Guided docking: First step to locate potential binding sites

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JULY 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10380 · Source: PubMed

2 AUTHORS, INCLUDING:

Paul A Bates
The Francis Crick Institute
179 PUBLICATIONS **7,305** CITATIONS

SEE PROFILE

PROTEINS: Structure, Function, and Genetics 52:28–32 (2003)

Guided Docking: First Step to Locate Potential Binding Sites

Paul W. Fitzjohn and Paul A. Bates*
Biomolecular Modelling Laboratory, Cancer Research U.K. London Research Institute, London, United Kingdom

ABSTRACT **The long-range electrostatic forces of the targets in round 2 of the Critical Assessment of PRediction of Interactions (CAPRI) experiment were examined and a simple guided docking method, based on these forces, was applied. The method described consists of calculating an initial rigid body trajectory and an optional final, fully flexible refinement stage. Although only limited success was found in predicting the final complexes, some interesting information was discovered. In particular, the long-range forces seem to give some insight into the unusual binding mode of target 4 while raising some questions about target 7, which warrant further investigation. Proteins 2003;52:28–32.**

Key words: long-range electrostatics; CAPRI; fully flexible refinement; protein complex prediction

INTRODUCTION

Many of the current docking algorithms that have been developed conduct rigid body searches of the ligand over the surface of the receptor protein.[1] Scoring schemes are then used to rank these complexes, with the aim of selecting the complex closest to the native structure as the top ranked solution. Most of these scoring schemes use shape complementarity as a component, which is known to be a poor indicator of the correct complex when unbound components are used.[2] The algorithms seek to get around this constraint by softening the potentials used for shape complementarity near the protein's surface[3] and using other information such as electrostatics in the determination of the best complex.[4]

Much work is being devoted to improving the scoring functions used, but the assumption that they can be improved to the required level has not been proved. The softening of potentials to try to take into account flexibility may hinder the goal of consistently discerning correct complexes from many possibilities.

In addition, there is a very real requirement to dock protein models. The amount of data in the Protein Data Bank (PDB),[5] is still small compared with the amount of sequence information available. Results from the EValuation of Automatic protein structure prediction (EVA)[6] suggest that even at high-sequence similarities, current modeling methods will have an average root-mean-square deviation (RMSD) from the native structure of around 2–4 Å. This level of error is likely to cause problems for many current docking methods.

Some useful information may be provided by looking at the pathway by which two proteins form a complex, even when the subunits are distorted from their final bound configurations. Brownian dynamics have been used to look at diffusional association rates of proteins,[7] and some of these ideas have been directly applied to protein–protein binding.[8] Recent work[9,10] has been quite successful in getting close to the native structure from starting positions whose RMSDs are as much as 10 Å away from the native. In this article we look at a simple, long-range, electrostatic guidance method applied to the targets in round 2 of the Critical Assessment of PRediction of Interactions (CAPRI) experiment, which give some insight into the initial phases of their complex formation.

MATERIALS AND METHODS

Preparation of Targets

The targets were first prepared by splitting the receptor and ligand chains into separate files. The chains were then centralized so that their C_α center of mass (COM) was placed at the origin. To ensure the original orientation of the targets would not influence the results, their direction of largest extent was aligned arbitrarily along the *y* axis.

Crystallographic waters were removed, but any ions associated with the proteins were retained. Finally, all hydrogen atoms were added to both chains, ensuring that the N- and C-termini were handled correctly.

Initial Starting Configuration

The ligand was placed into multiple starting positions or orientations around the receptor, which was accomplished by placing the ligand chain at the 12 vertices of an icosahedron centered on the receptor. In addition, at each of these points, the ligand was rotated so that the *y* axis of the ligand was aligned with the vertices of an icosahedron centered around the ligand itself. This resulted in 144 starting positions for each ligand-receptor pair. The size of the icosahedron was chosen so that there were ~100 Å between the two chains. For all of the CAPRI targets, a COM separation of 150 Å was found to be close to this.

*Correspondence to: Paul A. Bates, Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields, London WC2A 3PX, United Kingdom. E-mail: Paul.Bates@cancer.org.uk

Received 14 October 2002; Accepted 11 December 2002

Rigid Body Trajectory Calculation

The next phase of the method involved calculating the rigid body trajectory for each of the initial starting configurations. The calculations were conducted in vacuo by using the CHARMM22 all-atom force field.[11] Initial tests established that the final trajectories from this simple method were not significantly different when using a dielectric constant of 1, or a distance-dependent dielectric equal to the separation distance between the two atoms (results not shown). As a result, our analysis of the CAPRI targets was conducted with a dielectric constant of 1.

The trajectory was calculated by summing the forces on each atom in a chain, due to electrostatic and van der Waals (VDW) interactions with all atoms on the other chain. No distance cutoff was used. These forces were then converted into a force and torque on the COM of that chain. Internal forces were ignored because they result in zero net force or torque on the COM. The chains were then moved according to these forces and torques.

To simplify the analysis of the trajectories, the receptor movements were mapped onto the ligand so that the receptor remained in the same position and orientation at each stage. The whole system was critically damped so the velocity of all chains was taken to be zero at each step. This allowed the force to be converted to a distance by dividing by the mass of the atom and multiplying by a constant. This constant was chosen so that at each step the farthest an atom moved because of a translation or rotation around the COM was no more than 0.1 Å.

A maximum of 500 iterations were performed for each initial starting configuration. The trajectory was stopped early if the ligand had come to rest. The trajectory was saved every five iterations, and a typical trajectory took a few hours to compute on an 866 MHz Pentium III. A cluster of machines was used to compute 12 trajectories in parallel, and this stage typically took just over a day for all the 144 starting configurations of a particular ligand-receptor pair.

Trajectory Analysis

Once the trajectories were computed, the information was collated onto a trajectory map, which showed the COM of the ligand at each stage of the calculation. The energy of the final resting positions of the ligand was calculated by using CHARMM22 and an additional solvation term added, which was dependent on the solvent-accessible area of each non-hydrogen atom.[12]

The trajectories were clustered by hand, and some were selected for further flexible refinement (see below). This clustering was performed on trajectories before the ligand had contacted the receptor. This ensured that the incoming trajectory was analyzed and not the final resting position of the ligand, which was dictated heavily by local VDW interactions.

Flexible Refinement

In some cases, a final flexible refinement stage was conducted. This involved starting the ligand at a point on its rigid body trajectory where separation between the two chains was of the order of 10 Å. The forces on the individual atoms were again calculated but this time including contributions from atoms in the same molecule, and this was used to move the individual atoms, as well as transferring this force to the COM of the chain. The atom moves and COM moves were scaled in such a way that when a COM move resulted in a collision, atom moves became the predominant mode until the clash was resolved.

In the bound-unbound case of targets 4–6, the camelid antibody was kept rigid, and only the amylase was made flexible in this final refinement stage. In the unbound-unbound case of target 7, both chains were made flexible.

Final Results Selection

The best results for each of the targets were selected by using the energy calculation described in trajectory analysis above. This was performed separately for the final rigid trajectory positions and for the flexible refinement.

The flexible refinement method used normally resulted in considerable deformation of the protein chains. This often meant that flexible results had considerably worse internal energies than their rigid counterparts. For the CAPRI targets, no attempt was made to correct for this; however, this dictated that an automatic final submission policy, based on energy alone, could not be used.

The final set of five predictions were manually selected from the best rigid and flexible results and ranked primarily due to energy considerations.

RESULTS AND DISCUSSION

Targets 4–6 (Different Camelid Variable Domains and Porcine Pancreatic α-Amylase)

The first three targets from round 2 of CAPRI consisted of different camelid antibodies docked with the same pig α-amylase. The camelid antibodies coordinates were supplied by the crystallographers[13] via the CAPRI organizers. The pig α-amylase was unbound, and the coordinates used were those from 1PIF.[14] Figure 1(a) shows the trajectory paths for each of the three targets. All three targets can be seen to converge at the top half of the amylase near to the binding site, although the binding funnel is spread out. The effect of starting orientation differs for the three targets, having little effect on the trajectories of target 4 and a large effect for target 5, with target 6 somewhere in between. Our results for targets 4–6 can be seen in Tables I, II, and III, respectively.

The trajectories for targets 5 and 6 naturally resulted in the complementarity determining regions (CDRs) pointing toward the amylase. Our best prediction for target 5 was orientated well but positioned further into the binding groove than the actual, unusual position. Our best prediction for target 6 was in the correct place but orientated slightly incorrectly, not allowing it to fit as snugly into the groove as the final complex.

Target 4 trajectories consistently resulted in the bound β-sheet face, instead of the CDRs of the ligand, pointing toward the amylase. We assumed this was incorrect and restarted each of the trajectory clusters at a point ~10 Å

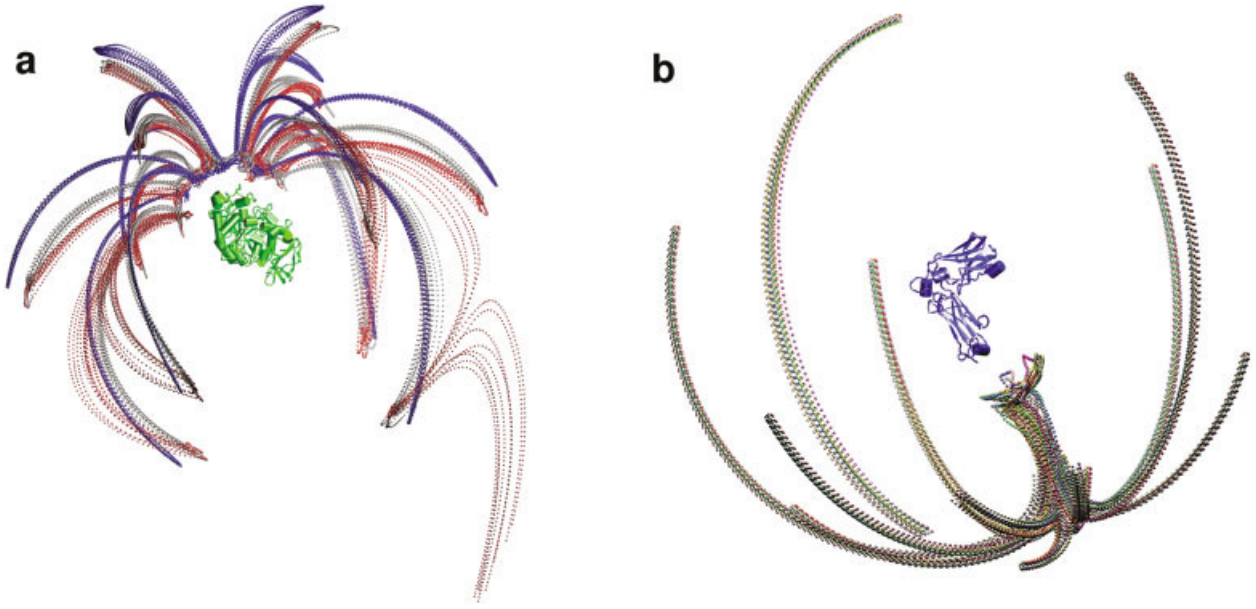

Fig. 1. Center of mass trajectory plots. **a:** The plot for targets 4–6 with the α-amylase shown in a green cartoon representation. Targets 4, 5, and 6 are shown by the blue, red, and gray trajectories, respectively. **b:** The plot for target 7 with the T-cell receptor shown in a blue cartoon representation with the trajectories colored by initial orientation This figure was prepared by using VMD.[21]

TABLE I. Evaluation of Target 4 Predictions

Prediction	Contacts	Interface	Residues	Theta	Distance	CC
1F	0/58	10/27	0/37	150.3	49.2	1
2R	0/58	10/27	0/37	162.4	48.1	1
3F	0/58	11/27	0/37	166.2	46.8	1
4F	0/58	4/27	0/37	116.7	43.6	2
5F	0/58	11/27	0/37	156.0	51.6	1

The first column shows the prediction rank and a single letter code which can be as follows: F, flexible refinement stage used; R, rigid body only; H, superposition on homologous complex. The second, third, and fourth column shows the number of contacts, ligand interface residues, and receptor interface residues correctly predicted. The theta and distance columns show the angle and distance from the correct ligand geometric center. The final column shows the number of close contacts. All results are from the CAPRI evaluators.

TABLE II. Evaluation of Target 5 Predictions

Prediction	Contacts	Interface	Residues	Theta	Distance	CC
1R	0/64	20/29	13/25	21.9	23.3	4
2R	0/64	9/29	0/25	124.9	46.5	1
3R	0/64	6/29	0/25	174.5	43.4	1
4F	0/64	13/29	2/25	74.9	28.9	4
5R	0/64	4/29	0/25	63.4	28.6	2

prior to contact with the amylase, having reoriented the ligand so that the CDRs pointed toward the amylase. The final complex revealed that this was a mistake and target 4 did indeed bind using its β-sheet. Although the trajectories indicated this, they did not identify the correct region of the amylase to which it actually bound.

Target 7 (Streptococcal Pyrogenic Exotoxin A1 and T-Cell Receptor)

The final target was a toxin combined with a T-cell receptor, whose coordinates were taken from PDB entries 1B1Z[15] and 1BEC,[16] respectively. The trajectory map for this target is shown in Figure 1(b) and the results in Table IV.

Our first prediction is very good; however, this consisted of a superposition onto the staphylococcal enterotoxin C3 (SEC3) in 1JCK,[17] whose similarity with streptococcal pyrogenic exotoxin A1 (SPEA) had already been documented[18] with software from our in-house comparative modeling server 3D-JIGSAW.[19]

The other results were predicted by using the trajectory method and are not particularly good. The trajectories for

TABLE III. Evaluation of Target 6 Predictions

Predictions	Contacts	Interface	Residues	Theta	Distance	CC
1R	0/65	14/29	14/37	168.6	10.0	2
2R	0/65	14/29	9/37	170.8	23.5	4
3R	0/65	12/29	14/37	120.3	18.2	2
4F	6/65	15/29	19/37	45.8	7.8	2
5F	0/65	14/29	19/37	130.0	5.1	3

TABLE IV. Evaluation of Target 7 Predictions

Predictions	Contacts	Interface	Residues	Theta	Distance	CC
1H	30/37	20/21	15/17	7.5	2.5	3
2F	0/37	12/21	11/17	43.8	6.5	6
3R	0/37	7/21	3/17	81.4	50.6	2
4R	1/37	12/21	4/17	145.1	22.1	4
5R	3/37	3/21	8/17	76.8	27.4	1

this target were well defined but converged on the wrong end of the T-cell receptor. Because we knew this to be incorrect, we formed our final submissions by starting the toxin within 10 Å of the known binding region. However, the attraction of the binding region was significantly weaker than that from the other end of the receptor, causing the trajectories to be perturbed as they homed in. Once the final complex was available (1L0X[20]), an investigation of the crystal contacts was performed, but this did not reveal any explanation as to why the strong association did not hinder complex formation.

CONCLUSIONS

Long-range electrostatic analysis does seem to provide some useful information. Although its ability to correctly predict complexes is not expected and certainly has not been shown, some of the information gleaned and questions raised by it may prove useful.

In the case of the first three targets considered, the trajectories produced do seem to have the correct orientation for the three antibodies in relation to the amylase. Looking at the nature of the long-range electrostatic field produced by the antibodies reveals them to be extremely polar (results not shown). In the case of target 6, which bound in the expected way, the polar axis runs along the length of the antibody through the CDRs. Target 5 has this axis shifted slightly over to one side, and target 4 has its dipole moment shifted by 90° compared to target 6.

The amylase also exhibited this polar property but was not as strongly defined as the antibodies. However, although it was less specific, it did create a general attraction of the camelid variable domains toward the top half of the amylase, which contains the binding groove. A more accurate study of the electrostatics for the amylase may provide more information. The deep groove along the amylase is likely to have a focusing effect on the electrostatic attraction, when the difference in dielectric between protein and solvent is taken into account.

Significantly less information was discovered for target 7 than for the other three targets. The long-range electrostatic forces seem to hinder the correct association of this complex. An understanding of what is happening in this case may be extremely useful. If the simplistic treatment of electrostatics is to blame, then this affects many docking methods that use fixed or distance-dependent dielectrics as part of their scoring function. The frequency of complexes similar to target 7 and, therefore, the overall problem this may pose is, as far as we know, unknown.

Although the simplistic treatment outlined here has only limited success in predicting complexes, some valuable information is still present. Further work obviously needs to be done, but hopefully such long-range information could be used, either as a possible filter on other methods to reduce false positives or possibly as a method to highlight the binding region, which would allow a more exhaustive local flexible search to be performed.

REFERENCES

1. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
2. Fernandez-Recio J, Totrov M, Abagyan R. Soft protein-protein docking in internal coordinates. Protein Sci 2002;11:280–291.
3. Smith GR, Sternberg MJ. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
4. Camacho CJ, Gatchell DW, Kimura SR, Vajda S. Scoring docked conformations generated by rigid-body protein-protein docking. Proteins 2000;40:525–537.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
6. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
7. Gabdoulline RR, Wade RC. Biomolecular diffusional association. Curr Opin Struct Biol 2002;12:204–213.
8. Camacho CJ, Kimura SR, DeLisi C, Vajda S. Kinetics of desolvation-mediated protein-protein binding. Biophys J 2000;78:1094–1105.
9. Camacho CJ, Vajda S. Protein-protein association kinetics and protein docking. Curr Opin Struct Biol 2002;12:36–40.
10. Camacho CJ, Vajda S. Protein docking along smooth association pathways. Proc Natl Acad Sci USA 2001;98:10636–10641.
11. MacKerell AD Jr, Bashford D, Bellot RL, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Matoos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M,

Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
12. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
13. 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.
14. Machius M, Vertesy L, Huber R, Wiegand G. Carbohydrate and protein-based inhibitors of porcine pancreatic alpha-amylase: structure analysis and comparison of their binding characteristics. J Mol Biol 1996;260:409–421.
15. Papageorgiou AC, Collins CM, Gutman DM, Kline JB, O'Brien SM, Tranter HS, Acharya KR. Structural basis for the recognition of superantigen streptococcal pyrogenic exotoxin A (SpeA1) by MHC class II molecules and T-cell receptors. Embo J 1999;18:9–21.
16. Bentley GA, Boulot G, Karjalainen K, Mariuzza RA. Crystal structure of the beta chain of a T cell antigen receptor. Science 1995;267:1984–1987.
17. Fields BA, Malchiodi EL, Li H, Ysern X, Stauffacher CV, Schlievert PM, Karjalainen K, Mariuzza RA. Crystal structure of a T-cell receptor beta-chain complexed with a superantigen. Nature 1996; 384:188–192.
18. Li H, Llera A, Mariuzza RA. Structure-function studies of T-cell receptor-superantigen interactions. Immunol Rev 1998;163:177–186.
19. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins 2001; Suppl 5:39–46.
20. Sundberg EJ, Li H, Llera AS, McCormick JK, Tormo J, Schlievert PM, Karjalainen K, Mariuzza RA. Structures of two streptococcal superantigens bound to TCR beta chains reveal diversity in the architecture of T cell signaling complexes. Structure (Camb) 2002;10:687–699.
21. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–8, 27–8.