Published in final edited form as:
*Proteins*. 2008 August 15; 72(3): 993–1004. doi:10.1002/prot.21997.

# Discrimination of near-native structures in protein-protein docking by testing the stability of local minima

**Dima Kozakov**[1], **Ora Schueler-Furman**[2], and **Sandor Vajda**[1]

[1] Department of Biomedical Engineering, Boston University, Boston, Massachusetts

[2] The Hebrew University of Jerusalem, Jerusalem, Israel

## Abstract

Fast Fourier Transform (FFT) correlation methods of protein-protein docking, combined with the clustering of low energy conformations, can find a number of local minima on the energy surface. For most complexes the locations of the near-native structures can be constrained to the 30 largest clusters, each surrounding a local minimum. However, no reliable further discrimination can be obtained by energy measures because the differences in the energy levels between the minima are comparable to the errors in the energy evaluation. In fact, no current scoring function accounts for the entropic contributions that relate to the width rather than the depth of the minima. Since structures at narrow minima loose more entropy, some of the non-native states can be detected by determining whether or not a local minimum is surrounded by a broad region of attraction on the energy surface. The analysis is based on starting Monte Carlo Minimization (MCM) runs from random points around each minimum, and observing whether a certain fraction of trajectories converge to a small region within the cluster. The cluster is considered stable if such a strong attractor exists, has at least 10 convergent trajectories, is relatively close to the original cluster center, and contains a low energy structure. We studied the stability of clusters for enzyme-inhibitor and antibody-antigen complexes in the Protein Docking Benchmark. The analysis yields three main results. First, all clusters that are close to the native structure are stable. Second, restricting considerations to stable clusters eliminates around half of the false positives, i.e., solutions that are low in energy but far from the native structure of the complex. Third, dividing the conformational space into clusters and determining the stability of each cluster, the combined approach is less dependent on a priori information than exploring the potential conformational space by Monte Carlo minimizations.

### Keywords

Fast Fourier Transform; Monte Carlo minimization; structure refinement; selection of near-native structures

## INTRODUCTION

The goal of protein-protein docking is to determine the structure of a complex in atomic detail, starting from the coordinates of the unbound component molecules.[1–3] Based on the thermodynamic hypothesis, at fixed temperature and pressure the Gibbs free energy of the macromolecule-solvent system reaches its global minimum at the native state of the complex. Thus, docking requires a computationally feasible free energy evaluation model and an effective minimization algorithm. Most of the current docking methods start with rigid body

*Corresponding author: Sandor Vajda, Department of Biomedical Engineering, Boston University, 44 Cummington street, Boston, MA 02215, USA, vajda@bu.edu, phone: 617-353-4757, fax : 617-353-6766.

docking which relies on simplified representations of the interaction energy, and perform the search in the 6D space of rotations and translations.[4] Although the rigid body assumption limits the applicability of the method, it is a good approximation for several classes of protein complexes. In particular, the transition from unbound to bound state in enzyme-inhibitor and antibody-antigen complexes mostly affects the orientation of side chains, and the changes in the backbone conformation are generally small, or are restricted to a few loop regions.

The Fast Fourier Transform (FFT) correlation approach, introduced in 1992 by Katchalski-Katzir and co-workers,[5] revolutionized rigid body protein-protein docking. The basic idea of the method is to represent one of the proteins (which will be identified as the receptor) on a fixed grid, the second protein (which will be referred to as the ligand) on a movable grid, and consider an interaction energy written in the form of a correlation function (or as a sum of a few correlation functions).[6] Since such energy functions can be efficiently calculated via Fast Fourier Transforms, one can exhaustively sample the conformational space of protein-protein complexes evaluating the energies for billions of conformations on the grids, and thus to dock proteins without any *a priori* information on the expected structure.[7,8] The original scoring function, introduced by Katchalski-Katzir et al.,[5] accounted only for shape complementarity, but was later extended to include additional terms representing electrostatic interactions,[9,10] or both electrostatic and solvation contributions.[11] More recently we have extended the method to be used with a scoring function that includes a pairwise interaction potential, thereby further increasing the number of near-native structures found.[6] In all scoring functions the shape complementarity term allows for overlaps, thereby accounting for the differences between bound and unbound (separately crystallized) structures.

Since the FFT correlation method performs exhaustive sampling on a dense grid, it necessarily samples near-native conformations, independent of the shape of the energy surface. However, due to the approximate nature of the energy function and the need for tolerating potential overlaps, the structures that are close to the native conformation do not necessarily have the lowest energies. In order to avoid eliminating such potentially useful conformations, it is necessary to retain a large number (usually 2,000 to 20,000) of low energy docked structures for further processing. The majority of these retained structures are false positives, i.e., conformations with low energy but far from the native. Thus, the initial docking yields a long list of candidate structures rather than a small number of models, and obtaining meaningful results requires some form of post-processing, which may include re-scoring of the docked conformations using a more accurate energy function, or refining the conformations followed by re-scoring.[12] These treatments usually improve discrimination and result in a number of near-native conformations among the 10 to 100 lowest energy structures, but in most cases are unable to eliminate all false positives.

Over the last few years we have developed a post-processing method that retains a number of low energy conformations, clusters them using pairwise RMSD as the distance measure, and then ranks the clusters according to their size, i.e., identifying conformations that have large numbers of neighbors.[13,14] The method is based on the observation that, in the free energy landscapes of partially solvated receptor-ligand complexes, the free energy attractor at the binding site generally has the greatest breadth among all local minima.[15] Hence, following the uniform sampling of the conformational space defined by translations and rotations of the ligand, the docked conformations that are below an energy threshold are expected to form the largest cluster around the native complex (Figure 1). This approach is confirmed by our results: analyzing the protein pairs in the Docking Benchmark Set,[16] we have shown that if the clusters are ranked by size (i.e., by the number of their members), then at least one near-native solution (with less than 10 Å RMSD from the native) is contained in the largest cluster for 31% of the ° complexes, in the 10 largest clusters for 74% of the complexes, and in the 30 largest clusters for 93% of the clusters.[13] Thus, adding conditions on the shape of the free energy surface

around the native complex can substantially reduce the number of candidate structures, but is still unable to eliminate all false positives.

Post-processing by clustering is based on the assumption that it is more likely to find near-native structures at minima with broad regions of attraction than at narrow minima with comparable depth. In this paper we go one step further, and for each retained local minimum investigate whether or not it is surrounded by a broad region of attraction. The analysis is similar to determining the stability of fixed points in dynamical systems. A fixed point is stable and it is called an attractor if all trajectories started in a neighborhood of the fixed point remain in the neighborhood. In good agreement with this analogy, we will examine the ”stability” of each local minimum by starting Monte Carlo simulations from random points around the cluster center in order to determine whether or not the trajectories converge to some small region within the region defined by the cluster. The idea is similar to the one introduced by Brutlag, Latombe, and co-workers,[17,18] who employed stochastic roadmap simulations to determine the number of Monte Carlo steps required for the escape of a ligand from various putative binding sites.

The algorithm we describe in this paper starts with rigid body docking based on the FFT correlation approach, identifies a number of low energy regions by clustering, and determines the ”stability” of these regions by Monte Carlo simulations. While the rigid body search is global but has to rely on simplified energy functions defined on a grid, the Monte Carlo analyses of stability are restricted to individual clusters but can involve more detailed energy functions and more thorough searches, possibly accounting for side chain flexibility. Exactly this type of Monte Carlo search has been implemented in the protein docking program Rosettadock by Baker and co-workers,[19] and hence we simply use RosettaDock for our stability analyses. Obviously RosettaDock can be and has been used for protein docking without any FFT-based global search. However, due to the side chain search and improved energy evaluation, RosettaDock requires extensive calculations, and the simulations can explore only limited regions of the conformational space on reasonable time scales. Although the Monte Carlo minimization (MCM) trajectories in RosettaDock can move ”uphill” and thus cross energy barriers, there is no guarantee that the search converges to the global minimum. In fact, RosettaDock and other Monte Carlo based docking methods such as ICM[20] include a first stage that uses simplified protein models and energy functions to explore the conformational space, and only then switch to simulations that involve models with more detailed geometry and more accurate energy functions.

The main advantage of the prediction scheme introduced in this work is that it combines a full low-resolution search with a high-resolution refinement, and that the quality of the refined structures are judged based on how far the high resolution model is from the low resolution start. This approach substantially reduces the chances that the search converges to a non-native structure with low energy, while all near-native structures are lost. Indeed, as we will show for a set of enzyme-inhibitor and antibody-antigen complexes, one can define stability criteria based on the convergence of relatively short Monte Carlo trajectories such that clusters with near-native states are always ”stable”, i.e., a certain fraction of trajectories converge to a smaller region within the region defined by the cluster. The lack of such convergence indicates that the cluster does not include near-native states, and hence can be excluded from further consideration, thereby reducing the number of potential complex structures. We will also show that the discrimination of clusters based on stability analysis can be more reliable than the discrimination based on relative energies.

# METHODS

## Rigid body docking

Docked conformations were generated using the ClusPro server[13] with the docking program ZDOCK.[11] ZDOCK is based of the Fast Fourier Transform correlation approach, and exhaustively samples the rigid body mutual orientations of the docking partners. The scoring function of ZDOCK is a weighted sum of energy terms representing shape complementarity (van der Waals energy), Coulombic electrostatics with the distance dependent dielectric permittivity $\varepsilon = 4r$, and a simplified implementation of the atomic contact potential score (ACP),[21] which essentially measures the solvation/desolvation contributions to the binding free energy. 2000 structures with the lowest values of the scoring function were retained for further evaluation. Note that we used ZDOCK 2.3, and newer versions of the program are currently available.

## Discrimination by clustering

The clustering of the retained conformations is based on the pairwise root mean square deviation of ligand structures, calculated for the atoms that are within 10 Å of any atom of the fixed receptor (to be referred to as ligand RMSD). We use a simple greedy algorithm to find the structures with the largest number of neighbors within a clustering radius $R_C$. As we described earlier,[14] the choice of $R_C$ depends on a clustering parameter $0 \leq \Delta \leq 1$, which is based on the histogram of pairwise RMSD values, and measures the depth of the separation between clusters. $\Delta = 1$ indicates perfect separation of inter-cluster and intra-cluster length scales. Such separation means that clustering is very easy, and that the use of the optimal radius $R_C$, calculated from the histogram of pairwise RMSD values,[14] substantially increases the number of near-native structures in the top clusters. The analysis of the docked structures for the proteins in the benchmark set[14] showed that for $\Delta \geq = 0.4$ the use of the optimal radius generally increases the number of the near-native predictions in the largest clusters. In contrast, for $\Delta < 0.4$ the choice of the calculated optimal radius does not necessarily improve the results, and hence it is better to use the default clustering radius of 9 Å. Once a clustering radius $R_C$ is selected, the structure with the highest number of neighbors within $R_C$ is considered as the center of the first cluster. The members of this cluster are removed, and we select the next structure with the highest number of neighbors from the remaining ligands until the set is exhausted, thereby generating 10 to 30 rank ordered clusters.[14]

## Optimization of model by RosettaDock Monte Carlo Minimization

The regions defined by each of the clusters as obtained in the previous section are explored using the Monte Carlo Minimization method implemented in the RosettaDock program of Gray et al.[22] First, the position of the ligand is perturbed by random translations and rotations, and the distance between the ligand and receptor is adjusted as to create a contact. Next, a fast MCM at low resolution optimizes the complex orientation with respect to features that do not depend on the explicit conformations of the side chains (e.g. amino acid propensity at the interface, amino acid pair preferences, etc;[23]). Finally, the side chains are added back, and an all-atom optimization locates the local minimum energy conformation. Each of the 50 MCM cycles includes the readjustment of the interface side chain conformations and the optimization of the rigid-body position.

The complete set of interface side chains is repacked every eight cycles: side chains are optimized combinatorially starting from a backbone dependent rotamer library which also includes the side chain conformations in the unbound proteins. The optimal combination of rotamers is found using a simulated annealing Monte Carlo search. Subsequently, off-rotamer conformations that further reduce the energy are sampled by rotamer trial minimization, where for each position minimization of the dihedral angles of all possible rotamers searches for even

lower energy conformations (rtmin, see[22]). In the remaining cycles, a faster procedure evaluates only replacement of single side chain conformations ("rotamer trial"[19,24]).

Once the side-chains have been re-packed, the rigid body displacement is optimized using a Davidon-Fletcher-Powell quasi-Newton minimization algorithm. After each move, side chain packing, and minimization, an energy score is calculated. The new position is kept or rejected according to the standard Metropolis acceptance criterion.

RosettaDock uses a detailed energy function which includes van der Waals interactions with a linear term serving as the repulsive part; a solvation term based on a pairwise Gaussian solvent exclusion model;[25] hydrogen bonding energies using an orientation-dependent empirical function;[26] a rotamer probability term; residue-residue atom pair interactions for charged residues, and a simple electrostatic term across the interface.[27] While the weights of most of the terms in the scoring function are of the same order of magnitude, the dominant contributions to discrimination are the van der Waals (packing) interactions, followed by solvation.[23] The only term of small contribution is the electrostatics across the interface; the effect of electrostatics is described mainly by the pair term. The scoring function used during minimization assigns a larger weight to repulsion in order to remove clashes in the structure. This weight is reduced in the final scoring.

## Cluster stability tests using Monte Carlo minimization

In order to discriminate local minima surrounded by broad regions of attraction from those that do not have such regions we perform "stability" tests for each of the retained clusters using RosettaDock. Starting from the structure defined as the cluster center we generate 1200 random perturbations of the ligand structure in the range of 7 Å RMSD. Each perturbation involves up to 5 Å random translation along the vector connecting the centers of mass of the two component proteins, up to 10 degrees rotation around this axis, and 10 degrees tilt. Each of the perturbed conformations serves as the starting point for low-resolution optimization followed by 50 full-atom steps of Monte Carlo minimization (MCM) by RosettaDock. From the 1200 simulation runs we retain 200 structures with the lowest energy scores. According to our experience, 1200 trajectories and the resulting 200 low energy conformations provide an adequate sampling of the free energy surface around the cluster center.

As described in the introduction, our goal is to study the convergence of MCM trajectories following the perturbations. To identify potential attractors, the 200 retained structures are clustered with a 3 Å RMSD cutoff radius. In order to distinguish the resulting clusters from the original cluster, the former will be referred to as subclusters. It is important to keep in mind that these "subclusters" are based on the 200 points from the Monte Carlo minimization (MCM) runs rather than simply re-clustering the points of the original cluster. Although we will refer to the "stability of a cluster", we actually investigate whether or not some region in the vicinity of the cluster center attracts a fraction of the convergent Monte Carlo minimization (MCM) trajectories. The analysis of stability of a given cluster will be based on the properties of the highest occupancy subcluster S. The cluster is considered stable if and only if:

- **A.** the highest occupancy subcluster S has at least 10 entries;
- **B.** the center of S (i.e., the lowest energy conformation in S) is less than 12 Å from the center of the cluster; and
- **C.** S contains at least one of the seven lowest energy structures from the MCM runs.

As shown in Figure 2A, to determine the stability of the cluster we find the small region S which contains the highest number of structures from the 200 MCM runs. The cluster is considered stable if such a strong attractor exists, has at least 10 entries, is relatively close to the original cluster center, and contains at least one low energy structure. The cluster is unstable

if no such subcluster can be found, i.e., the trajectories either diverge, converge outside the original cluster, or the convergent trajectories lead to high energy states. We note that since for stability we require convergence to a narrow region in the cluster, a small fixed rather than variable radius is used gto define the sub-clusters.

The stability criteria were based on the analysis of a subset of the enzyme-inhibitor complexes in the protein docking benchmark, and have been chosen such that all near-native clusters would be classified as stable. However, the same criteria also classified as stable all near-native clusters for the remaining enzyme-inhibitor complexes, as well as all near-native clusters for the antibody-antigen pairs in the benchmark set, indicating that the conditions have some general validity. As we will show, the criteria are somewhat conservative and classify about 50% of non-native clusters as stable, in addition to the ones that include near-native conformations. However, since the unstable clusters are always free of near-native conformations and hence can be removed from further considerations, the analysis generally yields substantial reduction in the number of candidate models, at least for cases where the backbone does not change significantly.

## RESULTS

Clustering the structures from the rigid body docking using the optimally selected cluster radius may yield up to 30 clusters.[13] However, as mentioned in the introduction, at least one of the 10 largest clusters includes near-native solutions (i.e., structures with less than 10 Å ligand RMSD from the native structure) for 74% of the complexes in the protein docking benchmark set. In order to apply the method described in this paper to as many complexes as possible while keeping the amount of computations at reasonable levels, we considered only enzyme-inhibitor and antibody-antigen complexes in Benchmark Set 1 that satisfied this condition, i.e., their $n$th cluster (where $n \leq 10$) was near-native. The restriction results in 26 complexes listed in Table I. Although Table I shows the Protein Data Bank (PDB) codes only for the target complexes, the docking and the stability analysis involve either unbound-unbound (i.e., separately crystallized) or bound-unbound proteins as given in the Benchmark Set.[16] We note that most enzyme-inhibitor test cases are in the unbound-unbound category, whereas the antibody-antigen cases involve the docking of an unbound antigen structure to the bound structure of the antibody.[16] Computational requirements were further reduced by restricting consideration to the $n$ largest clusters for each complex, where the $n$th is the first near-native cluster. For example, if the second largest cluster includes a near-native structure for a particular complex, then only clusters 1 and 2 were considered in this paper. While this restriction is acceptable for a validation study, in real applications we generally have to study the stability of up to 30 clusters.

The first goal of our analysis is to show that all near-native clusters (with a cluster center located less than 10 Å from the native complex structure) are stable. Table I shows several properties of the near-native clusters for the complexes in our test set. To describe these properties we consider the first row of the table, i.e., the complex of *a*-chymotrypsinogen and pancreatic secretory trypsin inhibitor (PDB code 1CGI) as an example. After docking the unbound component proteins (PDB codes 1CHG and 1HPT) and clustering the top 2000 docked structures as described in the Methods, the largest cluster (ranked as number 1) is near-native, with its center at 3.95 Å RMSD from the native structure. After performing 1200 Monte Carlo minimization runs from random points around the cluster center, selecting the 200 lowest energy structures, and clustering them using a 3 Å RMSD clustering radius, the largest subcluster S has only 22 members, but it includes the lowest energy structure of all the 200 structures retained. The center of this subcluster S is within 5.14 Å RMSD from the center of cluster 1. Thus, based on the properties of subcluster S, cluster 1 satisfies all three conditions for stability, and therefore it can be classified as stable in Table I. According to the last column,

the center of subcluster S (which in this case includes the structure with the lowest RosettaDock energy) is 5.50 Å from the native. Since the original cluster center is at 3.95 Å RMSD from the native, RosettaDock did not improve the accuracy of the solution for this complex, but demonstrated that it has a well-defined region of attraction.

The results in Table I show that the near-native cluster is stable for all the complexes studied. Clustering of the 200 best-energy structures from the Monte Carlo minimization runs yields in general larger subclusters than for 1CGI. In addition, the distance between the native structure and the center of the largest subcluster is generally smaller than the distance between the native structure and the center of the cluster, i.e., the Monte Carlo minimization by RosettaDock improves the structures obtained by rigid body docking. According to Table I, RosettaDock reduces the RMSD in 18 of the 26 complexes. For the purposes of this paper, however, the more important observation is that the near-native clusters are stable for all the 26 complexes.

Our second goal is to show that the stability criterion can discriminate between near-native and non-native clusters, and hence help to eliminate some of the “false positive” clusters that have low energy but are not close to the native. Therefore we performed the same simulations on the non-native clusters that are ranked higher than the first near-native one, in order to see whether some of them can be excluded from further analysis. It is expected that the Monte Carlo simulations will improve discrimination as we switch from the the rigid body energy function, which is defined on a grid and has limited accuracy, to the more accurate RosettaDock energy function. Detailed results are presented for the enzyme-inhibitor complex 1MAH (Table II) and the antibody-antigen complex 1NMB (Table III). These complexes were selected because the near-native cluster is preceeded by nine non-native clusters in both cases, and hence they provide good examples to study the outcome of Monte Carlo minimization calculations in these clusters.

Table II shows the results of our analysis for the enzyme-inhibitor complex 1MAH. These include the RMSD between the cluster center and the native complex for each of the top ten clusters obtained by the rigid body docking of the unbound component proteins (column 2), as well as the results of the MC minimizations for each cluster. As in Table I, we list the number of elements in the largest subcluster, the rank of the subcluster center (among the 200 structures retained from the MC minimization), and the RMS displacement of the subcluster center from the cluster center. Clusters 5, 7, 8, and 10 satisfy all three conditions for stability, and hence are considered stable. The other 6 clusters fail at least for one condition and hence are unstable. Cluster 3 is unstable, because the center of the largest subcluster is more than 12 Å from the cluster center, and with 7 entries the subcluster is too small. The highest occupancy subclusters are also too small in Clusters 2, 3, and 6, whereas in clusters 1, 6, and 9 the S subclusters do not include low energy structures and hence do not satisfy Condition C. Table II also shows the RMSD between the center of the most populated subcluster and the native structure, demonstrating that starting from a point in a non-native cluster, the short Monte Carlo minimizations alone are generally unable to substantially reduce the RMSD from the native structure. In fact, since the energy landscape is very rugged, the Monte Carlo protocol gets stuck pretty near to its starting point, and therefore no trajectory leading to the global minimum can be created if the starting structure is too far away.

Figure 3 shows the 200 lowest RosettaDock scores from the 1200 Monte Carlo minimization runs as a function of the RMSD from each cluster center. These plots reveal a funnel-like behavior within 10 Å RMSD from the cluster center for the stable clusters 5, 7, 8, and 10, and the lack of funnels leading to low energy within this RMSD range for the remaining 6 clusters. The energy plot for cluster 8 shows two or three funnels within the same cluster. However, the funnel located only at 2.5 Å RMSD from the cluster center represents the most populated

subcluster. Although this subcluster is very small, with only 10 members (see Table II), cluster 8 satisfies the conditions for stability. Cluster 3 has a well-defined funnel, but it is beyond the 12 Å stability threshold. Although the plot also shows a single structure with very low energy below the 10 Å mark, this structure is not surrounded by a populated subcluster, and cluster 3 is considered unstable. As shown in Table II, this results is correct, since the RMSD between the center of cluster 3 and the native complex is 31 Å.

Table III and Figure 4 show the results of stability analyses for the antibody-antigen complex 1NMB. As for 1MAH, four out of the ten clusters (clusters 1, 3, 7, and 10) satisfy all three conditions for stability. The comparison to the 1MAH results show a number of differences that appear to generally exist between enzyme-inhibitor and antigen-antibody complexes. First, for 1NMB all clusters except for cluster 10 are around 30 Å RMSD away from the native. This is in agreement with the observation that due to the smaller role of shape complementarity, rigid-body energy functions carry less information and usually yield less accurate results for antibody-antigen complexes than for enzyme-inhibitor complexes.[4] Second, the Monte Carlo simulations generally move farther from the cluster center than for 1MAH (see Table III). Indeed, the displacements are close to 10 Å, even for the native cluster 10. However, RosettaDock works very well for the near-native cluster of 1MNB, and the MCM simulations converge at 1.12 Å RMSD from the native structure (Table III). To illustrate the complexity of the global energy surface, Figure 5 shows the union of results for the 10 clusters (200 points for each), this time as a function of the RMSD from the native structure. According to this plot, there is a near-native funnel for this complex with a 1.12 Å RMSD from the native structure, and RosettaDock finds this solution in 41 of the 200 simulations started around the center of cluster 10 (Table III). However, Figure 5 reveals the existence of much lower energy structures at almost 30 Å RMSD from the native, emphasizing that even a detailed energy function, such as the one used by RosettaDock, is unable to eliminate all false positives, and that the chance for finding such low energy non-native structures increases as the search is extended beyond a neighborhood of the native state.

Monte Carlo minimization runs have been performed for all non-native clusters preceding a near-native cluster in order to identify and remove unstable clusters. As shown in Table IV, the largest cluster from the rigid body docking is near-native for 9 enzyme-inhibitor and 2 antibody-antigen complexes. As already shown in Table I, all near-native clusters are stable and hence remain ranked as number 1. For the remaining 15 complexes about half of the non-native clusters are classified as unstable. Since such clusters are always non-native, they can be removed from further consideration, thereby generally reducing the number of potential conformations by a factor of two. This reduction in the number of non-native clusters indicates that the limitations of the rigid body energy functions produce a substantial number of false positive solutions with low energy. In the examples considered half of these can be removed by switching to the more accurate RosettaDock energy and restricting consideration to the minima with broad regions of attraction. We note that in a ”blind-trial” applications, where 30 or so clusters would have to be considered, the fraction of non-native clusters that are unstable could be different. However, the main goal is reducing the number of non-native clusters that are larger than the first near-native one.

Table IV lists two RMSD values for each complex. The first is the RMSD between the native complex and the center of the first near-native cluster from the rigid body docking. The second is the RMSD between the native complex and the most populated subcluster obtained after the 1200 Monte Carlo minimizations by RosettaDock. According to these results, the application of RosettaDock to the rigid body docking results in the near-native clusters improves the predictions for 18 of the 26 complexes, in 11 cases reducing the RMSD by more than 2 Å. Thus, refining low resolution models using a higher resolution energy function definitely yields better accuracy. We should however keep in mind that we compare here the best RMSD

structures rather than the ones with the best energy. As we have shown for the complex 1NMB, the more global search by the combined FFT-RosettaDock can easily yield very low energy non-native structures (and thus the best-RMSD structure is not necessarily the one selected).

## Discussion

Protein-protein docking can be described as the problem of locating the global minimum of a scoring function that attempts to describe the free energy of the protein-solvent system. Although for some classes of interacting proteins it is feasible to assume an essentially rigid body association and hence the problem can be reduced to a search in the 6D space of translations and rotations, finding near-native structures of the complex remains di±cult for a number of reasons. First, even this so-called rigid body approximation must account for conformational changes due to the re-orientation of side chains in the interface. Thus, the problem is not simply matching two irregular shapes; one has to minimize energy-like scoring functions that account for electrostatic and chemical complementarity of the interacting surfaces. Second, the scoring functions are approximate, and hence one has to retain a number of sub-optimal solutions from the initial search for further evaluation. This implies that one has to explore the entire 6D conformational space, possibly generating billions of conformations, unless prior structural information is available. Third, the observation that is most disconcerting for a computational scientist, is that finding the lowest minima of the scoring functions does not necessarily imply finding near-native complex structures. In fact, none of the current scoring functions account for the loss of entropy upon the association of the two proteins, and hence a deep but narrow minimum, which necessarily leads to substantial entropy loss, is less likely to yield a physically meaningful solution than a broader albeit possibly not as deep alternative minimum.

Most methods of protein-protein docking that performed well at CAPRI (Critical Assessment of Predicted Interactions), the communitywide experiment devoted to protein docking[7] are based only on two approaches. These approaches are rigid body exhaustive search, most frequently involving Fast Fourier Transforms (FFT) for evaluating the scoring function, and Monte Carlo minimization. Both methods have distinct advantages and disadvantages. The Fast Fourier Transform (FFT) correlation approach can globally explore the conformational space of a protein-protein complex, evaluating the energies for billions of conformations on a grid. However, due to the approximate nature of the energy function and the need for tolerating potential overlaps, the method yields a large number of false positives, i.e., conformations with low energy that are located far from the native orientation. No further discrimination can be obtained by using the energy measures defined on the grid, because the differences in energy levels between the minima are comparable to the errors in the energy calculation. Docking algorithms based on Monte Carlo minimization (MCM) can use more detailed and hence more accurate energy functions. However, Monte Carlo is a statistical method, and due to the need for extensive calculations it can explore only limited regions of the conformational space on reasonable time scales. Thus, results provided by MCM methods may heavily depend on the initial points of the simulations. Finally, none of the methods account for the entropic consideration suggesting that the native complex structures are likely to be located at minima with broad regions of attraction.

The method described in this paper combines the above two approaches to docking. The first step of the method is globally sampling the conformational space by FFT-based rigid body docking. This step also includes the clustering of the docked conformations, generally resulting up to 30 clusters, each covering a neighbourhood of different local minima. In the second step, Monte Carlo minimization runs are carried out from 1200 random points within each retained cluster in order to test the ”stability” of the cluster, i.e., to determine whether the minimum has some region of attraction that prevents the divergence of Monte Carlo minimization trajectories

out of the region defined by the cluster. For the analysis of stability we retain the 200 lowest energy structures from the MCM runs, and cluster them with a 3 Å RMSD radius, resulting in a number of generally small subclusters. The cluster is considered stable if the highest occupancy subcluster S contains more than 10 convergent trajectories, the center of S is close enough to the original cluster center, and S contains at least one low energy structure.

We note that that it would be highly desirable to develop a meaningful estimate of the free energy including both enthalpic and entropic contributions. However, if the conformations have very similar energies, the ranking of conformations is primarily determined by the entropic terms. Since the energy function is approximate, the cluster size appears to be a better measure of entropy than the volume of the basin, calculated from the energy surface. In addition, our energy expression already includes entropic contributions in the solvation term. Similar results were shown by Ruvinsky and Kozintsev,[28] who developed free energy measures incorporating the volume of the energy basin, but none of the models derived performed better than the cluster size. The advantage of using cluster size for finding near-native structures in protein-protein docking has been confirmed by the results of a recent comparative study.[29]

We studied the stability of clusters for enzyme-inhibitor and antibody-antigen complexes in the Protein Docking Benchmark. The analysis yields three main results. First, the near-native clusters are always stable. Second, restricting considerations to stable clusters eliminates about half of the false positives, i.e., solutions that have low energy but are far from the native structure of the complex. Third, breaking up the docking problem into stability analyses of clusters of docked conformations makes the combined approach less dependent on a priori information than exploring the potential conformational space by Monte Carlo minimizations without the global search by rigid body docking.

We admit that the exact stability criteria were based on the analysis of some enzyme-inhibitor complexes. Although the same conditions turned out to be applicable to all enzyme-inhibitor and antigen-antibody complexes of the docking benchmark,[16] the generality of the conditions remains somewhat questionable, and may have to be adjusted as results for more complexes will become available. However, the general idea is very clear: the energy minima close to native states must be strong attractors. The existence of such attractors can be determined by examining the convergence of Monte Carlo minimization trajectories started from points around the local minimum, and this is the basis of the method described here. Selecting minima with broad regions of attraction increases the probability of finding near-native structures, as the minimization itself can lead to false positives, i.e., conformations that have low energy but are far from the native. We have found that the near-native clusters are stable for all complexes in this study (Table I), which means that the RosettaDock energy function is accurate enough to guarantee the convergence of Monte Carlo minimization trajectories within the regions defined by these clusters. We have also shown that the minimization improves the prediction for the majority of complexes (Table IV), again restricting considerations to trajectories started within near-native clusters.

It is far from simple to compare our two-step method with RosettaDock in terms of the required CPU time. The RosettaDock approach usually involves sampling around $10^4$ starting orientations in a global run, and then the local optimization of the best starting structures found. The rigid body docking in our procedure is certainly faster than the initial sampling in RosettaDock. However, we then generally explore the stability of up to 30 clusters, performing 1200 short MCM runs in each, most likely making the approach computationally as expensive as the original RosettaDock procedure.

## Acknowledgments

DK and SV thank David Baker (University of Washington, Seattle) for the RosettaDock program and for his help. This work has been supported by the grant GM061867 from the National Institutes of Health. For the CPU time used for this paper we are grateful to the Boston University Scientific Computing and Visualization Center for the opportunity of running the program on the Blue Gene/L supercomputer.

## References

1. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–43. [PubMed: 12001221]
2. Smith G, Sternberg M. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35. [PubMed: 11839486]
3. Camacho C, Vajda S. Protein-protein association kinetics and protein docking. Curr Opin Struct Biol 2002;12:36–40. [PubMed: 11839487]
4. Vajda S, Camacho C. Protein-protein docking: is the glass half full or half empty? Trends in Biotech 2004;22:110–116.
5. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A, Aflalo C, Vakser I. Molecular surface recognition - determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199. [PubMed: 1549581]
6. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: An FFT-based protein docking program with pairwise potentials. Proteins Nov;2006 65(2):392–406. [PubMed: 16933295]
7. Mendez R, Leplae R, De Maria L, Wodak S. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. Proteins 2003;52:51–67. [PubMed: 12784368]
8. Mendez R, Leplae R, Lensink M, Wodak S. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins 2005;60:150–69. [PubMed: 15981261]
9. Gabb H, Jackson R, Sternberg M. Modelling protein docking using shape complementarity, electrostatics, and biochemical information. J Mol Biol 1997;272:106–120. [PubMed: 9299341]
10. Mandell J, Roberts V, Pique M, Kotlovyi V, Mitchell J, Nelson E, Tsigelny I, Ten Eyck L. Protein docking using continuum electrostatics and geometric fit. Protein Eng 2001;14:105–113. [PubMed: 11297668]
11. Chen R, Li L, Weng Z. ZDOCK: An initial-stage protein-docking algorithm. Pro-teins 2003;52:80–87.
12. Li L, Cheng R, Weng Z. RDOCK: Refinement of rigid-body protein docking pre-dictions. Proteins 2003;53:693–707. [PubMed: 14579360]
13. Comeau S, Gatchell D, Vajda S, Camacho C. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 2004;20:45–50. [PubMed: 14693807]
14. Kozakov D, Clodfelter K, Vajda S, Camacho C. Optimal clustering for detecting near-native conformations in protein docking. Biophys J 2005;89:867–875. [PubMed: 15908573]
15. Camacho CJ, Weng Z, Vajda S, DeLisi C. Free energy landscapes of encounter complexes in protein-protein association. Biophys J Mar;1999 76(3):1166–1178. [PubMed: 10049302]
16. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52:88–91. [PubMed: 12784372]
17. Apaydin MS, Guestrin C, Varma C, Brutlag DL, Latombe J-C. Stochastic roadmap simulation for the study of ligand-protein interactions. Bioinformatics 2002;18:S18–S26. [PubMed: 12385979]
18. Apaydin MS, Brutlag DL, Guestrin C, Hsu D, Latombe JC, Varma C. Stochastic roadmap simulation: an e±cient representation and algorithm for analyzing molecular motion. J Comput Biol 2003;10(3–4):257–281. [PubMed: 12935328]
19. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol Aug;2003 331(1):281–299. [PubMed: 12875852]
20. Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci 2002;11:280–291. [PubMed: 11790838]

21. Zhang C, Vasmatzis G, Cornette J, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Molec Biol 1997;267:707–726. [PubMed: 9126848]

22. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci May;2005 14(5):1328–1339. [PubMed: 15802647]

23. Gray J, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl C, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Molec Biol 2003;331:281–299. [PubMed: 12875852]

24. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A Sep;2000 97(19):10383–10388. [PubMed: 10984534]

25. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins May;1999 35(2): 133–152. [PubMed: 10223287]

26. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol Feb; 2003 326(4):1239–1259. [PubMed: 12589766]

27. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93. [PubMed: 15063647]

28. Ruvinsky A, Kozintsev A. Novel statistical-thermodynamic methods to predict protein-ligand binding positions using probability distribution functions. Proteins Jan;2006 62:202–208. [PubMed: 16287127]

29. Lorenzen S, Zhang Y. Identification of near-native structures by clustering protein docking conformations. Proteins Jul;2007 68:187–194. [PubMed: 17397057]

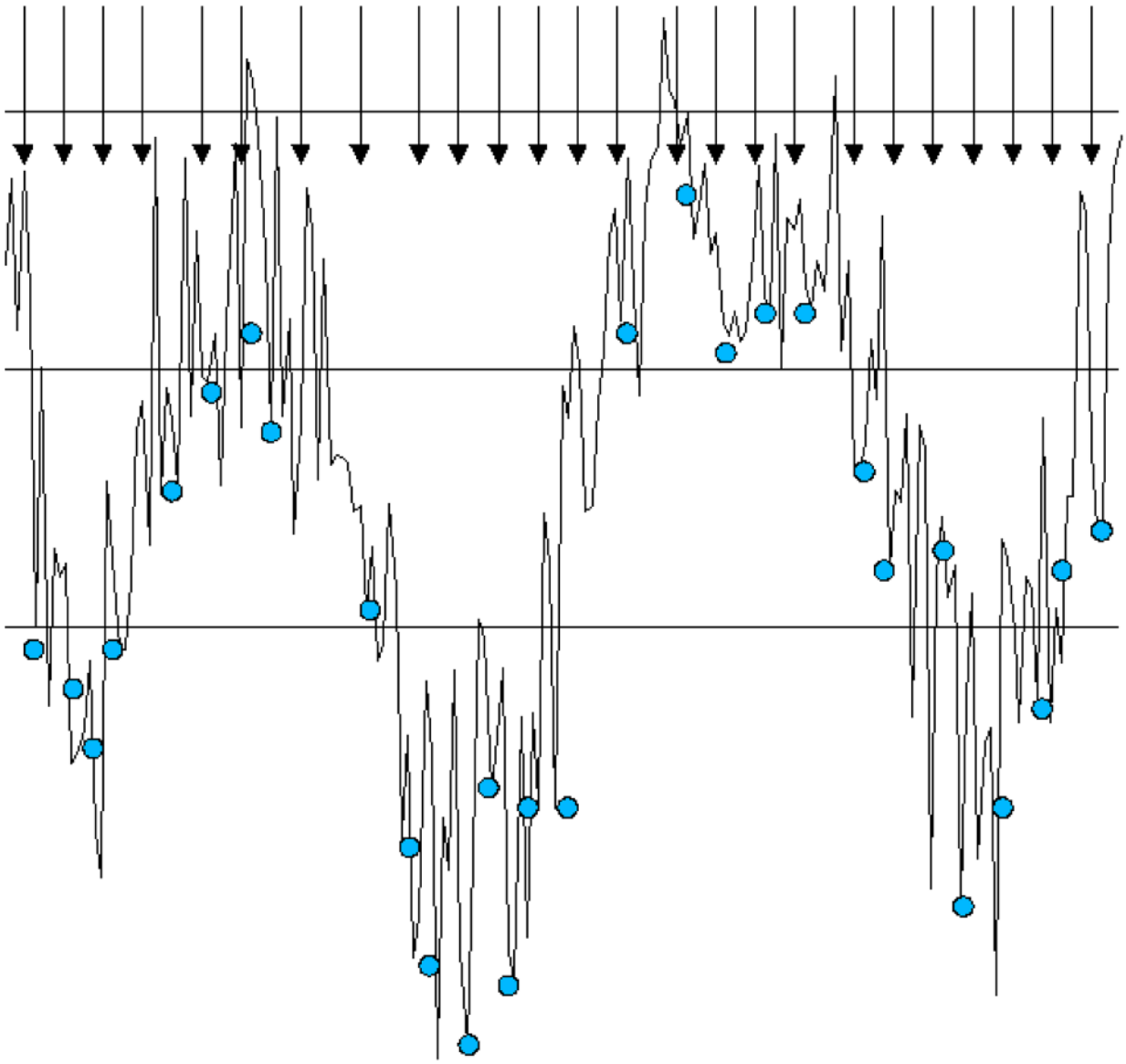

**Figure 1.**
Schematic representation of the energy surface along an arbitrary association coordinate. The arrows represent the sampled conformations, with small filled circles representing the corresponding energies. Retaining only the structures with energies below a threshold and clustering the retained structures using pairwise RMSD as the distance measure yields a number of clusters, where the large clusters correspond to minima that have broad regions of attraction.

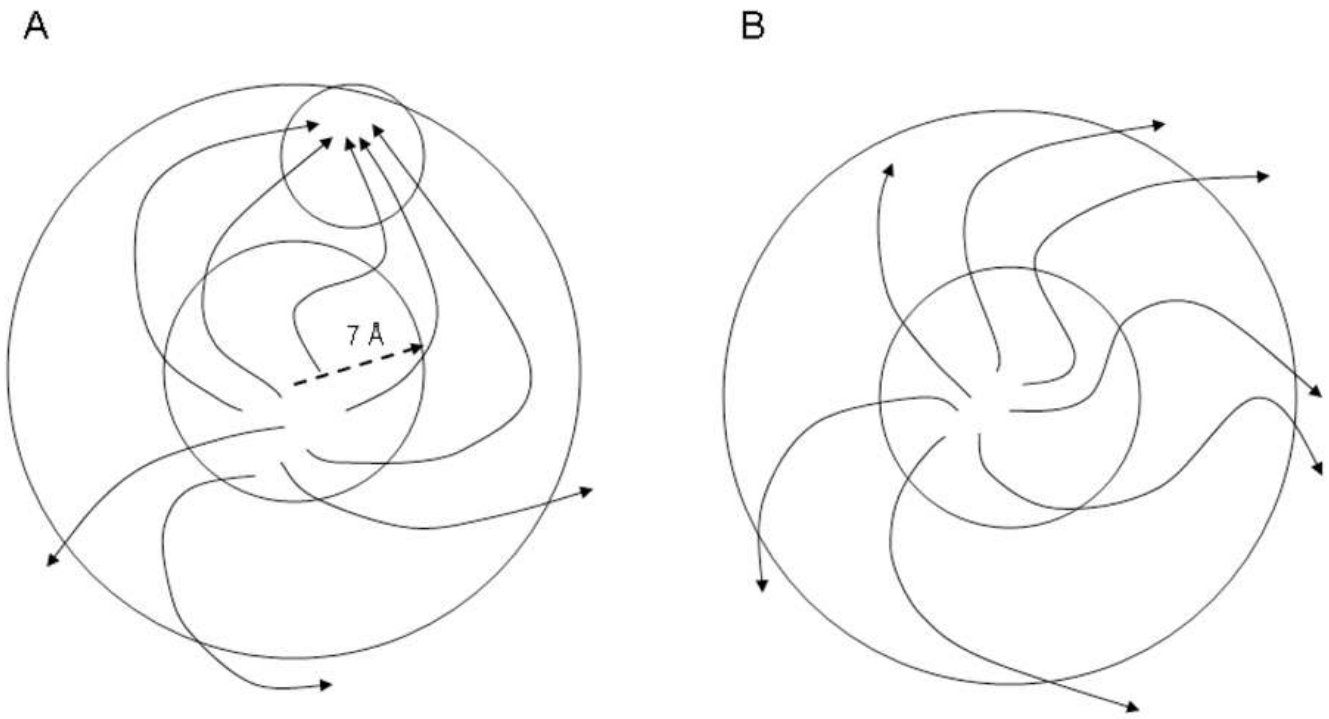


**Figure 2.**
Schematic representation of the Monte Carlo minimization trajectories. The large circle represents the cluster boundary, and the smaller circle around the cluster center shows the region in which the initial points are selected for the Monte Carlo minimization (MCM) runs. Panel A shows a stable cluster in which a number of MCM trajectories converge to a small region representing the subcluster S within the space defined by cluster. Panel B shows an unstable cluster with no convergent trajectories.

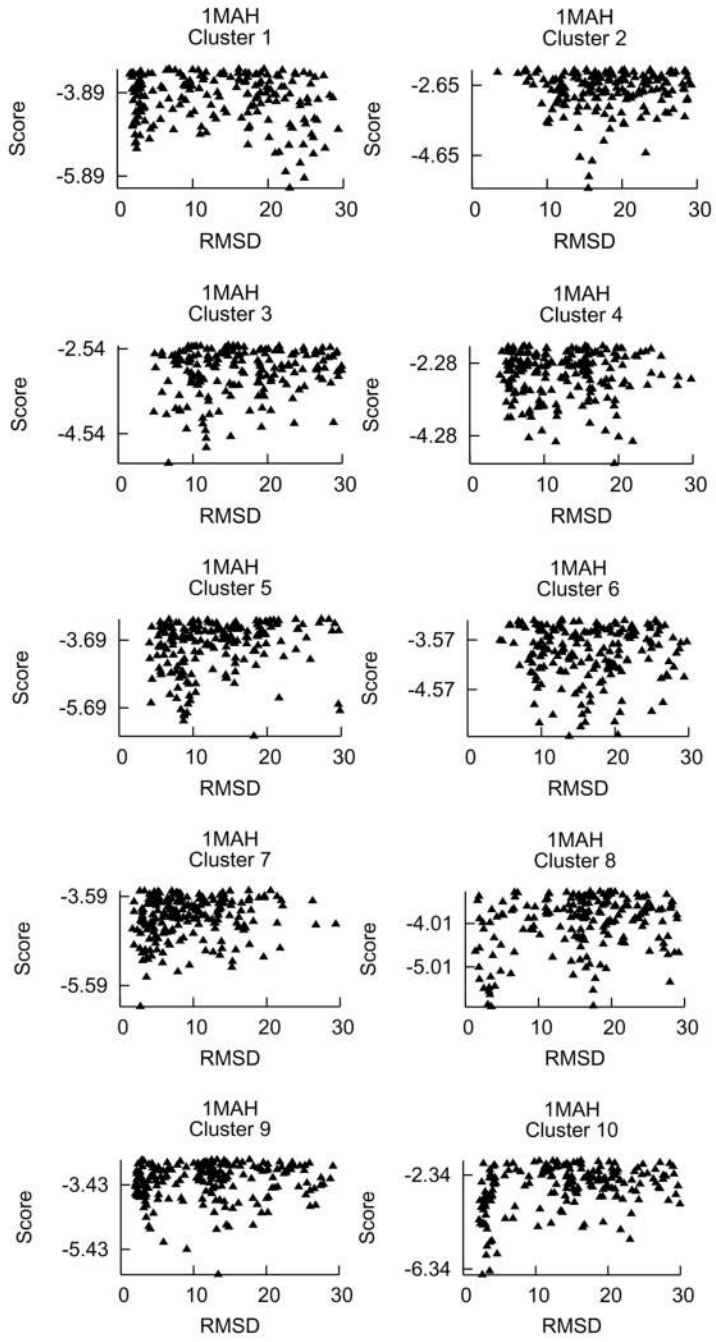


**Figure 3.**

The lowest 200 energy values selected from the Monte Carlo minimization (MCM) runs for each of the 10 top clusters obtained for the enzyme-inhibitor complex 1MAH. The energies across the interface are shown as functions of the RMSD from the respective cluster centers. Thus, a small RMSD implies that the MCM run does not substantially affect the conformation obtained by the rigid body docking.

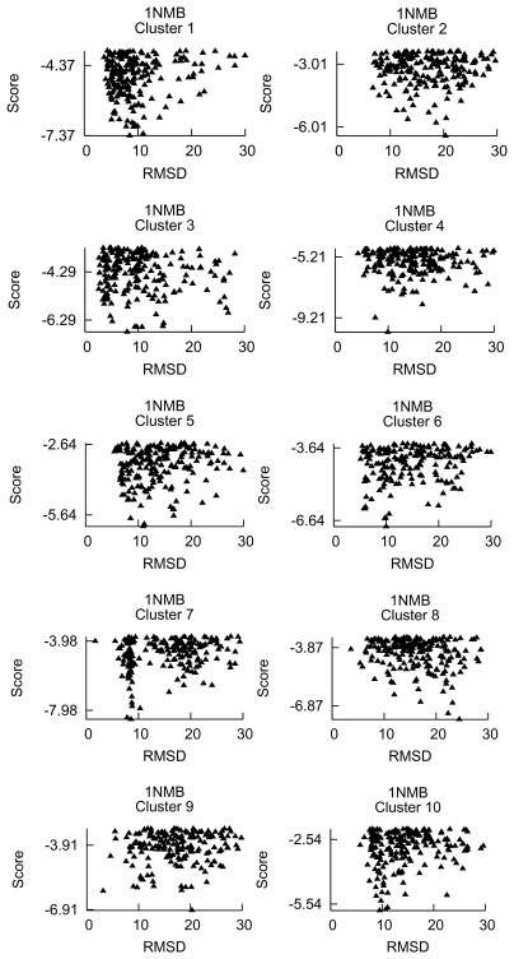


**Figure 4.**

The lowest 200 energy values selected from the Monte Carlo minimization (MCM) runs for each of the 10 top clusters obtained for the antigen-antibody complex 1NMB. The energies across the interface are shown as functions of the RMSD from the respective cluster centers. Thus, a small RMSD implies that the MCM run does not substantially affect the conformation obtained by the rigid body docking.

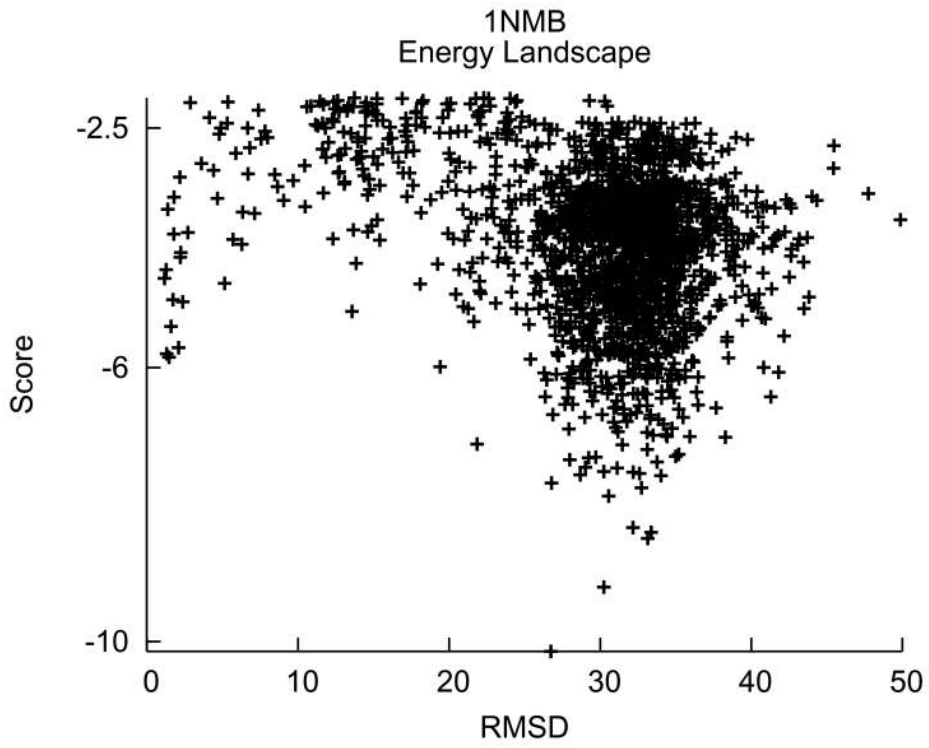


**Figure 5.**
Union of all Monte Carlo simulation results (2000 structures) for all the 10 clusters of the enzyme-inhibitor complex 1NMB shown in Figure 4. The energies are shown as functions of the RMSD from the native complex, so a small RMSD indicates a near-native structure. According to this Figure, a global search using the RosettaDock energy would yield structures with almost 30Å RMSD from the native complex.

**TABLE I**

Stability of near-native clusters

| Complex | Type[a] | Native cluster rank *b* | RMSD center to native[c] | Energy rank of subcluster[d] | Occupancy of top subcluster[e] | Subcluster displacement from center[f] | Stability[g] | RMSD top subcluster to native[h] |
|---|---|---|---|---|---|---|---|---|
| 1CGI | e-i | 1 | 3.95 | 1 | 22 | 5.14 | Y | 5.50 |
| 1CHO | e-i | 1 | 1.73 | 1 | 35 | 2.51 | Y | 1.75 |
| 2PTC | e-i | 3 | 6.93 | 1 | 30 | 11.10 | Y | 4.20 |
| 1TGS | e-i | 1 | 3.41 | 1 | 34 | 4.86 | Y | 1.69 |
| 2SIC | e-i | 2 | 2.53 | 1 | 67 | 4.64 | Y | 4.72 |
| 1CSE | e-i | 9 | 5.04 | 1 | 20 | 8.49 | Y | 8.48 |
| 1BRC | e-i | 1 | 9.80 | 2 | 25 | 6.14 | Y | 9.20 |
| 1ACB | e-i | 3 | 4.31 | 1 | 19 | 7.90 | Y | 4.90 |
| 1BRS | e-i | 5 | 9.05 | 2 | 20 | 8.20 | Y | 4.54 |
| 1UGH | e-i | 5 | 8.39 | 4 | 58 | 5.87 | Y | 2.99 |
| 1DFJ | e-i | 1 | 8.34 | 5 | 21 | 4.71 | Y | 10.03 |
| 1AVW | e-i | 1 | 3.77 | 2 | 15 | 7.90 | Y | 6.20 |
| 1PPE | e-i | 1 | 3.24 | 2 | 38 | 4.67 | Y | 2.39 |
| 1TAB | e-i | 11 | 5.59 | 1 | 77 | 3.82 | Y | 3.52 |
| 1UDI | e-i | 10 | 2.79 | 7 | 22 | 3.72 | Y | 1.87 |
| 1STF | e-i | 1 | 6.28 | 1 | 51 | 5.41 | Y | 1.34 |
| 2TEC | e-i | 1 | 5.15 | 1 | 56 | 5.91 | Y | 1.99 |
| 4HTC | e-i | 3 | 5.00 | 4 | 14 | 3.95 | Y | 4.30 |
| 1MAH | e-i | 10 | 4.09 | 1 | 43 | 2.65 | Y | 2.20 |
| 1AHW | a-a | 1 | 2.41 | 1 | 20 | 2.78 | Y | 4.61 |
| 1BQL | a-a | 6 | 5.65 | 5 | 40 | 4.90 | Y | 4.06 |
| 1BVK | a-a | 4 | 6.60 | 2 | 13 | 1.10 | Y | 6.12 |
| 2JEL | a-a | 6 | 5.93 | 1 | 34 | 6.45 | Y | 1.00 |
| 1MEL | a-a | 2 | 6.74 | 1 | 44 | 7.36 | Y | 2.29 |
| 1NMB | a-a | 10 | 9.73 | 1 | 16 | 9.52 | Y | 1.12 |
| 1NCA | a-a | 1 | 4.01 | 1 | 92 | 3.96 | Y | 1.01 |

[a] e-i: enzyme-inhibitor; a-a: antibody-antigen

[b] Rank of the first cluster with less than 10 Å RMSD from the native complex

[c] $C_\alpha$ RMSD between the cluster center and the native structure of the complex

[d] Energy rank of the cluster center (i.e., the lowest energy structure) of the most populated subcluster S

[e] Number of cluster members of the most occupant subcluster S

[f] RMSD between the cluster center and the center of the highest occupancy subcluster S

[g] Y indicates that the cluster satisfies all three conditions for stability

[h] RMSD between the subcluster center and the native structure of the complex

**TABLE II**

Stability of the top 10 clusters for the enzyme-inhibitor complex 1MAH

| Cluster[a] | RMSD[b] | Energy[c] | Occupancy[d] | Displacement[e] | RMSD to native[e] | Stability[f] |
|---|---|---|---|---|---|---|
| 1 | 24.91 | 12 | 27 | 2.37 | 24.10 | N |
| 2 | 20.20 | 87 | 3 | 9.94 | 21.60 | N |
| 3 | 31.00 | 2 | 7 | 12.30 | 35.50 | N |
| 4 | 19.34 | 8 | 26 | 5.13 | 16.60 | N |
| 5 | 47.00 | 3 | 14 | 8.78 | 52.80 | Y |
| 6 | 21.00 | 20 | 6 | 9.50 | 21.02 | N |
| 7 | 27.00 | 3 | 30 | 7.41 | 24.30 | Y |
| 8 | 18.60 | 4 | 10 | 2.52 | 17.60 | Y |
| 9 | 29.60 | 15 | 28 | 3.17 | 30.10 | N |
| 10 | 4.09 | 1 | 43 | 2.65 | 2.21 | Y |

[a] Rank of the cluster

[b] $C_\alpha$ RMSD from the cluster center to crystal structure

[c] Global energy rank of the lowest energy structure in the most occupied subcluster S.

[d] Number of cluster members in the most occupied subcluster

[e] RMSD between the center of the highest occupancy subcluster and the native structure of the complex

[f] Y indicates that the cluster satisfies all three conditions for stability

**TABLE III**

Stability of the top 10 clusters for the antigen-antibody complex 1NMB

| Cluster[a] | RMSD[b] | Energy [c] | Occupancy [d] | Displacement [e] | RMSD to native[e] | Stability[f] |
|---|---|---|---|---|---|---|
| 1 | 31.59 | 2 | 29 | 10.97 | 34.54 | Y |
| 2 | 33.84 | 57 | 4 | 19.30 | 30.42 | N |
| 3 | 27.58 | 5 | 22 | 5.14 | 36.14 | Y |
| 4 | 31.66 | 12 | 8 | 9.29 | 29.03 | N |
| 5 | 26.58 | 9 | 9 | 8.01 | 30.28 | N |
| 6 | 29.13 | 1 | 8 | 10.01 | 27.40 | N |
| 7 | 33.91 | 1 | 34 | 8.58 | 32.47 | Y |
| 8 | 32.03 | 2 | 7 | 22.30 | 34.75 | N |
| 9 | 31.83 | 8 | 5 | 8.99 | 30.90 | N |
| 10 | 9.73 | 1 | 16 | 9.52 | 1.12 | Y |

[a] Rank of the cluster

[b] $C_\alpha$ RMSD from the cluster center to crystal structure

[c] Global energy rank of the lowest energy structure in the most occupied subcluster S.

[d] Number of cluster members in the most occupied subcluster

[e] RMSD between the center of the highest occupancy subcluster and the native structure of the complex

[f] Y indicates that the cluster satisfies all three conditions for stability

*Proteins*. Author manuscript; available in PMC 2010 February 17.

*Proteins*. Author manuscript; available in PMC 2010 February 17.

**TABLE IV**

Improvement in the ranking of near-native clusters

| Complex | Type[a] | Rank before[b] | RMSD [c] | Rank after[d] | RMSD[e] |
|---|---|---|---|---|---|
| 1TAB | e-i | 11 | 5.59 | 5 | 3.52 |
| 1MAH | e-i | 10 | 4.09 | 4 | 2.20 |
| 1UDI | e-i | 10 | 2.79 | 8 | 1.87 |
| 1CSE | e-i | 9 | 5.04 | 4 | 8.48 |
| 1BRS | e-i | 5 | 9.05 | 1 | 4.54 |
| 1UGH | e-i | 5 | 8.39 | 4 | 2.99 |
| 4HTC | e-i | 3 | 5.00 | 1 | 4.30 |
| 2PTC | e-i | 3 | 6.93 | 1 | 4.20 |
| 1ACB | e-i | 3 | 4.31 | 2 | 4.92 |
| 2SIC | e-i | 2 | 2.53 | 2 | 4.72 |
| 1CGI | e-i | 1 | 3.95 | 1 | 5.50 |
| 1CHO | e-i | 1 | 1.73 | 1 | 1.75 |
| 1TGS | e-i | 1 | 3.41 | 1 | 1.69 |
| 1BRC | e-i | 1 | 9.80 | 1 | 9.20 |
| 1STF | e-i | 1 | 6.28 | 1 | 1.34 |
| 2TEC | e-i | 1 | 5.15 | 1 | 1.99 |
| 1DFJ | e-i | 1 | 8.34 | 1 | 10.03 |
| 1AVW | e-i | 1 | 3.77 | 1 | 6.20 |
| 1PPE | e-i | 1 | 3.24 | 1 | 2.39 |
| 1AHW | a-a | 1 | 2.41 | 1 | 4.61 |
| 1BQL | a-a | 6 | 5.65 | 3 | 4.06 |
| 1BVK | a-a | 4 | 6.60 | 1 | 6.12 |
| 2JEL | a-a | 6 | 5.93 | 4 | 1.00 |
| 1MEL | a-a | 2 | 6.74 | 1 | 2.29 |
| 1NMB | a-a | 10 | 9.73 | 5 | 1.12 |
| 1NCA | a-a | 1 | 4.01 | 1 | 1.01 |

[a] e-i: enzyme-inhibitor; a-a: antibody-antigen

[b] Rank of the near- native cluster based on occupancy

[c] RMSD between the cluster center and the native structure

[d] Rank of the cluster after removing the non-stable clusters

[e] RMSD between the center of the highest occupancy subcluster and the native structure