Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 8477–8482, July 1999
Biophysics

A systematic study of low-resolution recognition in protein–protein complexes

ILYA A. VAKSER*†, OMAR G. MATAR‡, AND CHAN F. LAM‡

*Department of Cell and Molecular Pharmacology and ‡Department of Biometry, Medical University of South Carolina, 171 Ashley Avenue, Charleston, SC 29425

Edited by Peter G. Wolynes, University of Illinois at Urbana-Champaign, Urbana, IL, and approved May 26, 1999 (received for review April 13, 1999)

ABSTRACT **A comprehensive nonredundant database of 475 cocrystallized protein–protein complexes was used to study low-resolution recognition, which was reported in earlier docking experiments with a small number of proteins. The docking program GRAMM was used to delete the atom-size structural details and systematically dock the resulting molecular images. The results reveal the existence of the low-resolution recognition in 52% of all complexes in the database and in 76% of the 113 complexes with an interface area >4,000 Å^2. Limitations of the docking and analysis tools used in this study suggest that the actual number of complexes with the low-resolution recognition is higher. However, the results already prove the existence of the low-resolution recognition on a broad scale.**

Protein–protein interactions play a central role in protein function. Because these interactions are determined by the structure of the components that form the complex as well as by the physicochemical properties of the environment, studies of these factors are important for better understanding of protein functions and for the subsequent application of this knowledge to protein engineering and drug design.

Computer modeling makes it possible to perform direct computational experiments to study fundamental principles of protein interactions in a way that often would be impossible in "real" experiments. What is the role of the large-scale structural motifs (e.g., the main-chain fold) in protein recognition? A direct experiment to determine this role would be to eliminate the small, atom-size structural elements and test the recognition properties of the remaining structure. Such an experiment (1–3) is clearly feasible only computationally (*in silico*). Studies of large-scale recognition factors include correlation of the antigenicity of surface areas with their accessibility to large probes (4), role of the surface clefts (5), automatic binding site identification based on geometric criteria (6, 7), study of the "low-frequency" surface properties (8), and "fuzzy" binding-site descriptors (9, 10). Several protein-recognition techniques use smoothed potential functions (11–14), which effectively are equivalent to the averaging of the contribution of neighboring atoms and, thus, to the "smoothing" of the local structural elements. Studies of protein binding (15, 16) and energy landscapes in protein folding (17) and protein interactions (18, 19) confirm the existence of nonlocal recognition preferences.

Progress in understanding the principles of protein recognition leads to better computational methods for protein docking (20–22). The principal drawback of the existing docking methodologies is sensitivity to structural inaccuracies. One example of such inaccuracies is conformational changes upon the formation of the complex (23, 24). A major obstacle to the docking of protein structures obtained with modeling is significant errors in these structures (25). This aspect is especially important in view of the current progress in genome sequencing. Most of the resulting protein structures will have to be modeled rather than determined experimentally (26). Thus, the structure-based functional studies will require computational techniques capable of docking large numbers of protein models of limited accuracy within reasonable computational time. In short, the docking methods needed for global, genome-scale studies have to be fast and have to tolerate structural inaccuracies on the order of a few angstroms, even at the expense of substantially lower precision in the docking results.

The program GRAMM (1, 27, 28) has been shown to adequately address these issues in a number of tests (2, 24, 29, 30). The procedure allows docking at variable "resolutions," depending on the accuracy of the structural components to be docked. The high-resolution docking yields high-precision results and is relatively slow (hours of computational time). The low-resolution docking is fast (several seconds of cpu time) and may tolerate structural inaccuracies on the order of 7 Å, which is a precision characteristic of many protein models (31–33). However, it can predict only complex's gross features, which may serve as a starting point for a more detailed study. The essence of the procedure is the reduction of protein structures to digitized images on a three-dimensional grid. The structural elements smaller than the step of the grid are not present in the docking. Thus, the procedure provides a convenient tool to eliminate smaller (e.g., atom-size) details. This feature is the source of tolerance to structural inaccuracies. At the same time, it makes possible the study of the role of the low-resolution recognition factors in protein complexes.

The low-resolution recognition was studied earlier with GRAMM on a limited number of protein complexes (2). The results show the existence of preferences to the correct structure of the complex even at the resolution of 7 Å. The limited number of test cases, however, did not allow broader conclusions to be drawn about the existence of such factors in general.

In the present study, a comprehensive nonredundant database of crystallized protein–protein complexes (I.V. and A. Sali, unpublished data) was used to determine the existence of the low-resolution recognition. This database provided an opportunity for a systematic study of protein recognition by using structural data presently available for this purpose. All details smaller than 7 Å were eliminated from the protein structures. GRAMM was able to determine the existence of the low-resolution recognition in 52% of the complexes with interface area >1,000 Å^2 and in 76% of the complexes with interface area >4,000 Å^2. Our inability to detect the low-resolution recognition in the remaining complexes does not

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

PNAS is available online at www.pnas.org.

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

†To whom reprint requests should be addressed at: Department of Cell and Molecular Pharmacology, Medical University of South Carolina, 173 Ashley Avenue, PO Box 250505, Charleston, SC 29425. e-mail: vakseri@musc.edu.

mean that these complexes do not have this property. The fact that GRAMM, like any other procedure, has limited capabilities suggests that the actual percentage of complexes with the low-resolution recognition is higher.

METHODS

The details of the GRAMM docking approach are described in refs. 1 and 27. The method involves (*i*) a projection of the two molecules on a three-dimensional grid; (*ii*) the calculation, using Fourier transformation, of a correlation function that assesses the degree of surface overlap and the penetration on relative shifts of the molecules in three dimensions; and (*iii*) a scan of the relative orientations of the molecules in three dimensions. The algorithm provides a list of correlation values that indicate the extent of geometric match between the surfaces of the molecules; each of these values is associated with six numbers describing the relative position (translation and rotation) of the molecules. The procedure is thus equivalent to a six-dimensional search but is much faster by design.

The overlap of the molecular images is equivalent to the intermolecular energy E calculated with a step-function potential (14).

$$E = \sum_{i,j} E(r_{ij}), \qquad E(r_{ij}) = \begin{cases} U, & 0 < r_{ij} \le R \\ -1, & R < r_{ij} \le 2R \\ 0, & r_{ij} > 2R \end{cases}$$

where E is the energy, U is the height of the repulsion part of the potential, R is the range of the potential (the grid step), and r_{ij} is the distance between atoms i (receptor) and j (ligand). Because the molecules are represented by grid images, no structural details smaller than the step of the grid are taken into account in the calculations. Thus, in the low-resolution docking, a sparse grid with ≈7 Å grid step eliminates all atom-size details.

RESULTS

Database of Complexes. The database included 475 complexes from the Protein Data Bank (34). A structure was considered a protein–protein complex if it consisted of more than one chain of 30 or more residues. For convenience, the larger and the smaller proteins within a complex were called "receptor" and "ligand," respectively. The database is nonredundant in that no complex has both the receptor and the ligand homologous to the receptor and the ligand of any other complex in the database. The criterion for the homology was 30% or greater sequence identity. The database had 631 complexes with physical contact between subunits. For the purpose of this study, only 475 complexes with interface area >1,000 Å^2 were taken. The protein pairs with smaller interfaces were not considered, in an attempt to minimize the number of complexes that are artifacts of crystallization and thus do not reflect biological functions (35–37).

Docking. Docking was performed by using GRAMM at low resolution. The procedure implemented an exhaustive grid search for the ligand–receptor structure matches. The docking parameters were: step of the grid, 6.8 Å; repulsion part of the potential, 6.5; and interval for rotations, 20°. For each complex, the 1,000 lowest-energy matches were analyzed. The values of the parameters were determined earlier (1) as

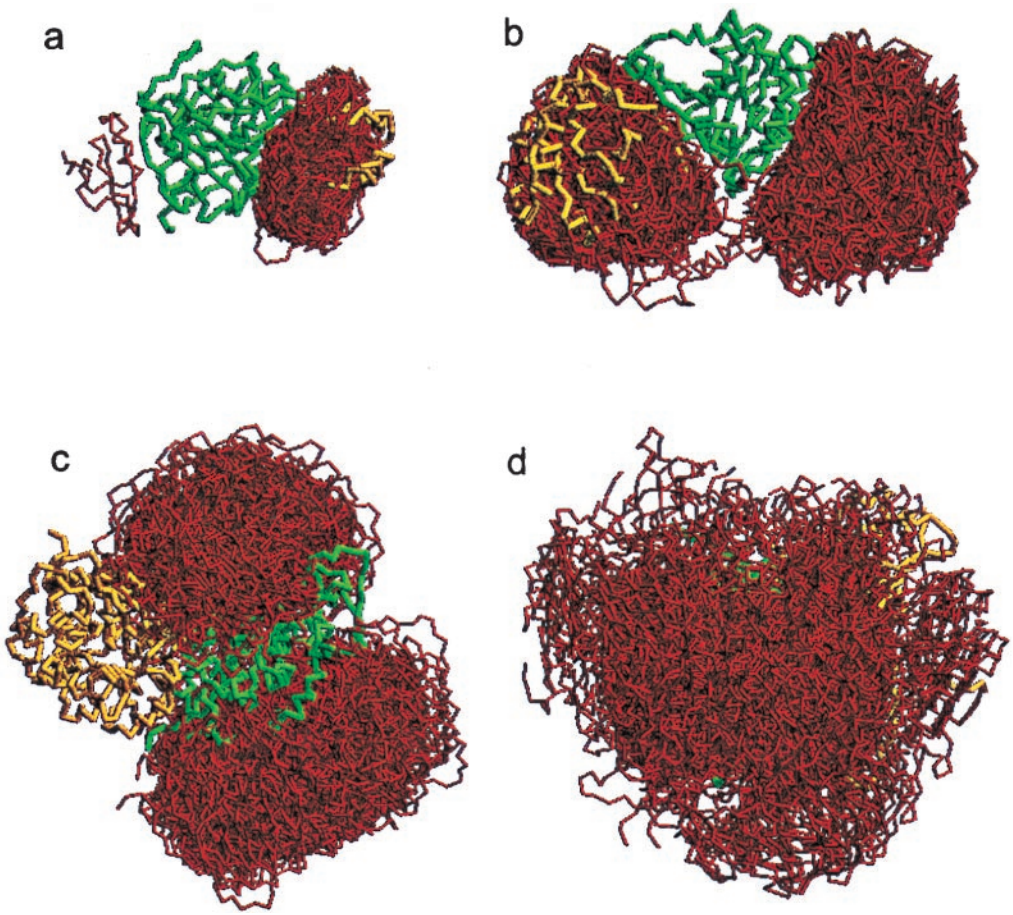

FIG. 1. Examples of the distribution of ligand positions. Receptors are shown in green and ligands in yellow, in the crystallographically determined position within the complex. The 100 lowest-energy ligand positions are shown in red. Matches are clustered primarily inside the binding area (*a*), inside and outside the binding area (*b*), outside the binding area (*c*), and not clustered (*d*).

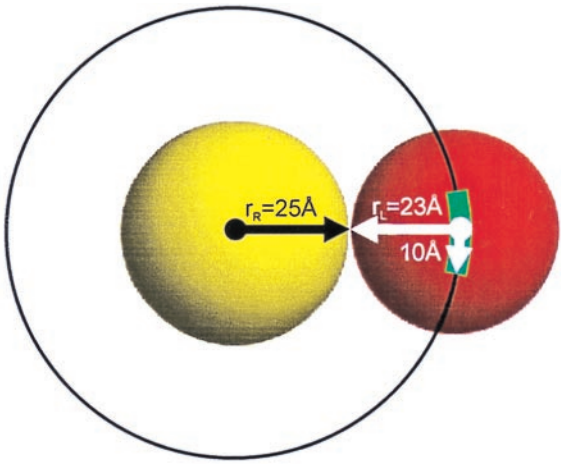

FIG. 2. Idealized representation of proteins. The receptor is shown in yellow and the ligand in red. Radii are calculated as the average distance of all atoms in a protein from its center of mass. The radii shown are the average radii of all receptors and ligands. The binding region is defined as the area within 10 Å of the crystallographic position of the ligand's center of mass and is shown in green.

quasi-optimal for the low-resolution docking, based on 10 receptor–ligand complexes. These values were obtained by largely empirical, nonquantitative considerations and were not optimized or modified in any other way for the criteria used in the present study.

The nature of the GRAMM approach dictates that all structural details smaller than the grid step (in this study, 6.8 Å) are deleted from the molecular images. It was shown earlier (14) that the difference between the low- and the high-resolution docking is that at high resolution, the low-energy positions of the ligand are dispersed around the receptor (what is usually referred to as "the multiple-minima problem"), whereas at low resolution they tend to cluster in the area of the global minimum (the binding site on the receptor). From the point of view of molecular shape, this effect has to do with smoothing smaller structural details, so that only the larger ones, usually associated with the binding site (e.g., deep cavity in the enzymes active sites) remain and attract ligand matches with different ligand orientation. From the point of view of the intermolecular energy, the transition to lower resolution means an increase in the potential range (14). This leads to a long-range, "mean force" potential that averages the contributions of multiple atoms. The potential corresponds to a smoother energy profile that leaves a smaller number of minima (ideally one), which leads to the clustering of the ligand positions in these minima (at the binding sites). Examples of the actual distribution of the ligand positions are shown in Fig. 1. In many cases, the clustering occurs in the areas that are not identified in the crystal structures as binding sites. Presently, it is not clear whether these clusters correspond to alternative binding sites.

Basic Assumptions in the Analysis of the Results. For the analysis of the docking results, we calculated the average distance of each atom from the center of mass for all proteins in the database. These average distances r_{Ri} and r_{Li} were considered, respectively, as the "radii" of the receptor and the ligand in the complex i. The average values of such radii for all receptors r_R and all ligands r_L were 25 Å and 23 Å, respectively (Fig. 2). In this study, for simplicity, we analyzed only the positions of the center of mass of the ligands. The low-resolution binding site on the receptor was defined as the area within 10 Å of the position of the ligand's center of mass in the crystal structure (Fig. 2).

The output of GRAMM docking is a list of ligand's positions sorted according to the score of the match. The score is proportional to the surface overlap (1, 27). At the same time, it is equivalent to the intermolecular energy calculated with a simplified potential (14). The basis of the analysis of the docking results was the assumption that if the ligand–receptor recognition exists, the low-energy matches are more populated

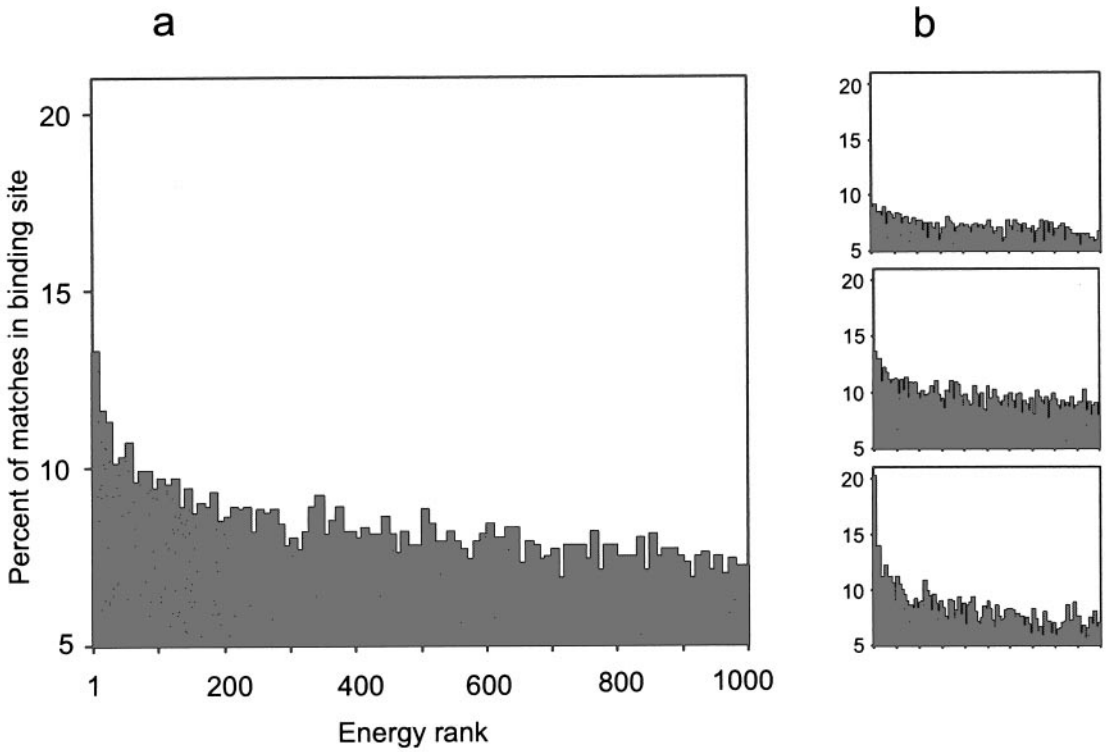

FIG. 3. Percent of matches inside the binding area according to the energy rank. The percent is based on the inside/total ratio of the matches of a given rank (see text). Energy rank is accumulated in the histogram in groups of 10. (*a*) All complexes. (*b*) Complexes with interface of 1,000–2,000 Å^2 (*Top*), 2,000–4,000 Å^2 (*Middle*), and >4,000 Å^2 (*Bottom*).

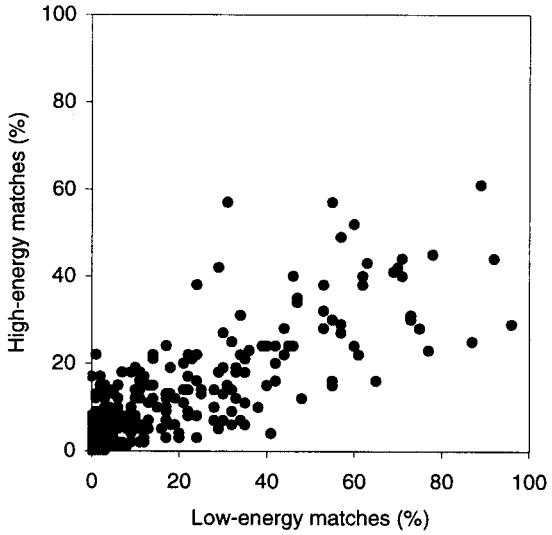

FIG. 4. Correlation of the percent of matches inside the binding area of low-energy (rank 1–100) and high-energy (rank 901–1,000) ligand positions. The percent values are calculated for every complex in the database.

in the binding site than the would-be random ones, whereas the high-energy matches are distributed randomly regardless of the position of the binding site.

A Trend Toward the Actual Structure of the Complex. GRAMM performs an exhaustive grid search, which reports all possible matches (within the accuracy of the grid), and outputs the list of matches sorted by energy. Thus, the evidence that the low-energy matches are better represented in the binding site than the high-energy ones would indicate a preference toward the actual structure of the complex. If the number of matches inside and outside the binding site is n_{in} and n_{out}, respectively, then the low-energy matches in the binding site are represented better than the high-energy ones, if $n^l_{in}/(n^l_{in} + n^l_{out}) > n^h_{in}/(n^h_{in} + n^h_{out})$, where l is low energy and h is high energy.

Fig. 3 shows the distribution of the percent of matches in the binding site $p = 100 \cdot n_{in}/(n_{in} + n_{out})$ for the entire database, according to the energy of the match. The distribution clearly shows a strong nonlinear correlation of this percent with the energy, resulting in an inside/total ratio of the low-energy matches significantly higher than the inside/total ratio of the high-energy ones. The other conclusion is that this difference in the low-energy and the high-energy matches depends on the area of the interface in the crystal structures (little difference for smaller interfaces and substantially bigger difference for larger interfaces).

As shown in Fig. 3, the trend to smaller inside/total ratio of the higher-energy matches continues through the entire energy spectrum (only the first 1,000 lowest-energy matches were analyzed for each complex). To assess the difference in the low-energy and the high-energy population objectively, it was useful to find out actually how high the high-energy values are. Fig. 4 shows a significant correlation between p^l based on 100 low-energy matches (rank 1–100) and p^h based on 100 high-energy matches (rank 901–1,000, highest-energy analyzed). This indicates that for a number of complexes, the highest-energy matches analyzed were still clustered in the binding site. Thus, for such complexes, the rank of the high-energy matches that are supposed to be distributed regardless of the binding site could be well beyond the first 1,000.

The Number of Complexes with Low-Resolution Recognition. The analysis of total values for the entire database reveals the general character of and trends in the low-resolution recognition. However, it does not answer one of the most intriguing questions, i.e., how many protein complexes follow

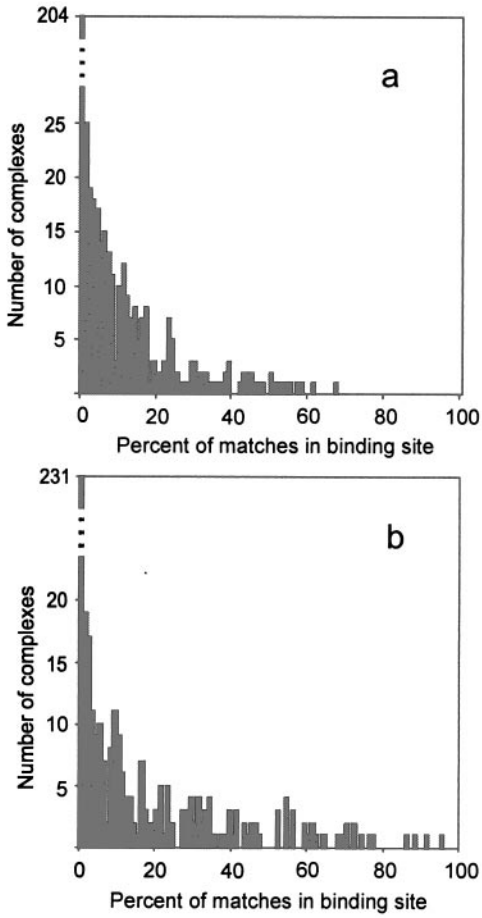

FIG. 5. Distribution of complexes according to the percent of matches inside the binding site. The total number of matches per complex is 1,000 (*a*) and 100 (*b*) (lowest energy matches).

the low-resolution recognition? Is it a universal feature or does it apply only to some proteins? To address this question, one has to look at the distribution of matches in individual protein complexes. The analysis of individual complexes in this study was based on an assumption that the absence of the low-resolution recognition corresponds to a random distribution of matches in the docking of low-resolution structures. Thus, detecting a significantly higher than the would-be random number of matches in the binding site would indicate the existence of low-resolution recognition.

An important aspect in such analysis is modeling of the random matches. The number of matches analyzed for each complex was 1,000, sorted from low to high energy. As shown

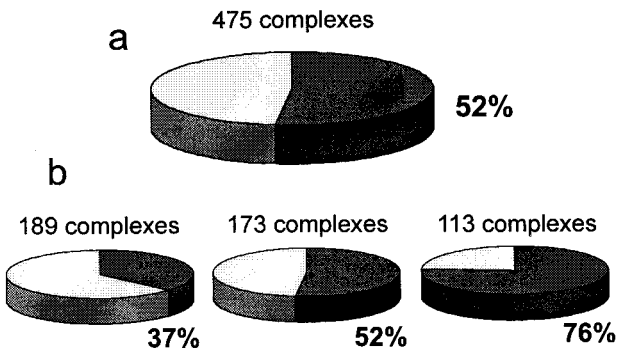

FIG. 6. Percent of complexes with detected low-resolution recognition. (*a*) All complexes. (*b*) Complexes with interface area 1,000–2,000 Å^2 (*Left*), 2,000–4,000 Å^2 (*Middle*), and >4,000 Å^2 (*Right*).

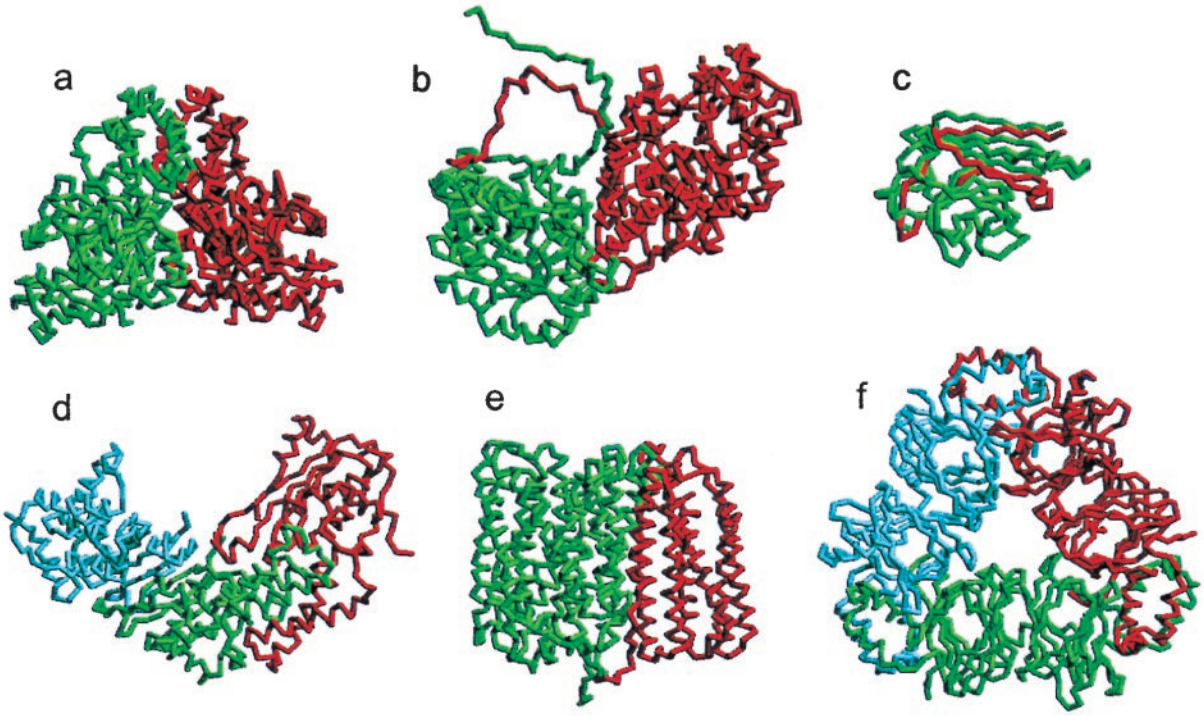

FIG. 7. Examples of complexes with and without detected low-resolution recognition. The receptor is shown in green and the ligand in red. All structures are in the cocrystallized positions. (*a*) A complex with established low-resolution. Complexes without detected low-resolution recognition: disordered termini that are part of the interface (*b*), interwoven chains (*c*), an alternative binding mode with the subunit identical to the ligand shown in blue (*d*), helix bundles with a cylinder-like low-resolution structure (*e*), and a ternary complex (*f*).

above, the highest-energy matches in this list, although distributed with smaller than the low-energy ones inside-binding-site/total ratio, in a number of complexes were still clustered in the binding area. Thus, they cannot be considered random. The option for modeling random matches that was found feasible, although far from being ideal, is based on two assumptions: first, the proteins may be roughly considered as spheres, with the radius equal to the average distance of all atoms from the center of mass (Fig. 2), and second, the random matches are uniformly distributed around the receptor. In such case, the area of the binding site is $S_b = \pi \cdot 10^2$ (Fig. 2), and the total area available for the matches (a match is positioned in the ligand's center of gravity) in a complex i is $S_{ti} = 4\pi(r_{Ri} + r_{Li})^2$. Fig. 2 shows the average values of r_{Ri} and r_{Li}. In our analysis, however, these radii were calculated and taken into account individually for each complex. The number of sites with the area equal to that of the binding site is $n = S_{ti}/S_b$. The probability of K matches in such site (e.g., the binding site) could be approximated by a Poisson distribution (38), with the mean number of matches $m = 1{,}000/n$ and SD $m^{1/2}$. The 2-SD confidence interval ($\approx$95% interval) is $m \pm 2m^{1/2}$. Thus, if the actual number of matches in the binding site K is larger than $m + 2m^{1/2}$, it is significantly larger than the random one and, consequently, the low-resolution recognition was considered detected for this complex.

The distribution of complexes according to the percent of matches in the binding site is shown in Fig. 5. As can be seen, a significant number of complexes have a very large percentage of matches in the binding site. At the same time, many complexes have no matches in the binding site. Both cases point to the nonrandom character of the match's distribution. In the case of no matches at the ligand–receptor interface, the matches were usually clustered at different sites, which may be an indication of alternative binding modes. The analysis of complexes based on the comparison with the distribution of the random matches (Fig. 6) determined 52% of all complexes to have the low-resolution recognition property (37%, 52%, and 76% of complexes with interface area 1,000–2,000 Å^2, 2,000–4,000 Å^2, and >4,000 Å^2, respectively). Obviously, like any computational approach, both GRAMM and the analysis procedure have limitations in terms of the algorithm, implementation, choice of parameters, etc. Thus, it is unrealistic to expect detection of all low-resolution recognition cases. The actual number of such cases may be significantly higher. However, we presently do not have a better estimate of this number.

Complexes Without Established Low-Resolution Recognition. Examples of complexes in which we did not succeed in detecting the low-resolution recognition are shown in Fig. 7. In most such cases, the factors that cause fewer matches in the crystallographically determined binding sites are clear (e.g., alternative binding mode, chain interpenetration, nonbinary complex). A deeper insight into such special configurations of complexes would allow one to increase the number of detected low-resolution recognition cases. Such study would require multiple sets of docking parameters and more sophisticated analysis tools. At this point, however, we chose a simple approach, that in a systematic way confirms the existence of low-resolution recognition on a broad scale and left a more comprehensive analysis for future study.

CONCLUSIONS

A comprehensive, nonredundant database of cocrystallized protein–protein complexes was used to study low-resolution recognition, which was reported in earlier docking experiments with a small number of proteins. The docking program GRAMM was used to delete the atom-size structural details and systematically dock the resulting molecular images. Analysis of the results revealed the following. (*i*) The distribution of matches in the entire database showed that inside-binding-site/total ratio for the low-energy matches is higher than that for the high-energy matches, indicating the existence of a general docking preference toward the actual binding mode

and, thus, showing the significance of the low-resolution recognition. (*ii*) Significantly higher than random number of matches in the binding area, indicating the existence of the low-resolution recognition, was detected in 52% of all complexes (in 37%, 52%, and 76% of complexes with interface area 1,000–2,000 Å^2, 2,000–4,000 Å^2, and >4,000 Å^2, respectively). Limitations of the docking and analysis tools used in this study suggest that the actual number of complexes with low-resolution recognition is higher. However, the results already prove the existence of the low-resolution recognition on a broad scale.

The authors thank Dan Knapp and John Hildebrandt for reading the manuscript and for helpful comments. This work was supported by the National Science Foundation Computational Biology Activities grant and the South Carolina/National Science Foundation Experimental Program to Stimulate Competitive Research Cooperative Agreement.

1. Vakser, I. A. (1995) *Protein Eng.* **8,** 371–377.
2. Vakser, I. A. (1996) *Biopolymers* **39,** 455–464.
3. Vakser, I. A. (1996) *Protein Eng.* **9,** 741–744.
4. Novotny, J., Handschumacher, M., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A. & Rose, G. D. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 226–230.
5. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996) *Protein Sci.* **5,** 2438–2452.
6. Peters, K. P., Fauck, J. & Frommel, C. (1996) *J. Mol. Biol.* **256,** 201–213.
7. Ho, C. M. W. & Marshall, G. R. (1990) *J. Comput. Aided Mol. Des.* **4,** 337–354.
8. Duncan, B. S. & Olson, A. J. (1993) *Biopolymers* **33,** 231–238.
9. Fetrow, J. S. & Skolnick, J. (1998) *J. Mol. Biol.* **281,** 949–968.
10. Fetrow, J. S., Godzik, A. & Skolnick, J. (1998) *J. Mol. Biol.* **282,** 703–711.
11. Pappu, R. V., Marshall, G. R. & Ponder, J. W. (1999) *Nat. Struct. Biol.* **6,** 50–55.
12. Trosset, J.-Y. & Scheraga, H. A. (1998) *Proc. Nat. Acad. Sci. USA* **95,** 8011–8015.
13. Robert, C. H. & Janin, J. (1998) *J. Mol. Biol.* **283,** 1037–1047.
14. Vakser, I. A. (1996) *Protein Eng.* **9,** 37–41.
15. Berg, O. G. & von Hippel, P. H. (1985) *Annu. Rev. Biophys. Biophys. Chem.* **14,** 131–160.
16. McCammon, J. A. (1998) *Curr. Opin. Struct. Biol.* **8,** 245–249.
17. Panchenko, A. R., Luthey-Schulten, Z., Cole, R. & Wolynes, P. G. (1997) *J. Mol. Biol.* **272,** 95–105.
18. Zhang, C., Chen, J. & DeLisi, C. (1999) *Proteins* **34,** 255–267.
19. Camacho, C. J., Weng, Z., Vajda, S. & DeLisi, C. (1999) *Biophys. J.* **76,** 1166–1178.
20. Sternberg, M. J. E., Gabb, H. A. & Jackson, R. M. (1998) *Curr. Opin. Struct. Biol.* **8,** 250–256.
21. Kuntz, I. D., Meng, E. C. & Shoichet, B. K. (1994) *Acc. Chem. Res.* **27,** 117–123.
22. Vajda, S., Sippl, M. & Novotny, J. (1997) *Curr. Opin. Struct. Biol.* **7,** 222–228.
23. Dixon, J. S. (1997) *Proteins,* Suppl. 1**,** 198–204.
24. Vakser, I. A. (1997) *Proteins,* Suppl. 1, 226–230.
25. Dunbrack, R. L. J., Gerloff, D. L., Bower, M., Chen, X., Lichtarge, O. & Cohen, F. E. (1997) *Fold. Des.* **2,** R27–R42.
26. Sanchez, R. & Sali, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13597–13602.
27. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 2195–2199.
28. Vakser, I. A. & Aflalo, C. (1994) *Proteins* **20,** 320–329.
29. Chang, Y.-T., Stiffelman, O. B., Vakser, I. A., Loew, G. H., Bridges, A. & Waskell, L. (1997) *Protein Eng.* **10,** 119–129.
30. Bridges, A., Gruenke, L., Chang, Y.-T., Vakser, I. A., Loew, G. & Waskell, L. (1998) *J. Biol. Chem.* **273,** 17036–17049.
31. Martin, A. C. R., MacArthur, M. W. & Thornton, J. M. (1997) *Proteins,* Suppl. 1**,** 14–28.
32. Marchler-Bauer, A., Levitt, M. & Bryant, S. H. (1997) *Proteins,* Suppl. 1, 83–91.
33. Lesk, A. M. (1997) *Proteins,* Suppl. 1, 151–166.
34. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. L. & Weng, J. (1987) in *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Commission of the International Union of Crystallography, Bonn, Germany), pp. 107–132.
35. Janin, J. & Rodier, F. (1995) *Proteins* **23,** 580–587.
36. Carugo, O. & Argos, P. (1997) *Protein Sci.* **6,** 2261–2263.
37. Tsai, C.-J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996) *J. Mol. Biol.* **260,** 604–620.
38. Papoulis, A. (1965) *Probability, Random Variables and Stochastic Processes* (McGraw-Hill, New York).