J. Mol. Biol. (1991) **219**, 79–102

"Soft Docking": Matching of Molecular Surface Cubes

Fan Jiang[1,3]† and Sung-Hou Kim[2,3]

[1] *Graduate Goup in Biophysics*

[2] *Department of Chemistry*

[3] *Lawrence Berkeley Laboratory*
University of California, Berkeley, CA 94720, U.S.A.

(Received 26 June 1990; accepted 30 November 1990)

Molecular recognition is achieved through the complementarity of molecular surface structures and energetics with, most commonly, associated minor conformational changes. This complementarity can take many forms: charge–charge interaction, hydrogen bonding, van der Waals' interaction, and the size and shape of surfaces. We describe a method that exploits these features to predict the sites of interactions between two cognate molecules given their three-dimensional structures. We have developed a "cube representation" of molecular surface and volume which enables us not only to design a simple algorithm for a six-dimensional search but also to allow implicitly the effects of the conformational changes caused by complex formation. The present molecular docking procedure may be divided into two stages. The first is the selection of a population of complexes by geometric "soft docking", in which surface structures of two interacting molecules are matched with each other, allowing minor conformational changes implicitly, on the basis of complementarity in size and shape, close packing, and the absence of steric hindrance. The second is a screening process to identify a subpopulation with many favorable energetic interactions between the buried surface areas. Once the size of the subpopulation is small, one may further screen to find the correct complex based on other criteria or constraints obtained from biochemical, genetic, and theoretical studies, including visual inspection.

We have tested the present method in two ways. First is a control test in which we docked the components of a molecular complex of known crystal structure available in the Protein Data Bank (PDB). Two molecular complexes were used: (1) a ternary complex of dihydrofolate reductase, NADPH and methotrexate (3DFR in PDB) and (2) a binary complex of trypsin and trypsin inhibitor (2PTC in PDB). The components of each complex were taken apart at an arbitrary relative orientation and then docked together again. The results show that the geometric docking alone is sufficient to determine the correct docking solutions in these ideal cases, and that the cube representation of the molecules does not degrade the docking process in the search for the correct solution.

The second is the more realistic experiment in which we docked the crystal structures of uncomplexed molecules and then compared the structures of docked complexes with the crystal structures of the corresponding complexes. This is to test the capability of our method in accommodating the effects of the conformational changes in the binding sites of the molecules in docking. For this, an uncomplexed trypsin inhibitor (4PTI in PDB) and an uncomplexed trypsin (3PTN in PDB) were used in one case, and an uncomplexed lysozyme (1LYZ in PDB) and the antibody portion of a lysozyme–antibody complex (2HFL in PDB) in the other (the crystal structure of this antibody alone is not known).

Our results verify the importance of both the geometric and energetic complimentarity in docking. In both cases the correct solutions were found among the top 500 solutions (out of over 10,000 to 30,000 solutions) saved from the first stage of geometric docking alone. These solutions were then screened in the second stage of the method, based on favorable and unfavorable energetic interactions. Only a small number of solutions showed overall favorable interactions and the correct docking solution was among them (within the top 4 solutions for 4PTI and 3PTN, and within the top 12 solutions for 1LYZ and 2HFL variable domain). Thus, our "soft docking" procedure was able to reduce drastically the search space resulting in a small number of candidate solutions while tolerating the conformational changes associated with complexing processes in these two tested examples.

Keywords: Molecular docking; soft docking; molecular surface complementarity; cube representation of molecular surface; molecular recognition

† Present address: Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, MA 02138, U.S.A.

0022-2836/91/090079-24 $03.00/0

1. Introduction

Functions of many biological molecules depend on correct molecular recognition among interacting partners. It is now clear that molecular recognition is achieved through the complementarity of molecular surface structures and energetic properties, a fact that has been supported unequivocally by high-resolution X-ray crystallographic structures of several molecular complexes, including antigen–antibody complexes. An integral part of molecular recognition between two molecules is molecular docking, which has been the subject of many studies for more than a decade (Levinthal *et al.*, 1975; Salemme, 1976; Wodak & Janin, 1978; Kuntz *et al.*, 1982; Kuhl *et al.*, 1984; DesJarlais *et al.*, 1986, 1988; Warwicker, 1989). Molecular docking has proved to be a difficult problem, both in terms of developing computational methods and understanding the primary determinants. For example, recent studies show that the surface of a protein is not merely a simple two-dimensional object but has properties that are characteristic of fractal surfaces (Mandelbrot, 1983) with a fractal dimension ranging between 2 and 3 (Lewis & Rees, 1985; Pfeifer *et al.*, 1985; Aqvist & Tapia, 1987). This fractal dimension has been correlated with regions of surfaces involved in complex formation or active sites, and hence may be used in studying molecular interactions or even docking. Despite the apparent difficulties, a good docking procedure may have practical applications in rational drug-design, identification of functionally related surface structures, prediction of molecular complexes, computer simulation of molecular association between substrate and enzyme, and so on.

The docking problem is really a search problem consisting of two steps: first, generating all the potential solutions and, second, eliminating the improbable or incorrect solutions. The variables used for describing the solution space depend on the co-ordinate system in which a molecular surface structure is represented. Two commonly used systems are an internal co-ordinate system and a Cartesian co-ordinate system. In an internal co-ordinate system, molecular surface structures may be represented by graphs (Kuntz *et al.*, 1982; Kuhl *et al.*, 1984), and the docking problem becomes that of finding the optimal match or association between two graphs. Consequently, searching in the corresponding solution space can be achieved with combinatorial algorithms. For example, given a set of criteria defining whether two nodes from two graphs are matched, the optimal docking geometry of two molecules may be derived from the maximum clique found in the matched nodes between the two graphs. This approach has been taken for its computational efficiency. This is true when an existing algorithm can be adapted to the combinatorial search, and when the number of nodes needed to represent a molecule and its surface structure is small, so that the total number of combinations is small. Furthermore, computational efficiency can be improved by designing a good function or criterion that can prune the search trees using only partial solutions, and thus limit the search depth. The combinatorial approach using graph representations has another advantage; namely, it can easily incorporate special features of a surface structure such as charge and hydrophobicity by redefining the properties of graph nodes. However, one of the problems with the combinatorial approach is that, because of the need to maintain computational efficiency, the search depth is limited and there is a great chance of missing potentially correct solutions. Therefore, the needs for computational efficiency and for completeness in searching the solution space have to be delicately balanced. As a result, graph representations in the previous work tend to use oversimplified models for molecular surface structures.

A Cartesian co-ordinate system, on the other hand, is different from an internal co-ordinate system in representing the surface of a molecule and in searching the solution space. Six variables are needed to describe a docking solution, namely, the rigid body rotation and translation, in a Cartesian co-ordinate system. A complete six-dimensional search is usually considered computationally prohibitive, and the early attempts had to use highly simplified models to describe molecular surface structures (Wodak & Janin, 1978). But representing molecular surface structures directly in a Cartesian co-ordinate system has its own advantages: the corresponding physical model is obviously simple and intuitive and, in some aspect, can even be realistic. The search in solution space can be well controlled to be either partial or complete, and the region searched can be easily associated with physical models. Furthermore, the "cube representation" of molecular surface that we have developed can incorporate both the geometric properties and the implicit interaction energies into the determination of the complementarity between two molecular surfaces (see Methods). This type of representation using a Cartesian co-ordinate system, above all, allows direct computer simulation of molecular docking, and therefore provides a useful tool for studying molecular recognition.

After the potential solutions are generated in a chosen solution space in the first stage of our "soft docking" procedure, the goodness of matching between two surface structures is evaluated in the second stage from the energetic viewpoint. This step utilizes a detailed knowledge of the atomic interactions involved in molecular association, much of which has been obtained from examples of the high resolution X-ray crystallographic structures of molecular complexes. The most important fact about the structure of a molecular complex is that it has an interface of (nearly) perfect complementarity. This complementarity can take many forms: charge–charge interaction, hydrogen bonding, van der Waals' interaction and geometric surface complementarity. The interactions listed above can, in principle, be calculated in terms of energies,

provided precise positions of all the atoms are known, by using the empirical force field constants and energy potentials, as implemented in computer program packages such as AMBER (Weiner & Kollman, 1981) and CHARMM (Brooks *et al.*, 1983).

The collective effect of all the interactions determines the size and shape of a surface structure and the volume of a molecule. This effect has been observed and described as the complementarity in size and shape of the interface and the absence of steric hindrance in the interface between the molecules in a complex (Chothia *et al.*, 1985; Schulz & Schimer, 1979; Rebek, 1987). As further shown by the X-ray crystallographic structures of macromolecular complexes such as the complex of lysozyme and antibody (Sheriff *et al.*, 1987), the effect of this type of complementarity results in a complete or almost complete exclusion of solvent molecules from the interface and provides high specificity in molecular recognition. In contrast to the atomic interactions included in the empirical potential energy, there are many difficulties in estimating quantitatively the free energy attributed to the complementarity in size and shape, although this energy has been calculated through the change of entropy generated by the exclusion of solvent molecules upon complex formation (Pratt & Chandler, 1977; Richmond, 1984; Chothia, 1974; Tanford, 1979). Furthermore, this effect is, in general, not very sensitive to the precise positions of individual atoms but is sensitive only to the collective change or movement of the atoms.

Here, we describe a computational procedure ("soft docking") that exploits the geometric and energetic complementarity of molecular surfaces, allowing small conformational changes implicitly, to solve the docking problem without the operator's subjective intervention. In the present method, the surface of a molecule is represented by surface dots with the attached surface normals. Then, the molecule is digitized in a grid space; consequently, the surface dots are converted into surface cubes and the molecular volume into volume cubes. This "cube representation" enables us not only to design a fast algorithm to perform a six-dimensional search of the solution space but also to accommodate implicitly the effects of the conformational changes associated with complex formation. This is because the cube representation effectively "softens" the surface of a molecule; that is, two molecules are allowed to penetrate each other when their surface cubes are matched, and the amount of softening is determined by the cube size. Furthermore, the volume cubes of two molecules are also allowed to overlap to a certain extent to accommodate large conformational changes and/or less than perfect fit due to the approximation in the cube representation and coarse sampling in the rotation space. In addition, the cube representation that we have developed can incorporate both the geometric properties and the interaction energies in evaluating the complementarity between two molecular surfaces (see Methods).

In the first stage of docking, only the geometric complementarity, i.e. the complementarity in size and shape, and the absence of steric hindrance are used in selecting a population of solution candidates. In this stage a complete six-dimensional search is conducted. The solutions from the first stage are then screened in the second stage of docking to identify those with many favorable energetic interactions.

We have tested the present procedure in two ways. The first is a control experiment in which we separated the components of a molecular complex of known crystal structure at an arbitrary relative orientation and then docked them together. We found that geometric docking alone is sufficient to restore the original complex, and that the cube representation of molecules does not degrade the search for the correct solution. The second is a more realistic test in which we docked uncomplexed molecules whose crystal structures have been solved separately. We found that under these more realistic conditions the geometric docking combined with the screening, by scoring the favorable interactions between the buried surfaces in a docked complex, were able to select a subpopulation of a small number of solutions that contains the correct complex.

2. Methods

(a) *Flow chart of the docking procedure*

The soft docking procedure is summarized here in 6 steps:

Step 1: Designate the smaller of the 2 molecules to be the probe and the other the target. The probe is rotated and translated to dock to the target. The centers of mass of both molecules are moved to the origin of the coordinate system.

Step 2: Generate the analytical molecular surfaces of the probe and the target and generate 2 kinds of dot surface, one at a low density (0·3 dots/Å^2; 1 Å = 0·1 nm) and the other at a high density (1·0 dots/Å^2), respectively.

Step 3: Set the docking parameters, which include the cube size and cone angle for testing the geometric local complementarity of the matched surface cubes, the volume cut-off for eliminating bad contacts, and the parameters used in the cluster analysis and the scoring of the total interaction. Sample evenly the entire rotation space.

Step 4: Find the best translation vectors using the surface dots generated at the low density, and then refine them using the surface dots generated at the high density. Repeat the translation search for all the rotations sampled in step 3.

Step 5: Find the clusters of docking solutions in all the solutions saved from step 4. The average rotation and translation are calculated for each cluster found, as well as the average number of matches. Sort the averaged solutions in descending order according to the averaged number of matches.

Step 6: Score the total interaction between the probe and the target surface cubes for each docking solution saved from step 5. This is usually repeated with different cube sizes. Save only those docking solutions that consistently have positive total interactions and therefore are overall favorable.

(b) *Analytical molecular surface and its dot representation*

The solvent accessible surfaces commonly used to represent a molecular surface are defined in 2 ways. One definition was originated by Lee & Richards (1971), implemented by many workers (Richards, 1977; Shrake & Rupley, 1973; Alden & Kim, 1979; Richmond & Richards, 1978; Richmond, 1984) and used in many studies related to molecular surfaces (Chothia *et al.*, 1985; Lewis & Rees, 1985; Miller *et al.*, 1987). In this definition, a solvent accessible surface is the trajectory of the center of a probe sphere rolling around a molecule. This is very much like adding the probe radius to the van der Waals' radius of each atom and then calculating the exposed van der Waals' spheres of all the atoms of the molecule. The other definition (Connolly, 1983*a*) is based on the contact and re-entrant surfaces of a molecule (Richards, 1977). The contact surface is the trajectory of the contact points between a probe sphere and the atoms of a molecule, and the re-entrant surface is a surface patch of the probe where there is no direct contact between the probe and the atoms. The combination of these 2 types of surface results in a smooth molecular surface, which can be represented analytically by patches of convex, concave and saddle surfaces (Connolly, 1983*b*). In our docking procedure, the latter definition is used, as solvent molecules must be excluded from the interface of a molecular complex for it to have close packing and complementarity.

Molecular surfaces are initially represented by uniformly distributed surface dots, each with a surface normal attached. This is done by first calculating the analytical molecular surface (Connolly, 1983*b*), which is a set of surface patches, and then generating dots on each surface patch at a given dot density. The advantage of using the analytical molecular surface is that the dot representation can be calculated at different dot densities easily. The dots on a surface patch can also be positioned to achieve optimal uniformity on the basis of the size and shape of the patch given by the analytical parameters, and the number of dots required can be calculated from the given dot density and the area of that patch. Surface patches with very small areas are ignored to decrease the total number of dots, and thereby the computing time in docking.

(c) *A cube representation of molecular surface in a grid space*

The dot representation of a molecular surface is further simplified with a "cube representation" in a grid space. In this cube representation, the position of a dot is replaced by 3 integer co-ordinates corresponding to the center of the cube enclosing the dot in an equally spaced grid space, while the surface normal remains attached to the dot. So the dots in the same cube with the same integer co-ordinates are still distinguished from each other by their surface normals. In practice, the grid size was chosen so that on the average a "surface cube" may be shared by 2 or 3 dots, where a surface cube is defined as a cube that contains surface dots. At the same time, cubes that do not contain any surface dots and are inside the surface of a molecule are defined as the "volume cubes" of that molecule. Volume cubes will be used to calculate the amount of overlap between 2 molecules.

For each rotation, the probe surface dots and atoms are rotated about the center of mass (the origin of the co-ordinate system) and then converted to surface and volume cubes, respectively. Surface and volume cubes are mutually exclusive and, when they coincide with each other, the former override the latter. In Fig. 1(a), a probe surface cube with 2 dots, after they have been rotated, is matched with a target surface cube with 3 dots.

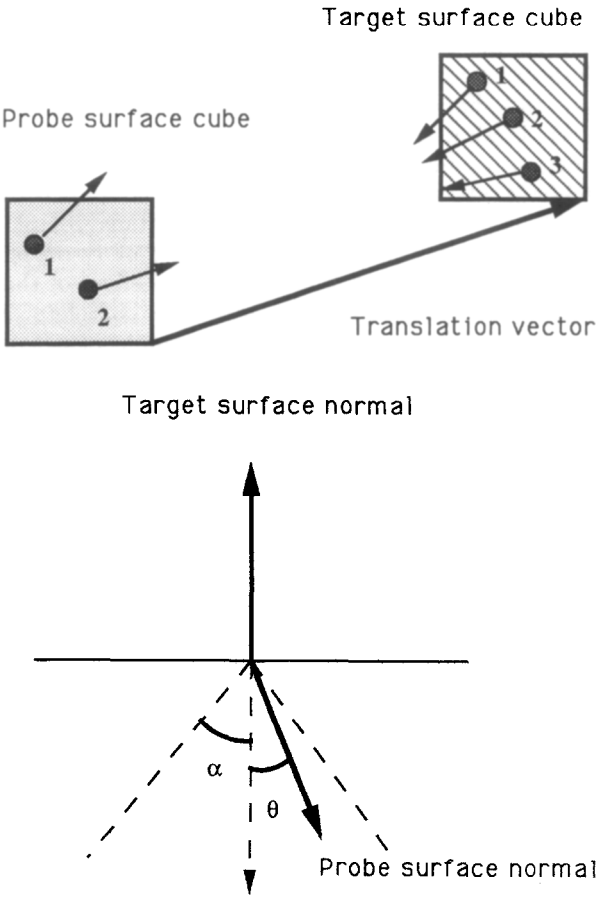


Figure 1. Matching of a pair of probe and target surface cubes. (a) A probe surface cube containing 2 surface dots and a target surface cube containing 3 surface dots are shown; the arrows attached to the dots are their local surface normals. Here surface cubes are shown as squares, i.e. the projection of the surface cubes in a 3-dimensional co-ordinate system to the x–y plane for the purpose of illustration. If the probe surface cube is translated by the vector depicted as a long heavy arrow, it will coincide with the target surface cube. The number of matches for this pair of surface cubes is then determined by how the probe surface normals match with the target surface normals. If all pairs of matched surface normals should satisfy the condition to be defined in (b), the number of matches would be 6, the maximum possible for this pair of surface cubes. (b) The local complementarity of matched surface cubes is ensured by requiring that the matched surface normals be approximately opposite to each other. A quantitative definition of this condition is that the angle θ, defined as the angle between a probe surface normal and the inverse of a matched target surface normal, must be no greater than a given cone angle α. If this condition is applied to the match in (a) with the cone angle shown in (b), out of all 6 pairs of matched surface normals, the pair between dot 1 in the probe surface cube and dot 3 in the target surface cube is the only pair that does not satisfy the condition. Therefore, the total number of matches between this pair of surface cubes for the translation vector shown in (a) is 5.

(d) *Rotation search*

Since the molecule used in our docking procedure is represented in a Cartesian co-ordinate system, an exhaustive search of the solution space means a complete sampling of the rotation and translation space. This is accomplished by fixing one molecule as the target while applying rotations and translations to the other molecule, i.e. the probe. So, the probe is rotated by all the possible rotations, and then, after each rotation, it is translated to match the target to find the best docking.

In sampling the rotation space, polar angles are used as the variables. It is important to sample the rotation space as uniformly as possible, not only to reduce the total number of rotations that have to be searched but also to make it possible to cluster the docking solutions in rotation space as well as in translation space. To ensure the uniformity, i.e. to maintain a constant sampling density, we need to define an "angle distance" similar to that (Euclidean distance) in translation space because the sampling with equal grid spacing of the polar angles will not result in a uniform distribution of rotations. The angle distance, χ, between 2 rigid body rotations R_1 and R_2 is given by the following formula:

$$1 + 2\cos\chi = \text{Trace } R_1 R_2^{-1}, \quad \text{where } 0 < \chi < 180°.$$

All the rotations to be searched can be generated by starting from an identity rotation, in which $\chi = 0$, and gradually moving to rotations with larger and positive χ angles. The number of rotations to be sampled becomes larger and larger as the angle χ increases. To avoid undersampling the rotation space, an angle distance cut-off is used to determine whether two rotations sampled are neighbors: they are neighbors if the angle distance between them is not greater than the angle distance cut-off. Each rotation sampled must have at least 4 neighbors; thus, the total number of rotations sampled is determined by this cut-off angle. Table 1 shows an example of such a sampling of the rotation space, in which the angle distance cut-off used is 9°.

(e) *Translation search*

In the translation search, the program examines only the translation vectors that produce at least 1 pair of matched surface cubes, by calculating the difference vectors between the probe and the target surface cubes and the corresponding number of dots in matched cubes.

Table 1
Sampling of the rotation space

No. of χ (deg.)	Rotations[a]	No. of χ (deg.)	Rotations
9	14	108	1843
18	85	117	2018
27	187	126	2227
36	290	135	2446
45	417	144	2674
54	566	153	2820
63	788	162	3064
72	1003	171	3284
90	1454		
99	1651	Total	28,043

[a] The number of sampling in the rotation space for the angle distance cut-off of 9°.

After sorting the translation solutions according to the number of matches, the program counts the number of overlapping volume cubes starting from the top solutions. The goodness of a translation solution is the number of surface cube matches minus the number of volume cube overlaps.

A match between a probe surface cube and a target surface cube must satisfy 2 conditions. One is that the probe surface cube, after being translated, must overlap with the target surface cube. The other is that the probe surface normals must be approximately in the opposite directions of the target surface normals for the matched dots between the 2 surface cubes, which ensures the local complementarity of the matched surfaces. To be exact, whether 2 surface normals are in opposite directions is determined by an input cone angle, as shown in Fig. 1(b). The number of matches for a pair of probe and target surface cubes is the number of matched probe and target dots that satisfy the 2 conditions.

An example of how the number of matches is determined for a pair of matched surface cubes is shown in Fig. 1, in which a probe surface cube with 2 dots will be overlapped with a target surface cube with 3 dots (if the probe surface cube is translated by the vector depicted as a big heavy arrow). The number of matches for this pair of surface cubes, if there were no condition on surface normals, would be 6 (the total number of pairs of probe and target surface dots that can be matched). To satisfy the 2nd condition for a match, the surface normal of a probe surface dot must lie within the cone defined by the inverse of the surface normal of a target surface dot paired with the probe surface dot, as shown in Fig. 1(b). If the cone angle in Fig. 1(b) is used to define the cone, the pair between the probe surface dot 1 and the target surface dot 3 in Fig. 1(a) does not satisfy the condition on the surface normals, while the remaining 5 pairs do satisfy the condition. Therefore, the number of matches between this pair of surface cubes for the given translation vector as shown in Fig. 1(a) should be 5, not 6.

In Fig. 1(a), it may be noted that the translation vector applied to the probe surface cube is actually equal to the difference vector between the positions of the 2 surface cubes. This suggests the following algorithm for calculating the matches for all translation vectors. First, the difference vectors are calculated for all possible pairs between surface cubes of the probe and those of the target. Then, the matches between probe cubes and target cubes are counted after translating the probe by each difference vector. Next, all the distinct difference vectors along with their numbers of matches are stored in a 3-dimensional array mapping the translation space. This algorithm is more efficient than an algorithm that applies translation vectors to the probe surface cubes and then checks to see if they match with any target surface cube.

The result of the above translation algorithm is a 3-dimensional array with non-zero elements corresponding to the translation vectors with a positive number of matches. The 3-dimensional array is then searched for peaks, local maxima of the numbers of matches, to find the best translation vectors. There are 26 translation vectors in the nearest neighbor of each translation vector, from which they are different by ± 1 grid along any of the 3 axes of the co-ordinate system. The number of matches of a translation vector is a local maximum if it is equal to or greater than the numbers of matches of all its neighbors. All the translation vectors corresponding to local maxima are saved and sorted in descending order according to the number of matches. Then, for each translation vector starting from the top of

the sorted list of translation vectors, the number of overlaps between the probe and target volume cubes is calculated by applying the translation vector to the probe volume cubes and checking to see if they coincide with any target volume cubes. Overlaps between surface and volume cubes do not count. If the number of overlapping volume cubes for a translation vector is greater than a given cut-off value, V_l or V_h in Table 2, the translation vector will be discarded from the list. This is to exclude solutions that result in large bad contacts but have not been removed by the conditions on matching surface normals.

(f) *Two stages of translation search*

The translation search is carried out in 2 stages, each using a different set of parameters: dot density, grid size, cone angle and volume overlap cut-off. In the 1st stage, surface dots are generated at a low density and digitized in a large grid space. The resulting best translation vectors are refined locally in the 2nd stage by using surface dots generated at a high density and digitized in a small grid size. The rest of the relevant parameters for each stage are listed in Table 2. We found that when the dot density is below 0·3 dots/Å^2, the total number of dots generated for a molecular surface remains approximately constant, which means that at this density the number of surface dots is equal to the number of surface patches, as defined in the analytical molecular surface. The cube representation at this dot density seems to be optimal for initial screening of the docking solutions because not only does it preserve the main features of a molecular surface but it also reduces the number of dots to a minimum possible. We found that a high density of 1·0 dots/Å^2 is ideal for generating the surface dots for the high-density translation search. For docking the components of a molecular complex of known crystal structure, as will be shown in Results, the correct solution after this high-density translation search has the highest number of matches that is clearly distinguishable from that of the incorrect solutions (see Table 5). For docking the crystal structures of uncomplexed molecules, the correct solution is among the top solutions; but many incorrect solutions have matches that are not distinguishable from the correct solutions. Nevertheless, it is necessary to refine the translation vectors in the 2nd stage for them to be used in the scoring of the total interaction between the molecules in a docked complex.

(g) *Cluster analysis of docking solutions*

The translation vectors found in all the rotations sampled may be clustered; that is, a group of docking solutions may be close to each other in both rotation and translation space. Clusters are determined by applying cut-off values for the translation vector and rotation angle distances. Only the clusters with enough number of solutions, more than a cut-off value (N_c in Table 2), are saved. For each cluster, the translations and rotations, as well as the number of matches, are averaged, and the averaged rotation and translation are used as the solution for that cluster. This averaging has proved to give more reliable results because it can reveal some of the incorrect docking solutions that accidentally have an isolated high number of matches.

(h) *Scoring of favorable and unfavorable interactions*

So far, we have described how the present procedure uses the complementarity in size and shape of molecular surfaces, close packing, and the absence of steric hindrance in the interface to evaluate the goodness of a geometric docking solution as the 1st stage of the method. For docking the crystal structures of uncomplexed molecules, in which conformational changes at the binding sites are significant, the energetic complementarity provides an additional constraint for screening the solution population. To eliminate further the incorrect solutions and enhance the signal of the correct solutions, the goodness of a docking solution is estimated by scoring the favorable and unfavorable interactions between the

Table 2
Parameters used in docking

Complex (Target/Probe)	Low-density dots[a] C_l (Å)	ϑ_l (deg.)	V_l (cubes)	High-density dots[b] C_h (Å)	ϑ_h (deg.)	V_h (cubes)	N_s[c]	N_{ang}[d]	ϑ_c[e] (deg.)	N_c[f]	T_{dist}[g] (Å)	ϑ_{dist}[h] (deg.)	C_r[i] (Å)
DHFR/NADPH	2·8	50	80	1·4	30	150	10	8280	15				
DHFR/MTX	2·8	50	50	1·4	30	120	10	8280	15				
(DHFR+NADPH)/MTX													
2PTCE/2PTCI	3·2	60	80	2·0	30	220	8	8789	12				
3PTN/4PTI	3·2	60	80	2·0	30	220	8	28043	9	4	2·0	12·0	1·6
2HFLV/1LYZ	3·2	60	80	2·0	30	220	10	14974	10	5	2·0	12·0	1·6

[a] C_l, ϑ_l and V_l are, respectively, the cube size, the cone angle and the volume overlap cut-off used in the 1st stage of translation search with the low-density dot representation of molecular surface.

[b] C_h, ϑ_h and V_h are, respectively, the cube size, the cone angle and the volume overlap cut-off used in the 2nd step of translation search with the high-density dot representation of molecular surface.

[c] N_s is the number of solutions saved for each rotation. This number does not have to be too large to save the correct solutions.

[d] N_{ang} is the total number of rotations sampled in the rotation space.

[e] ϑ_c is the angle distance cut-off used to determine the sampling density in the rotation space so that each rotation has at least 4 neighbors.

[f] N_c is the minimum number of solutions that a cluster must have for it to be saved in the cluster analysis.

[g] T_{dist} is the cut-off distance in the translation space used in the cluster analysis.

[h] ϑ_{dist} is the angle distance cut-off in the rotation space used in the cluster analysis.

[i] C_r is the cube size used in calculating the total interaction surface. The translation vector space was surveyed to find all the local maxima for the total interaction.

Table 3

Favorable and unfavorable interactions classified by atom types

	+	−	H-donor	H-acceptor	Polar	Hydrophobic
+	U	F	U	F	F	U
−	F	U	F	U	F	U
H-donor	U	F	U	F	F	U
H-acceptor	F	U	F	U	F	U
Polar	F	F	F	F	F	U
Hydrophobic	U	U	U	U	U	F

Atoms or atomic groups of the 20 amino acids are classified into 6 types: (1) positively charged (+) NH_2^+ in Lys and NH_3^+ in Arg; (2) negatively charged (−) COO^- oxygen atoms in Asp and Glu; (3) hydrogen donors (H-donor) e.g. NH in main chain and His, and NH_2 in Arg, Asn and Gln; (4) hydrogen acceptors (H-acceptor) e.g. CO oxygen atoms in main chain (peptide), Asn and Gln; (5) polar groups (which can act as both hydrogen donor and acceptor) e.g. OH in Ser, Thr and Tyr, and NH (H-donor when protonated and H-acceptor when deprotonated) in His; (6) hydrophobic atoms (which include the rest of the atoms) e.g. C in main chain, and CH, CH_2 and CH_3 in all side-chains. Fs and Us represent interactions that are favorable and unfavorable, respectively, and each pair of atom types is assigned either a favorable or an unfavorable interaction.

buried surface areas of the probe and the target in the interface of the corresponding docked complex.

The interactions are calculated only within each pair of matched surface cubes between the probe and the target. The type of interaction, either favorable or unfavorable, is determined by the types of the atoms that contribute the surface areas involved. A simple classification is used in this study, as shown in Table 3. There are 6 types of atom or atomic group: positively charged (+), negatively charged (−), hydrogen donor, hydrogen acceptor, polar (which can be either hydrogen donor or acceptor) and hydrophobic. This classification of interaction types is based on an intuitive knowledge about the atomic interactions. For example, the interaction between hydrophobic and hydrophobic surfaces should be favorable, as well as that between hydrogen donors and acceptors. On the other hand, the interaction between polar and hydrophobic surfaces is unfavorable, and more so between the surface areas of atoms of like charge.

Within a pair of matched surface cubes, the dots from the probe and target are paired and each pair of dots is assumed to interact with each other. The magnitude of the interaction is defined, for simplicity, as the sum of the surface areas represented by the paired dots, and for each sum a sign is assigned according to the list in Table 3, positive for favorable interaction pairs and negative for unfavorable ones. The energetic interaction for a pair of matched surface cubes is, therefore, the algebraic sum of interaction surface areas represented by all combinations between the probe and the target surface dots in that pair of matched surface cubes. The total interaction is then the signed sum of the surface areas represented by dots in all the matched surface cubes.

We found that the total interaction scored with this method was able to eliminate most of the incorrect solutions. We also tried several different cube sizes to score the total interaction and select only those solutions whose total interactions are consistently positive. Whenever the cube size is changed, the translation vector of a docking solution is refined locally, moving fewer than 2 steps in any direction, to optimize the total interaction.

(i) *Testing of "soft docking" procedure*

The present docking procedure has been tested in 2 ways. All the molecules used in testing are crystal structures extracted from the Protein Data Bank (PDB†; Bernstein *et al.*, 1977). In the 1st way of testing, we separated the components of a molecular complex of known crystal structure, reoriented them, and used our procedure to dock them back together, and then compared the docked complex with the crystal structure of the complex. The molecular complexes used are: (1) ternary complex (3DFR in PDB; Matthews *et al.*, 1978) of dihydrofolate reductase (DHFR in PDB) with methotrexate (MTX in PDB) and NADPH, and (2) a trypsin–trypsin inhibitor complex (2PTC in PDB; Marquart *et al.*, 1983). In the 2nd way of testing, we docked the crystal structures of uncomplexed molecules, and then compared the docked complex with the crystal structure of the corresponding complex found in the PDB. The crystal structures of uncomplexed molecules used are: (1) trypsin (3PTN in PDB; Marquart *et al.*, 1983) and trypsin inhibitor (4PTI in PDB; Wlodawer *et al.*, 1987), and (2) hen egg-white lysozyme (1LYZ in PDB; Johnson & Phillips, 1965; Diamond, 1974; Levitt, 1974). The crystal structure of a lysozyme–antibody complex (2HFL in PDB; Sheriff *et al.*, 1987) was also used to check our results.

The probe radii used for generating the analytical molecular surfaces of the molecules used in testing are shown in Table 4. The low and high densities used for generating surface dots are 0·3 and 1·0 dots/Å^2, respectively. The corresponding numbers of surface dots are also

† Abbreviations used: PDB, Protein Data Bank; r.m.s., root-mean-square; c.p.u., central processor unit.

The abbreviations for the crystal structures of the molecules (in parentheses): dihydrofolate reductase (DHFR); reduced nicotinamide adenine dinucleotide phosphate (NADPH); methotrexate (MTX); a trypsin–trypsin inhibitor complex (2PTC); the trypsin component (2PTCE) of 2PTC; and the trypsin inhibitor component (2PTCI) of 2PTC; free trypsin inhibitor (4PTI); free trypsin (3PTN); free lysozyme (1LYZ); a lysozyme–antibody complex (2HFL); the antibody component used in docking (2HFLV), which is the variable domain including residues L1 to L110 and H1 to H115. These structures are extracted from the Protein Data Bank (Bernstein *et al.*, 1977). See Results for references to individual structures.

Table 4

Parameters used in the calculation of molecular surfaces and dots

Molecule names	Number of atoms	Probe radius[a] (Å)	Number of dots (low)[b]	Number of dots (high)[c]
DHFR	1293	1·7	2534	6540
NADPH	48	1·7	190	491
MTX	33	1·7	135	350
2PTCE	1629	2·0	3248	7745
2PTCI	454	2·0	1203	2665
3PTN	1628	2·0	2469	7087
4PTI	454	2·0	1013	2796
2HFLV	1748	2·0	3074	9078
1LYZ	1001	2·0	1799	5126

[a] The probe radii used in calculating the analytical molecular surfaces.
[b] The number of surface dots generated at the low density.
[c] The number of surface dots generated at the high density.

shown in Table 4. The parameters used in docking are listed in Table 2 for each pair of the docked molecules.

3. Results

(a) *Docking of the components of a molecular complex*

The results of docking the components of molecular complexes are shown in Table 5. For each docking experiment, the first entry, indicated by Expt., is the rotation and translation derived from the crystal structure of the corresponding complex to be compared with the docking solutions. In column 2, Solution no. gives where a solution is ranked after the geometric docking. Solution nos. 1 and 2 correspond to the best and the second best solutions from the geometric docking, respectively. The average number of matches and the standard deviation for all the docking solutions saved from the geometric docking are, respectively, 74 and 23 for NADPH to DHFR, 78 and 20 for MTX to (NADPH + DHFR), 76 and 23 for MTX to DHFR. The best solution of the first docking, NADPH to DHFR, is practically identical with the correct solution within the errors of the representation, i.e. the angle distance for sampling the rotation space and the cube size for digitizing the translation space. The second best solution has a matching score substantially lower (by $3{\cdot}4\sigma$) than that of the first, and thus, is easily distinguishable. Similar results were found in docking MTX to DHFR with NADPH bound as in the crystal structure of the complex. The best solution corresponds to the correct solution and has a matching score much

Table 5

Docking of the components of molecular complexes

Complex name[a]	Solution no.[b]	Rotation[c] (deg.)			Translation[d] (grids)			Number of matches[e]	Overlap[f] (cubes)	S/N[g] (σ)
DHFR/NADPH	Expt.[h]	104	46	88	7	1	−3	365	52	
	1	107	47	85	7	1	−3	357	100	8·0
	2	286	66	135	0	6	−3	255	75	4·6
(DHFR+NADPH)/ MTX	Expt.	77	43	162	−2	6	−1	360	35	
	1	77	47	165	−2	6	−1	338	37	11·2
	2	251	133	175	−2	6	−1	267	42	7·4
DHFR/MTX	1	123	66	155	9	0	−2	259	71	4·9
	2	265	85	125	5	2	−4	251	74	4·4
2PTCE/2PTCI	Expt.	260	167	102	1	11	−5	623	29	
	1	280	163	99	1	11	−5	575	35	6·2
	2	167	116	99	0	−1	11	444	98	2·4

[a] The complex names are given as Target/Probe, e.g. DHFR/NADPH.
[b] The number in this column gives the ranking of the solutions after the geometric docking, namely the position of the solution in the sorted list of docking solutions saved from the geometric docking.
[c] The rotations applied to the probe are given in polar angles.
[d] The translation vectors are in grids. The grid size is the same as the smaller cube size used at the 2nd step of translation search, in which the translation vectors from the geometric docking were refined locally.
[e] The number of matched cubes listed in this column is after the local translational refinement.
[f] The number of overlapping volume cubes listed here is also after the local translational refinement.
[g] The signal-to-noise ratio of a docking solution from the geometric docking.
[h] The Expt. entry gives the experimentally observed solution derived from the crystal structure of the corresponding molecular complex.

higher (by 3·8σ) than that of the next best solution. The results of docking MTX to DHFR without NADPH bound are quite interesting. Neither the best nor the second best solution corresponds to the correct solution. Furthermore, their matching scores are similar and close to the matching scores of the second best solutions in the previous two docking experiments. This may be explained by the fact that there is a large contact area between MTX and NADPH in the ternary complex structure of DHFR/NADPH/MTX. The loss of this contact area due to the absence of NADPH in docking MTX to DHFR decreases the matching score for the correct solution to such a level that it is no longer distinguishable from the incorrect solutions. Nevertheless, this result is consistent with the observation that there is a co-operative binding of MTX and NADPH to DHFR (Williams *et al.*, 1979; Birdsall *et al.*, 1978).

The results of docking 2PTCI to 2PTCE are shown as the last test in Table 5. The average number of matches and the standard deviation for all the docking solutions are 222 and 51, respectively. The rotation and translation of the best solution is clearly close to that of the correct solution, and the best matching score is substantially higher (by 3·8σ) than that of the next best solution. The rotation of the best docking solution is different from the correct rotation by about ten degrees of rotation, which is less than the angle distance cut-off used for sampling the rotation space. These solutions can be easily refined to produce the corresponding crystal structure.

(b) *Docking of a free trypsin inhibitor and a free trypsin*

The ability of the present docking procedure to deal with conformational changes in the interface upon complex formation is tested by docking the crystal structures of uncomplexed molecules. The crystal structures of an uncomplexed trypsin (3PTN) and an uncomplexed trypsin inhibitor (4PTI) have been solved separately in different crystal forms (Marquart *et al.*, 1983; Wlodawer *et al.*, 1987). A comparison of their structures with the corresponding components of a complex, for example, 2PTC, has indicated that significant conformational changes have occurred (Marquart *et al.*, 1983). The change in trypsin is relatively small; the root-mean-square (r.m.s.) distance between the superimposed, complexed and uncomplexed, trypsin structures, for the 241 atoms in the receptor site (defined by including the atoms within 7 Å of the bound trypsin inhibitor), is 0·68 Å. This is the same as the r.m.s. distance for all the atoms in trypsin. A superposition of the receptor site atoms in trypsin is shown in Figure 2(a), in which there is no significant conformational change visible. However, there are significant conformational changes in trypsin inhibitors. The r.m.s. distance between the superimposed, complexed and uncomplexed, trypsin inhibitors, is 1·70 Å for all 454 atoms, and 1·80 Å for the 107 atoms in the binding site (defined by including the atoms within 7 Å of the bound trypsin). Most of the conformational changes found in the trypsin inhibitors are located on the surface, including residues around the N and C termini. But the most interesting changes are in the binding loops, which include residues Pro13 to Ile19 and Gly37 to Arg39. As shown in Figure 2(b) by superimposing the binding site atoms of the two structures, the C^{α}–C^{β} dihedral angles of both Lys15 and Arg17 have clearly adopted different positions, respectively, and Arg39 has shifted its side-chain atoms considerably.

The parameters used for generating molecular surfaces and surface dots and for docking are the same for docking 4PTI to 3PTN as those for docking 2PTCI to 2PTCE in the previous test, as listed in Tables 2 and 4. Thus, the results from these two tests may be compared to see the effects of conformational changes and test the sensitivity of the parameters used.

The results of docking are shown in Table 6, in which nine docking solutions are listed along with the correct solution in the Expt. entry. The solutions in the Table are sorted in descending order according to the numbers of matches, and Solution no. in the second column of Table 6 represents the position of a solution in this sorted list of solutions from the geometric docking alone. The average number of matches and the standard deviation for all the docking solutions from the geometric docking are 321 and 43, respectively. The first point to notice is that the best solution is no longer significantly better than the rest of the solutions, and it does not correspond to the correct solution. After the cluster analysis, 200 solutions were selected for the calculation of the total interaction in the second stage of the procedure. The calculation was repeated with different cube sizes and only nine solutions were found to have total interactions that are consistently favorable; they are labeled I1 to I9. Solution no. 51 (corresponding to I6) in Table 6 is very close to the correct solution, different from each other by about 15 degrees of rotation. The correct rotation and translation (the Expt. entry in Table 6) was obtained by first superimposing 4PTI over 2PTCI and 3PTN over 2PTCE, respectively, and then, calculating the difference transformation. Thus, of over 8700 possible interaction complexes, this procedure could select nine complexes, of which one was correct within the resolution used.

To see to what the remaining eight solutions correspond, we have visually inspected the complexes derived from solutions I1 to I9 using interactive computer graphics and found that most of the incorrect solutions correspond to a match between a patch of relatively smooth surface of 4PTI to a similarly smooth surface patch of 3PTN. Almost all the incorrectly docked complexes involve, in the interface, the same region of the surface formed by the two β-strands in 4PTI, namely residues Ala16 to Ala25 and Gly28 to Gly36, whereas the contact regions on the surface of 3PTN scatter over many different locations, as long as they are relatively flat and smooth. To illustrate

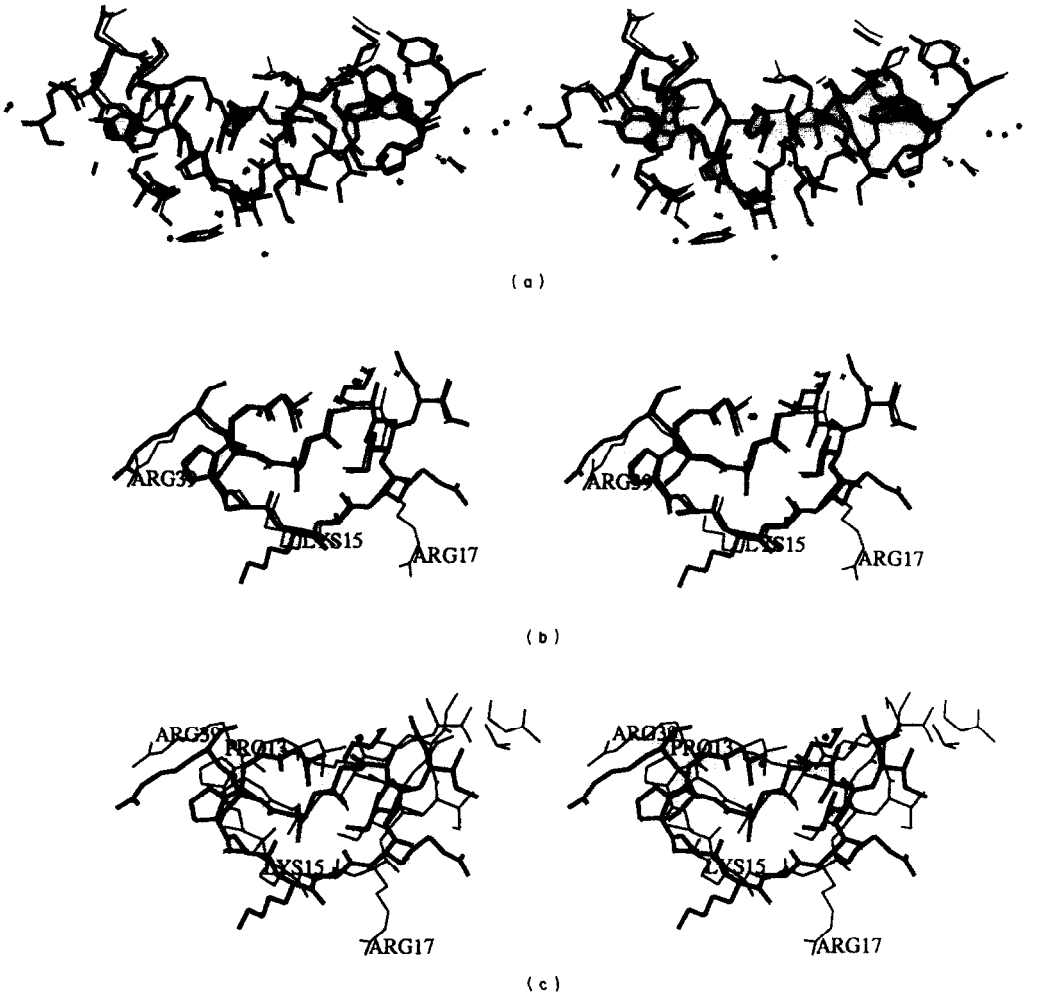


Figure 2. (a) Comparison of the contacting atoms of trypsin in complexed and uncomplexed states in stereo. The receptor site atoms from 3PTN (Marquart *et al.*, 1983), an uncomplexed trypsin, shown in thick lines, are superimposed over those from 2PTCE (Marquart *et al.*, 1983), a complexed trypsin, in thin lines. The atoms included in the Figure are within 7 Å of the trypsin inhibitor (2PTCI) in the complex (2PTC) and the resulting r.m.s. distance is 0·68 Å. One can see that the 2 structures are almost identical and there are no significant changes in the shape of the surfaces. (b) Comparison of the contacting atoms of trypsin inhibitor in complexed and uncomplexed states. The binding site atoms (within 7 Å of trypsin) from 4PTI (Wlodawer *et al.*, 1987) of an uncomplexed trypsin inhibitor, in thick lines, are superimposed over those from 2PTCI (Marquart *et al.*, 1983), in thin lines; and the r.m.s. distance for these atoms is 1·80 Å. It is not difficult to see that the conformational changes are on the surface: the most significant changes have occurred in residues Lys15, Arg17 and Arg39, resulting in a dramatic change in the shape of the surface in contact with trypsin. In Arg39, the change is due to the large movement of the side-chain atoms, while in Lys15 and Arg19, the changes are caused by the 2 residues adopting different C^{α}–C^{β} dihedral angles in the complexed and uncomplexed structures. (c) The superposition of the binding site atoms of the docked trypsin inhibitor and the complexed trypsin inhibitor. The trypsin inhibitor (2PTCI) in the complex (2PTC) is shown in thin lines in the same view as in (b). The docked model, shown in thick lines, was obtained by applying the rotation from I6 (solution no. 51) in Table 6 to 4PTI and then superimposing the center of mass of the thick line atoms (of 4PTI) over that of the thin line atoms (of 2PTCI). The resulting r.m.s. distance is 2·56 Å.

this, the C^{α} atom plots of the docked complexes based on solutions I1 to I9, as well as the crystal structure 2PTC, are shown in Figure 3. To show how the atoms are matched, we applied the rotation from the correct docking solution (solution no. 51 in Table 6) to 4PTI and then superimposed the center of mass of the binding site atoms of 4PTI over that of 2PTCI, as shown in Figure 2(c). The resulting r.m.s. distance is 2·56 Å, which is reasonably good considering that the r.m.s. distance derived from least-square fitting is 1·80 Å. Further adjustment will be needed to this model before it can be used as a starting model for energy minimization or molecular dynamics refinement. The large r.m.s. value is caused mainly by the coarse sampling of the rotation space. To show how the surfaces match each other in the interface of the crystal structure, the derived complex from least-square fitting, and the docked complex, similar sections of the matched surfaces are shown in Figure 4 for comparison.

Table 6

Docking of a free trypsin inhibitor to a free trypsin

Complex name	Solution no.	Rotation (deg.)			Translation (grids)			Number of matches	Overlap (cubes)	Total interaction[a]
3PTN/4PTI	Expt.	258	56	101	6	14	4	391	58	82
I1	5	236	109	171	1	14	6	519	10	34
I2	12	347	79	162	5	−12	−1	489	56	110
I3	22	306	34	171	−9	−11	−7	460	9	104
I4	36	5	70	126	5	−9	10	436	100	135
I5	37	342	142	108	11	10	−1	435	12	51
I6*	51	256	46	99	5	13	2	421	54	55
I7	52	22	79	72	3	−16	−1	419	29	47
I8	60	112	81	153	−4	10	13	408	36	8
I9	65	15	146	171	13	−7	4	402	21	43

[a] The total interaction corresponds to the sum of all the areas of favorable interactions minus that of unfavorable interactions in all the matched surface cubes. The units are in Å^2 rounded off to nearest integers. See Methods, section (h), for detailed descriptions.

The other columns are similar to their counterparts in Table 5. 3PTN and 4PTI are the crystal structures of free trypsin and free trypsin inhibitor, respectively. Solution no. 51 (I6 in Fig. 3), marked with an asterisk in the 1st column, is the correct solution, and is very close to the rotation and translation derived from the superposition of 4PTI over 2PTCI (the typsin inhibitor in the complex structure 2PTC) by least-squares fitting, which is given in the Expt. entry.

(c) *Docking of a free lysozyme and a complexed lysozyme antibody*

As another example of the second way of testing, we docked the crystal structure of an uncomplexed hen egg-white lysozyme (Johnson & Phillips, 1965; Levitt, 1974; Williams *et al.*, 1979) to the antibody portion of a lysozyme–antibody complex (2HFL), which had been recently solved (Sheriff *et al.*, 1987) because the crystal structure of the corresponding free antibody is not known. An analysis of several crystal structures of antigen–antibody complexes (Davies *et al.*, 1988, 1990) suggests that the conformational changes due to the binding of lysozyme to

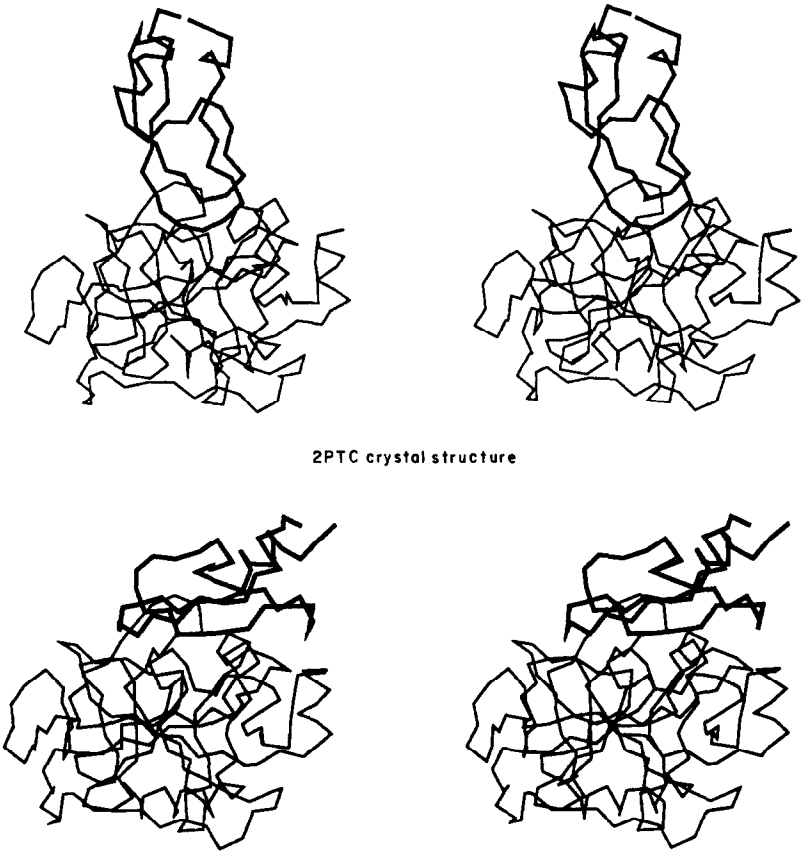
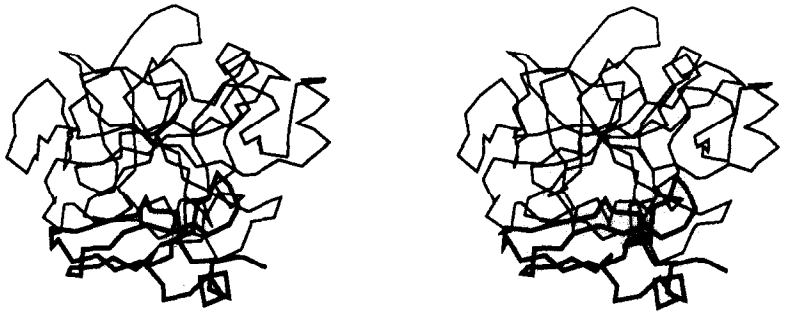
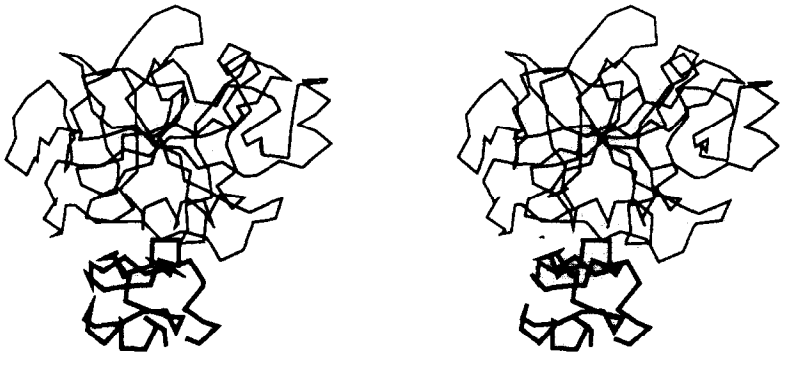


Fig. 3.

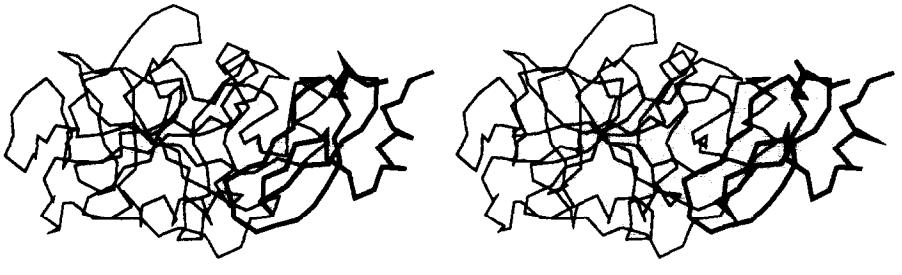


I2

I3



I4



I5

Fig. 3.

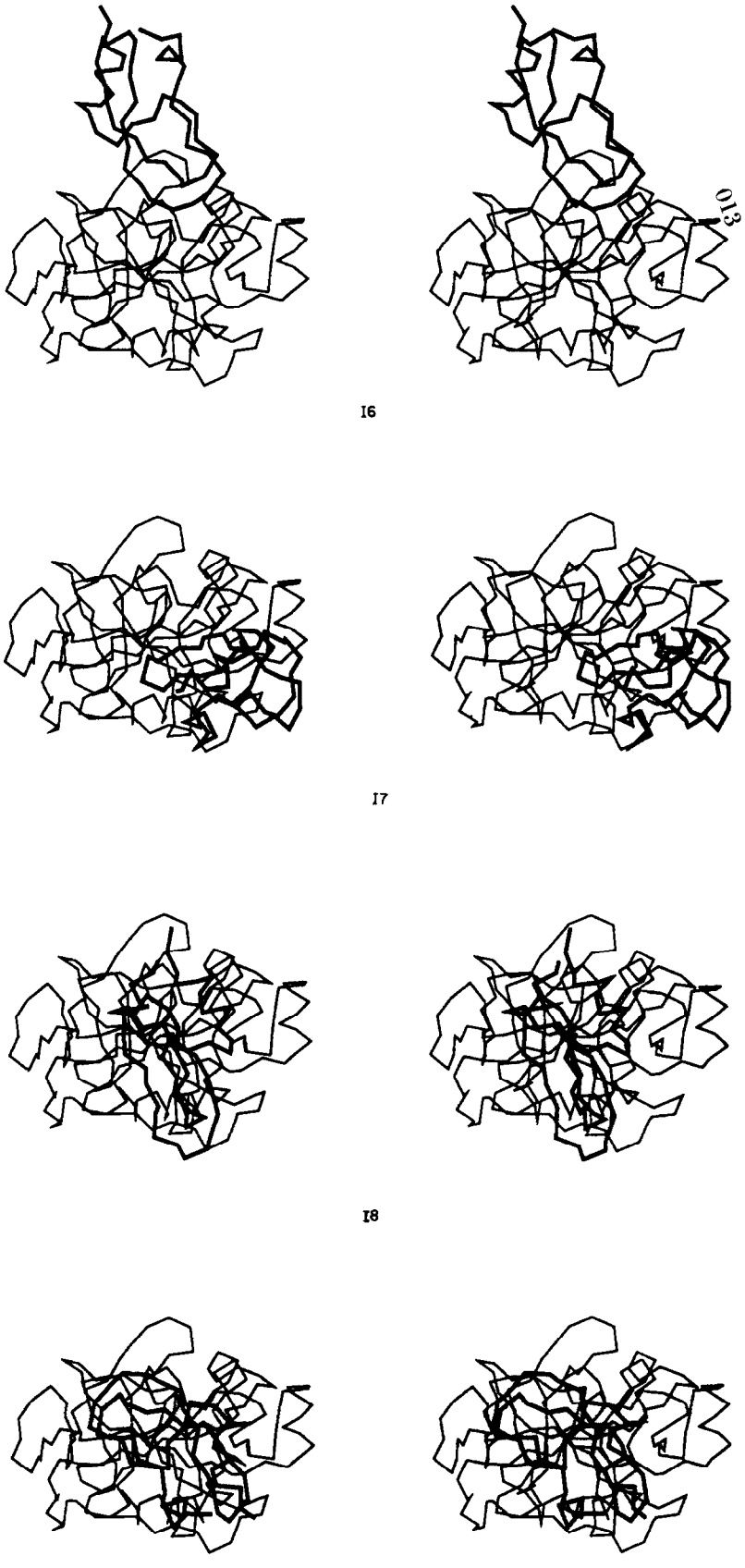


Figure 3. The C^{α} atom plots in stereo of the docking solutions in Table 6. The plot of crystal structure 2PTC corresponds to the Expt. entry, and those labeled I1 to I9 correspond to the docking solutions from the 2nd to the last entry in Table 6. The C^{α} atom chains of trypsin are shown in the same orientation in all plots in thin lines. Those of trypsin inhibitor are in thick lines. Plot I6 corresponds to the correct docking solution.

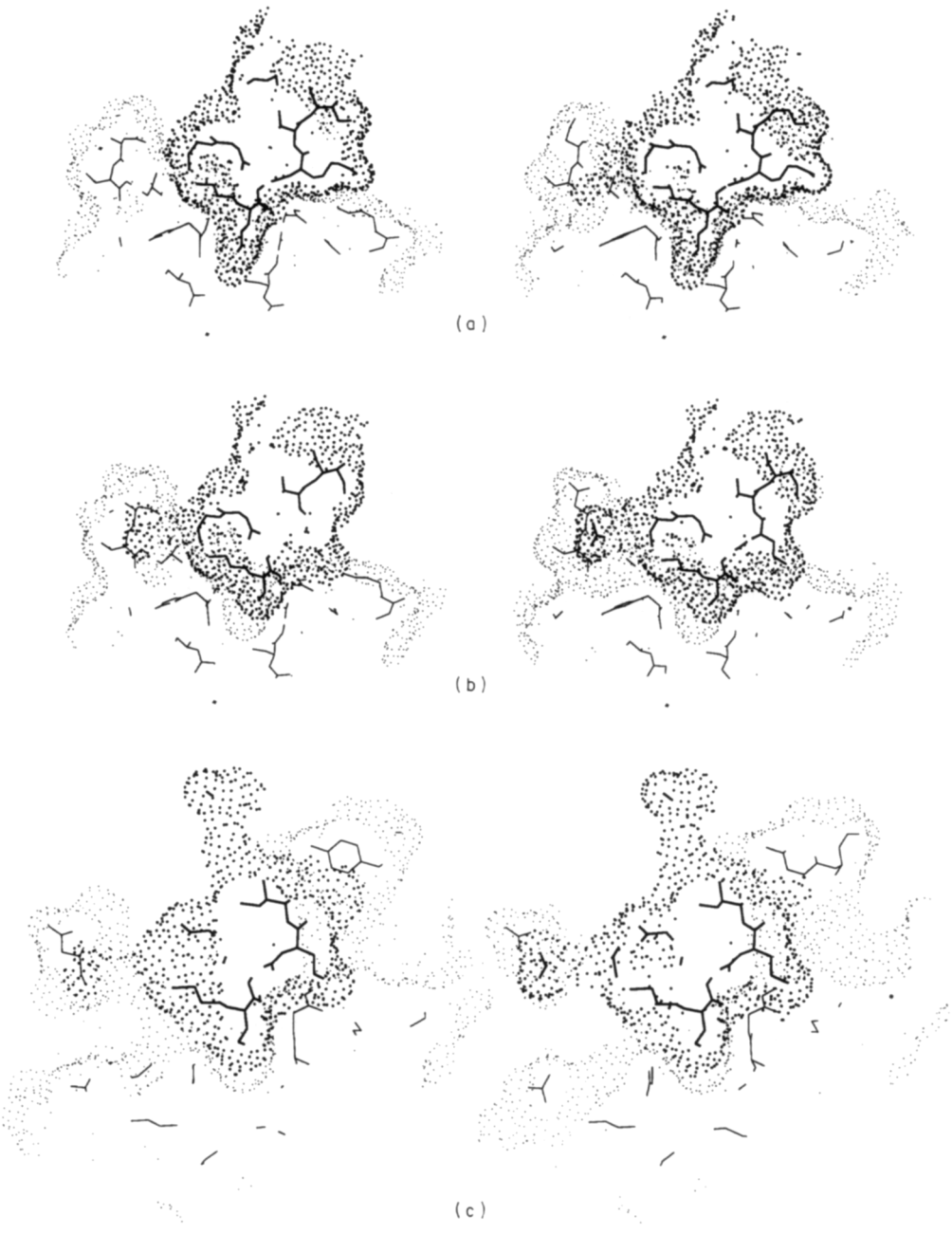


Figure 4. Sections of the interfaces of trypsin–trypsin inhibitor complexes in stereo. All 3 sections are from the comparable regions of three complexes obtained differently. The atoms and surface dots of trypsin inhibitor are shown in thick lines and heavy dots, respectively, while those from trypsin in thin lines and light dots. (a) A section of the interface from the crystal structure complex (2PTC). The residue shown to fit snugly into the pocket in trypsin is Lys15 of trypsin inhibitor. (b) A similar section of the interface from the complex of 3PTN with 4PTI, which is superimposed over 2PTCI by least-squares fitting. Lys15 of 4PTI no longer points downward because of the difference in its C^{α}–C^{β} dihedral angle; Arg19 to the right of Lys15 is protruding into the surface of 3PTN, also because of the difference in its C^{α}–C^{β} dihedral angle. (c) A similar section of the interface from the complex of 3PTN with the docked 4PTI, as described in Fig. 2(c). The surface of Lys15 is shown to fit slightly better than that in (b), but the overlaps between the 2 molecules are obvious, and are accepted by the present docking procedure.

lysozyme Fab affect mostly the positions of the side-chain atoms involved in the binding site, while the movement of backbone atoms is relatively small. A superposition of the binding site atoms (within 7 Å of the antibody component of the complex 2HFL) of the complexed lysozyme (2HFLY) over those in the uncomplexed lysozyme (1LYZ) shows that significant differences in atomic positions are in two loops, residues Gln41 to Gly49 and Asn65 to Ser72, and the largest changes are in the side-chains of Arg45, Thr47, Arg68 and Pro70, as shown in Figure 5(a). The r.m.s. distance for the 165 atoms in the

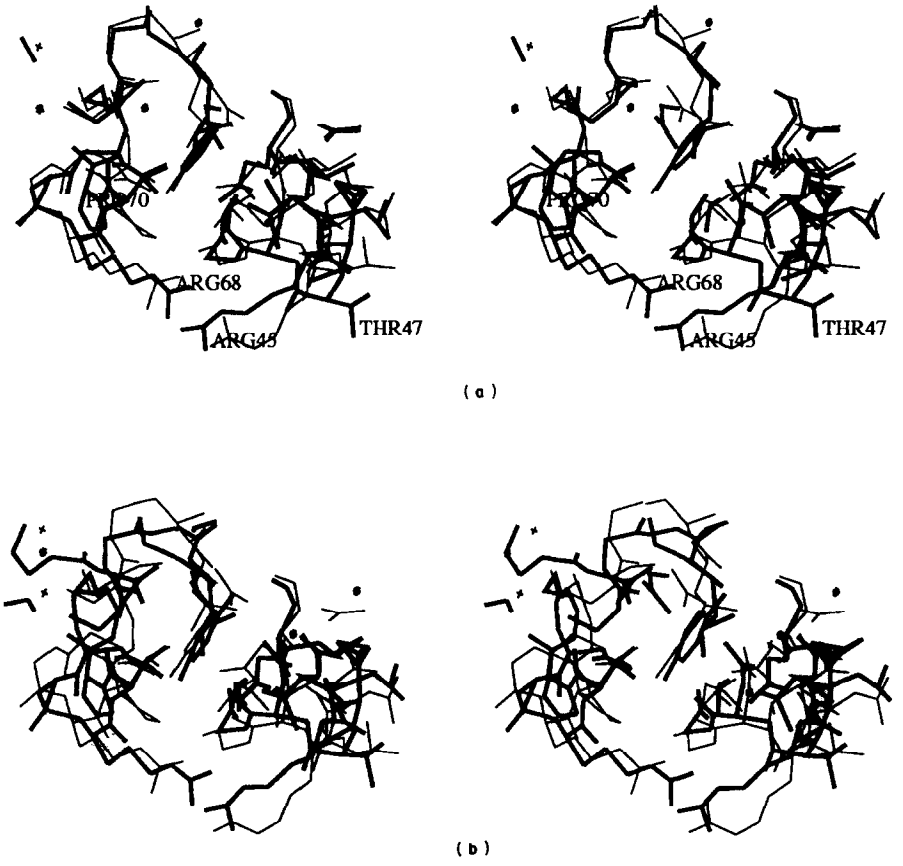


Figure 5. (a) Comparison of the contacting atoms of lysozyme in complexed and uncomplexed states in stereo. The binding site atoms of 1LYZ in thick lines (Johnson & Phillips, 1965), an uncomplexed lysozyme, is superimposed over those of 2HFLY in thin lines, (Sheriff *et al.*, 1987), a complexed lysozyme. The r.m.s. distance for the atoms within 7 Å of the interface is 1·33 Å. The most significant conformational changes are in residues Arg45, Thr47, Arg68 and Pro70, as well as some large movement in the main-chain atoms around Thr47. (b) The superposition of the binding site atoms of the docked lysozyme and the complexed lysozyme. The complexed lysozyme, 2HFLY, is shown in thin lines, the same as in (a). The docked lysozyme, shown in thick lines, was obtained by applying the rotation from solution no. 235 to 1LYZ and then superimposing the center of mass of the thick line atoms (of 1LYZ) over that of the thin line atoms (2HFLY). The resulting r.m.s. distance is 1·55 Å.

binding site is 1·33 Å, while for all the atoms in lysozyme it is 0·96 Å; so, it is obvious that most of the changes occur in the binding site. Since there is no crystal structure of the corresponding uncomplexed lysozyme antibody, we used the antibody portion of the complex as a surrogate of free antibody structure. To make a realistic test, the whole surface of the two variable domains (residues H1 to H115 and L1 to L110) was used in the docking, including the surface buried in the junction between the variable and constant domains.

About 500 solutions were selected from the cluster analysis and subjected to the scoring of the total interaction. The average number of matches and the standard deviation for all the docking solutions from the geometric docking are 364 and 66, respectively. After scoring the total interaction with different cube sizes, 15 solutions were found to have positive (favorable) interaction. They are listed in Table 7, and labeled Y1 to Y15. Among them, solutions nos. 235 and 347 (labeled Y8 and Y12, respectively) are closest to the correct solution. The correct solution (the Expt. entry in Table 7) was determined by superimposing the uncomplexed lysozyme 1LYZ to the complexed lysozyme 2HFLY with least-squares fitting. Solutions Y8 and Y12 differ from the correct solution by 9 and 13 degrees of rotation, respectively, and from each other by 15·5 degrees, which is slightly bigger than the angle distance cut-off (ϑ_{dist} in Table 2), 12 degrees. Therefore, although they can be considered to be the same solution, they are listed separately. The r.m.s. distance for the binding site atoms, between the uncomplexed lysozyme 1LYZ, rotated with the rotation of solution no. 235, and the complexed lysozyme 2HFLY, and with their centers of mass superimposed (Fig. 5(b)), is 1·55 Å. This is very close to that calculated with least-squares fitting, which is 1·33 Å. Similar sections of the interface from the crystal structure of the lysozyme–antibody complex (2HFL), from the derived complex based on least-squares fitting, and from the docked complex based on solution no. 235 are shown in Figure 6. The complementarity of the interface from the docked

Table 7

Docking of a free lysozyme to a lysozyme antibody

Complex name	Solution no.	Rotation (deg.)			Translation (grids)			Number of matches	Overlap (cubes)	Total interaction
2HFLV/1LYZ	Expt.	344	137	158	5	4	19	546	24	15
Y1	58	114	111	145	−9	15	−3	557	30	63
Y2	82	278	90	75	−7	12	−7	549	39	130
Y3	100	270	21	55	−10	11	−6	543	32	138
Y4	122	16	36	65	−13	8	−10	538	23	145
Y5	144	135	131	35	7	−16	−2	531	14	62
Y6	204	14	135	155	−15	9	−7	521	34	109
Y7	223	105	106	145	−7	15	−1	517	71	138
Y8*	235	338	135	155	4	5	19	515	24	47
Y9	254	136	75	155	−7	12	−9	513	54	20
Y10	272	83	135	105	−6	15	−3	511	145	55
Y11	282	262	84	115	−11	11	−7	509	17	134
Y12*	347	337	141	165	5	4	19	501	12	96
Y13	350	278	81	125	−8	11	−7	501	54	63
Y14	352	340	27	65	−12	8	−11	500	28	152
Y15	357	42	42	95	−6	11	−8	500	46	89

1LYZ is the crystal structure of a free lysozyme molecule in uncomplexed state. 2HFL is the crystal structure of a lysozyme–antibody complex, and 2HFLV is the variable domain of the antibody in the complex, which includes residues L1 to L110 and H1 to H115. See Tables 5 and 6 for what each column represents. Solution nos. 235 and 347 (Y8 and Y12, respectively, in Fig. 7), marked with asterisks in the 1st column, are the correct docking solutions, and they are close to the rotation and translation derived from the superposition of 1LYZ over 2HFLY (the lysozyme in the complex structure 2HFL) by least-squares fitting. The latter is given in the Expt. entry.

complex (Fig. 6(c)) is as good as that from the derived complex (Fig. 6(b)). The docked complexes corresponding to solutions Y1 to Y15, as well as the crystal structure 2HFL, are plotted in Figure 7, with C^{α} atoms only. The contact surfaces derived from the incorrect solutions involve either the buried surface in the junction of the variable and constant domains, or the surface of the rather smooth β-barrels of the variable domains. This is similar to the case for the incorrect solutions of docking the free typsin inhibitor to the free trypsin structures. Thus, in both cases, visual examination of the final solutions could easily identify most or all of the wrong solutions, leaving one or only a few as the best candidates for the correct solution.

4. Discussion

The computing time required by the present procedure is approximately proportional to the product of the number of the probe surface dots and the number of the target surface dots. A typical computing time required can be estimated from one of the docking experiments, e.g. docking trypsin inhibitor 4PTI to trypsin 3PTN, in which 1013 and 2469 low-density surface dots were used for the probe and the target, respectively, and 28,043 rotations were sampled. The total c.p.u. time is equivalent of about five days of c.p.u. on a VAX/785 computer. The computer memory required is determined by the sizes of the molecules in docking; it is approximately proportional to the volume of the larger one of the two molecules. Fortunately, all the tests presented can be run on a computer with an eight megabyte memory and with faster computers such as IBM 3090/XA, which requires 22 c.p.u. hours for the above example.

How well the present docking procedure works for a particular pair of molecules depends to a certain extent on the choices of parameters used in calculating molecular surfaces, surface dots and cubes, and in testing the local complementarity. As the most intensive part of the computing, the translation search should use as few dots as possible without losing the essential features of the surface structures needed for distinguishing the potentially correct from the obviously incorrect solutions. In the first step of a translation search, the parameters that worked for the tests we have done are 0·3 dots/Å^2 for the low density, 0·5 Å^2 for the area cut-off used in discarding the small surface patches, and 3·2 Å for the cube size, which may be varied within a range of 0·4 Å. The corresponding optimal cone angle for testing the local complementarity ranges between 50° and 60°. In the second step of translation search, a dot density of 1·0 dots/Å^2 is sufficiently high to give reasonably good results in refining the translation vectors; so, using a higher dot density is not justified because of the increase in the computing time. The corresponding optimal cube size and cone angle range from 1·8 to 2·0 Å and from 30° to 40°, respectively.

The goodness of a docking solution from the geometric docking alone, the first stage of the method, is defined as the number of matches between the probe and the target surface cubes minus the number of overlaps between their volume cubes. The former measures the degree of complementarity between the matched surface cubes and the latter the number of bad contacts. A stringent

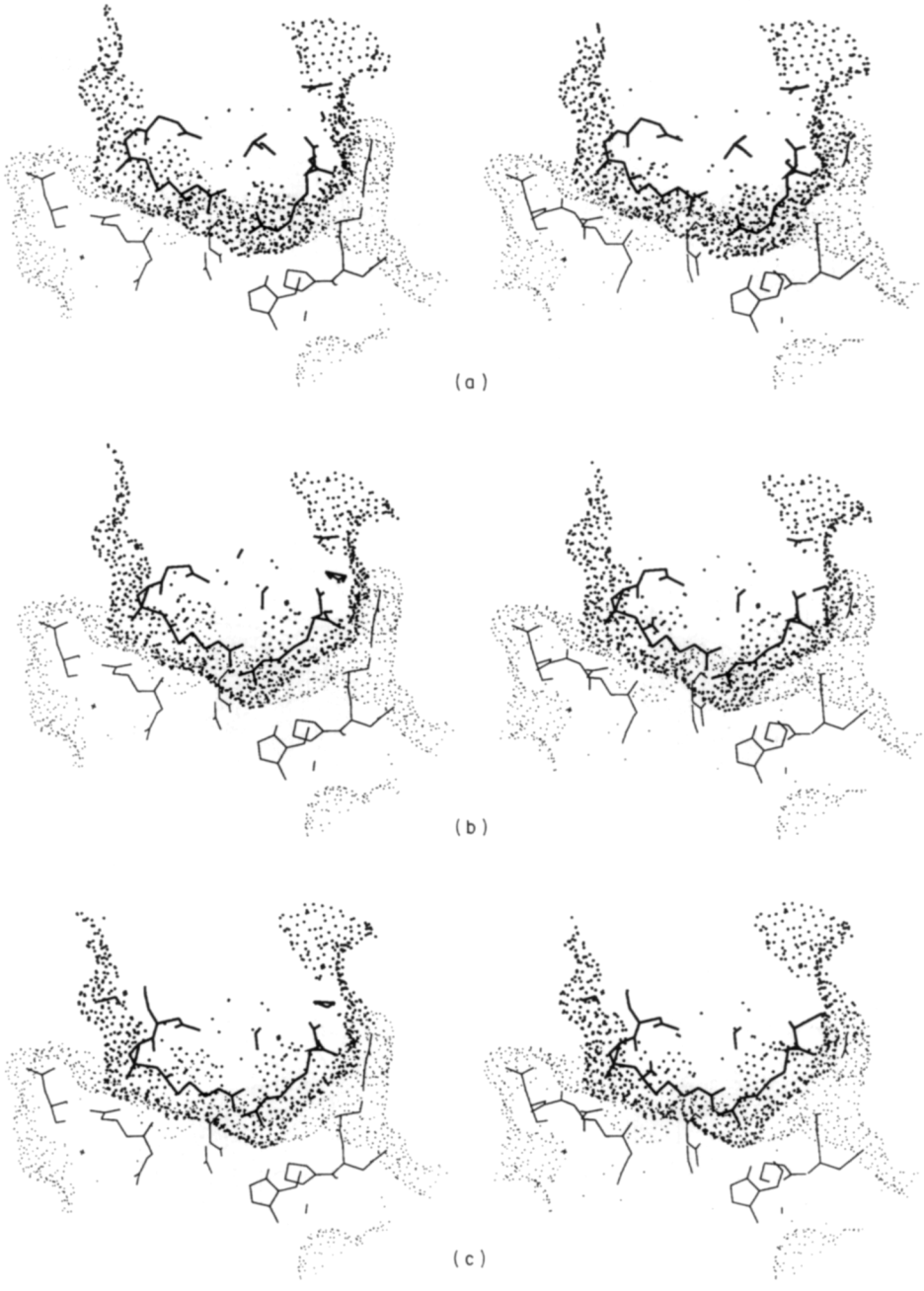
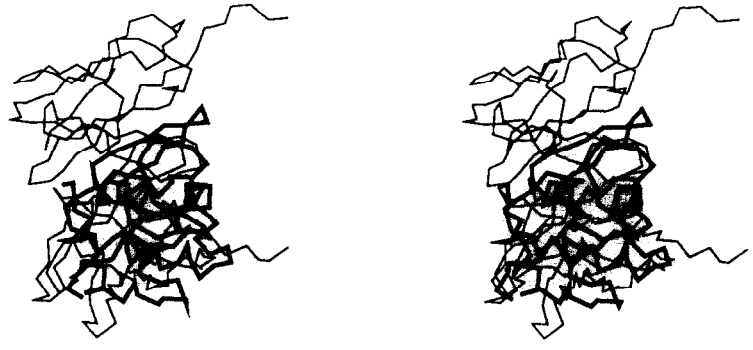


Figure 6. (a) A section of the interface in the crystal structure of a lysozyme–antibody complex in stereo. The 2 residues shown in the complementary interface are Arg68 (left) and Arg45 (right). (b) A similar section of the interface from the complex of 2HFLV with 1LYZ, which is superimposed over 2HFLY by least-squares fitting. One can see the overlaps of the surfaces caused by the downward movement of Arg68 and the upward movement of Arg45. (c) A similar section of the interface from the complex of 2HFLV and the docked 1LYZ, which was obtained as described in (b). The overlap between the 2 surfaces is reduced in the docked complex by adjusting the relative orientation of 1LYZ to 2HFLV to achieve a better match between the surfaces.

directional constraint on matching surface normals can rule out most solutions that would result in a large number of bad contacts. But the complementarity test using surface dots is, after all, a local constraint confined within a pair of matched surface cubes and, therefore, does not confer any restriction on how two molecules should be situated relative to one another as rigid bodies. Thus, it is necessary to use a cut-off for the maximum number of overlaps between the volume cubes to exclude solutions with a large number of bad contacts. From our experience, a good cut-off value is about one third of the highest matching score. This value usually eliminates about a fourth of the top solutions found by considering the number of matches alone.

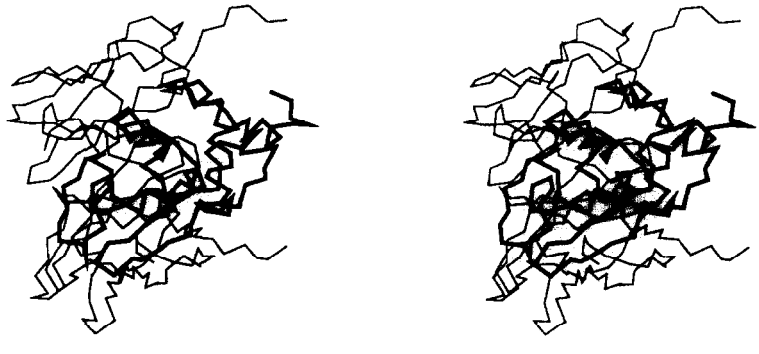
The effect of conformational changes can be seen

2HFL crystal structure



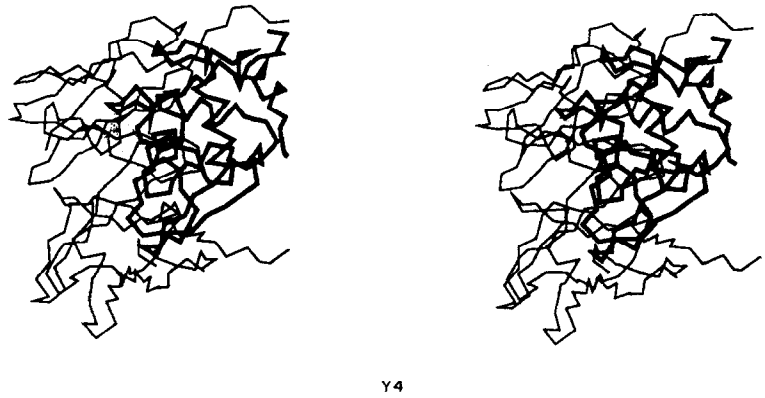
Y1

Y2

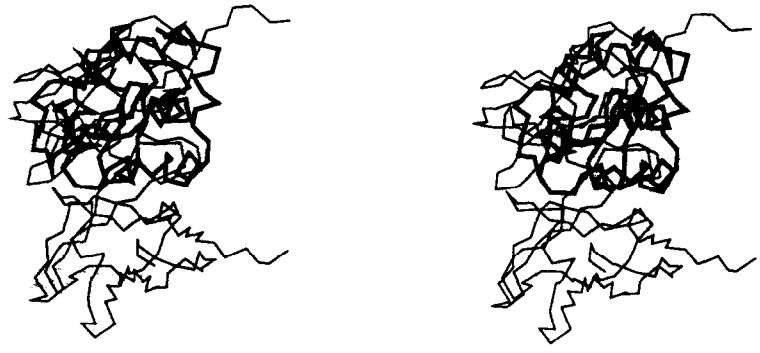


Y3

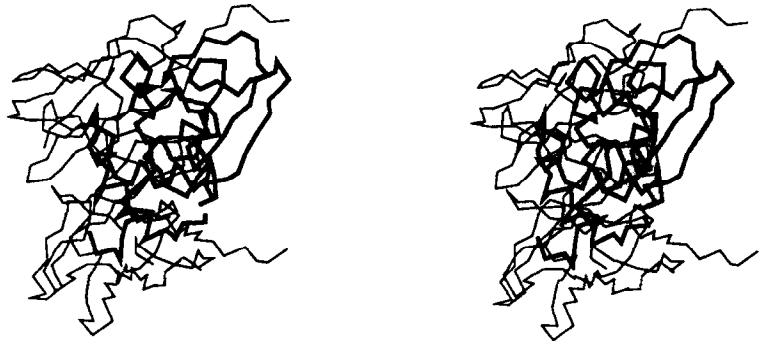
Fig. 7.



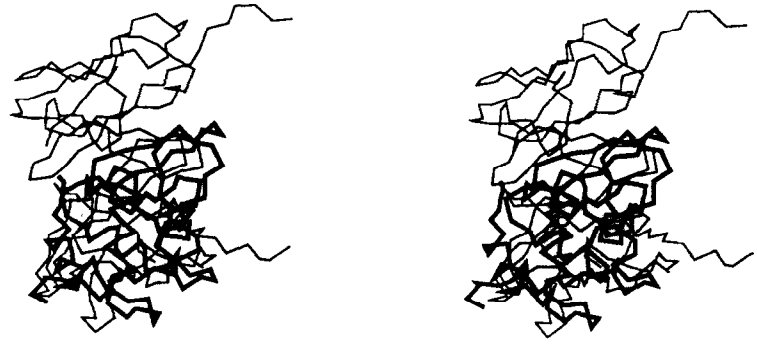
Y4



Y5



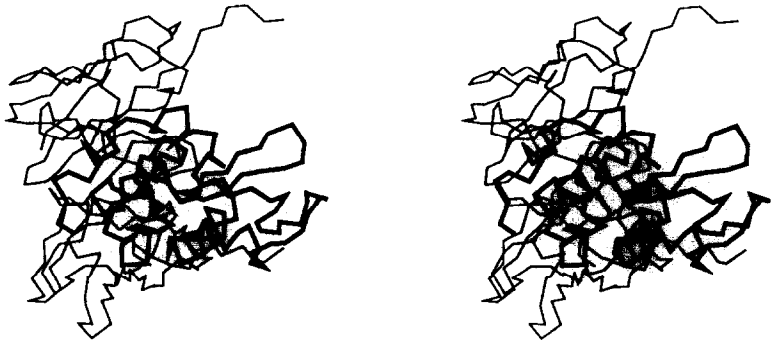
Y6



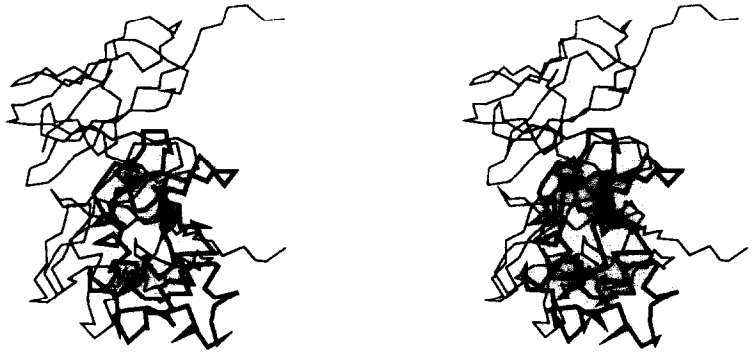
Y7

Fig. 7.

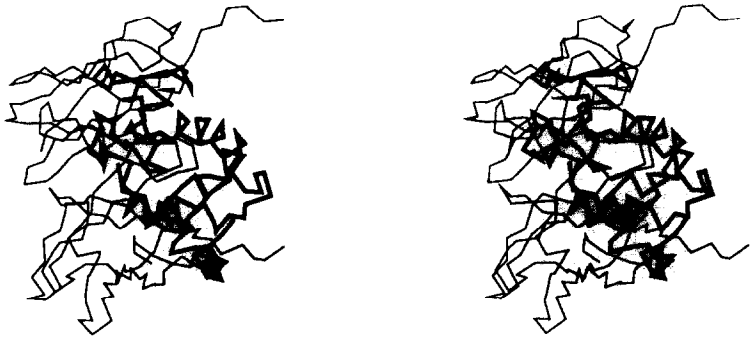
Y8



Y9



Y10



Y11

Fig. 7.

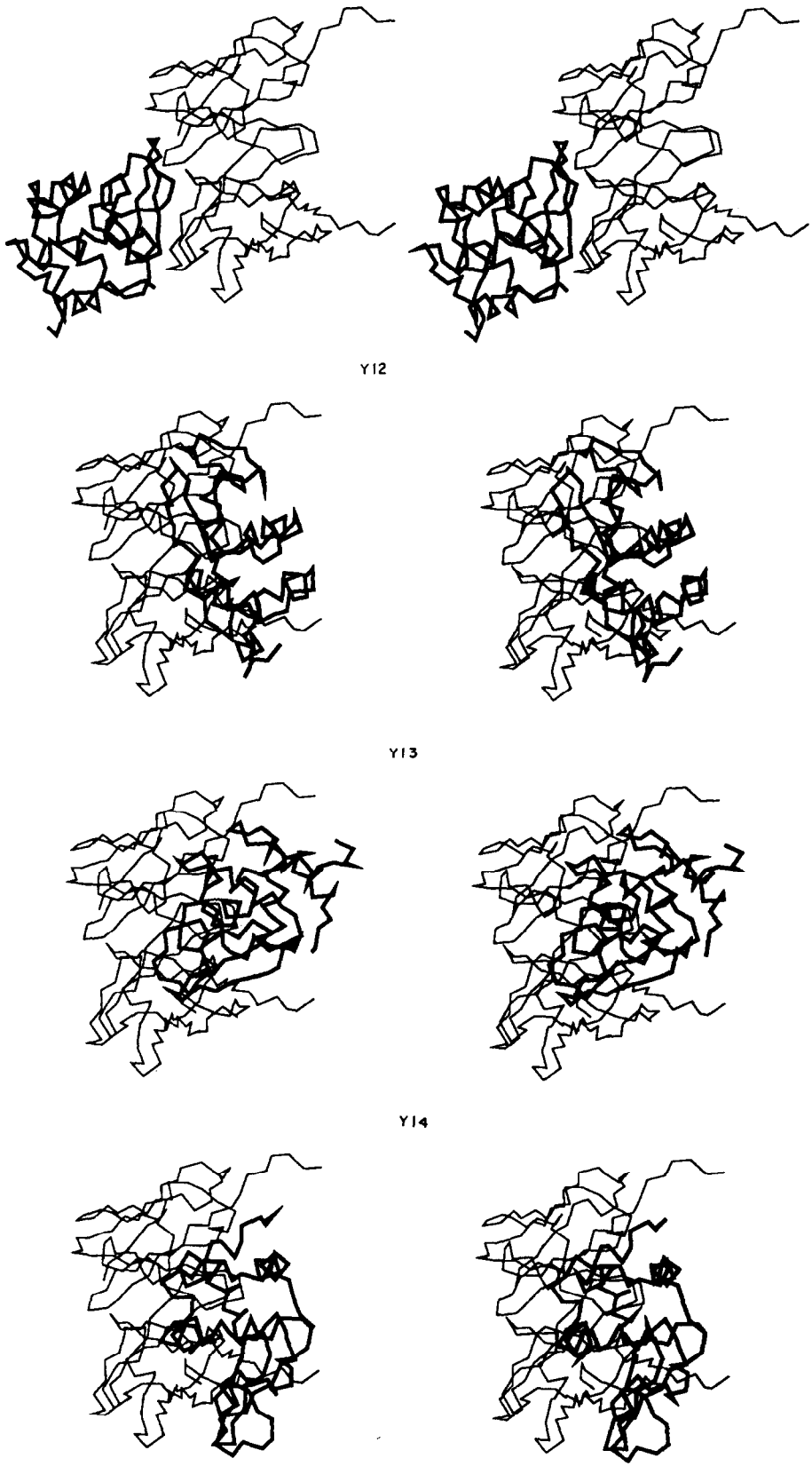


Figure 7. The C^{α} atom plots in stereo of the docking solutions in Table 7. The plot of crystal structure 2HFL corresponds to the Expt. entry, while the rest, labeled Y1 to Y15, correspond to the docking solutions from the 2nd entry to the last in Table 7. The C^{α} atom chains of lysozyme are shown in thick lines and those of the antibody variable domain in thin lines. Both Y8 and Y12 have a complex geometry similar to that of the crystal structure. Notice that most of the wrong solutions give rise to complexes in which the lysozyme binds outside of the binding pocket of the antibody.

from the results of the second way of testing, i.e. docking the crystal structures of uncomplexed molecules. It is clear that geometric complementarity alone is no longer sufficient, as expected, to discriminate the correct solutions, and the energetic complementarity is necessary. In the present procedure, the latter is assessed by the summation of the surface areas within each pair of matched surface cubes. We have also tried to use the product of the surface areas, or to include the interactions between the matched surface areas and the surface areas in the nearest-neighbor surface cubes. But neither worked as well as the summation in discriminating the correct solutions from the noise solutions.

Occasionally, the total interaction for a given docking solution becomes positive just by chance for a given cube size. The size of surface patches that have similar characteristics (such as hydrophobic or polar) is expected to be about 4 Å in proteins. Therefore, a docking solution corresponding to a complex with true complementarity should have a positive total interaction, independent of the cube size, whereas accidental matches are dependent on the cube size. All the solutions listed in Tables 6 and 7 were found to have positive total interactions consistently for different cube sizes (such as 1·6, 2·0 and 2·4 Å) with changes in magnitude only (results not shown).

It is interesting to compare the results of two tests on docking the crystal structures of free molecules. Although the r.m.s. distance for the binding site atoms between the trypsin inhibitors, 4PTI and 2PTCI, is not much bigger than that between the two lysozyme structures, 1LYZ and 2HFLY, the former being 1·80 and the latter 1·33 Å, the effect of the conformational changes in the trypsin inhibitors on docking seems to be much more significant. This is because residues Arg15 and Lys17 of 4PTI adopt C^{α}–C^{β} dihedral angles very different from those in 2PTCI (Fig. 2). As a result, the surface shape of the binding site changes dramatically and the large bumps generated by these two residues are completely dislocated. Matching the distorted surface of these bumps to their intended pockets in the receptor site would give rise to a docking solution that could be quite far away, especially in orientation, from the geometry derived from the superposition by least-squares fitting. In fact, the correct docking solution (solution no. 51 in Table 6) fits the surfaces slightly better than the superposition, because it forces residue Lys15 into the corresponding pocket in the receptor site of 3PTN. The interfaces of the superimposed and the docked complex are shown in Figure 4(b) and (c), respectively. In contrast, the main conformational changes in the lysozyme structures, 1LYZ and 2HFLY, involve only the shift of side-chain atoms without significant changes in dihedral angles. Therefore, the shape of the binding site does not change very much, except for some translational movement made by residues Arg45, Thr47, Arg68, and Pro70 on the surface (Fig. 5). This type of conformational change can be accommodated by the present docking procedure without any problems, and the results show that the correct docking solutions are within the sampling errors. In Figure 6 are shown the interfaces of the crystal structure of the complex, the superimposed complex, and the docked complex.

After the docking solutions with negative total interaction are eliminated, there are still some noise solutions left that have high scores as the correct ones. We can approach this problem in the following ways. First, we may single out the correct solutions from a prior knowledge of the approximate location of the binding site on either the ligand or receptor molecule derived from biochemical, molecular biological, or genetic studies. For example, solutions resulting in complex formation outside the binding sites of the variable domain of an antibody can be easily eliminated. Second, we can try to improve the accuracy of the calculation of the total interaction for a docked complex. Our preliminary tests using the atomic solvation parameters (Eisenberg & McLachlan, 1986) indicate that the calculated free energy is not very effective in ruling out incorrect solutions, although it accepts the correct solutions. It is not clear whether this is caused by errors in the geometry of the docked complex or conformational changes in the uncomplexed molecules. Another way of improving the current calculation of the total interaction is to determine empirically a set of weights for each type of interaction in the current summation by studying known complex structures.

We should also point out that the docking solutions from our procedure may not yet be accurate enough to be used as an initial model for energy minimization or molecular dynamics simulation. And, from the above discussion, we have seen that it is very important for a docking procedure to be able to deal with the type of conformational changes involving side-chain dihedral angles. Therefore, improvement on the present method can be made by considering the conformational changes explicitly to optimize the geometry of a docked complex to achieve better and tighter fit, so that the docked complex can be used as a starting model for the application of molecular dynamics simulation techniques to predict the true complex structure.

One important implication of this study is that a description of a molecular surface structure should include the following aspects. First, the local features of a molecular surface structure are best described at the size of about 4 to 5 Å, which seems to be the optimal size for exhibiting the characteristics of a group of atoms on or near the surface. Second, the organization of the local features determines the overall shape and size of a surface structure, which is analogous to the landscape of a surface. Third, the volume effect, namely the short-ranged strong repulsion due to the steric hindrance and close packing of atoms, is an essential part of the description. With these features in mind, one could improve methods that will speed up the present docking procedure. Another implication is that an exhaustive search in rotation and trans-

lation space may not be necessary for most docking problems, but it is certainly very fruitful to have done an exhaustive search and learned all the possibilities. It is now possible to apply the Monte Carlo procedure or other techniques (Metropolis *et al.*, 1953) to optimize the sampling of the solution space and thus improve the efficiency of the present docking procedure.

At present there are very few experimental data available on the dynamic process of molecular association. Many have tried to study diffusion controlled reactions with molecular dynamic or Monte Carlo simulations (McCammon *et al.*, 1987), but so far the systems simulated have been mostly those in which electrostatic interaction dominates the kinetics of the association process (Northrup *et al.*, 1988; Allison *et al.*, 1988; Sharp *et al.*, 1987). Success of the present docking procedure suggests that the docking solutions found may be used as a starting model for theoretical studies of the dynamic process of molecular recognition and association on a variety of systems using simulation techniques. Electrostatic interaction is an important aspect of complementarity in a molecular complex (Warwicker, 1989), and since it is the only long-range non-bonded interaction, compared with van der Waals' interaction and hydrogen bonding, it has been indicated as the sole driving force in the process of molecular recognition and association, at least kinetically and dynamically if not thermodynamically, but it remains to be seen whether this is a universal mechanism.

Several systems have been tested in previous work on docking. Macromolecular–ligand systems tested by Kuntz *et al.* (1982) are the heme–myoglobin interaction and the binding of thyroid hormone analogs to prealbumin. In their method, the binding sites of an enzyme are first identified through the clusters of spheres on the surface, and then the ligand spheres are matched with the receptor spheres. It was shown that approximately correct structures could be readily recovered and identified as feasible solutions by examining many binding geometries and evaluate them in terms of steric overlap. Wodak & Janin (1978) tried to use a simplified model in which each residue is replaced by one interaction center to analyze protein–protein interactions. This simplified model was applied to a trypsin–trypsin inhibitor complex of known crystal structure, and all possible modes (orientations and contact distances) of interaction between the inhibitor and the active center of the enzyme were generated systematically (5 variables because the binding site of the inhibitor is constrained to be near the binding site of the enzyme through an intermolecular vector). Nine structures selected by their procedure have non-bonded interaction energies and buried surface areas similar to those of the native complex (crystal structure), which is among the nine structures. The prediction of protein complexes has also been attempted by Connolly (1986), who developed a computational method based on matching complementary patterns of knobs and holes on a molecular surface. Although both complexes used in testing are crystal structures, this method succeeded in predicting the association of the α and β subunits to form the $\alpha\beta$ dimer corresponding to the $\alpha_1\beta_1$ interface in the hemoglobin tetramer, but failed to predict the trypsin–trypsin inhibitor interface. More recently, Warwicker (1989) tried to use a reduced electrostatic model to predict protein–protein interactions and found that the crystal structures of the complex between trypsin and trypsin inhibitor (Marquart *et al.*, 1983) and the complex between lysozyme and antibody (Sheriff *et al.*, 1987) correspond to configurations with maximum electrostatic potential energies. However, when tested on an uncomplexed trypsin inhibitor (docked to the trypsin in the complex), this reduced model of electrostatic interactions failed to give maximum peaks to configurations corresponding to the crystal structure complex, presumably because the calculation is sensitive to side-chain orientations that differ in the binding sites of the two inhibitor molecules. In comparison, our procedure appears to be more robust. And it worked in many systems including DNA–drug and protein dimer association (our unpublished results), and more importantly, in docking uncomplexed molecules, in which there are significant conformational changes.

In conclusion, our "soft docking" method consisting of two steps, geometric docking and interaction scoring, is able to accommodate conformational charges associated with the molecular complexing process, and to identify the correct complex model or a small number of complex models, which can be further screened visually or based on other experimental results to distinguish the correct complex model.

References

Alden, C. J. & Kim, S.-H. (1979). *J. Mol. Biol.* **132**, 411–434.

Allison, S. A., Bacquet, R. J. & McCammon, J. A. (1988). *Biopolymers,* **27**, 251–269.

Aqvist, J. & Tapia, O. (1987). *J. Mol. Graphics,* **5**, 30–34.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Birdsall, B., Burgen, A. S. V., de Miranda, J. R. & Roberts, G. C. K. (1978). *Biochemistry,* **17**, 2102–2110.

Brooks, B. R., Bruccoler, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comp. Chem.* **2**, 187–217.

Chothia, C. (1974). *Nature (London),* **248**, 338–339.

Chothia, C., Novotny, J., Bruccoleri, R. & Karplus, M. (1985). *J. Mol. Biol.* **186**, 651–663.

Connolly, M. L. (1983*a*), *Science,* **221**, 709–713.

Connolly M. L. (1983*b*). *J. Appl. Crystallogr.* **16**, 548–558.

Connolly, M. L. (1986). *Biopolymers,* **25**, 1229–1247.

Davies, D. R., Sheriff, S. & Padlan, E. A. (1988). *J. Biol. Chem.* **263**, 10541–10544.

Davies, D. R., Padlan, E. A. & Sheriff, S. (1990). *Annu. Rev. Biochem.* **59**, 439–473.

DesJarlais, R. L., Sheridan, R. P., Dixon, S., Kuntz, I. D. & Venkataraghavan, R. (1986). *J. Med. Chem.* **29**, 2149–2153.

DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, S., Kuntz, I. D. & Venkataraghavan, R. (1988). *J. Med. Chem.* **31**, 722–729.

Diamond, R. (1974). *J. Mol. Biol.* **82**, 371–391.

Eisenberg, D. & McLachlan, A. D. (1986). *Nature (London)*, **319**, 199–203.

Johnson, L. N. & Phillips, D. C. (1965). *Nature (London)*, **206**, 761–763.

Kuhl, F. S., Crippen, G. M. & Friesen, D. K. (1984). *J. Comp. Chem.* **5**, 24–34.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). *J. Mol. Biol.* **161**, 269–288.

Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.

Levinthal, C., Wodak, S. J., Kahn, P. & Dadivanian, A. K. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 1330–1334.

Levitt, M. (1974). *J. Mol. Biol.* **82**, 393–420.

Lewis, M. & Rees, D. C. (1985). *Science*, **230**, 1163–1165.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*, W. H. Freeman and Company, New York.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). *Acta Crystallogr. sect. B*, **39**, 480–490.

Matthews, D., Alden, R. A., Bolin, J. T., Filman, D. J., Freer, S. T., Hamlin, R., Hol, W. G. J., Kisliuk, R. L., Pastore, E. J., Plante, L. T., Xuong, N. & Kraut, J. (1978). *J. Biol. Chem.* **253**, 6946–6954.

McCammon, J. A., Bacquet, R. J., Allison, S. A. & Northrup, S. H. (1987). *Faraday Discuss. Chem. Soc.* **83**, 213–222.

Metropolis, M., Rosenbluth, M., Rosenbluth, A., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). *J. Mol. Biol.* **196**, 641–656.

Northrup, S. H., Boles, J. O. & Reynolds, J. C. (1988). *Science*, **241**, 67–70.

Pfeifer, P., Welz, U. & Wippermann, H. (1985). *Chem. Phys. Lett.* **113**, 535–540.

Pratt, L. R. & Chandler, D. (1977). *J. Chem. Phys.* **67**, 3683–3704.

Rebek, J., Jr (1987). *Science*, **235**, 1478–1484.

Richards, F. M. (1977). *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richmond, T. J. (1984). *J. Mol. Biol.* **178**, 63–89.

Richmond, T. J. & Richards, F. M. (1978). *J. Mol. Biol.* **119**, 537–555.

Salemme, F. R. (1976). *J. Mol. Biol.* **102**, 563–568.

Schulz, G. E. & Schimer, R. H. (1979). *Principles of Protein Structure*, Springer-Verlag, Berlin, Heidelberg and New York.

Sharp, K., Fine, R. & Honig, B. (1987). *Science*, **236**, 1460–1463.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987). *Proc. Nat. Acad. Sci., U.S.A.* **84**, 8075–8079.

Shrake, A. & Rupley, J. A. (1973). *J. Mol. Biol.* **79**, 351–371.

Tanford, C. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 4175–4176.

Warwicker, J. (1989). *J. Mol. Biol.* **206**, 381–395.

Weiner, P. K. & Kollman, P. A. (1981). *J. Comp. Chem.* **2**, 287–303.

Williams, J. W., Morrison, J. F. & Duggleby, R. G. (1979). *Biochemistry*, **18**, 2567–2573.

Wlodawer, A., Deisenhofer, J. & Huber, R. (1987). *J. Mol. Biol.* **193**, 145–156.

Wodak, S. J. & Janin, J. (1978). *J. Mol. Biol.* **124**, 323–342.

Edited by R. Huber