Protein docking algorithms: simulating molecular recognition

Jacqueline Cherfils and Joël Janin

Université Paris-Sud, Orsay, France

Docking algorithms simulate protein–protein association in molecular assemblies such as protease–inhibitor or antigen–antibody complexes by reconstituting the complexes from their component molecules. They not only efficiently retrieve native structures but also select a number of non-native structures with structural and physicochemical features that were assumed to be unique to the native complexes. Some of these 'false positives' may deserve further examination in experimental studies of protein–protein recognition.

Current Opinion in Structural Biology 1993, **3**:265–269

Introduction

The docking problem can be formulated thus: given the three-dimensional structure of two molecules, find whether or not they can associate, and predict the structure of the complex. Docking one protein onto another protein, a DNA sequence or a small molecule is a way of simulating recognition, an essential function of proteins. Although the physical principles that govern protein–protein association are probably the same as for other ligands, docking algorithms designed for protein–protein association differ from those intended for small ligands, which are used in drug design (reviewed in [1]). Work on protein docking has been very active recently, with new algorithms and applications to new protein–protein complexes dealing with such important questions as the mechanism of antigen–antibody recognition through the prediction of antigen–antibody complexes.

Docking proteins as rigid bodies

Docking algorithms proceed by bringing two molecules in contact and giving a score to the contact (Fig. 1). In doing so, the molecules are kept rigid and have only six degrees of freedom: three rotations and three translations. Positions with high scores are regarded as candidate complexes and retained for further analysis. The rules for docking are derived from the observation of complexes of known X-ray structure, which comprise mostly protease–inhibitor and antigen–antibody complexes [2]. The area of the protein surface buried in contacts, or interface area, is ~ 1500 Å^2. Atoms are close-packed and water is excluded from this surface. Compared with the rest of the protein surface, interfaces are neither more hydrophobic nor enriched in groups bearing electric charges. In addition to numerous van der Waals contacts and to the stabilizing hydrophobic effect resulting from water exclusion, interfaces contain between eight and 13 hydrogen bonds, which sometimes (but not always) include salt bridges.

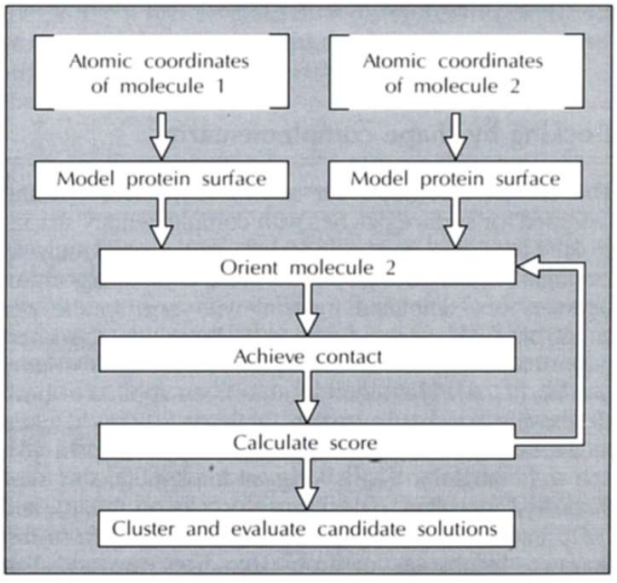

Fig. 1. Flowchart of docking algorithms.

The rigid-body approximation is an important feature of docking algorithms. Internal degrees of freedom can be handled for small molecules, but in proteins, the number is excessively large. This approximation is justified by comparing the X-ray structures of complexes with those of their free components [2]. Associating pro-

Abbreviations

BPTI—bovine pancreatic trypsin inhibitor; **MBP**—maltose-binding protein.

teins are generally observed to behave as rigid bodies. But counter-examples are known. For example, the small protease inhibitor hirudin undergoes a large structural change upon complexation with thrombin [3•]. In other complexes, the polypeptide chain rarely moves by more than 1Å. Larger but localized rearrangements at the protein surface are nevertheless frequent, especially for large flexible amino acid side chains. These movements are accounted for in 'soft docking' algorithms by ignoring or smoothing details of the protein surface and allowing some overlap of atoms at interfaces. Atomic properties can still be re-introduced after candidate complexes have been selected.

Even with rigid bodies, the systematic sampling of all orientations and positions requires a large number of steps. Combining all patches of the surface of one protein molecule with all patches of the second molecule takes of the order of 10^7 trials. Monte Carlo simulated annealing is then often used. Monte Carlo energy minimization proceeds along a random path through parameter space, taking downwards steps in energy. So that the minimization may overcome energy barriers, upwards steps are also allowed, but with a Boltzmann-type probability of $e^{-\Delta E/kT}$, which depends on the energy change ΔE and on a 'temperature' T. In the annealing procedure, T is progressively lowered until the system freezes in an energy minimum [4]. This state should be the global minimum if enough steps were performed, but it is often more practical to explore all deep minima by iterating the whole procedure.

Docking by shape complementarity

The earliest algorithms for automatic protein docking searched for surface patches with complementary shapes [5,6]. The Greer–Bush algorithm was applied only to hemoglobin subunits [6]. The Wodak–Janin algorithm operates on 'simplified protein' with one sphere per amino acid; the score function is based on the interface area, and includes a repulsive term to limit volume overlap [5,7,8]. This algorithm has been applied to both the bovine pancreatic trypsin inhibitor (BPTI), docked in all orientations onto the active site of trypsin [5], and to hemoglobin [7,8]. A recent implementation used simulated annealing to dock three protease–inhibitor and three lysozyme–Fab complexes [9••]. Starting from dissociated complexes or from free lysozyme and Fab HyHEL5, high-score native-like solutions were found, together with many 'false positives' with similar or better scores. When the latter were subjected to conformational energy minimization, up to 12 hydrogen bonds and/or salt bridges formed at the interface, with few buried polar groups remaining unpaired. Electrostatic complementarity was thus achieved in several false positives just as in the native complex.

Similar results have appeared as several different docking algorithms have been tested almost simultaneously on the same systems. The Shoichet–Kuntz algorithm [10••] is adapted from one originally designed for small ligands [11]. A set of spheres fills ridges on the surface of one protein, another set fills grooves on the other protein. Shape complementarity is achieved by matching clusters of spheres in both sets and comparing internal distances. The Jiang–Kim algorithm [12•] models proteins with cubes. For each orientation, it finds translations that bring into coincidence surface cubes from the two molecules, rejects those that create too many volume overlaps, and derives a score from the number of surface overlaps. In both algorithms, native complexes were retrieved from their dissociated components and also from some of the free proteins, together with false positives that could not be eliminated by applying various established criteria, e.g. conformational energies, electrostatic interactions and Eisenberg–MacLachlan solvation free energies [13]. An example is shown in Fig. 2.

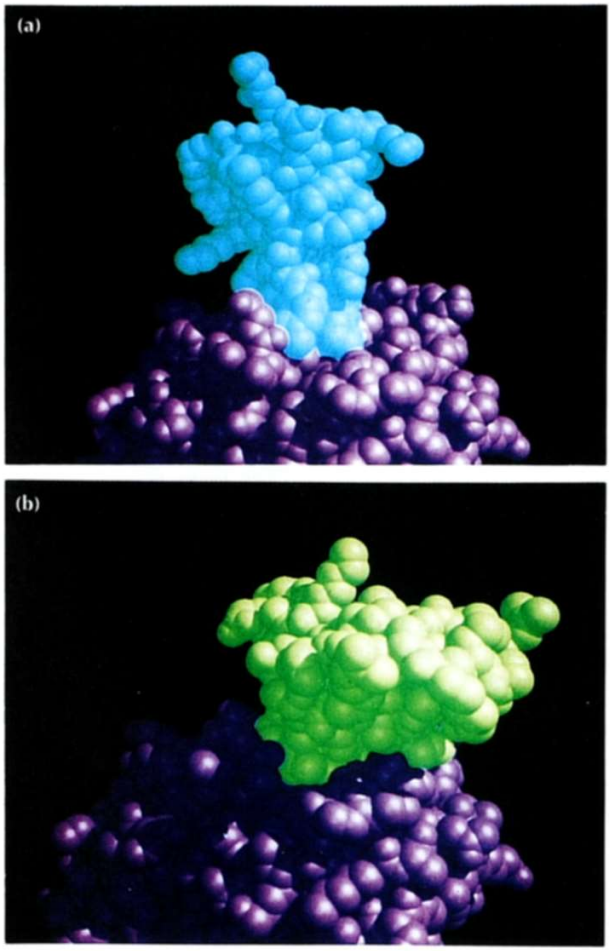

Fig. 2. Two solutions to the trypsin–BPTI docking problem. Docking was performed by superposing a set of spheres filling the inhibitor onto another set, filling grooves on the trypsin surface like a cast. (**a**) A solution almost coincident with the X-ray structure of the complex. (**b**) A 'false positive' with BPTI tilted to 90°; trypsin (bottom) has the same orientation as in (**a**). Both configurations have extensive, highly complementary interfaces. Reproduced with permission from [10••].

The procedures described above [9••,10••,12•] are of comparable computing efficiency: they take a few hours to complete searches with no more than a few million steps, carried out on small computer workstations. It has

been suggested that the translational search on a cubic grid can be accelerated using fast Fourier transform, but with little overall gain [14]. Another approach is to use a much more powerful computer and to apply brute force. A systematic search has been carried on a parallel machine with 4096 processors [15]. As in the early work of Greer and Bush [6], the protein surface was described in reference to a plane; the score function included a soft Lennard–Jones potential. When applied to lysozyme–Fab complexes, the procedure retained a few thousand out of 310 million trials. Native-like solutions were found among the candidates when starting from the dissociated complexes, but not when starting from free lysozyme and a modelled Fab fragment.

Docking by pattern recognition of surface features

The algorithms described above achieve an overall fit by matching patches of protein surface. Other algorithms assume that interaction sites contain patterns of distinctive features, and identify those features on each protein before pairing them. They can be electric charges, as in an early attempt to model the association of two cytochromes [16], or steric features such as 'knobs and holes' [17]. Connolly [18] found that four pairs of knobs and holes uniquely determined the αβ-subunit interface in hemoglobin, but he failed to dock BPTI onto trypsin in this way. An improved procedure based on the same principle retrieved the native trypsin–BPTI complex together with many false positives [19•]. Pairing individual knobs and holes and refining candidate complexes by simulated annealing gave similar results [20].

Other features of interest can be ligand field potentials calculated by placing atoms of different types at grid points on the protein surface and evaluating their energy of interaction with protein atoms [21]. Stored potentials are summed over atoms of a ligand molecule. This procedure has been used to dock small flexible ligands [22], peptides [23] and even proteins [24,25]. A mode of interaction of the *Escherichia coli* maltose-binding protein (MBP) with the aspartate receptor has been predicted by peptide docking [26•]: two nine-residue peptides, selected from mutational studies and docked separately on the receptor, found locations that were compatible with the structure of MBP, thus assigning a position to the whole protein.

Polar interactions in pattern recognition and docking

Electrostatic interactions play a major role in specific recognition, and have provided the basis of several docking algorithms. In Warwicker's procedure [27], electric charges on one molecule move in the field created by the other and the electrostatic energy is calculated. This procedure found the trypsin–BPTI complex and a lysozyme–Fab complex at local energy minima when the dissociated components were used, but failed when starting from free BPTI, presumably because of side-chain movements in lysine and arginine residues. Bacon and Moult [28] also calculated an electrostatic energy of interaction. They retrieved a *Streptomyces* protease–ovomucoid complex starting from both the dissociated and the free components; however, only three degrees of freedom were explored, the position of the interaction region on the protease surface being fixed. With other complexes, the same procedure found non-native solutions with electrostatic energies lower than those of the native complex. Pellegrini and Doniach [29] reconstituted lysozyme–Fab complexes using coarse potentials to select 10 candidate solutions. After minimization, those closest to the native structures had the lowest electrostatic interaction energy, but only 10 000 initial positions were tested, and reconstitution from free lysozyme was not attempted.

As most charged groups are carried by mobile surface side chains, the electrostatic complementarity observed in a complex may not preexist in free proteins. This effect should be less severe for hydrogen bonding. The matching of two sets of hydrogen-bond donors and acceptors has been attempted [30]. Their location on the surface of each protein was described by graphs, and maximal subgraphs were superimposed. Native complexes were retrieved, but a remarkable result of this study is that the problem had several solutions, even though one set of hydrogen-bonding groups was kept fixed, severely restricting the number of possible combinations.

Conclusion

The apparent simplicity of the rigid-body docking is deceiving. Published algorithms reconstitute protease–inhibitor and lysozyme–Fab complexes in reasonable time and with acceptable precision (a few degrees in orientation, 1 or 2 Å in position). Yet the native solution never appears alone when all rigid-body parameters are explored. Moreover, tests performed on the free proteins often fail, though some complexes, that of lysozyme with Fab HEL5 for example, behave consistently better than others. Small conformation changes and side-chain movements, when they do not prevent correct docking altogether, severely diminish the contrast between the native and the false-positive solutions. Relaxing internal degrees of freedom by energy minimization clearly does not solve the problem.

The selection of candidate solutions is the critical step in docking algorithms. It relies on the score function. Those based on a coarse criteria, the interface area, perform no worse than more elaborate ones involving conformational energies or electrostatic interactions. All known complexes have large interfaces, and the selective power of this criterion is illustrated in Fig. 3. In contrast, many false positives achieve as many electrostatic interactions as the native complexes, and their calculated energy

is often lower. This suggests that the physicochemical basis of specificity is still weak. Yet false positives with correct geometry and interactions could be more than just computational artifacts. They could be telling us that the specificity of protein–protein recognition is less absolute than often assumed. We compared false positive trypsin–BPTI complexes produced by different algorithms, and found some that occur repeatedly. Could they be alternative modes of binding? This should be tested in experiments where the native mode is destabilized by site-directed mutagenesis. The selection by monoclonal antibodies of particular epitopes on the lysozyme surface could also be explored and compared with the results of docking experiments.

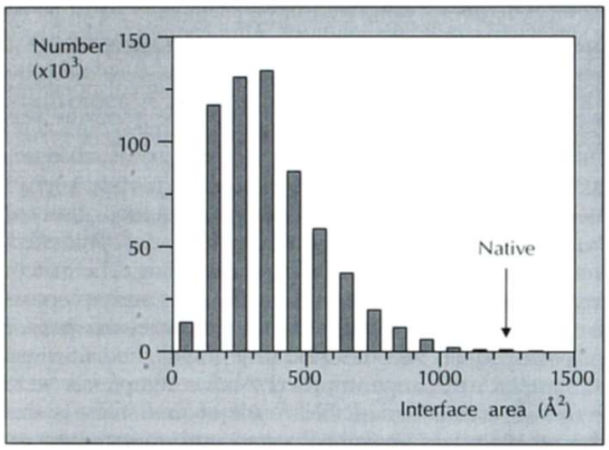

Fig. 3. Interface areas achieved by docking BPTI onto trypsin. BPTI in all orientations is docked near the active site of trypsin using the Wodak–Janin algorithm [6,9••]. The interface area of 620 000 complexes formed in this way is shown in a histogram. With 1250 Å^2, a 'native' orientation very close (five degrees) to that observed in the X-ray structure is among the top 0.1%. Yet 97 artificial complexes achieve interfaces 100–200 Å^2 larger. Note that interface areas calculated on 'simplified protein' models with one sphere per residue are about 10% smaller than with all atoms present.

Unlike the tests described herein, which were performed only to calibrate algorithms, the prediction of an unknown complex will make use of information coming from biochemical or genetic studies. This would reduce the scope of the docking search, and a unique solution may appear. In the best cases, no docking is needed: Blow *et al.* [31] correctly predicted the mode of binding of BPTI to trypsin before the X-ray study of the complex was completed. The recently determined structure [32••] of a complex between cytochrome *c* peroxidase and cytochrome *c* shows that an earlier prediction [33], based on the complementarity of surface charges as in [16], was less successful. External information can be taken into account at various stages of docking. Information on which residue interacts with which, possibly derived from NMR studies, is easily incorporated into harmonic distance constraints added to the score function ('NMR docking' [34]). The power of such constraints, tested on several complexes by simulated annealing [35•], is such that the correct structure is retrieved even when the constraints are approximate, as they will be when attempting a prediction.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. BURT SK, HUTCHINS CW, GREER J: **Predicting Receptor–Ligand Interactions.** *Curr Opin Struct Biol* 1991, 1:213–218.

2. JANIN J, CHOTHIA C: **The Structure of Protein–Protein Recognition Sites.** *J Biol Chem* 1990, 265:16027–16030.

3. • RYDEL TJ, TULINSKY R, BODE W, HUBER R: **Refined Structure of the Hirudin–Thrombin Complex.** *J Mol Biol* 1991, 221:583–601.

Leech hirudin has a flexible carboxy-terminal tail which becomes immobilized in a groove on the surface of the blood protease. A very interesting X-ray structure, relevant here only as a counter-example to rigid-body association.

4. KIRKPATRICK S, GELATT CD JR, VECCHI MP: **Optimization by Simulated Annealing.** *Science* 1983, 220:671–680.

5. WODAK S, JANIN J: **Computer Analysis of Protein–Protein Interactions.** *J Mol Biol* 1978, 124:323–342.

6. GREER J, BUSH BL: **Macromolecular Shape and Surface Maps by Solvent Exclusion.** *Proc Natl Acad Sci U S A* 1978, 75:303–307.

7. JANIN J, WODAK S: **A Reaction Pathway for the Quaternary Structure Change in Hemoglobin.** *Biopolymers* 1985, 24:509–552.

8. WODAK S, DE CROMBRUGGHE M, JANIN J: **Computer Studies of Interactions between Macromolecules.** *Prog Biophys Mol Biol* 1987, 49:29–63.

9. •• CHERFILS J, DUQUERROY S, JANIN J: **Protein–Protein Recognition Analyzed by Docking Simulation.** *Proteins* 1991, 11:271–280.

Protease–inhibitor and lysozyme–Fab complexes are docked by the Wodak–Janin procedure, screened by simulated annealing on the basis of their interface area, and refined for energy. Non-native solutions are found along with the native ones; their significance is discussed.

10. •• SHOICHET BK, KUNTZ ID: **Protein Docking and Complementarity.** *J Mol Biol* 1991, 221:327–246.

A careful and well documented application to protease–inhibitor complexes of a docking algorithm originally designed for small molecules. The discussion of false positives and the final 'reprise' on protein docking are of unusual quality.

11. KUNTZ ID, BLANEY JM, OATLEY SJ, LANDGRIDGE R, FERRIN TE: **A Geometric Approach to Macromolecule–Ligand Interactions.** *J Mol Biol* 1982, 161:269–288.

12. • JIANG F, KIM SH: **"Soft Docking": Matching of Molecular Surface Cubes.** *J Mol Biol* 1991, 219:79–102.

'Soft docking' of complementary surfaces is done by drawing proteins on a cubic grid of adjustable coarseness and superposing the cubes. The method is applied to the trypsin–BPTI and lysozyme Fab HyHEL5 complexes.

13. EISENBERG D, MCLACHLAN AD: **Solvation Energy in Protein Folding and Binding.** *Nature* 1986, 319:199–203.

14. KATCHALSKI-KATZIR E, SHARIV I, EISENSTEIN M, FRIESEM AA, AFLALO C, VAKSER IA: **Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques.** *Proc Natl Acad Sci U S A* 1992, 89:2195–2199.

15. WALLS PH, STERNBERG MJE: **A New Algorithm to Model Protein–Protein Recognition Based on Surface Complementarity: Application to Antibody–Antigen Docking.** *J Mol Biol* 1992, 228:277–297.

16. SALEMME FR: **An Hypothetical Structure for an Intermolecular Electron Transfer Complex of Cytochromes *c* and *b5*.** *J Mol Biol* 1976, 102:563–568.

17. LEE RH, ROSE GD: **Molecular Recognition. Automatic Identification of Topographic Surface Features.** *Biopolymers* 1985, 24:1613–1627.

18. CONNOLLY ML: **Shape Complementarity at the Hemoglobin α1β1 Subunit Interface.** *Biopolymers* 1986, 25:1229–1247.

19. • CONNOLLY ML: **Shape Distribution of Protein Topography.** *Biopolymers* 1992, 32:1215–1236.
A lively presentation of mathematical descriptors of the protein surface and of pattern recognition methods applied to hemoglobin and the trypsin–BPTI complex.

20. WANG H: **Grid-Search Molecular Accessible Surface Algorithm for Solving the Protein Docking Problem.** *J Comput Chem* 1991, 12:746–750.

21. GOODFORD PJ: **A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Molecules.** *J Med Chem* 1985, 28:849–857.

22. GOODSELL DS, OLSON AJ: **Automated Docking of Substrates to Proteins by Simulated Annealing.** *Proteins* 1990, 8:195–202.

23. CAFLISCH A, NIEDERER P, ANLIKER M: **Monte-Carlo Docking of Oligopeptides to Proteins.** *Proteins* 1992, 13:223–230.

24. HART TN, READ RJ: **A Multiple Start Monte-Carlo Docking Method.** *Proteins* 1992, 13:206–222.

25. MENG EC, SHOICHET BK, KUNTZ ID: **Automated Docking with Grid-Based Energy Evaluation.** *J Comput Chem* 1992, 13:505–524.

26. • STODDARD L, KOSHLAND DE JR: **Prediction of the Structure of a Receptor–Protein Complex Using a Binary Docking Method.** *Nature* 1992, 358:774–776.
A bold attempt to predict a complex between a periplasmic *E. coli* protein involved in chemotaxis and its membrane receptor by docking two short peptides. Experimental tests should follow soon.

27. WARWICKER J: **Investigating Protein–Protein Interaction Surfaces Using a Reduced Stereochemical and Electrostatic Model.** *J Mol Biol* 1989, 206:381–395.

28. BACON DJ, MOULT J: **Docking by Least-Squares Fitting of Molecular Surface Patterns.** *J Mol Biol* 1992, 225:849–858.

29. PELLEGRINI M, DONIACH S: **Computer Simulation of Antibody Binding Selectivity.** *Proteins* 1993, in press.

30. KASINOS N, LILLEY GA, SUBBARAO N, HANEEF I: **A Robust and Efficient Automated Docking Algorithm for Molecular Recognition.** *Protein Eng* 1992, 5:69–75.

31. BLOW DM, WRIGHT CA, KUKLA D, RÜHLMANN A, STEIGEMANN W, HUBER R: **A Model for Association of Bovine Pancreatic Trypsin Inhibitor with Chymotrypsin and Trypsin.** *J Mol Biol* 1972, 69:137–144.

32. •• PELLETIER H, KRAUT J: **Crystal Structure of a Complex between Electron Transfer Partners, Cytochrome *c* Peroxidase and Cytochrome *c*.** *Science* 1992, 258:1748–1755.
The X-ray structure of the complex between two heme proteins reveals a possible electron-transfer pathway. A model had been built on the basis of the electrostatic complementarity between the protein surfaces and of biochemical data [33]. It had parallel hemes; in the X-ray structure, the hemes are at an angle of 60°, which may be critical to the mechanism of electron transfer.

33. POULOS TL, KRAUT J: **A Hypothetical Model of the Cytochrome *c* Peroxidase–Cytochrome *c* Electron Transfer Complex.** *J Biol Chem* 1980, 255:10322–10330.

34. WEBER DJ, GITTIS AJ, MULLEN GP, ABEYGUNAWARDANA C, LATTMAN EE, MILDVAN AS: **NMR Docking of a Substrate into the X-Ray Structure of Staphylococcal Nuclease.** *Proteins* 1992, 13:275–287.

35. • YUE S-Y: **Distance Constrained Molecular Docking by Simulated Annealing.** *Protein Eng* 1990, 4:177–184.
Protease–inhibitor complexes are reconstituted by docking under harmonic distance constraints; five to seven distances suffice determining the solution.

J Cherfils and J Janin, Laboratoire de Biologie Structurale, Bât. 433, Université Paris-Sud, 91405-Orsay, France.