Docking of protein models

ANDREI TOVCHIGRECHKO,[1] CHRISTOPHER A. WELLS, AND ILYA A. VAKSER[1]

Department of Cell and Molecular Pharmacology, Medical University of South Carolina, Charleston, South Carolina 29425, USA

(RECEIVED NOVEMBER 29, 2001; FINAL REVISION May 8, 2002; ACCEPTED May 8, 2002)

Abstract

An adequate description of entire genomes has to include information on the three-dimensional (3D) structure of proteins. Most of these protein structures will be determined by high-throughput modeling procedures. Thus, a structure-based analysis of the network of protein–protein interactions in genomes requires docking methodologies that are capable of dealing with significant structural inaccuracies in the modeled structures of proteins. We present a systematic study of the applicability of our low-resolution docking method to protein models of different accuracies. A representative nonredundant set of 475 cocrystallized protein–protein complexes was used to build an array of models of each protein in the set. A sophisticated procedure was created to generate the models with RMS deviations of 1, 2, 3, . . ., 10 Å from the crystal structure. The docking was performed for all the models, and the predictions were compared with the configurations of the original cocrystallized complexes. Statistical analysis showed that the low-resolution docking can determine the gross structural features of protein–protein interactions for a significant percent of complexes of highly inaccurate protein models. Such predictions may serve as starting points for a more detailed structural analysis, as well as complement experimental and computational data on protein–protein interactions obtained by other techniques.

Keywords: Protein recognition; structural genomics; bioinformatics; protein structure; protein interactions

Protein interactions are the basis of life processes at the molecular level. Most of the protein interactions are with other proteins. Thus, the efforts to recreate the network of protein–protein interactions are important for the interpretation of the information encoded in genomes. The number of protein–protein interactions is significantly larger than the number of individual proteins. Thus, high-throughput methods are needed for studies of these interactions on a genome scale. The existing methodologies, both experimental and computational, may be roughly separated into methods detecting direct physical interactions between proteins (e.g., two-hybrid analysis, mass spectrometry, etc.) and the function-assigning methods (e.g., correlation of mRNA levels, method of phylogenetic profiles, fusion pattern method, sequence alignment, and fold comparison). The outline of "the postgenomic" methods is presented in several reviews (Eisenberg et al. 2000; Oliver 2000; Skolnick et al. 2000; Vukmirovic and Tilghman 2000).

The only computational approaches that directly model physical interactions between proteins are docking (Vajda et al. 1997; Sternberg et al. 1998) and binding simulations (McCammon 1998). Docking approaches, as opposed to binding simulations, are not concerned with modeling of real binding pathways, but rather focus on the final configuration(s) of the complex. This makes docking computationally efficient and potentially suitable for high-throughput "first approximation" structural analysis on a genome scale.

Because proteins are 3D objects, the importance of the direct 3D analysis of protein–protein interactions is obvious. Such analysis is necessary for the prediction of these interactions, their adequate study, and for further applications (e.g., structure-based drug design). The direct experimental approaches (primarily, X-ray crystallography) are

Reprint requests to: Ilya A. Vakser, Bioinformatics Laboratory, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600, USA; e-mail: vakser@ams.sunysb.edu; fax: (631) 632-8490.

[1]Present address: Bioinformatics Laboratory, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600, USA.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.4730102.

developing fast. However, they are capable of determining only a fraction of all protein structures. Thus, the structures of most of the individual proteins in genomes have to be modeled by high-throughput modeling approaches (Burley 2000; Sanchez et al. 2000). The growing availability of the experimentally determined structures of representative protein folds makes the template-based modeling of the majority of proteins in genomes quite realistic in the near future. The limitations of the direct experimental techniques are even more evident in the case of the structures of protein–protein complexes, which are, in general, more difficult to determine than the structures of individual proteins. However, even the first consideration only—the fact that almost all individual protein structures will be models—makes the computational docking approaches indispensable for the direct 3D analysis of protein–protein interactions in genomes.

The number of potential protein–protein interactions and the nature of protein structures to be docked impose strong requirements on the docking techniques. Because of the large number of proteins to dock, the docking has to be fast. At the same time, because the majority of individual protein structures in a genome will be models, the docking has to be capable of predicting complexes of modeled proteins. The major difference between an experimental (X-ray) protein 3D structure and a model, in general, is a substantially lower accuracy of the latter (Jones and Kleywegt 1999; Murzin 1999; Orengo et al. 1999). The accuracy of the protein models may vary significantly, based on the availability of the structural templates and the degree of target-template similarity, from ~1 Å RMSD (high sequence similarity to templates) to >6 Å RMSD (low sequence similarity to templates, no templates). Thus, the docking procedure has to be capable of tolerating very significant structural inaccuracies.

Obviously, docking cannot yield greater precision than the precision of the participating protein structures. However, even the low precision of ~10 Å displacement of the *ligand* (the smaller protein in the complex) relative to the *receptor* (the larger protein in the complex) results in meaningful predictions of the binding interfaces and the gross structural features of the complex (Vakser et al. 1999). Our procedure GRAMM was shown to adequately address the variable resolution docking of protein structures, by performing fast, approximate docking of low-resolution molecular images and slower, precision docking of more accurate molecular representations (Katchalski-Katzir et al. 1992; Vakser and Aflalo 1994; Vakser 1995). These studies suggested the possibility of docking inaccurate protein models.

In our earlier work (Vakser et al. 1999), we reported the application of GRAMM at low resolution to X-ray protein structures form our nonredundant database of 475 cocrystallized protein–protein complexes. The results of the study were further analyzed in our subsequent report (Tovchigrechko and Vakser 2001), using various statistical models. In the present report, we apply the same techniques to the docking of protein models of different accuracies. To simulate the precision of protein models, all proteins in the protein–protein database were structurally modified in the range of 1–10 Å RMSD, with 1 Å interval. A sophisticated procedure was specifically designed and implemented for that purpose. All resulting models of the proteins were docked. The statistical significance of the docking was analyzed, and the results were correlated with the precision of the models. The data showed that even highly imprecise protein models (>6 Å RMSD) may still yield structurally meaningful docking results, that are accurate enough to predict binding interfaces and to serve as starting points for further structural analysis. The study demonstrated the applicability of existing docking techniques to genome-wide modeling of protein–protein interactions.

Docking tools

Docking algorithm

The docking was performed by our program GRAMM. The details of the docking approach are described elsewhere (Katchalski-Katzir et al. 1992; Vakser 1995). The docking algorithm predicts the structure of a complex by maximizing the geometric match of the molecular images. The digitized images are obtained by projecting the 3D atomic structures of the molecules on a 3D grid. The algorithm is based on the correlation between the digitized molecular images, using fast Fourier transformation. The approach was later reformulated in terms of atom–atom potentials and energy landscapes (Vakser 1996a).

The procedure performs an exhaustive 6D search on a grid and outputs all intermolecular matches with the energy below a set level. An important implication of the grid representation of molecules is that no structural details smaller than the grid step are present in the molecular images. Thus, large grid steps (e.g., 6–7 Å) make it possible to ignore smaller structural inaccuracies. The high-resolution protein docking yields a broad distribution of low-energy positions of the ligand, corresponding to the multiple-minima character of the intermolecular energy landscape. The low-resolution docking, which smoothes the energy landscape, usually results in clustering of the low-energy minima in the area of the binding site (Vakser 1996a; Vakser et al. 1999), corresponding to the position of the binding funnel in the intermolecular energy landscape (Tsai et al. 1999; Shoemaker et al. 2000). The smoothing approach is to a certain degree similar to the concept of Scheraga and associates (Piela et al. 1989), which utilizes an alternative, diffusion-equation formalism. Application of potential smoothing algorithms to protein docking is described in several reports (Vakser 1996a; Trosset and Scheraga 1998;

Pappu et al. 1999). Although the atom-precision docking, naturally, becomes impossible in a low-resolution representation, the docking still predicts the gross features of the complex (an approximate orientation of the molecules, the binding residues, etc.). The validity of the low-resolution docking was confirmed in a large number of studies (see, e.g., Vakser 1996b, 1996c, 1997; Chang et al. 1997; Bridges et al. 1998; Vakser et al. 1999).

Set of cocrystallized proteins

The database of protein models used in this study was generated from a representative set of cocrystallized protein–protein complexes. The set of cocrystallized proteins (Vakser and Sali, http://guitar.rockefeller.edu and http://reco3.ams.sunysb.edu) was built by the following procedure. A pair of chains was considered belonging to the complex if the chains had the same PDB four-character structure identifier and a different chain identifier. A family of complexes was defined as a set of complexes with homologous (>30% sequence identity) receptors and homologous ligands. The set of representative complexes (one complex per family) included 475 complexes with >1000 Å^2 interface area.

Analysis of docking

The analysis of large-scale low-resolution (6.8 Å grid step) docking is described in detail elsewhere (Vakser et al. 1999; Tovchigrechko and Vakser 2001). The results based on large data sets are statistical in nature. Thus, a certain number of correct docking predictions could be obtained just by randomly placing proteins next to each other. Because of that, we define the docking of two proteins as successful if the probability of obtaining the same or better results by a random procedure is <0.05. The random model is based on two assumptions: first, the proteins *A* and *B* in the complex may be roughly considered as spheres, and second, the random matches are uniformly distributed around the receptor (Fig. 1). Assuming that the atoms are homogeneously packed inside the spheres, the radii of the spheres (R_A and R_B) are such that the average distance of the atoms from the sphere's center of mass is the same as it is in the real protein. These radii were calculated and taken into account individually for each complex. The random procedure places the center of mass of a ligand on a sphere with the radius $R_A + R_B$. For both docking and random sets, we calculate the number of matches within "the binding site," defined as <10 Å distance from the correct (crystallographic) position of the ligand's center of mass. The analysis of the low-resolution docking of 475 cocrystallized proteins in our database indicated that 52% of the protein pairs were docked successfully (Vakser et al. 1999).

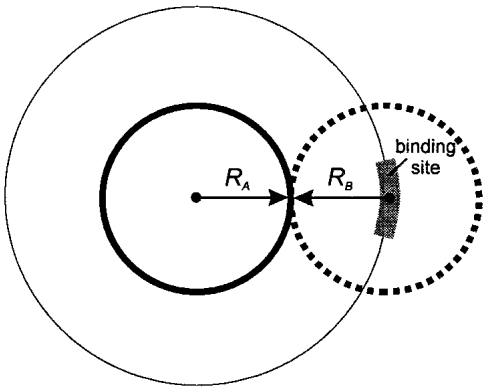

Fig. 1. Model used to calculate the statistical significance of the docking results. Proteins are approximated by spheres. Matches, represented by ligand's center of mass, are placed around the receptor using the uniform random distribution for each match. The binding site is shown in gray. The size of the proteins, shown relative to the size of the binding site (<10 Å), approximately corresponds to the average R_A and R_B values in the database.

Results and Discussion

Because most protein 3D structures in genomes will be computational models, adequate docking techniques are needed for genome-wide structural studies of protein interactions. Large-scale structural features are less affected by inaccuracies in protein modeling. Thus, the low-resolution approach is the appropriate technique for docking these models. The low-resolution docking is applicable to cases of structural inaccuracies in general, not necessarily caused by modeling. GRAMM was applied in the low-resolution mode to the benchmark set of 54 protein–protein complexes (http://zlab.bu.edu/~rong/dock/ benchmark.shtml), in which protein structures were determined by X-ray crystallography in both bound and unbound conformation. The results revealed no statistically significant difference between the docking of bound and unbound structures (the difference appears only at high, atomic-level, resolution). The similarity of bound and unbound protein structures at low resolution is natural because the structural differences, in general, are smaller than the 6.8 Å grid resolution. However, the case of modeled structures is much more complicated, because of potentially large degrees of structural inaccuracies. A number of docking experiments, reported earlier, support the applicability of our GRAMM approach to modeled structures (Vakser 1996b, 1996c; Chang et al. 1997; Bridges et al. 1998). The goal of this study is to provide a *systematic* evaluation of the GRAMM procedure, applied in the low-resolution mode to a representative set of protein models.

Such a set has to satisfy three basic conditions: (a) to be suitable for the validation purpose, the correct docking mode for each structure in the set has to be known; (b) to provide statistical significance to the evaluation results, the set has to be large; (c) to reflect different accuracies of

protein models in genomes, the accuracy of the structures in the set has to vary. Our dataset of 475 cocrystallized protein–protein complexes satisfies conditions (a) and (b). To satisfy condition (c), for each protein in the dataset, we need to build a set of models of different accuracies. We define the accuracy of a model as the RMS deviation of C^{α} atoms from the crystal structure at the best superposition. To be representative, the array of models for each protein has to have at least 10 models, with the accuracy distributed uniformly in the range of 1 to 10 Å. The concept of the whole database is shown in Figure 2.

A natural approach to building the array of models for each protein in the database would be to use a template-based modeling procedure (e.g., homology modeling). However, these procedures do not provide direct control of the RMS deviations. The goal of these procedures is to minimize the RMS deviation from the template structure(s). It is impossible to find a set of templates for each protein in a large database, such that the templates would uniformly cover the required RMSD range for each target protein.

Thus, we chose to simulate the modeled structures by designing a procedure that would distort the experimentally determined structure of a protein to a required degree of inaccuracy. Protein structures in artificially generated conformations, "protein decoys," have long been built by researchers for testing force fields in protein modeling. There are publicly available libraries of the decoys on the Internet (http://prostar.carb.nist.gov, http://dd.stanford.edu). These libraries, however, do not contain any significant number of structures that can be used in our set (i.e., the structures that satisfy the conditions formulated above). Thus, our task was to build decoy-like structures, based on our dataset of protein–protein complexes.

In addition to the requirement of control over RMSD, we formulated the following criteria for such structures:

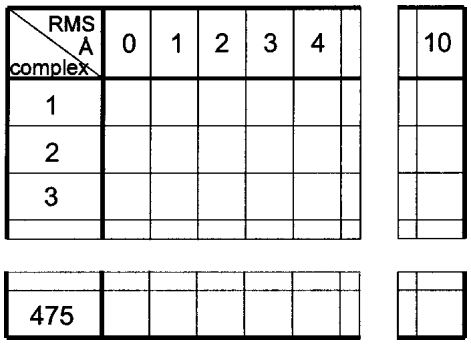

Fig. 2. The concept of the database of protein models for validation of docking. Each of 475 protein–protein complexes is complemented by 10 models of both protein subunits. The accuracy of the models ranges from 1 to 10 Å, with a 1 Å interval.

1. The structures must be in densely packed conformations. The packing has to be similar to the packing in the experimentally determined conformations.
2. The conformations have to include the secondary structure elements—helices and β-strands.
3. Because the structures will be used for low-resolution docking, the atom-size details are not critical. In particular, the structures can have a certain number of stereochemical clashes between atoms, as long as it does not involve significant interpenetration of large structural blocks (e.g., secondary structure elements).
4. The total number of generated models has to be: (475 complexes) × (2 chains in a complex) × (10 RMSD levels) = 9500 (actually, a little less, because some protein chains participate in more than one complex). Thus, the procedure for generating these structures has to be computationally efficient.

None of the existing methods of structure randomization or decoy building could satisfy these criteria. Our computational experiments showed that a naïve approach of randomly shifting atoms positions in the crystal structure, at RMSD >2 Å, destroys the packing (condition 1) and the secondary structure (condition 2). Molecular dynamics runs are computationally very expensive (violating condition 4). The decoy-building method of Park and Levitt (1996), faces the problem of combinatorial explosion for medium and large-size proteins (violating condition 4). Thus, we had to design a new approach to this problem.

Generating protein models

The method that we developed and applied is close to the one of Park and Levitt (1996). First, we obtained the secondary structure assignment in the crystal structure, using the DSSP algorithm. At that point, Park and Levitt choose up to 10 flexible residues in the areas between secondary structure elements, and allow each flexible residue to have four possible conformations. They then enumerate all the conformations, filtering out protein structures with large gyration radii or a large number of clashes. In the case of large proteins, however, 10 flexible residues are not enough to obtain tight packing for large RMSD. If we increased the number of flexible residues, the complete enumeration soon would lead to the combinatorial explosion. Because of that, we adopted another approach, outlined in Figure 3. We marked the φ and ψ angles at all C^{α} atoms outside the secondary structure elements (helices and β-strands) as flexible. All other internal degrees of freedom were "frozen." Then the protein's PDB coordinates were converted into internal coordinates in z-matrix representation, with free coordinates corresponding to the selected angles (typi-

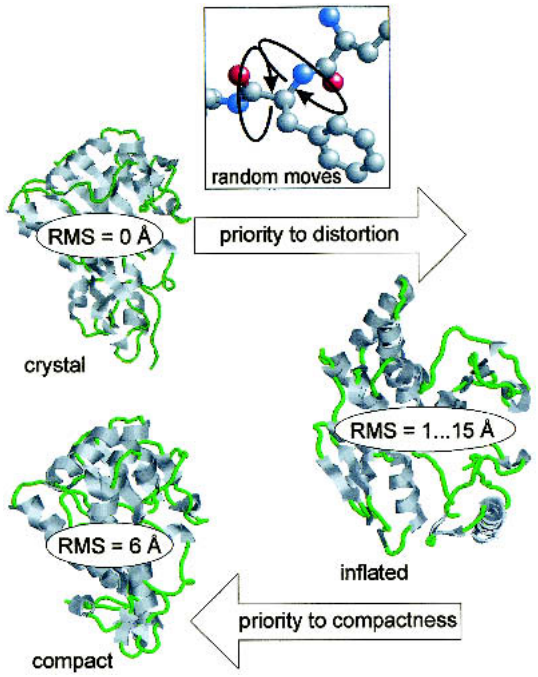

Fig. 3. The outline of the algorithm for generating protein structures with the predefined RMSD from the native structure. Areas with flexible φ and ψ angles (see magnified *inset*) are shown in green. Secondary structure elements, shown in gray, are treated as rigid bodies. See text for details.

cally, >200 degrees of freedom). The coordinates could change continuously. The following Monte Carlo-like procedure was applied to obtain a set of distorted structures in two stages.

Distortion stage.

The procedure performs a random change (move) of each coordinate, randomly choosing the coordinate to change. After each move, it evaluates the objective function, which favors the increase in RMSD and gives a penalty for clashes. The move is accepted with the probability based on the value of the objective function. Because we are interested in faster movement in configurational space and not in obtaining configurations from any particular thermodynamic ensemble, the procedure generates values of random moves from the uniform distribution. While the RMSD gradually increases, the program stores one structure for each RMSD interval 1–2 Å, 2–3 Å, . . ., 9–10 Å, . . ., 14–15 Å.

Compression stage.

The purpose of this stage is to optimize the packing of the distorted structures obtained at the first stage. For each stored structure, we perform the same Monte Carlo-like procedure as above, with the objective function that favors reduction of the gyration radius and penalizes for clashes. The coordinates are dynamically ranked by their "effectiveness": those that recently led to a larger reduction of the gyration radius are used more often. The compressed structures are stored in a table, with 10 elements corresponding to RMSD intervals 1–2 Å, 2–3 Å, . . ., 10–11 Å (Fig. 2). If a new structure with a smaller gyration radius than the one of the structure already stored is obtained, the procedure stores the new one instead. If the procedure went through all structures obtained at the distortion stage and did not fill all the elements in the table, it goes to the first stage starting with RMSD 1 Å below the lowest missing value.

The main reason for splitting the algorithm into two stages is that increasing RMSD and keeping the gyration radius from growing are two tasks that work in opposite directions, when started from a tightly packed crystal structure. Thus, the whole procedure can be viewed as heating the system, to jump over the energy barriers at the distortion stage, and then cooling it and falling into some distant minimum, at the compression stage.

The algorithm was implemented in Fortran using modules from TINKER molecular modeling toolkit (Pappu et al. 1998). The procedure proved to be computationally efficient. On average, it took <30 min on an SGI Octane workstation to generate 10 structures with RMSD 1, 2, . . ., 10 Å for one protein.

The application of the procedure to the set of 475 cocrystallized protein complexes yielded 8285 structures. The procedure failed to generate structures with high RMSD values for several smallest proteins because the movement of secondary structure elements in such proteins would never put them far enough from the original structures. By its nature, the algorithm did not work for any protein chain composed entirely from a single α-helix.

Analysis of generated structures

Visual inspection of the generated structures revealed good preservation of the secondary structure elements, even at RMSD = ~10 Å, although, at large RMSD, β-sheets tend to be destroyed due to a relative movement of the strands.

We analyzed how the distortion of native protein structures affected the binding site area, defined as the original interface area between two proteins in a complex. We assign the residues to the interface if C^{β}–C^{β} (C^{α} for Gly) distance between two residues from different proteins is <7 Å.

The quantitative analysis showed that most of the original interface surface residues (the interface residues in the native structures) remain on the surface in generated structures of different accuracies (Fig. 4). We further calculated the d-RMSD values, which characterize the average change of C^{α}–C^{α} distance between the residues, for the surface patch at the interface area and for the surface patches of the same size outside the interface (Fig. 5). The data show a slight trend toward relatively larger distortion of noninterface ar-

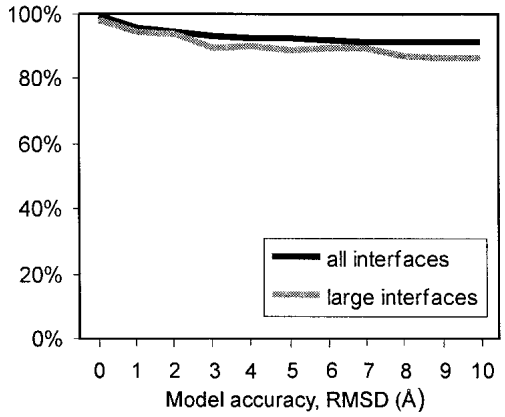

Fig. 4. Percent of original interface residues that remain on the surface in model structures of different accuracies. The surface residues were defined as residues with the side-chain accessible surface >7% of the total accessible surface of the side chain for the residue type (Mizuguchi et al. 1998). The accessible surface was calculated by PSA (Sali and Blundell 1993). The data is averaged over all 475 complexes and over 110 complexes with large interfaces (>4000 $Å^2$).

eas in low-accuracy models. However, comparing with the large absolute values of d-RMSD in these models, the difference is not significant for the purposes of low-resolution docking.

An example of the evolution of the interface at the increasing RMSD values is shown in Figure 6. In the protein chain shown, the geometry of the binding site is still preserved at 4 Å RMSD and is completely destroyed at 10 Å RMSD, with the 6 Å RMSD model being an intermediate case.

Docking of models

For each complex in the database, we performed the low-resolution docking of the original receptor and ligand, then

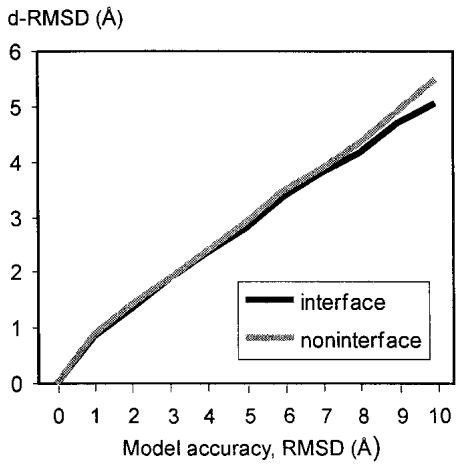

Fig. 5. Average change of distance between residues (d-RMSD) in model structures of different accuracies. The noninterface data is the average over the interface-size surface patches outside the interface.

the docking of 1 Å RMSD receptor model with 1 Å RMSD ligand model, 2 Å RMSD receptor model with 2 Å RMSD ligand model and so on, up to 10 Å RMSD models. The docking protocol (see Docking tools) was the same as in our previous study of cocrystallized structures (Vakser et al. 1999). As in that earlier work, the presence of low-resolution protein–protein recognition was detected by estimating statistical significance of the number of predictions found in the correct location of the binding site (see Docking tools). To evaluate the predictions for a complex of two distorted structures, we assumed the "correct" orientation to be the one defined by the original cocrystallized complex, with the models fitted to the crystal structures (by minimizing the C^{α} RMSD).

An example of the docking results is shown in Figure 7. In this example, the correct binding site is easily identified in the case of the crystal structures by a large cluster of matches. The cluster is still present in the case of a low-accuracy (6 Å RMSD) model.

For a systematic evaluation of the docking results for all complexes, at each level of models accuracy (RMSD = 1, 2, 3, . . ., 10 Å), we calculated the percent of complexes with the correctly predicted low-resolution docking mode. As discussed above, the "correct prediction" means a significantly larger number of matches in the binding site area than in the case of a random distribution of matches. The results (Fig. 8a) show that the ability of the docking procedure to predict the correct structure of a complex declines with the decrease of the protein model's accuracy. However, the percent of correct predictions is still large for significantly inaccurate models—38% for 6 Å RMSD models. Moreover, the percent is still ~30% for models distorted up to 10 Å RMSD. Although the 10 Å RMSD level implies highly inaccurate protein models, the deviation from the original crystal structure may be created by a relative movement of domains. The domains responsible for the interaction may have a lower level of inaccuracy than the structure in general. That may explain the presence of a significant number of successfully docked complexes at the 10 Å RMSD level.

A further analysis of these results helped in their interpretation. In a number of complexes, the position of the clusters of matches (clusters appear because of the smoothing of the intermolecular energy landscape, see Docking Tools) is unrelated to the crystallographically determined binding mode. An important reason for such "incorrect" clusters is the alternative binding modes of the proteins in the complex (Vakser et al. 1999). We analyzed the results to determine the degree to which the docking success rates may be attributed to the clustering itself, rather than to the protein–protein recognition in the correct configuration of the complex. For that purpose, in each complex of crystal structures we placed the binding site in a random place on the receptor. The application of our statistical criterion

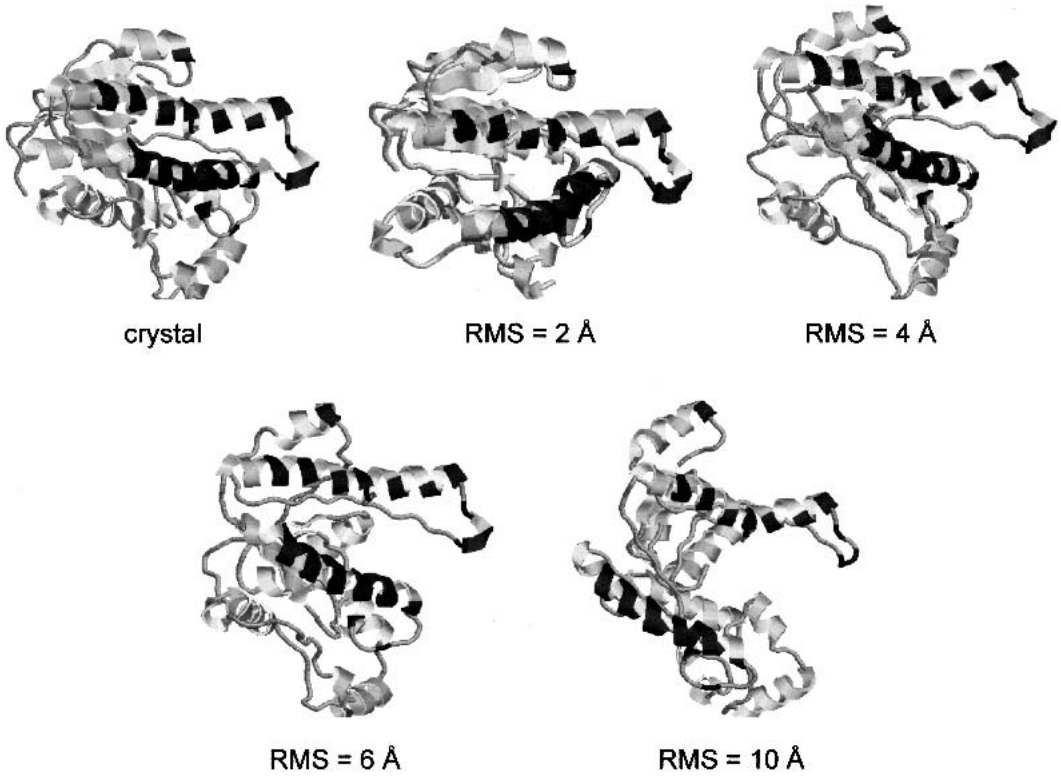

Fig. 6. Evolution of the interface in a set of models with increasing RMSD from the native structure. The structure shown is 1cyd, subunit D. The interface is with the subunit C. The definition of the interface is in the text. The interface residues are in black.

(same as for the real binding sites) to the random binding sites resulted in 18% of complexes placed in the "correctly" predicted category (horizontal line in Fig. 8). Thus, for example, we may estimate that, at the 10 Å RMSD level, approximately half of the correct predictions may come from the clustering of matches only and, thus, may be considered random. However, it is important to emphasize that this is the upper estimate of such randomness. One cannot claim that in a given complex the cluster of matches is positioned near the binding site strictly due to coincidence, and not due to the nature of the interaction.

Because our docking is based on shape complementarity, it is natural to expect that the docking of proteins with large interfaces, on average, will be more successful than the docking of proteins with small interfaces. Indeed, it is the case for the crystal structures (Vakser et al. 1999). To find out whether the same effect exists for inaccurate models, we determined the percent of correctly docked complexes for proteins with small interfaces (1000–2000 $Å^2$, 189 complexes) and large interfaces (>4000 $Å^2$, 110 complexes). The results (Fig. 8b,c) confirm that the proteins with small interfaces are docked less reliably, over the entire range of model accuracy (except for what appears to be a fluctuation at 10 Å RMSD level). The subset with large interfaces was docked more reliably than average. In particular, the complexes of models with 6 Å RMSD and large interfaces were docked successfully in about 50% of cases.

Conclusions

Knowledge of 3D protein structures is important for an adequate description of genomes. Most of these protein structures will be determined by high-throughput modeling procedures. Thus, a structure-based analysis of the network of protein–protein interactions in genomes requires docking methodologies that are capable of dealing with significant structural inaccuracies in the modeled structures of proteins. We present a systematic study of the applicability of our low-resolution docking method to protein models of different accuracies. A representative nonredundant set of 475 cocrystallized protein–protein complexes was used to build an array of models of each protein in the set. A sophisticated procedure was created to generate the models with RMS deviations of 1, 2, 3, . . ., 10 Å from the crystal structure. The docking was performed for all the models, and the predictions were compared with the configurations of the original cocrystallized complexes. Statistical analysis showed that the low-resolution docking can determine the gross structural features of protein–protein interactions for a significant percent of complexes of highly inaccurate pro-

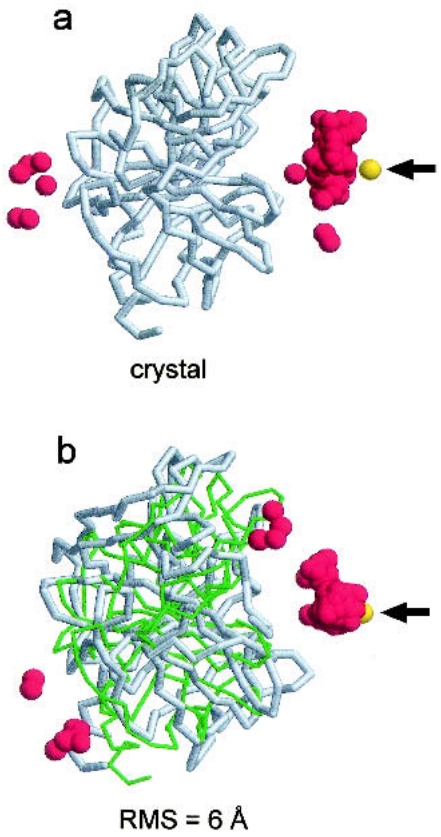

Fig. 7. Results of the low-resolution docking of trypsin and its protein inhibitor BPTI. (*a*) Experimental structures. (*b*) Low-resolution models (RMS = 6 Å, both trypsin and BPTI). The red spheres are the BPTI center of mass in 100 lowest energy positions. The yellow sphere (indicated by an arrow) is the BPTI center of mass in the cocrystallized complex. For comparison, the experimental structure of trypsin, in green, is overlapped with the model. The docking of the models clearly preserves the cluster of correct predictions in the area of the binding site.

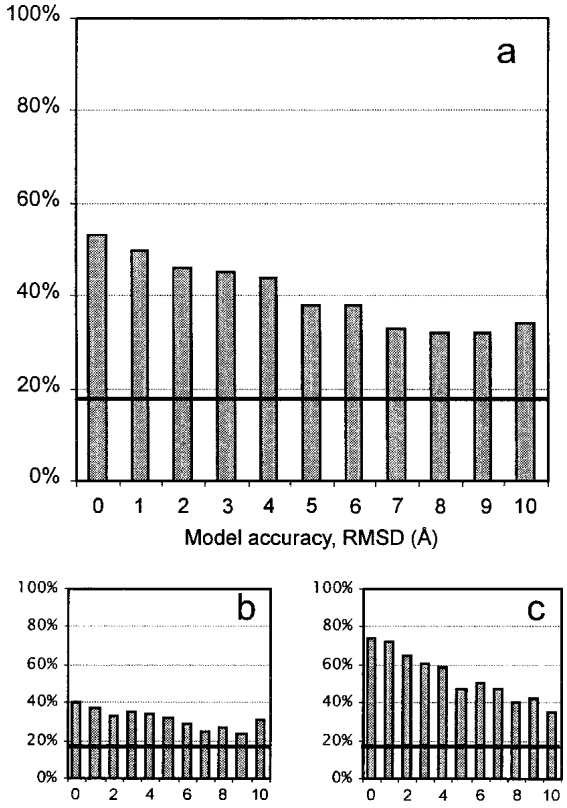

Fig. 8. Percent of correctly predicted complexes for protein models of different accuracies. (*a*) All 475 complexes. (*b*) Complexes with small interfaces (1000–2000 Å^2, 189 complexes). (*c*) Complexes with large interfaces (>4000 Å^2, 110 complexes). The black horizontal line indicates the estimated upper limit on the percent of complexes where the correct prediction could occur by chance, due to the clustering of matches only. See text for the definition of the correct prediction and other details.

tein models. Such predictions may serve as starting points for a more detailed structural analysis, as well as complement experimental and computational data on protein–protein interactions obtained by other techniques. Our current efforts focus on improving the reliability of the large-scale docking of protein models by taking advantage of physicochemical preferences and knowledge-based information.

Acknowledgments

The study was supported by NIH R01 GM61889-01, NSF DBI-9808093, and South Carolina Commission on Higher Education grants.

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

References

Bridges, A., Gruenke, L., Chang, Y.-T., Vakser, I.A., Loew, G., and Waskell, L. 1998. Identification of the binding site on cytochrome p450 2b4 for cytochrome b5 and cytochrome p450 reductase. *J. Biol. Chem.* **273:** 17036–17049.

Burley, S.K. 2000. An overview of structural genomics. *Nat. Struct. Biol.* **7:** 932–934.

Chang, Y.-T., Stiffelman, O.B., Vakser, I.A., Loew, G.H., Bridges, A., and Waskell, L. 1997. Construction of a 3D model of cytochrome p450 2b4. *Protein Eng.* **10:** 119–129.

Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. 2000. Protein function in the post-genomic era. *Nature* **405:** 823–826.

Jones, T.A. and Kleywegt, G.J. 1999. CASP3 comparative modeling evaluation. *Proteins* **Suppl. 3:** 30–46.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. 1992. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.* **89:** 2195–2199.

McCammon, J.A. 1998. Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.* **8:** 245–249.

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S., and Overington, J.P. 1998. JOY: Protein sequence-structure representation and analysis. *Bioinformatics* **14:** 617–623.

Murzin, A.G. 1999. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins* **Suppl. 3:** 88–103.

Oliver, S. 2000. Guilt-by-association goes global. *Nature* **403:** 601–603.

Orengo, C.A., Bray, J.E., Hubbard, T., LoConte, L., and Sillitoe, I. 1999. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* **Suppl. 3:** 149–170.

Pappu, R.V., Hart, R.K., and Ponder, J.W. 1998. Analysis and application of

potential energy smoothing and search methods for global optimization. *J. Phys. Chem. B* **102:** 9725–9742.

Pappu, R.V., Marshall, G.R., and Ponder, J.W. 1999. A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Biol.* **6:** 50–55.

Park, B. and Levitt, M. 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* **258:** 367–392.

Piela, L., Kostrowicki, J., and Scheraga, H.A. 1989. The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.* **93:** 3339–3346.

Sali, A. and Blundell, T.L. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779–815.

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M.A., Madhusudhan, M.S., Mirkovic, N., and Sali, A. 2000. Protein structure modeling for structural genomics. *Nat. Struct. Biol.* **7:** 986–990.

Shoemaker, B.A., Portman, J.J., and Wolynes, P.G. 2000. Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proc. Natl. Acad. Sci.* **97:** 8868–8873.

Skolnick, J., Fetrow, J.S., and Kolinski, A. 2000. Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.* **18:** 283–287.

Sternberg, M.J.E., Gabb, H.A., and Jackson, R.M. 1998. Predictive docking of protein–protein and protein–DNA complexes. *Curr. Opin. Struct. Biol.* **8:** 250–256.

Tovchigrechko, A. and Vakser, I.A. 2001. How common is the funnel-like energy landscape in protein–protein interactions? *Protein Sci.* **10:** 1572–1583.

Trosset, J.-Y. and Scheraga, H.A. 1998. Reaching the global minimum in docking simulations: A Monte Carlo energy minimization approach using Bezier splines. *Proc. Nat. Acad. Sci.* **95:** 8011–8015.

Tsai, C.-J., Kumar, S., Ma, B., and Nussinov, R. 1999. Folding funnels, binding funnels, and protein function. *Protein Sci.* **8:** 1181–1190.

Vajda, S., Sippl, M., and Novotny, J. 1997. Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* **7:** 222–228.

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

Vakser, I.A. 1996a. Long-distance potentials: An approach to the multiple-minima problem in ligand-receptor interaction. *Protein Eng.* **9:** 37–41.

———. 1996b. Low-resolution docking: Prediction of complexes for underdetermined structures. *Biopolymers* **39:** 455–464.

———. 1996c. Main-chain complementarity in protein–protein recognition. *Protein Eng.* **9:** 741–744.

———. 1997. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin–antibody complex. *Proteins* **Suppl. 1:** 226–230.

Vakser, I.A. and Aflalo, C. 1994. Hydrophobic docking: A proposed enhancement to molecular recognition techniques. *Proteins* **20:** 320–329.

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

Vukmirovic, O.G. and Tilghman, S.M. 2000. Exploring genome space. *Nature* **405:** 820–822.