PROTEINS: Structure, Function, and Genetics 52:33–40 (2003)

Finding Needles in Haystacks: Reranking DOT Results by Using Shape Complementarity, Cluster Analysis, and Biological Information

Dennis S. Law,[1] Lynn F. Ten Eyck,[1] Omer Katzenelson,[1] Igor Tsigelny,[1] Victoria A. Roberts,[2] Mike E. Pique,[2] and Julie C. Mitchell[1]*

[1]*San Diego Supercomputer Center, University of California at San Diego, La Jolla, California*
[2]*The Scripps Research Institute, Department of Molecular Biology, La Jolla, California*

ABSTRACT **We present an evaluation of our results for the first Critical Assessment of PRedicted Interaction (CAPRI). The methods used include the molecular docking program DOT, shape analysis tool FADE, cluster analysis and filtering based on biological data. Good results were obtained for most of the seven CAPRI targets, and for two systems, submissions having the highest number of correctly predicted contacts were produced. Proteins 2003;52: 33–40.** © **2003 Wiley-Liss, Inc.**

Key words: CAPRI; DOT; FADE; cluster analysis; molecular docking

INTRODUCTION

The objective of molecular docking is the prediction of a bound complex between two molecules given their individual three-dimensional structures. Docking prediction with unbound coordinates presents a larger problem than prediction with bound coordinates due to the possibility of conformational changes in the proteins on binding. Correctly accounting for flexibility in the proteins is often too computationally intensive to be completed in a reasonable amount of time with current methods. If the proteins are modeled as rigid bodies, the computational time decreases substantially and the energy of interaction can be evaluated more quickly.

Here, we present results for the first Critical Assessment of PRedicted Interactions (CAPRI) prediction exercise. Our approach uses the molecular docking program DOT,[1] the shape analysis tool FADE,[2] and cluster analysis, followed by visual inspection. Although the results suggest ways in which the methods can be improved, they compared favorably with those of other groups. In particular, the methods returned some correct contacts in our first prediction for all three CAPRI round 1 systems (targets 1–3), and they achieved the highest number of predicted contacts for two of the four CAPRI round 2 systems (targets 4–7). The analysis will show that DOT is able to generate good predictions for most CAPRI systems, although the program is not always able to rank the best results as its top choice. Shape complementarity and cluster analysis both made a positive contribution to the post-DOT analysis, although neither can be used exclusively to deduce the best DOT results.

MATERIALS AND METHODS

This section will give detailed descriptions of the methods used in making our predictions. In brief, we used the following procedures:

- Interaction energies were estimated by using the molecular docking program DOT. The top 100,000 DOT results were saved as a starting point for additional analysis.
- Geometric complementarity analysis was performed for the top 1500 DOT results by using the FADE program.
- Proximity analysis was performed whenever it was appropriate. For antibody-antigen systems, the amount of contact with residues in the complement determining region (CDR) was analyzed. Two methods were used for the proximity analysis. The first was a distance cutoff filter, and the second used FADE to determine the amount of contact and shape complementarity between regions of interest. The first filter was applied to the top 100,000 DOT results and second to the top 1500.
- Cluster analysis was performed on the top 1000 results returned by DOT. For targets 4–6, additional clustering was done by using results that passed the strict CDR filtering based on the CDR distance cutoff.
- The results were examined visually to exclude those that are clearly not possible due to unfavorable charge–charge interactions. Those with more favorable interactions, including charge–charge, hydrogen bonding, nonpolar–nonpolar were selected.

This outline will hopefully serve as a central point of reference for the various methods used in our search for good predictions, and we now provide more details on each procedure.

Grant Sponsor: National Science Foundation; Grant number: NSF DBI-9911196; Grant sponsor: U.S. Department of Energy, Grant number: DE-FG03-01ER25497; Grant sponsor: Proctor and Gamble Company.

*Correspondence to: Julie C. Mitchell, San Diego Supercomputer Center, University of California at San Diego, 9500 Gilman Dr. MC0527, La Jolla, CA. 92093-0527. E-mail: mitchell@sdsc.edu

Received 1 November 2002; Accepted 11 December 2002

The DOT Molecular Docking Program

Initial analysis was performed by using the program DOT, which was previously described in detail.[1,3] The DOT energy model consists of an electrostatic component and a nonbonded contact term. DOT calculates the electrostatic energy by convolving the potential field of the first molecule and a rotated charge distribution of the second. This allows rapid determination of billions of possible relative orientations between the two molecules. The nonbonded contact term is computed by counting the number of atoms of one molecule that fall within the "surface layer" of the other molecule, assigning an energy contribution of -0.1 kcal/mol for each atom. The nonbonded contact term is also calculated by using convolution integrals.

Atomic coordinates for the seven targets were made available by the CAPRI management committee. For each CAPRI target, water from both components was removed, polar hydrogen atoms were added to both proteins with Insight II, and partial charges for both proteins were assigned according to an AMBER united atom parameter set. Prosthetic groups were modeled with Insight II. The electrostatic potential grid for all targets was generated by solving the linearized Poisson–Boltzmann equation with UHBD.[4] The potential was evaluated on a $128 \times 128 \times 128$ grid for targets 2, 4, 5, 6, and 7 and on a $256 \times 256 \times 256$ grid for targets 1 and 3. A grid spacing of 1 Å, a solvent dielectric of 78.0, a protein dielectric of 3.0, an ionic radius of 1.4 Å, and a solvent radius of 1.4 Å were used for all systems.

The time to complete a run is proportional to $N(2 \log N + 1)$, where N is the number of grid points.[1] Doubling the grid extent in each of the three coordinate directions (an 8-fold increase in the number of grid points) results in approximately a 10-fold increase in the computational time. By using a 128 Å^3 electrostatic grid, a DOT run can be completed in about 5 h with 40 processors on a Sun Ultra HPC 10000, whereas for a 256 Å^3 grid, it takes about 50 h. The amount of time to complete a run also depends on the computer system used. The Compaq DS20 was about twice as fast per processor as the Sun Ultra HPC, but it has only two processors, whereas the Sun machine has 64.

Shape Complementarity Analysis With FADE

The Fast Atomic Density Evaluator (FADE) is a method for analyzing macromolecular shape and shape complementarity. The radial density computations used by FADE are based on fast Fourier transform methods and convolution integrals.[2] A point that lies inside a crevice, and is thus surrounded by atoms, typically has a higher local density of atomic neighbors than one near a protrusion or flat edge. The atomic density exponent, λ, is defined in such a way that the number of atomic neighbors surrounding a point varies as r^{λ} in the range $0 \leq r \leq 10$ Å. In theory, we expect the median "flat edge" density exponent to be approximately 3.0. In practice, distributions found in proteins return a median density exponent of $\lambda_0 \approx 2.8$. Density exponents that are larger or smaller than λ_0 tend to occur in regions near crevices and protrusions, respectively.

To deduce shape complementarity, the density exponents for each molecule can be combined into a local shape complementarity score. If x is a point in the interface between the two molecules, and λ_1 and λ_2 are the density exponents relative to the first and second molecule, the formula

$$C_i = (\lambda_1 - \lambda_0)\,(\lambda_2 - \lambda_0) \tag{1}$$

gives a local measure of shape complementarity near the point x. If the value of C_i is negative, then one of $(\lambda_1 - \lambda_0)$, $(\lambda_2 - \lambda_0)$ is positive and the other negative, which suggests a crevice-protrusion match. If, on the other hand, the value of C_i is positive, the shape features are the same on both sides of the interface. That is, we have a crevice–crevice or protrusion–protrusion mismatch.

For each of the seven CAPRI targets, we ran FADE on the top 1500 results returned by DOT. FADE returns three valuable pieces of information. The number of interface points indicates how many grid points (at 1 Å spacing) lie within 3 Å of both molecules. Although not identical to a buried surface area calculation, this measure indicates the size of the protein–protein interface. The total complementarity score is the sum of local complementarity scores over all interface points. This measure indicates the level of shape complementarity across the entire interface. The average complementarity score is the total score normalized by the number of interface points. This measure indicates the level of shape complementarity in a way that is independent of the interface size.

Proximity Analysis

Because it is often difficult to solve the protein-docking problem by using only computational methods, filtering on any biochemical data can limit the search space significantly. One approach to the problem is by filtering out results that contradict known behavior. For targets 2–6 we assumed that the CDR should be in contact because this is typically the case for antibody-antigen systems.

Filtering based on CDR contact was done in two ways. One filter used very strict criteria, requiring that all the C_α atoms of the CDR region be within 10–12 Å of an atom in the other protein. A second filter used FADE to analyze shape complementarity between the receptor and the antibody CDR. This filter was useful in estimating the amount of surface contact between the receptor and CDR residues, as well as the level of shape match at the interface. Combined with FADE information for the whole ligand, it was possible to estimate what fraction of the surface contact was centered on the CDR.

Cluster Analysis of Docking Results

The goal of cluster analysis is to reduce the number of results to the point that each cluster is functionally different. The advantage of clustering is that it allows one to group a large collection of results into sets, each of which contains multiple configurations having high similarity to one another and little similarity to configurations in other clusters.

The rigid body transformation from the reference configuration to a transformed configuration can be represented

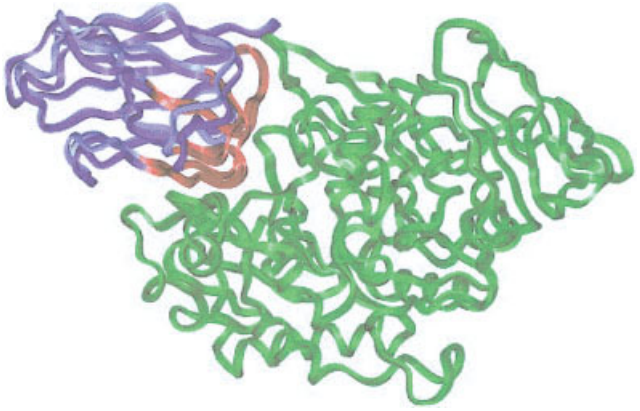

Fig. 1. Prediction 6-2 is shown with the crystal structure for target 6. α-Amylase is shown in green and the camelide antibodies in blue, with the crystal structure given a slightly lighter shade of blue. The antibody CDR regions are shown in red. The two structures are remarkably close, with prediction 6-2 correctly identifying 60 of 64 contacts.

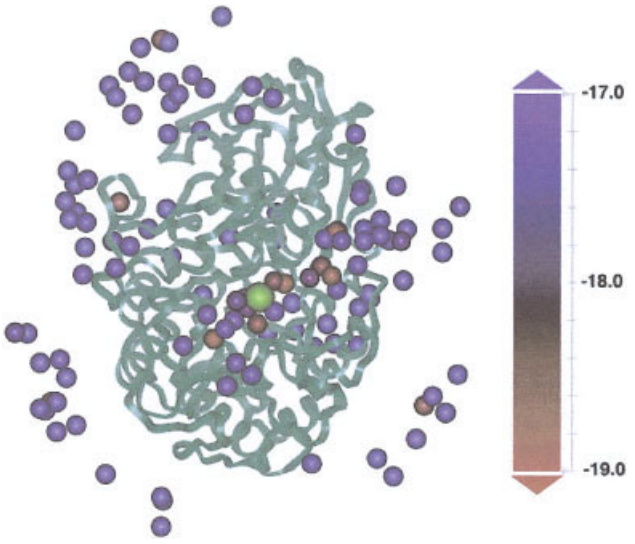

Fig. 2. The spatial distribution of cluster representatives is shown for target 6, relative to α-amylase. Spheres are drawn at the center of geometry of each cluster and colored according to its Boltzmann mean energy, with blue indicating a higher mean energy and red a lower one. The center of geometry for the crystal structure is shown in yellow. The top 1000 DOT configurations resulted in 118 distinct clusters, and clusters with lowest Boltzmann mean energies were concentrated near the correct solution.

in terms of a translation vector and a rotation matrix. First, the rotation matrix is applied to the atoms in one molecule (often called the "moving" molecule); then the atoms are shifted according to the translation vector. This provides a means of positioning the moving molecule relative to the center of coordinates used for the "stationary" one. To cluster configurations, a metric to define distances between them must be used. We define a six-dimensional distance (SDD) between two configurations as

$$SDD = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2 + (r \cdot \theta_{A,B})^2} \quad (2)$$

where (x_a, y_a, z_a) and (x_b, y_b, z_b) are the translation vectors of the two configurations; A and B are the rotation matrices of the two configurations; r is the average distance of surface atoms of the moving molecule from their center of coordinates; and $\theta_{A,B} = \arccos(\frac{1}{2}(\mathrm{Trace}([A]^{-1}[B]) - 1))$. This metric takes into account the difference in centers of mass and the difference in orientations.

For a collection of docking configurations, the pairwise distances (with respect to SDD) between configurations are stored in a matrix. Matrix entries are then set to 0 or 1, according to whether the distance is larger or smaller than a specified threshold. After setting the distance threshold, almost any clustering algorithm will produce satisfactory results. We have used the "maximum clique" method.[5] The advantage of using the maximum clique over other methods is that it requires no prior knowledge of the number of clusters. A clique is a set of interconnected points, and each pair of points in the set is closer than the threshold value. The maximum clique algorithm traverses all possible cliques by using a combinatorial search and chooses the largest one. The method then eliminates elements of that clique from the list of configurations and repeats the process exhaustively.

RESULTS

As is typical, our methods performed well on some systems while results for other systems were less satisfactory. We outline the computational results and analysis on a system by system basis, starting with our most successful efforts. Throughout this section, we refer to our n-th submitted prediction for the m-th system as prediction m–n (e.g., our first prediction for CAPRI target 6 is prediction 6-1.)

Target 6 (α-Amylase-Camelide Antibody-VH Domain 3)

Our best overall prediction in all seven targets was for target 6, where 60 of 65 contacts and an interface RMSD of 1.78 Å were obtained for prediction 6-2. This result had more correct contacts than any other prediction submitted for CAPRI target 6. DOT ranked this result as its top choice. Thus, out of several billion configurations, this result produced a better DOT energy than any other. It is also worth noting that within the top 1500 DOT results for target 6, there were only two results with >75% of contacts correctly predicted, and both of these configurations were submitted among our predictions.

In addition to being DOT's top answer, prediction 6-2 fared well with respect to our filtering criteria. The CDR was assumed to be in contact, and the crystal structure showed this assumption to be correct (Fig. 1). Prediction 6-2 thus passed the strict CDR contact filter, and it returned a good score with respect to FADE's CDR complementarity analysis. Prediction 6-2 was part of the second largest cluster of DOT results for target 6. This configuration was contained in a cluster having 59 members, of which 31 were within $SDD < 10$ Å of the crystal structure. Overall, there were 118 clusters found in the top 1000 DOT results (Fig. 2). The CDR-filtered set contained eight distinct clusters.

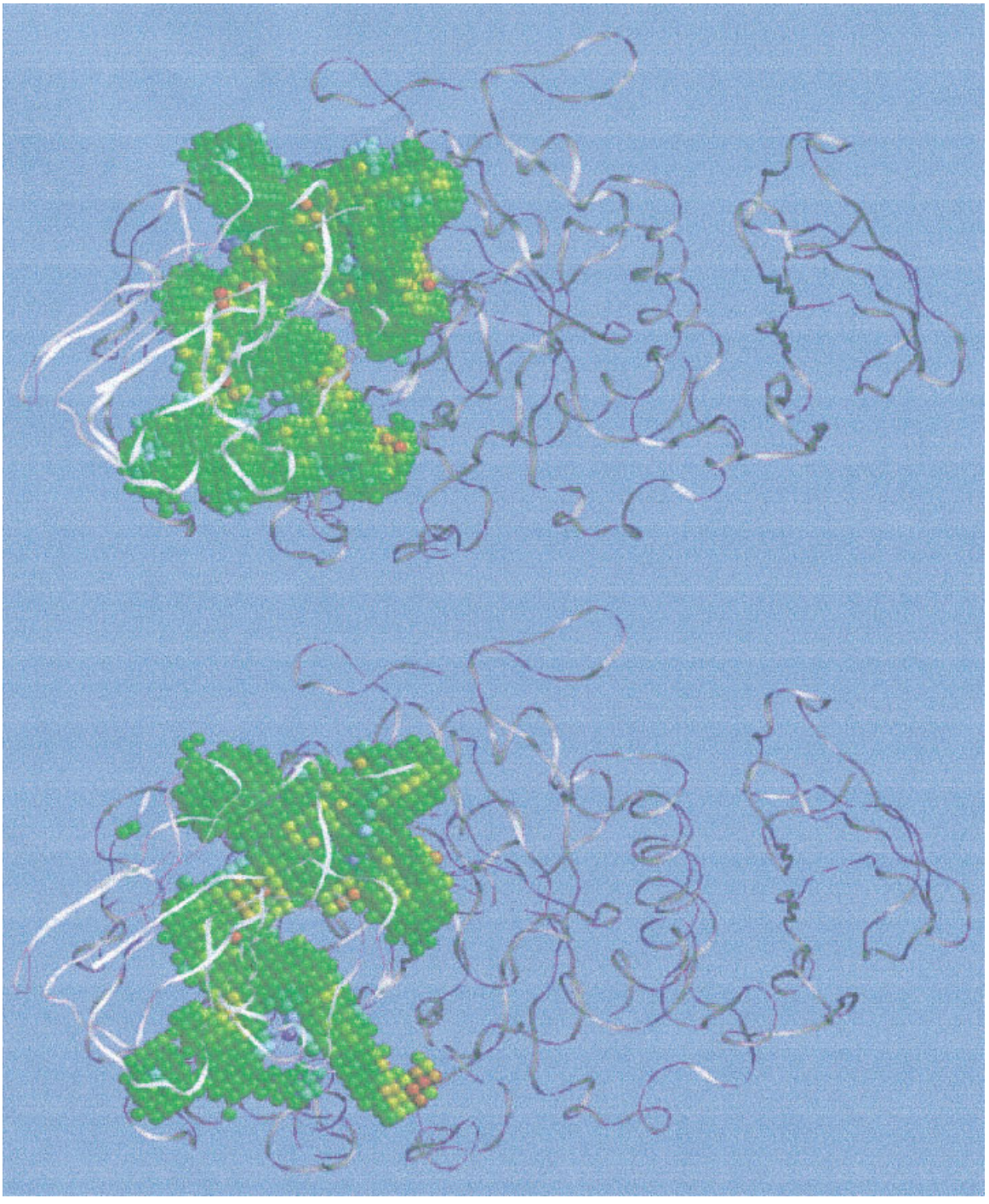

Fig. 3. FADE complementarity data are shown for prediction 6-2 (top) and the target 6 crystal structure (bottom). Points in the interfaces are colored according to their local shape complementarity scores, with red indicating a highly complementary region and blue a mismatch. The complementarity markers (red-orange regions) are similarly located in both cases, suggesting that prediction 6-2 correctly deduced the essential "knobs to holes" matches present in the crystallographic complex.

Shape complementarity for protein–protein interfaces is illustrated in Figure 3. The interface of the target 6 bound crystal structure is shown along with the interface of our best prediction for this system. The prediction has an interface RMSD of <2 Å, and the shape complementarity results are very similar to those of the crystallographic solution.

The number of FADE interface points was higher for prediction 6-2 than for any other result in the top 1500 DOT results, and the total and average complementarity scores were very similar to those obtained for the crystal structure. However, although the shape complementarity data were useful in making predictions, the total and average complementarity scores for prediction 6-2 were

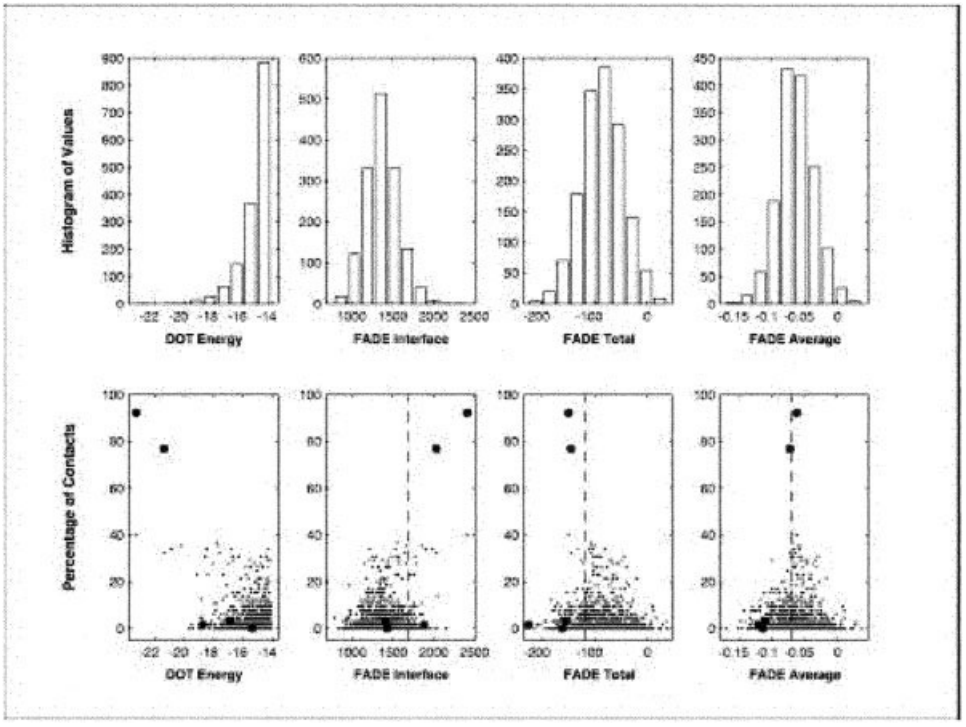

Fig. 4. The top row shows histograms for the following quantities obtained during analysis of the top 1500 DOT results for target 6: DOT energy, number of FADE interface points, total FADE complementarity score, and average FADE complementarity score. The bottom graphs plot these results against the percentage of correct contacts obtained for each configuration. The large dots indicate scores for our target 6 predictions, and the dashed vertical line represents the FADE data obtained for the crystal structure. For the number of FADE interface points, a larger score is regarded as more favorable, whereas a more negative score is considered better for the other quantities. The analysis produced many results with larger interfaces and better complementarity than the crystal structure, and the best DOT results had similar average complementarity scores as the correct solution. Comparable behavior was seen for the other CAPRI systems (not shown.) Thus, it appears that geometric complementarity cannot deduce the best solution, but it is still very good at eliminating many incorrect results.

TABLE I. Results for the Best Prediction[†]

Target	Prediction	Contacts	Int-1	Int-2	Theta angle	Distance (Å)
1	1	6/52	12/27	11/29	66.85	10.91
2	1	20/52	23/27	24/27	45.00	11.34
3	1	2/63	15/32	27/35	63.36	25.23
4	1	0/58	21/27	1/37	146.6	51.09
5	2	10/64	15/29	20/25	76.52	10.80
6	2	60/65	27/29	35/37	9.68	1.78
7	1	0/37	0/21	0/17	43.09	72.25

[†]Measured in terms of correct contacts, the best prediction was submitted for each CAPRI target are shown. The rank order of the submission (column 2); the number of correct residue–residue contacts (column 3), the number of interface residues (columns 4 and 5), the theta angle distance to the crystal structure (column 6), and the distance of the center of coordinates from the crystal position (column 7) are shown. The best overall prediction was for target 6, where 60 of 65 contacts were identified.

not necessarily better than those obtained for other DOT results (Fig. 4). In fact, our prediction 6-1 was chosen as the first submission based on better shape complementarity, when in fact it was less correct.

Target 2 (Bovine Rotavirus VP6-Fab)

Prediction 2-1 correctly identified the interface residues for both molecules (Table I), but this prediction was somewhat translated and rotated relative to the crystallographic complex. Filtering based on shape complementarity and CDR contacts was useful at arriving at this result. However, it was expert analysis that really made the difference in our ability to make a good prediction.

A review of the DOT results indicated that the top ranked structures had significant uncompensated charge that was buried in the interface between the VP6 capsid protein and the antibody. This is electrostatically unfavorable, but DOT does not assign a penalty term for this

situation. By redefining the attractive surface layer of several charged residues to compensate for this error, more biologically realistic results were obtained with DOT.

Target 5 (α-Amylase-Camelide Antibody-VH Domain 2)

Prediction 5-2 had 10 of 64 contacts, which was the most obtained in any of the CAPRI submissions for this target. Although correctly predicting 10 of 64 contacts is not remarkable, this result is notable in having accurately predicted the interface regions (Table I). The prediction was obtained by filtering the top 100,000 DOT configurations. The CDR was partly in contact in target 5; hence, the CDR filtering was useful in highlighting some of the best DOT results.

Of the predictions submitted for target 5, prediction 5-2 had the most shape complementarity, although it was not the most complementarity interface within the top 1500 DOT results. Because there were no near-native structures in the configurations that were clustered, this target did not produce any clusters near the crystallographic solution. Overall, there were 147 clusters found within the top 1000 DOT results. Within the CDR filtered results, two clusters were found.

Target 1 (Lactobacillus HPr Kinase-*B. subtilis* HPr)

In target 1, the correct area was identified as the binding site, but HPr was rotated approximately 180° from the crystallographic orientation. Filtering of results based on the known fact that HPr can be phosphorylated on Ser 46 by HPr kinase-phosphatase[6] was helpful at arriving at this result, but this information was insufficient to yield the crystallographic solution. The top 1500 DOT results yielded only a few configurations with >10% correct contacts, which is likely due to the significant conformational changes seen upon binding for this system.

Target 3 (Flu Hemagglutinin-Fab HC63)

Prediction 3-1 is rotated roughly 180° from the crystallographic solution, but it does a good job at predicting the interface regions of both molecules (Table I). Predictions for target 3 were based primarily on shape complementarity and CDR contact. The crystal structure reveals that two Fab HC63′s bind to the flu hemagglutinin. In our model, there is only room for one Fab HC63 to bind. The possibility of multiple Fab HC63′s binding was not considered, and this may be the reason why the structure submitted was displaced from the crystallographic position.

Target 4 (α-Amylase-Camelide Antibody-VH Domain 1)

The assumption that the CDR was in contact was shown to be false for target 4, and this led to predictions that were far from the crystal structure. Our retrospective analysis indicates that the incorrect biological assumption was very damaging in this case. In particular, the top 1500 DOT results contained structures with nearly 80% of the contacts correctly predicted, and there were four results having >50% correct contacts.

Filtering the top 1000 DOT results resulted in 67 distinct clusters. The second largest cluster was close to the crystallographic position, but it was not considered because of our assumption that the antibody CDR would be in contact with the antigen. Also clustered were the results of filtering the top 100,000 DOT configurations, which produced four clusters having no similarity to the crystallographic solution. Shape complementarity analysis also highlighted several near-native configurations, but these were discarded on the basis of our false assumption of CDR contact.

Target 7 (T-Cell Receptor β-Chain-Streptococcal Pyrogenic Exotoxin A)

No experimental information was used in target 7, and the predictions submitted were far from the crystallographic location. Some good answers were found about 50,000 entries into the DOT results, but this is clear only in hindsight. None of the top 1500 DOT results had >12% correct contacts. The reason for this is as yet unclear. Without good hits from DOT, neither the shape complementarity or cluster analysis yielded good predictions.

DISCUSSION

Need for Flexibility

DOT's energy function is only an approximation, but it is sufficient to produce correct predictions in many cases. The necessity of generating millions or even billions of possible complexes and evaluating each makes it difficult to model the energy function accurately. However, because DOT models molecules as rigid bodies, any conformational changes induced on binding are not accounted for. This can make it difficult to predict the true binding site. Although it is possible to partially address this problem by allowing some interpenetration between molecules, this approach is not sufficient in all cases.

The lack of flexible docking was a clear disadvantage in CAPRI target 1. Because bound structures were provided for the antibodies in targets 2–6, flexibility was less of an issue for these systems. To most accurately predict bound complexes in future CAPRI experiments, better methods for handling flexible molecules will be needed.

Effectiveness of DOT

The CAPRI assessors regarded as notable those docking predictions having at least 30% of residue–residue contacts correct. Table II shows the DOT rank of the result with the highest percent of correct contacts for each of the seven CAPRI targets. For target 6, the best solution was ranked 1st by DOT and had 92% of the contacts correctly predicted. The best result achieved for target 4 had 78% of the contact correct, although it was ranked 746th by DOT. The results for target 2 contained a solution, ranked 465th by DOT, with 42% of contacts correct. Thus, for three systems, DOT's top 1500 solutions contained excellent results, although these were not always among the configurations selected as predictions.

TABLE II. Rank Results from the First 1500 DOT Results[†]

CAPRI system	DOT energy	FADE interface	FADE total	FADE average	Percentage of contacts
1	714	21	20	133	24.56
2	465	370	1162	1177	42.31
3	488	709	228	216	25.40
4	746	205	393	570	77.78
5	161	731	165	132	27.87
6	1	1	82	752	92.31
7	547	479	571	681	10.81

[†]Measured in terms of correct contacts, the best result was analyzed in how it ranked before and after analysis with use of FADE. For each CAPRI target, the rank of this result is given relative to the DOT energy, number of FADE interface points, FADE total complementarity score, and FADE local complementarity score. The percentage of correct contacts is also listed. FADE improved the rankings in three of seven targets, and in two cases the best result ranked similarly for the scoring functions used by DOT and FADE. A lower rank was seen for two systems.

For targets 1, 3, and 5, the best DOT results had ≈25% of correct contacts, ranked 714th, 488th, and 161st, respectively. This is relatively close to the 30% mark, suggesting that minor improvements to DOT could produce hits of this quality. For target 7, the best DOT result within the top 1500 had only 11% correct contacts. It is unclear why this system produced such unsatisfactory results.

Effectiveness of FADE

Exhaustive search methods, such as DOT, typically return a large collection of candidate structures, many of which are false-positive results. Shape complementarity analysis with FADE is one way the top DOT structures were filtered. The goal of this filter is to locate structures with a large binding interface and good geometric match. Although FADE was generally helpful in analyzing candidate configurations, our analysis also indicates that shape complementarity cannot be used as the sole criterion for selecting near-native structures.

The percentage of correct contacts was computed for the first 1500 DOT results of each target. The result with the most correct contacts was analyzed in terms of its ranking for DOT's energy function and the shape information returned by FADE (Table II). A review across all seven systems indicates that the crystal structure and best results did not stand out with respect to any of FADE's shape-based criteria, suggesting that shape cannot be used as the sole filtering criterion. This can be seen in Table II and Figure 4. For targets 1, 3, and 4, the shape complementarity information ranked the best result more favorably than DOT's affinity measure. Targets 5 and 7 had roughly comparable scores when measured by DOT or FADE.

Although shape complementarity analysis was generally helpful, there were two systems for which the shape complementarity information was misleading. With target 6, DOT ranked the best result as its number one pick, whereas FADE scored it less favorably. In target 2, the DOT rank was significantly better than the FADE rank for the best systems. This may be due to the existence of highly complementary candidate configurations with uncompensated buried charge. A detailed model for complementarity that includes electrostatics is necessary to properly identify the correct answer in this case.

Effectiveness of Proximity Filtering

Proximity filtering was useful in cases in which the criteria were appropriate. A large number of results are filtered out, leaving few to consider after clustering. Unfortunately, this filtering also removed the best predictions in some cases because of an overly strict distance cutoff value. In fact, the crystal structures themselves would not have passed the 10–12 Å cutoffs used for the proximity filtering. Selecting antibody configurations based on CDR contacts was useful in targets 2–3 and 5–6. This filter was a detriment in target 4 because the crystal structure shows very little contact between the antibody CDR and the antigen.

Effectiveness of Cluster Analysis

Cluster analysis was only used for targets 4–7, so it is difficult to draw extensive conclusions. The clustering was useful in target 6, where a large cluster with favorable Boltzmann averaged energy existed near the crystallographic solution. The set of configurations clustered for targets 5 and 7 did not contain any configuration near the crystallographic solution, and so the analysis was not helpful in this cases.

For targets 4 and 6, in which the clustered results did contain numerous configurations near the correct solution, cluster analysis was able to clearly identify them. For target 4, the clustering indicated the correct answer, but this result was discarded on the basis of false assumptions about the behavior of antibody-antigen systems. Although firm conclusions cannot be drawn on the basis of these cases, it is clear that cluster analysis provides a useful complement to the other types of analysis used herein. In particular, clustering significantly reduced the number of distinct configurations that had to be analyzed by other means.

CONCLUSIONS

Results for the first CAPRI suggest that the methods described herein are capable of producing competitive

docking predictions. The DOT molecular docking program obtained good hits for most systems within its top 1500 results. The use of cluster analysis, proximity filtering, and shape complementarity significantly improves predictive abilities over DOT alone.

The analysis also suggests that there is room for improvement both in DOT and in the postprocessing used to filter its results. We are currently working on tuning DOT's scoring function so that the crystallographic position appears as one of the energetically best results, while maintaining the amount of time it takes to complete a calculation. Certain parameters in DOT were chosen so that the program is more likely to include false positives than to exclude true positives. This choice was made under the assumption that the correct solutions would stand out from the incorrect ones. The results of the CAPRI experiment indicate that this balance should be better tuned in an effort to reduce the abundance of false-positive solutions.

Shape complementarity combined with other measures is a useful discriminant, despite the fact that shape cannot be used exclusively to deduce the best results. Although we found many interfaces that were more complementary than the crystal structure for each of the CAPRI targets, shape complementarity is useful in eliminating results having an interface that is not well packed.

Clustering results using the SDD metric must be regarded as preliminary because there are some obvious deficiencies in this distance measure. Work is presently being performed with improved measures. In any case, cluster analysis appears to be a promising tool for postprocessing. Cluster analysis will become more valuable as the number of false positives is reduced to the point that more near-native solutions appear in the top 1000 results. Clique finding is an NP-hard problem; hence, it cannot be performed on arbitrarily large data sets.

The proximity analysis is useful in reducing the number of possibilities to be considered. However, in this case, our criteria were much too strict, and the crystal structures did not pass the filtering criteria. The number of retained configurations grows rapidly with increasing cutoff distance. We are presently running tests to determine whether an optimal cutoff exists.

ACKNOWLEDGMENT

The CAPRI blind experiment was extremely valuable to the development of our docking methods. We thank the CAPRI management committee for this opportunity and the assessors for their hard work in evaluating the predictions. We also thank the experimental groups (Refs. 7 and 8 and J. Janin, F. Rey, and M. Knossow) for providing targets for this prediction exercise. DOT is distributed by the Computational Center for Macromolecular Structures (CCMS): http://www.sdsc.edu/CCMS; FADE is distributed jointly by the CCMS and the Molecular Docking and Shape Analysis (MDSA) group at the San Diego Supercomputer Center: http://biology.sdsc.edu/MDSA

REFERENCES

1. Mandell J, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. Protein docking using continuum electrostatics and geometric fit. Protein Eng 2001;14:105–113.
2. Mitchell JC, Kerr R, Ten Eyck LF. Rapid atomic density methods for molecular shape characterization. J Mol Graph Model 2001;19: 324–329.
3. Ten Eyck LF, Mandell J, Roberts VA, Pique ME. Surveying molecular interactions with DOT. In Proceedings of Supercomputing '95. IEEE, 1995. http://www.supercomp.org/sc95/proceedings/636.LTEN/SC95.HTM.
4. Davis ME, Madura JD, Luty BA, McCammon JA. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. Comp Phys Comm 1991;62:187–197.
5. Johnson DS, Trick MA, editors. DIMACS series in discrete mathematics and theoretical computer science, Vol. 26. Providence, RI: American Mathematical Society; 1996.
6. Márquez JA, Hasenbein S, Koch B, Fieulaine S, Nessler S, Russell RB, Hengstenberg W, Scheffzek K. Structure of the full-length HPr kinase/phosphatase from Staphylococcus xylosus at 1.95 Å resolution: mimicking the product/substrate of the phospho transfer reactions. Proc Natl Acad Sci USA 2002;99:3458–3463.
7. Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C. Three camelid VHH domains in complex with porcine pancreatic alpha-amylase. inhibition and versatility of binding topology. J Biol Chem 2002;277:23645–23650.
8. Li H, Sundberg EJ, Mariuzza RA. Crystal structure of two superantigen from Streptococcus pyogenes complexed with their respective T cell receptor ligands reveal novel superantigen-T cell receptor binding modes and a unique T cell signaling complex. Manuscript to be published (PDB Code 1L0X), 2002.