no false positives. It is also interesting to note that we find an

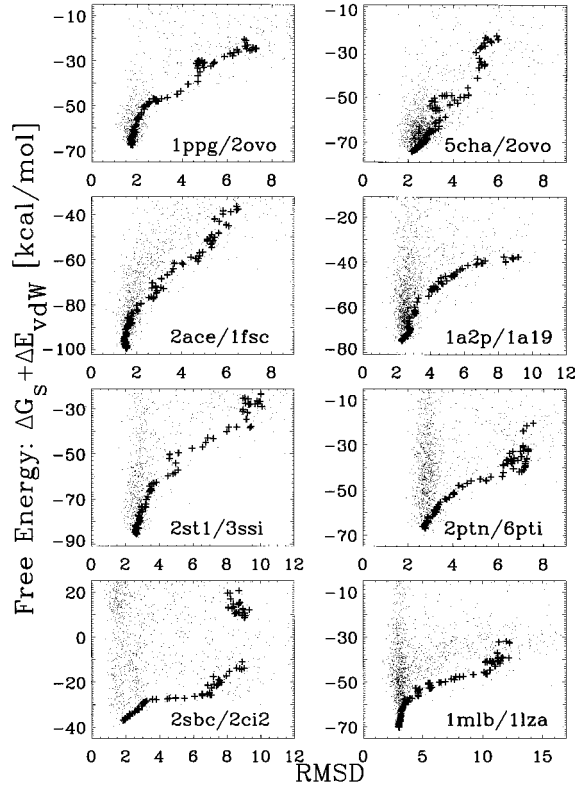

Fig. 5. Free energy as a function of rmsd. The + symbols correspond to the average free energy of the best 10 ranked structures. Dots denote the free energy of the sampled structures.

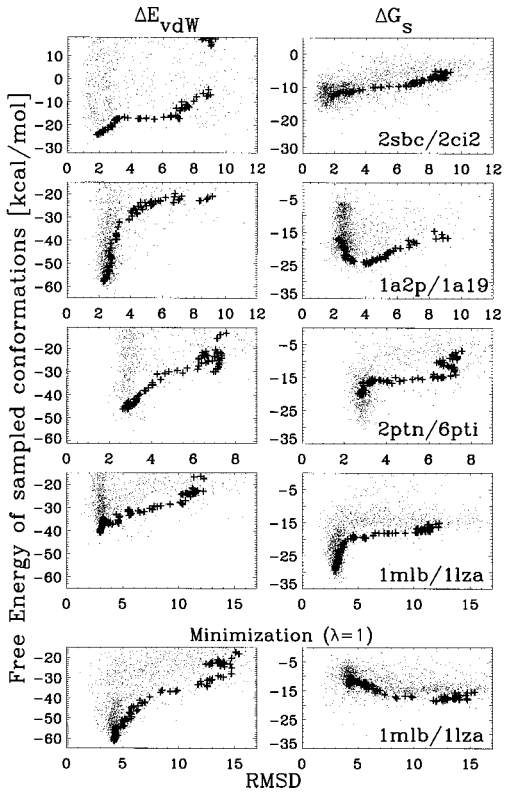

Fig. 6. Free energy components as a function of rmsd. Symbols are as in Fig. 5. The vdW energy is represented by the Lennard-Jones potential. The *Top four* panels correspond to the decomposition of the free energy for the same data shown in Fig. 5. The *Bottom* panel corresponds to a rerun of the docking algorithm for 1mlb/1lza using the full free energy ($\lambda = 1$) as the target function.

energy gap of around 10 kcal/mol between our predictions and structures with rmsds larger than 4 Å.

The breakdown of the free energy to its two main components, ΔG_s and ΔE_{vdW}, is shown for four cases in Fig. 6. Most cases behave as for 2sbc/2ci2, where the smooth component of the free energy ΔG_s is a good discriminator between 4 to 10 Å rmsd away from the complex, whereas ΔE_{vdW} is typically better below 4 Å rmsd. The only exceptions were found for 2ptn/6pti and 1mlb/1lza, where the level of discrimination of ΔG_s is overall weak. These two complexes show a sharp decrease of ΔG_s in the final stage of the minimization. However, this feature does not discriminate between high and low rmsd structures; instead, it reflects only the close proximity of some charge groups. Another interesting case is 1a2p/1a19, which shows that ΔG_s is a good discriminator between 4 to 10 Å rmsd, but below this range it has a negative correlation. These observations are consistent with the fact that these three complexes show very little chemical affinity if the bound protein conformations are replaced by the unbound structures. Indeed, it has been shown that, before association, these complexes undergo side chain rearrangements that significantly improve their affinity (21). Because our method does not include any preprocessing of the crystal structures, it is not surprising that these cases resulted in docked conformations with the highest rmsd.

We have also checked that a straightforward minimization of the full free energy does not necessarily improve the predictions. A good example of this is shown at the bottom of Fig. 6, where fixing $\lambda = 1$ for 1mlb/1lza leads to a different minimum with an rmsd of 4 Å, as well as a very different behavior of ΔG_s and ΔE_{vdW}. The two minimizations, varying λ or fixing it, highlight

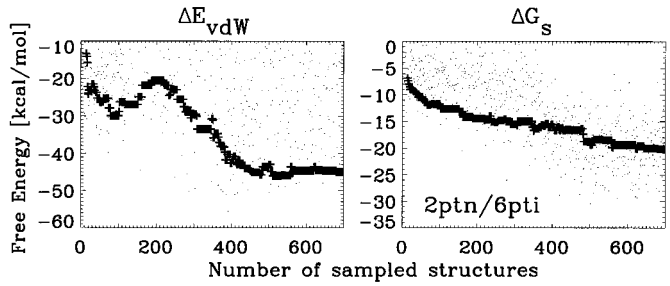

Fig. 7. Free energy components as a function of the number of sampled structures for 2ptn/6pti. Symbols are as in Fig. 5.

how the improvement of ΔG_s (or ΔE_{vdW}) can sometimes come at the expense of ΔE_{vdW} (or ΔG_s). In all likelihood, this behavior is artificially (21) enhanced by the poor positioning of side chains at the binding interface.

The thermodynamic analysis provided by Figs. 3, 5, and 6 might be somewhat misleading when analyzed independently of the dynamic process in which λ varies between 0 and 1. For example, for 2ptn/6pti, ΔG_s does not seem to be a good discriminator as a function of rmsd. However, as a function of the number of sampled structures (or "time" in a sense analogous to Monte Carlo time steps), ΔG_s works very well driving the system to lower free energies and lower rmsd structures (see Fig. 7). On the other hand, a full free energy minimization might trap the docking process in the high rmsd minimum apparent in the plot of ΔE_{vdW}.

For all eight of the test cases, we predict a complex structure with around 2 Å rmsd from the unbound "native-like" ligand. The worst prediction is the antibody-antigen 1mlb/1lza system, which predicts a 3 Å rmsd structure. Most predictions fare somewhat better when compared directly with the C-α chain of the native structure located within 10 Å of the interface of the cocrystallized complex structure. Table 2 shows, for the top ranked predictions, the rmsd with respect to the "native-like" unbound ligand, and the rmsd of both the predicted and the "native-like" unbound ligands with respect to the relevant C-α atoms of the native structure near the interface. Because the backbone of the unbound structures is kept rigid, even the best alignment of the unbound ligand to the bound structure yields around 1 Å rmsd. Strikingly, some of our predictions are within 1 Å or less of the binding site. However, for 1lza and 2ci2, the rmsd increases by as much as 0.5 Å when restricting consideration to the region within 10 Å of the interface. In these cases, the binding site in the unbound ligand structure is distorted by

Table 2. rmsd of predicted complex structures with respect to the unbound ligand aligned as in the complex crystal structure

Recep./ligand PDB codes	Prediction rmsd, Å	Prediction rmsd*, Å	Opt.Unb.lig. rmsd*, Å
1ppg/2ovo	1.85	1.36	1.19
5cha/2ovo	1.58	0.86	1.13
2ace/1fsc	1.59	1.78	0.97
1a2p/1a19	2.52	2.58	0.41
2st1/3ssi	2.79	1.12	0.83
2ptn/6pti	2.69	1.50	0.36
2sbc/2ci2	1.28	1.92	1.13
1mlb/1lza	3.03	3.51	0.88

For comparison, we list the rmsd of our predictions, both for all backbone atom and only for the ones in the interface, as well as for the optimal alignment of the unbound ligand to the native structure. rmsd* denotes the rmsd restricted to the C-α atoms within 10 Å of the native interface.

large misfolded side chains (Arg-45 and Met-40 in 1lza and 2ci2, respectively), thereby increasing the free energy of all near-native conformations. The rmsd for 1mlb/1lza is particularly large (3.51 Å for the restricted set of residues), but the rmsd of 3.03 Å for all C-α atoms indicates that, even for this large local deviation, the relative orientation of the two molecules in the complex is essentially correct. We also note that, even for this most difficult case, the method was able to refine the initial cluster of 12 Å rmsd structures to around 3 Å.

To study the convergence of the method and the robustness of the results, some of the calculations have been repeated by using different clusters of initial structures. We have checked that the predictions of the method are robust in the sense that small changes on the initial set of structures do not change the final rmsd by more than 0.5 Å. More generally, the details of the sampling method are not crucial for our results as long as the local character of the random sampling is preserved. However, the method fails to converge if the initial structures are too far apart from each other (around 20 Å rmsd). Docking tests of structures with cluster centers more than 15 Å away from the native often converge to unique structures, with free energies consistently higher than the docked structures near the binding region.

As already mentioned in *Methods*, the cluster centers of the systems studied here are typically around 7 Å rmsd away from the native. Reducing 7 Å to about 2 Å as seen in our results is a nontrivial problem, and the modified simplex method clearly performs very well. Assuming that at least some of the cluster points are in the region of attraction of the global minimum, the simplex method will move the entire cluster toward the native state. However, convergence to a false minimum can occur if, along this pathway, the cluster becomes small enough to be fully accommodated by the basin of attraction of a minimum. Three different mechanisms are used to reduce the possibility of such premature convergence. First, restricting consideration to electrostatic and solvation terms in the early stages of the search smoothes the target function, removing most of the local minima that could serve as traps. Second, the randomization in step *iii* of the algorithm substantially expands the region sampled by the method. Finally, it is important to select initial structures that are on the order of 10 Å from each other. For the small proteins considered in this paper, 10 Å appears to be the length scale of the region of attraction of the native state because of electrostatic and desolvation interactions. As the above discussion implies, despite the various countermeasures, convergence does not necessarily indicate that the global minimum has been found, and thus the reliability of results can be improved by performing repeated minimizations from different initial clusters.

Discussion

We focus on the prediction of bound complexes from independently crystallized (unbound) receptor and ligand structures starting from a set of structures within the basin of attraction of the binding site (10 Å rmsd from the native structure). Despite the structural differences between the cocrystallized complex and unbound structures, our method successfully predicts low rmsd structures, on the order of 2 Å, from the native for eight receptor-ligand systems. We emphasize that exactly the same method was used in all 8 complexes, and that the only presumption is that the initial set of structures are within the basin of attraction of the binding site (i.e., around 10 Å rmsd).

We provide strong evidence suggesting that it is possible to efficiently trail down the free energy binding funnel. Specifically, the method captures the basic mechanism of protein recognition as well as the short range tuning of the surface complementarity in the high affinity complex. The former is governed by the electrostatic and desolvation components of the free energy, which captures the partial affinity of the macromolecules when properly aligned near their binding site (8). As the binding funnel narrows (see standard deviation in Fig. 3), the vdW interactions become the dominant driving force on the biasing field, improving the surface complementarity between molecules. The robustness of our target function can be seen in Figs. 3 and 5, which show how the chemical affinity and vdW interactions complement each other to obtain the best possible fit, and is confirmed by the fact that exactly the same method yields reasonably accurate predictions for all of the eight systems considered here.

We argue that the evolution of the average top 10 ranked structures (+ symbols in the figures) should, to some degree, reflect the typical behavior of productive binding pathways. In accord with Fig. 2, some systems show a characteristic peak on the vdW energy at $\approx$5 Å rmsd when plotted against "time" (see Fig. 7). This transition peak is also seen sometimes in the full free energy. The peak corresponds to the transition point between the soft binding regime, controlled by the "smooth" chemical affinity, and the tight regime, controlled by the "rugged" vdW interactions. Because these interactions forcefully move the receptor and the ligand to close proximity, rotational and translational entropy plays no role on our docking method.

From a kinetic point of view, Figs. 2 and 7 are consistent with the view that the recognition process is dominated by a weakly bound intermediate whose affinity is mostly determined by the smooth free energy terms. This process can take place in a reasonably fast time scale ($\approx 10^{-8}$ s), in accord with recent Brownian dynamic simulations (10) and nonspecific aggregation data (22). On the other hand, the steric barriers arising from the snug fit of the interface should lead to longer time scales for docking to the final complex structure.

We are grateful to D. Gatchell for providing the DOT-generated structures. This research has been supported by Grants DBI-9904834 from the National Science Foundation and GM61867-01 from the National Institutes of Health.

1. Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988) *Proteins Struct. Funct. Genet.* **4,** 19–30.
2. Vajda, S., Sippl, M. & Novotny, J. (1997) *Curr. Opin. Struct. Biol.* **7,** 222–228.
3. Lazaridis, T. & Karplus, M. (1999) *Proteins Struct. Funct. Genet.* **35,** 132–152.
4. Camacho, C., Gatchell, D., Kimura, S. & Vajda, S. (2000) *Proteins Struct. Funct. Genet.* **40,** 525–537.
5. Jackson, R., Gabb, H. & Sternberg, M. (1998) *J. Mol. Biol.* **276,** 265–285.
6. Vakser, I. & Aflalo, C. (1994) *Proteins Struct. Funct. Genet.* **20,** 320–329.
7. Vakser, I., Matar, O. & Lam, C. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 8477–8482.
8. Camacho, C., Weng, Z., Vajda, S. & DeLisi, C. (1999) *Biophys. J.* **76,** 1166–1178.
9. Chothia, C. & Janin, J. (1975) *Nature (London)* **256,** 705–708.
10. Camacho, C., Kimura, S., DeLisi, C. & Vajda, S. (2000) *Biophys. J.* **78,** 1094–1105.
11. Camacho, C., DeLisi, C. & Vajda, S. (2001) in *Thermodynamics of the Drug-Receptor Interactions*, ed. Raffa, R. (Wiley, London).
12. Wawak, R. J, Pillardy, J., Liwo, A., Gibson, K. D & Scheraga, H. A. (1998) *J. Phys. Chem. A* **102,** 2904–2918.
13. Andricioaei, I. & Straub, J. (1998) *J. Comput. Chem.* **19,** 1445–1455.
14. Dill, K., Phillips, A. & Rosen, J. (1997) *J. Comp. Biol.* **4,** 227–239.
15. Eyck, L. T, Mandell, J., Roberts, V. & Pique, M. (1995) in *Proceedings of the 1995 ACM/IEEE Supercomputing Conference*, eds. Hayes, A & Simmons, M. (ACM Press, New York).
16. McCammon, J., Wolynes, P. & Karplus, M. (1979) *Biochemistry* **18,** 927–942.
17. Zhang, C., Vasmatzis, G., Cornette, J. & DeLisi, C. (1997) *J. Mol. Biol.* **267,** 707–726.
18. Miyazawa, S. & Jernigan, R. (1985) *Macromolecules* **18,** 534–552.
19. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4,** 187–217.
20. Nelder, J. A. & Mead, R. (1964) *Comput. J.* **7,** 308–313.
21. Kimura, S., Brower, R., Vajda, S. & Camacho, C. (2001) *Biophys. J.* **80,** 635–642.
22. Northrup, S. & Erickson, H. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 3338–3342.
23. Schreiber, G. & Fersht, A. (1996) *Nat. Struct. Biol.* **3,** 427–431.
24. Janin, J. (1997) *Proteins Struct. Funct. Genet.* **28,** 153–161.