Title:

GAPDOCK: A genetic algorithm approach to protein docking in CAPRI round 1

Authors:

Eleanor J. Gardiner*

Department of Information Studies and Department of Chemistry

Peter Willett

Department of Information Studies

Peter J. Artymiuk

Department of Molecular Biology and Biotechnology

Krebs Institute, Sheffield University, Sheffield S10 2TN, United Kingdom

*All correspondence to be addressed to Dr Eleanor Gardiner, Department of Information Studies, Sheffield University, Regent Court, 221 Portobello Street, Sheffield S1 4DP, United Kingdom.
Phone: 0114 2222674. Fax: 0114 2780300. E-mail e.gardiner@sheffield.ac.uk.

Key words: docking competition; protein-protein interaction; protein-protein docking; shape complementarity; molecular recognition.

Abstract

As part of the first **C**ritical **A**ssessment of **PR**otein **I**nteractions, round 1, we predict the structure of two protein-protein complexes, using a genetic algorithm, GAPDOCK, in combination with surface complementarity, buried surface area, biochemical information and human intervention. Amongst the five models submitted for target 1, HPr phosphocarrier protein (*B. subtilis*) and the hexameric HPr kinase (*L. lactis*), the best correctly predicts 17 / 52 inter-protein contacts, whilst for target 2, bovine rotavirus VP6 protein / monoclonal antibody, the best model predicts 27 / 52 correct contacts. Given the difficult nature of the targets, these predictions are very encouraging and compare well with those obtained by other methods. Nevertheless it is clear that there is a need for improved methods for distinguishing between 'correct' and 'plausible but incorrect' complexes.

Introduction

Protein docking is an extremely complex problem and the inherent difficulty is compounded by the lack of available test systems. Suitable protein complexes, where, ideally, the atomic coordinates of both native proteins and of the complex are deposited in the Protein Data Bank (PDB), are relatively few and are not especially variable. Most docking programs have therefore been developed in tests on two main classes of complex: enzyme/ inhibitor and antibody/antigen complexes. However, in cell biology, there are a wealth of more diverse protein-protein interactions which it is desirable to predict. Docking competitions such as CAPRI are therefore of great importance to the protein docking community, but provide a difficult challenge as the complexes to be predicted may be quite different to those on which the programs have been developed.

There have been two previous docking challenges. In the Alberta docking challenge, six groups correctly predicted the structure of a TEM-1 β-lactamase / β-lactamase inhibitory protein complex[1]. In a CASP protein docking test, no groups were able accurately to predict the structure of a hemagglutinin / antibody complex although a deliberately low resolution approach was able to predict parts of the binding sites[2,3].

The past decade has seen the development of many methods for protein-protein docking[4]. Our program, GAPDOCK, is a genetic algorithm for rigid-body protein-protein docking. In tests using native proteins wherever possible, GAPDOCK was able to generate at least one complex, in a list of 100 complexes, which resembled the crystal complex in 30 of 34 cases. However, for CAPRI, only five submissions are allowed. Our method is therefore to use GAPDOCK to generate a set of solutions and then to use all available biochemical information, in conjunction with other buried surface area and surface complementarity statistics, and also visual inspection, to select a list of five submissions. Here we give a very brief overview, followed by a more detailed description of the exact methods applied to each of the two targets we entered. We then analyse the results for both complexes and discuss the implications.

Materials and Methods

GAPDOCK, which has been described in detail elsewhere[5], is used to generate rotations of the smaller protein relative to the larger protein surface, which is held static. A fitness function counts matches between complementary Connolly surface points[6,7] of the fixed and rotated proteins whilst including a penalty for overlap of the proteins' interiors. Two parameters can be altered, J, a penalty multiplier and N, the angular tolerance on matching surface normals. For each J,N combination we obtain 100 potential solutions.

General selection procedures (applied to both targets).

For each of the targets we ran GAPDOCK 20 times with each of the following parameter combinations:- J=1.5, N=160; J=1, N=165; J=3, N=160; J=3, N=150, these being the parameter combinations which had performed best in our previous tests[5]. Thus we obtained 400 predicted complexes for each target. The disadvantage with this approach is that the GAPDOCK scores obtained using different parameters are not directly comparable. However, we 'normalized' the scores by ordering them, and then scaling them linearly to those from the J=1.5, N=160 complexes. The solutions generated are not usually all distinct and we have clustered the complexes produced, using a simple clustering program written by EJG.

We used a shape correlation program, Sc[8], from the CCP4 program suite[9]. It gives a single number score between 0 and 1 for shape correlation, low being poor. On the basis of previous work in our department (EJG, unpublished results), we rejected those approximately 20% of complexes with Sc score < 0.2.

Our earlier work had suggested that a small buried surface area at the interface also indicates that a complex is unlikely to be correct, but what counts as 'small' is complex-dependent[10]. We therefore rejected complexes whose buried surface area (calculated using the CCP4 program AREAIMOL[9]) seemed lower than the generality of the complexes for that target.

An undesirable feature of the surface matching used by GAPDOCK is that it tends to position proteins too closely together even when producing approximately correct solutions[5]. However, even allowing for this, some predicted complexes were very close. We rejected complexes with more than one Cα-Cα distance closer than 2Å or one Cα-Cα distance closer than 1.3Å.

We also applied biochemical information specific to the target proteins under consideration, as detailed below.

Detailed docking of HPr/HPr Kinase

The HPr kinase is a homo-hexamer, and in the native structure residues 241 – 252 were missing[11]. Our method does not, at present, include any allowance for mobility of side chains. We therefore deleted the neighbouring residues 235-240 and 253, 254 as they had very high B factors and appeared likely to be highly mobile. In addition a large N-terminal domain of HPr kinase was not present in the structure and was therefore necessarily omitted.

We used the structure of HPr from *Streptococcus faecalis* (PDB code 1PTF) rather than that from *B. subtilis* (PDB code 1SPH) because the former has greater sequence homology with the *Lactobacillus casei* HPr, *L. casei* being the source organism of the available HPr kinase structure[11]. We used the entirety of the HPr in docking. For the kinase, although we were interested in a particular site and used it in our screening (see below) we did not wish this to influence the docking procedure. However, as it was pointless to use the entire surface of the homo-hexamer (because of the duplication of surfaces) we, by inspection, selected all surface residues with any atoms with coordinates having $x > 42$, $y > 12$ and $z > 32$ to produce a relatively non-redundant surface.

We wished to provide a variety of solutions. One possibility was that several HPrs could bind at once – if this were to occur then a region at the centre of each trimer would be inaccessible. We therefore also performed docking with the residues previously selected minus all residues within 12Å of residue C269 (approximately the apex of the trimer). Any HPr positioned by these dockings should not prevent two further HPrs from binding. Thus, for the HPr/Kinase complex we actually generated 800 initial solutions, in two sets.

Clustering reduced these to sets of 218 (single HPr allowed) and 188 (multiple HPr allowed) complexes respectively.

Our main selection criterion was biochemical. In their structure report on the native kinase Fieulaine et al. pointed out a similarity between the kinase and adenylate kinase[11]. However a search of the PDB for structural homologues of the kinase using our program PROTEP[12] revealed a large and seemingly more significant area of similarity between the docking target and Phospoenolpyruvate Carboxykinase (PCK, PDB code 1aq2[13]) which was crystallized with ATP bound. As HPr kinase catalyses the ATP-dependent phosphorylation of Ser46 in HPr[11], we therefore modelled ATP and pyruvate molecules into our HPr kinase in the corresponding position and looked for predicted complexes with HPr Ser46 within 5Å of any pyruvate atom or the PG atom of the ATP. We note that the similarity between PCK and the HPr kinase structure has now been independently reported elsewhere[14]. The mean buried surface area for complexes which passed the biochemical selection was 1727Å^2. We rejected complexes with interfaces smaller than this.

The five final submissions were then chosen by visual inspection from amongst the solutions which had fewest inter-protein atomic clashes. The main criterion used was that the orientations of the HPr should be as different as possible in each of the submissions. Models TO1_P27.3.A and TO1_P27.5.A were produced when restricting the binding residues so that multiple HPrs could bind; the three remaining models, TO1_P27.1.A, TO1_P27.2.A and TO1_P27.4.A used all binding site residues. N.B. TO1_P27.1.A etc. are the labels given to our submissions for target one by the competition organisers[15].

Detailed docking of rotavirus VP6/Fab

GAPDOCK has not been designed for, or tested on, docking two such large proteins. We therefore decided to use binding sites in both proteins. This then left the problem of which 'binding sites' to choose. Matthieu et al. (2001)[16] designate three amino acids (A172, C305 and C306), all close together in the 3D structure, as being crucial to the binding of either type I or type II subgroup-specific antibodies. Although

we did not know that the antibody was subgroup-specific, in the absence of other information, we decided to assume that this region was part of the binding site. We therefore selected all residues within 25Å of A172. For the Fab, we first rotated the protein so that its long axis was parallel to the z-axis, then selected All residues with z > 80Å, giving 98 light chain residues in the range 1-103 and 91 heavy chain residues in the range 1-104. After docking and clustering 341 potential complexes remained. We selected all dockings which placed the Fab atoms within 5Å of the virus residues A 172, C 305 or C 306. From these we then chose those dockings which made more than 10 contacts (within 7Å) between the Fab and the virus. Inspection of a large number of protein/antibody complexes in the PDB revealed that, in almost all cases, the antibody is 'end on' to the antigen. Therefore, we next eliminated those dockings which did not place the Fab more or less 'end-on' to the virus. Then, by visual inspection, we selected complexes which buried one of A172, C305 or C306 in the interface. Of these remaining complexes, five were submitted, selected as follows: models TO2_P27.1.HL, TO2_P27.2.HL were selected by visual inspection; TO2_P27.3.HL had the highest normalized GAPDOCK score; TO2_P27.4.HL had high GAPDOCK score, Sc score and buried surface area, and TO2_P27.5.HL had the highest Sc score.

Results

After submission of predictions, the coordinates of the X-ray structures of the HPr/HPr kinase complex [17] and of the Fab/VP6 complex were released. This has permitted retrospective comparisons to be made by the CAPRI organizers[15] whose findings with respect to our submissions we summarize here.

HPr kinase/HPr

Of the five models submitted for this complex, one is significantly better than the rest. Model TO1_P27.5.A has 17/52 correct residue/residue contacts. This model (and also the next best, TO1_P27.3.A with 10/52 correct residue/residue contacts) was generated on the assumption that more than one HPr can bind simultaneously. We note that in the coordinates of the complex[17] an HPr is indeed bound to each kinase subunit. Figure 1 shows the structure of the complex together with TO1_P27.5.A. Serine 46 of

the modeled and crystal HPr's are shown space-filled to illustrate that they are close in 3-D space.

We have also compared our unsubmitted solutions with the correct docking. TO1_P27.5.A has 17/52 correct contacts. We found several solutions with 18 – 20 correct contacts which are clearly only slightly better than our submitted predictions. All our submissions for this target have many very close contacts. These are almost all between the mobile kinase residues (C235-240 and C253, C254) and the HPr. These contacts occur because we generated the dockings after deleting these residues. As we do not have a modeling element as part of our docking program suite, we merely left the residues in their positions in the native kinase rather than guess to which positions they might move. A comparison of the positions of these kinase residues in the crystal complex with those in the native kinase reveals that they do indeed move by up to 8Å in the process of complex formation. Thus, in order for GAPDOCK to perform as it did, their removal prior to docking was certainly necessary.

Deleted: [INSERT FIGURE 1 ABOUT HERE]¶

VP6/Fab

Of the five models we submitted for this target, TO2_P27.3.HL is clearly the best, with 27/52 correct inter-protein residue-residue contacts. Figure 2 shows the Cαs of the crystal structure with this model superposed. Although there are clear differences between the two Fab positions, the similarity is also apparent. It is clear that in the crystal complex, residue A172 of the virus is indeed surrounded by antibody residues, and also that the Fab is very 'end on' to the virus, justifying our selection criteria in this case.

Again we have examined our unsubmitted models to see if we found a better complex which we did not select. One complex had 35/52 correct contacts but was not selected because it has an Sc score of less than 0.2.

Discussion

Participating in CAPRI has been very instructive for us. In our previous docking experiments[5] we considered GAPDOCK to have succeeded in a test if a native-like complex was found in the top 100 solutions. Our most successful previous test results were obtained on enzyme/inhibitor systems, results for antibody / antigen complexes being somewhat less good. For CAPRI we were required to select only five predictions. Nevertheless, in terms of number of correct interactions predicted, our best HPr/HPr kinase prediction was the best submitted in the competition, and our Fab/VP6 prediction was second best. It is nevertheless clear that neither was perfect and that there is much scope for improvement in techniques.

Target 1, HPr /Kinase

Our best submitted result predicted 17 / 52 correct inter-protein contacts. GAPDOCK found a few solutions a little better than this which were eliminated by visual inspection and not by our buried surface area or Sc score criteria. It is not surprising that GAPDOCK was unable to find any complexes which were closer to the crystal structure, as it does not incorporate any mechanism for dealing with large conformational changes, such as the movement of the C-terminal helix.

Biochemical information was crucial to choosing predictions with correct contacts between the kinase and the HPr molecules. We eliminated many incorrect solutions because they did not place Ser 46 of the HPr within 5Å of our modelled ATP/pyruvate moiety. The need for expert human intervention was also demonstrated. All the predictions we generated with 10 or more correct inter-protein contacts were produced after we considered the possibility that multiple HPr's might be able to bind simultaneously.

Target 2, VP6 / Fab

We were fortunate that the binding site which we chose for the virus turned out to be the correct one. It also seems that the requirement the Fab fit squarely onto the antigenic site

is a reasonable one in this case. However, whilst our best submitted model was a reasonable one, we eliminated a much better one which failed our Sc score test.

Subjectivity vs. automation

We have demonstrated both success and failure for both subjective and automatic assessment of potential complexes. For target 1, our very best predictions (which were not *very* good) were de-selected in favour of a slightly poorer one, TO1_P27.5.A, which did not appear worse on inspection. However, subjectivity won out for target 2, where our automatic methods rejected a very good model, and the best prediction we submitted was the best remaining for visual inspection.

Conclusion

It is clear that we need better methods for scoring complexes generated by GAPDOCK. At best, buried surface area and Sc score only serve to reject a fairly small proportion of incorrect complexes, and at worst, they may reject correct ones. At present, the application of biochemical information seems to be by far the best method for choosing complexes. The ultimate goal of fully automated protein-protein docking seems unlikely to be realised in the immediate future.

Acknowledgements

We thank the BBSRC/EPSRC for support of this work and the Royal Society and Wolfson Foundation for provision of computing facilities. We thank Dr K. Linda Britton for useful discussions / suggesting that we should consider the possibility that multiple HPr's could bind simultaneously to one HPr kinase.

References

1 Strynadka NCJ, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F, Olson A, Duncan B, Rao M, Jackson R, Sternberg M, James MNG: Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. Nature Struct Biol 1996;3:233-239.

2 Dixon JS: Evaluation of the CASP2 docking section. Proteins 1997;Supplement 1:198-204.

3 Vakser IA: Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins 1997;Supplement 1:226:230.

4 Smith GR, Sternberg MJE: Prediction of protein-protein interactions by docking methods. Current Opinion in Structural Biology 2002;12:28-35.

5 Gardiner EJ, Willett P, Artymiuk PJ: Protein docking using a genetic algorithm. Proteins 2001;44:44-56.

6 Connolly ML: Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:708-713.

7 Connolly ML: Analytical molecular-surface calculation. J Appl Cryst 1983;16: 548-558.

8 Lawrence MC, Colman PM: Shape complementarity at protein-protein interfaces. J Mol Biol 1993; 234 : 946-950

9 Bailey S: The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr1994; 50: 760-763.

10 Elsom J: An investigation of two screening methods for protein-protein docking. MSc dissertation. University of Sheffield. 1999.

11 Fieulaine S, Morera S, Poncet S, Monedero V, Gueguen-Chaignon V, Galinier A, Janin J, Deutscher J, Nessler S: X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain. Embo J. 2001; 20: 3917-3927.

12 H.M. Grindley, P.J. Artymiuk, D.W. Rice & P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J Mol Biol 1993; 229: 707-721.

13 Tari LW, Matte A, Goldie H, Delbaere LTJ: Mg(2+)-Mn2+ clusters in enzyme-catalyzed phosphoryl-transfer reactions. Nature Structure Biology 1997; 4 : 990-994.

14 Russell RB, Marquez JA, Hengstenberg W, Scheffzek K. Evolutionary relationship between the bacterial HPr kinase and the ubiquitous PEP-carboxykinase. FEBS Letts 2002; 517: 1-6.

15 CAPRI: Critical Assessment of PRediction of Interactions [http://capri.ebi.ac.uk/round1/round1.html]. Site visited 27/08/02.

16 Mathieu M, Petitpas I, Navaza J, Lepault J, Kohli E, Pothier P, Prasad BVV, Cohen J, Rey FA: Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. Embo J. 2001; 20:1485-1497.

17 Fieulaine, S., Morera, S., Poncet, S., Galinier, A., Janin, J., Deutscher, J., Nessler, S.:X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. PNAS 2002; 99:13437-13441.

18 Kraulis PJ. Molscript: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991; 24:946-950.

19 Merritt EA, Murphy MEP. Raster 3D version 2.0: a program for photorealistic molecular graphics, Acta Crystallogr D Biol Crystallogr 1994; 50:869-873.

Figure 1 Target 1

The Cα trace of the crystal complex is shown with that of the predicted HPr from model TO1_P27.5.A superposed. The HPr kinase is shown in red, the crystal HPr in cyan and TO1_P27.5.A in brown. The serine 46 residues of both HPr's are space-filled. All figures have been generated using the software packages Molscript[20] and Raster 3D[21].

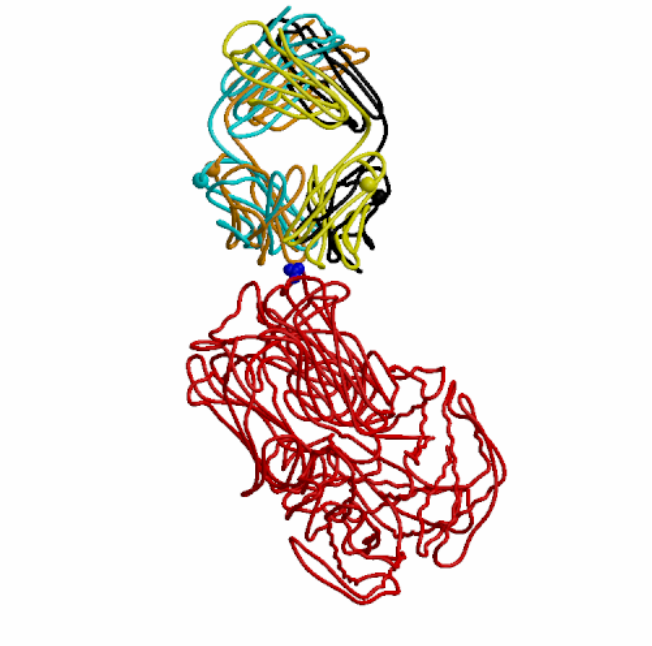

Figure 2 Target 2

The Cα trace of the crystal complex is shown with that of the predicted Fab from model TO2_P27.3.HL superposed. The virus is shown in red. The Fab light chains are shown in cyan (crystal complex) and brown (TO2_P27.3.HL) with the heavy chains in yellow (crystal complex) and black (TO2_P27.3.HL). The virus residue A172 is shown in blue in the centre of the complex interface. The alpha-carbon atoms of residues 17 of both crystal and modelled, light and heavy chains are shown in a space-filling representation to show that the crystal and predicted Fab positions superpose quite well.