PROTEINS: Structure, Function, and Genetics 25:120–129 (1996)

Molecular Docking Using Surface Complementarity

Vladimir Sobolev,[1] Rebecca C. Wade,[2] Gert Vriend,[2] and Marvin Edelman[1]
[1]Department of Plant Genetics, Weizmann Institute of Science, Rehovot 76100, Israel; and [2]Department of Molecular Structure, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

ABSTRACT **A method is described to dock a ligand into a binding site in a protein on the basis of the complementarity of the intermolecular atomic contacts. Docking is performed by maximization of a complementarity function that is dependent on atomic contact surface area and the chemical properties of the contacting atoms. The generality and simplicity of the complementarity function ensure that a wide range of chemical structures can be handled. The ligand and the protein are treated as rigid bodies, but displacement of a small number of residues lining the ligand binding site can be taken into account. The method can assist in the design of improved ligands by indicating what changes in complementarity may occur as a result of the substitution of an atom in the ligand. The capabilities of the method are demonstrated by application to 14 protein–ligand complexes of known crystal structure.**

Key words: molecular recognition, ligand binding, drug design, binding site

INTRODUCTION

Knowledge of the atomic details of a ligand–receptor complex is important for understanding the nature of the interaction as well as the function of the complex. Drug design is a prominent example in which such knowledge is of great value. If the structure of the complex is known, modification of the ligand and/or receptor can be addressed in a rational manner. Unfortunately, experimental determination of the three-dimensional coordinates of a ligand–protein complex is often a laborious process. Consequently, computational procedures for docking are required to estimate the location of a ligand in a complex whose structure is not yet solved.

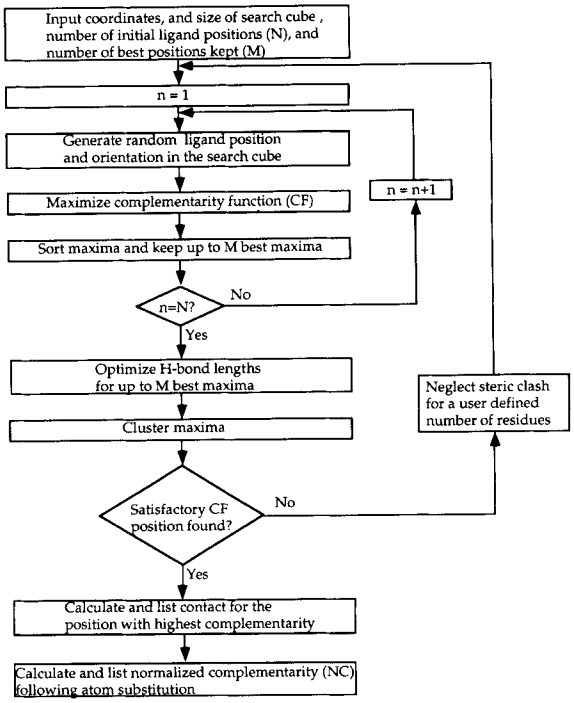
Docking of a ligand into a receptor can conceptually be divided into two parts: generation of the possible structures and determination of their fitness. Most existing methods use shape complementarity,[1–4] geometric constraints,[5–7] force fields,[8–20] or some combination of the above[21–23] to predict the structure of ligand–receptor complexes. Many of these docking methods have been reviewed recently.[24,25] Most methods treat ligand and receptor as rigid objects, while some allow for partial flexibility.[15,19,20] All assume that the receptor coordinates are known.

Methods based on shape complementarity can successfully predict the structures of molecular complexes when the docked molecules are large,[26–31] since the occurrence of a few unfavorable contacts can be compensated by a large number of favorable ones. For smaller ligands, however, shape complementarity alone is often not sufficient, and a more detailed description of the ligand is required.[32] Physically realistic force fields can provide such detail, but predictions can be very sensitive to the accuracy of the force field parameters and receptor coordinates and are often hampered by the lack of available parameters for many atom types and atom-atom interactions.

We previously described[33] a docking method that addresses many of the aforementioned problems. This method is based on maximization of surface complementarity. Here we describe a set of improvements to this method. We introduce a higher level of detail in the assignment of atom classes; hydrogen bond geometry is explicitly optimized, and limited protein flexibility is emulated. Besides better docking results, the higher level of detail in the atomic class assignment allows for a detailed analysis of the complex and can be used, in a fully automated fashion, to suggest better binding ligands. The method is tested on 14 protein–ligand complexes from the Brookhaven Protein Data Bank (PDB)[34] selected to contain examples of problems one can encounter during computational ligand docking. To test the method in realistic situations, we dock the ligands into aporeceptors. We also dock the ligands into holoreceptors from which the ligand is removed to estimate the upper limit of precision that can be expected.

Abbreviations: CF, complementarity function; NC, normalized complementarity; PDB, Protein Data Bank; DHFR, dihydrofolate reductase; MTX, methotrexate; RMSd, root-mean-square difference between the predicted ligand coordinates and those in the crystal structure.

Received August 8, 1995; revision accepted December 5, 1995.
Address reprint requests to Author responsible for correspondence Dr. Vladimir Sobolev, Department of Plant Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.



Scheme 1. Flow chart of LIGIN program.

METHODS

The docking algorithm is implemented in a program called LIGIN. A flowchart is shown in Figure 1, and the details are described below. In summary, a large number of ligand positions are generated at or near the putative binding site of the receptor. These starting positions are generated using the full six degrees of freedom available to the ligand, except that the center of geometry of the ligand is required to fall within a user defined cube. Each of these starting positions is subjected to a complementarity optimization procedure. The docked positions obtained are further refined by optimizing the lengths of hydrogen bonds formed with the ligand. If no satisfactory docking solutions are found, the process can be repeated with one or more side chains of the receptor left out of the calculations to emulate protein flexibility.

Complementarity Function

The concepts *complementarity function* (CF) and *contact surface between atoms* have been introduced previously.[33] The CF is used to characterize the fit of a ligand in a receptor binding site, and is given here by:

$$\mathrm{CF} = S_l - S_i - E \tag{1}$$

where S_l and S_i are the sums of "legitimate" and "illegitimate" contact surfaces between ligand atoms and atoms in the receptor binding site, and

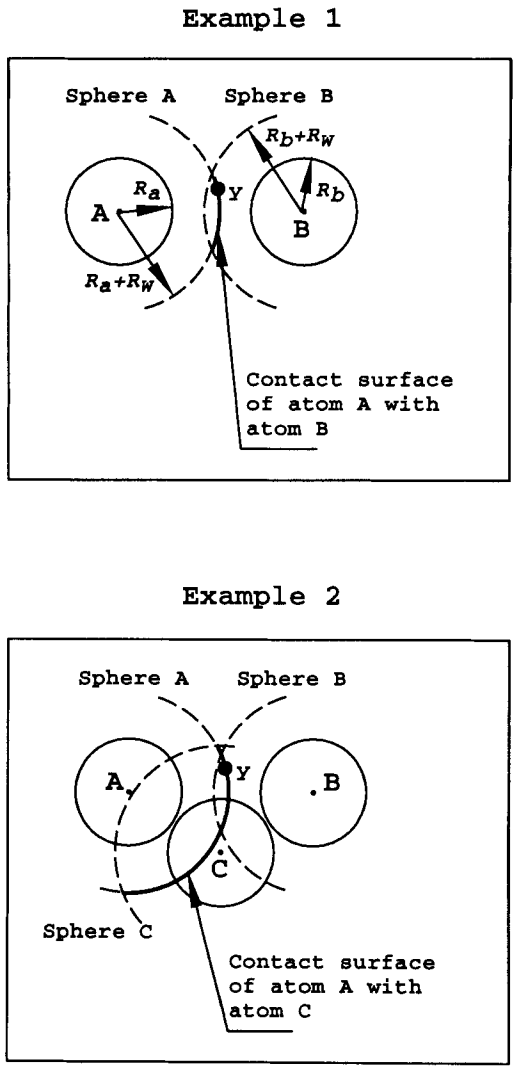


Fig. 1. Definition of atomic contact surfaces. R_a and R_b represent the van der Waals radii of atoms A and B, respectively, while R_W is the van der Waals radius of the solvent molecule (1.4 Å). Considering the contacts of atom A: If a solvent molecule at point *Y* on sphere A penetrates several atoms, this point is assigned to the contact surface of the nearest atom only. Therefore, while the distance between atoms A and B in examples 1 and 2 is the same, in example 1 point *Y* will belong to the contact surface with atom B, and in example 2, it will belong to the contact surface with atom C. (Adapted from Sobolev and Edelman.[33])

$$E = \sum_a \sum_b E_{ab} \tag{2}$$

where

$$E_{ab} = \begin{cases} 0 & \text{if } R_{ab} \geq R_0 \\ K(1/R_{ab}^{12} - 1/R_0^{12}) & \text{if } R_{ab} < R_0 \end{cases} \tag{3}$$

with $R_0 = 0.9(R_a + R_b)$ and $K = 10^6$ Å 14. R_a and R_b are the van der Waals radii[35,36] of the contacting atoms, and R_{ab} is the distance between them.

Contact surfaces are defined in Figure 1. Our definition implies that receptor atoms, up to a distance $R_a + R_b + 2R_W$ from any ligand atom, may con-

TABLE I. Legitimacy of Contact Between Atoms of Different Classes

LIGIN class*	Legitimate (+) or illegitimate (−) contact							
	Hydrophilic	Acceptor	Donor	Hydrophobic	Aromatic	Neutral	Neutral-donor	Neutral-acceptor
I. Hydrophilic	+	+	+	−	+	+	+	+
II. Acceptor	+	−	+	−	+	+	+	−
III. Donor	+	+	−	−	+	+	−	+
IV. Hydrophobic	−	−	−	+	+	+	+	+
V. Aromatic	+	+	+	+	+	+	+	+
VI. Neutral	+	+	+	+	+	+	+	+
VII. Neutral-donor	+	+	−	+	+	+	−	+
VIII. Neutral-acceptor	+	−	+	+	+	+	+	−

*I. Hydrophilic, N and O that can donate and accept hydrogen bonds (e.g., oxygen of hydroxyl group of Ser or Thr); II. Acceptor, N or O that can only accept a hydrogen bond; III. Donor, N that can only donate a hydrogen bond; IV. Hydrophobic, Cl, Br, I, and all C atoms that are not in aromatic rings and do not have a covalent bond to a hydrophilic atom; V. Aromatic, C in aromatic rings; VI. Neutral, C atoms that have a covalent bond to at least one atom of class I or two or more atoms from class II or III; N if it has covalent bonds with 3 carbons; S and F in all cases; VII. Neutral-donor, C that has a covalent bond with only one atom of class III; VIII. Neutral-acceptor, C that has covalent bond with only one atom of class II.

tribute to CF. In general, energetically favorable contacts (e.g., hydrophobic–hydrophobic) are classified as legitimate contact, whereas energetically unfavorable contacts (e.g., hydrophobic–hydrophilic) are classified as illegitimate ones.[33]

Several schemes for the division of atoms into atom classes according to their chemical properties were tried. The best results were obtained using the atomic classes and 'legitimate' and 'illegitimate' contact assignments shown in Table I. They permit the specificity of an intermolecular contact to be quantified. The classification is similar to those presented by Jiang and Kim[3] and Kasinos et al.[37]; however, in our scheme we treat carbon atoms that are bonded to polar oxygen or nitrogen atoms differently. The rationale behind this is simple. In a hydrogen bond, where the donor and acceptor form a favorable short contact, any atom covalently bonded to the donor (for example) is by sheer necessity also in close contact with the acceptor. To avoid such forced close contacts from contributing negatively to the CF (1), this type of carbon atom is classified as neutral. Neutral atoms can never contribute unfavorably to the CF. In our scheme, we also consider carbon atoms in aromatic rings as able to form favorable contacts with polar groups, in agreement with energy estimations for such systems.[38,39]

The absolute value of the CF depends on ligand size. We have normalized the CF by dividing by the solvent accessible surface of the uncomplexed ligand. The maximum normalized complementarity (NC) is 1.0 and corresponds to the case in which the ligand is 100% buried in its receptor and all atomic contacts are legitimate. Typically, NC values for small molecules in their experimentally observed positions in crystal structures fall between 0.45 and 0.75.

Generating Ligand Starting Positions

The docking procedure involves determining a ligand's position in six-dimensional space (three translational and three rotational degrees of freedom). While it is possible to consider docking of a ligand from positions around the whole protein, computation is much reduced if a limited region of the protein surface is searched. This region can often be defined on the basis of experimental data, such as the identification of specific residues involved in binding of the ligand or its analog. For our tests, we generated ligand starting positions within a $5 \times 5 \times 5$ Å^3 cube centered at the center of geometry of the ligand in the crystal structure. The program generates 500 randomly distributed ligand positions within this cube and maximizes the complementarity as a function of the six degrees of freedom of the ligand. No constraints are placed on the position of the ligands during optimization, and the center of geometry of the docked ligand can thus fall outside the initial cube. The size of the cube and number of starting points were determined empirically and can be altered by the user when needed. With the above parameters, most maxima are obtained more than once from different starting positions (Table II).

Complementarity Optimization

Since the CF (1) cannot be differentiated, gradient methods are not used for optimizing ligand positions. Instead, the flexible polyhedron search method[40] is used. To reduce central processing unit (CPU) time, maximization is not continued until convergence but rather performed in two consecutive stages of 100 function evaluations. After the first stage, the parameter step sizes are reset to the initial values (0.5 Å for translational degrees of freedom and 30° for rotational ones). This two-stage optimization procedure allows as much space to be searched as a one-stage optimization procedure carried out for a larger number of starting conformations but is less CPU time intensive.

After maximization, the maxima are clustered

TABLE II. Docking Performance of the LIGIN Program*

Example no.	Protein receptor		Ligand		Normalized complementarity†			No. of hits in cluster
	Name	PDB file	Name	PDB file	Docked	Crystal	RMSd‡ (Å)	
1	Dihydrofolate reductase	4dfr	Methotrexate	4dfr	0.53	0.46	0.8 (0.8)	4
					0.50		10.2	2
2	Dihydrofolate reductase	3dfr	Methotrexate	3dfr	0.47	0.46	0.4 (0.5)	4
					0.37		7.6	2
3	Dihydrofolate reductase + NADPH	3dfr	Methotrexate	3dfr	0.50	0.48	0.4 (0.3)	5
					0.44		1.3 (1.7)	2
4	Aconitase	7acn	Isocitrate	7acn	0.89	0.64	0.4 (0.4)	12
					0.83		4.0	9
5	Thermolysin	1tlp	Phosphoramidon	1tlp	0.64	0.54	0.6 (0.9)	8
					0.46		7.5	1
6	Thermolysin	2tmn	P-*Leu/NH_2	2tmn	0.80	0.65	1.1 (1.3)	2
					0.77		1.2 (1.1)	7
7	Penicillopepsin	1ppm	Cbz-Z-Z-L(P)-(O)PMe	1ppm	0.59	0.48	0.3 (0.7)	4
					0.40		11.5	3
8	Carbonic anhydrase II	1cim	PTS	1cim	0.83	0.76	0.3 (0.7)	9
					0.80		5.4	5
9	HIV-1 protease	4hvp	MVT-101	4hvp	0.52	0.44	0.4 (0.8)	6
					0.32		1.3 (1.4)	1
10a	Adipocyte lipid binding protein	1lif	Stearic acid	1lif	0.62	0.56	0.7 (0.7)	11
					0.59		1.5 (1.5)	5
10b	Adipocyte lipid binding protein§	1lib	Stearic acid	1lif	0.60		1.0 (0.8)	4
					0.59		9.9	1
11a	Retinol binding protein	1erb	N-ethyl retinamide	1erb	0.82	0.55	3.0	3
					0.81		1.0 (1.6)	6
11b	Retinol binding protein§	1hbq	N-ethyl retinamide	1erb	0.79		1.4 (1.3)	3
					0.79		2.4	3
12a	Ricin	1fmp	Formycin-5′-monophosphate	1fmp	0.76	0.66	0.9 (1.1)	10
					0.69		2.9	5
12b	Ricin§	2aai	Formycin-5′-monophosphate	1fmp	0.68		4.6	2
					0.68		5.7	2
					0.63		1.4 (1.7)	1
13a	Met repressor	1cmc	Corepressor s-adenosyl-methionine	1cmc	0.59	0.51	0.6 (0.7)	7
					0.57		2.6	4
13b	Met repressor¶	1cmb	Corepressor s-adenosyl-methionine	1cmc	0.56		2.1	3
					0.53		0.4 (0.6)	1
14	Streptavidin	1stp	Biotin	1stp	0.69	0.66	0.8 (1.0)	3
					0.68		7.0	5

*In all cases docking was to the holoreceptor (with ligand removed), except in Examples 10b, 11b, 12b, and 13b, in which docking was to the aporeceptor. These complexes can be viewed and their PDB coordinates obtained by accessing www page http: swift.embl-heidelberg.de/ligin/.

†Normalized complementarity (NC) is defined as the complementarity function (CF) divided by the solvent-accessible surface of the uncomplexed ligand. NC is given for the best docked positions, and for the ligand in the crystal.

‡Values represent the root-mean-square difference between the original crystallographic position of the ligand and its position following the full docking procedure, i.e., CF maximization + H-bond length optimization. Intermediate RMSd values (in parentheses) are the result of CF maximization alone.

§Motion of one protein side chain was allowed for during docking.

¶Motion of two protein side chains was allowed for during docking.

and sorted. This is done by taking the structure with the highest complementarity and defining all hits having a root-mean-square difference (RMSd) < 1 Å from this structure as belonging to one cluster. The program then continues with the best structure from those remaining and forms the next cluster using the same procedure. The number of structures in the cluster is recorded.

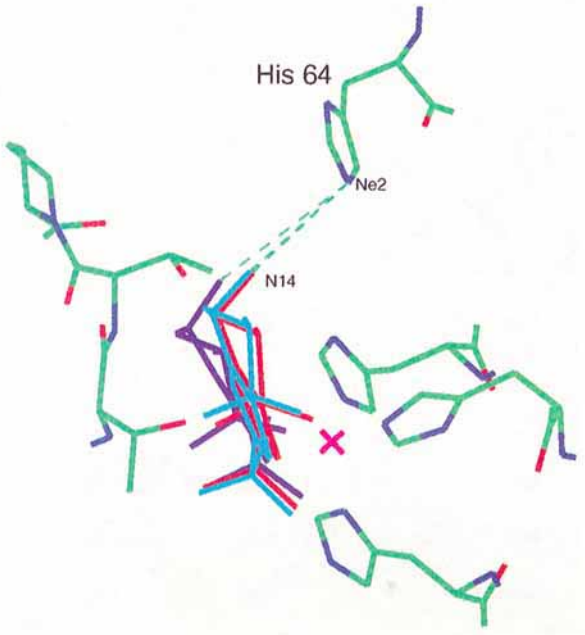


Fig. 2. Docking the PTS inhibitor in the binding site of carbonic anhydrase II. Only residues that have atoms in the hydrogen bond list generated by the LIGIN program are shown. The experimental ligand position is shown in red. The docked positions of the ligand before (cyan) and after (blue) hydrogen-bond length optimization have RMSd of 0.7 Å and 0.3 Å, respectively. The improvement on optimization is due to shortening of the Ne2 His-64–N14 contact from 4.7 Å to 3.9 Å. This contact distance is too long for a normal hydrogen bond but shows how the program can optimize favorable, polar interactions as well as hydrogen bonds. The Zn ion is shown by a pink cross.

Optimization of hydrogen bond lengths

Following CF maximization, a list of candidate hydrogen bonds is generated. All donor–acceptor pairs that share a contact surface are considered. These contacts are not necessarily hydrogen bonds, but for ease of nomenclature, and since most of them are in practice real hydrogen bonds, we will call them such. If the ligand–receptor complex is tightly packed (as in Fig. 1, example 2), only short hydrogen bonds are obtained, whereas in less tightly packed complexes (as in Fig. 1, example 1), very long hydrogen bonds are included. This formalism allows us to consider implicitly the presence of water-mediated hydrogen bonds and other short-range, favorable polar interactions.

The ligand position is refined by optimizing the lengths of the hydrogen bonds present in the candidate list. New hydrogen bonds are not added, nor are existing ones removed, during this procedure. Optimization is performed by minimizing the function:

$$Q = HB + E \tag{4}$$

with respect to ligand position in six-dimensional space. The term E is the same as in Eq. (1) and (2), and HB is defined as:

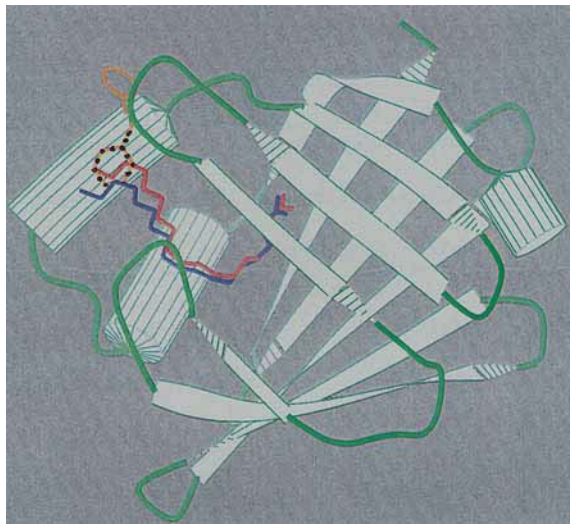


Fig. 3. Docking of stearic acid in adipocyte lipid binding protein. The position of the ligand in the structure of the holoprotein complex is shown in red, while that of the ligand docked in the apoprotein is shown in blue. Docking of the ligand in the apoprotein could only be achieved when allowance for the motion of Phe-57 out of the binding site was made. The crystal structures show that Phe-57 does indeed reorient on ligand binding and its positions in the holoprotein (orange) and apoprotein (orange dotted) crystal structures are shown.

$$HB = \sum_a \sum_b HB_{ab} \tag{5}$$

where

$$HB_{ab} = \begin{cases} 0 & \text{if } R_{ab} \le D_0 \\ 0.5(D_0^2 - R_{ab}^2) & \text{if } R_{ab} > D_0 \end{cases} \tag{6}$$

and $D_0 = 2.9$ Å. This function does not provide a complete description of the geometry of the hydrogen bonds but merely allows their lengths to be optimized. In practice, this fine tuning usually improves the accuracy of the final structure of the complex but rarely leads to large conformational changes (Table II).

Flexibility

Most proteins in the Brookhaven PDB (Release #69, July 1994) have very similar conformations in the complexed and the uncomplexed form. When differences between the apo- and the holoreceptor are observed, they are normally small, concerted displacements of a stretch of residues. In a few cases, the side chains of one or two residues in the binding pocket adopt a different rotamer upon ligand binding. To consider such a case, the program can be instructed to neglect the contribution to the CF (1) of one or more side chains. The program firstly estimates the degree of overlap between the ligand and the side chain of every residue that would move out of the binding site if the residue were to adopt another rotamer. The CF is then calculated for all atoms of all residues excluding the one (or more) res-

idues with the highest overlap(s). For these residues, only the backbone and C_β atoms are taken into account. The maximum number of residues treated in this way is user defined and is usually set ≤ 2. Residues thus excluded are listed for every docked conformation. The user is advised to inspect visually if flexibility of these residues is a reasonable assumption.

At present, more extensive receptor motion and ligand flexibility are not yet taken into account.

Improving Ligand Binding Properties

After docking, all atomic contacts between the ligand and the protein are listed along with their contribution to the CF (1). A table is automatically provided showing what the normalized complementarity would be after reclassifying each atom into an atom from each of the seven other classes. The NC is re-calculated for each one-atom change without refinement of the position of the ligand in the binding site. Improved complementarity indicates greater affinity for the altered, versus original, docked ligand.

Implementation

The LIGIN program is sufficiently fast to allow for the evaluation of the binding of a large number ligands to a protein. CPU time depends mainly on the number of atoms and size of the search cube. For example, starting from 500 points, a search for a ligand consisting of 20 atoms requires about 20 minutes on a Silicon Graphics Indigo Workstation (150 MHZ IP22 Processor) or 3 minutes on a DEC (TURBO-LASER) SERVER 8200 5/300. The LIGIN program is interfaced to the WHAT IF molecular modeling and molecular display software package[41] for easy visualization of results.

RESULTS AND DISCUSSION

The results of docking 14 receptor–ligand complexes are shown in Table II. All results were obtained with a single set of parameters and atom classes. The docking procedure in apoproteins includes an option that permits atomic overlap for a small, specified, number of protein residues. For each protein–ligand pair, the RMSd characterizing the accuracy of the docked ligand positions is given for the two (or three) positions with the best NC. For reference, the NC for the ligands in their original crystallographic positions in the holoprotein is also given. The examples illustrate specific aspects addressed by our method: the importance of contact specificity; hydrogen bond length optimization; protein side chain flexibility; use of poorly defined atom types (e.g., metal ions); clarification of features involved in complex stabilization; and improvement of ligands. In the following sections, we demonstrate each of these features.

Importance of Atom Classification and Contact Specificity

Sometimes it is possible to dock a ligand into a protein binding site by only optimizing geometric shape complementarity. In general, however, and particularly when the binding pocket is spacious, many possible ligand orientations will be found when only a geometric criterion is used. In these cases, the chemical complementarity of the contacting atoms must be considered to distinguish the correct ligand binding position. Computation of the CF is an effective way to do this and provides a simple quantification of contact specificity.

A pertinent example is the *Escherichia coli* dihydrofolate reductase-methotrexate (DHFR-MTX) complex (PDB-id = 4DFR),[42] which, on the basis of geometry alone, allows for many, seemingly equally good ligand orientations.[13,17,21,32] With our approach, in which docking is performed with eight classes of atom types, a single and correct orientation is readily obtained (Table II, example 1). The crucial factor here proved to be the introduction of multiple neutral and aromatic atom classes (Table I) in the scheme. Likewise, our scheme of eight atom classes is able to dock MTX into the *Lactobacillus casei* DHFR (PDB-id = 3DFR)[42] in both the presence and absence of NADPH (Table II, examples 2, 3). In both cases the MTX positions so obtained are very close to that of the ligand in the crystal structure (RMSd of 0.4 Å).

All atoms that occur in polypeptide chains can be assigned to one of the eight atom classes in Table I. Other kinds of atoms can, however, be involved in protein-ligand interactions. For example, metal ions can play an essential role in ligand binding. Since it is not generally clear to which atom class metal ions should be assigned, we have omitted them from the calculation of complementarity. Nevertheless, ligands that have metal ions in their binding pocket are usually docked correctly by our method. It seems that neglecting an interaction is to be preferred over including it incorrectly. For example, in the aconitase-isocitrate complex (PDB-id = 7ACN)[43] the whole 4Fe-4S cluster was neglected from the complementarity calculation. Nevertheless, the structure with the best complementarity has a RMSd of only 0.4 Å (Table II, example 4).

In contrast to the binding site in aconitase, which is completely buried in the protein,[16] the Zn ion-containing binding site in thermolysin is exposed and thus provides a more rigorous test case for a docking program. Docking phosphoramidon (Table II, example 5) or P-*Leu/NH^2 (Table II, example 6) into the thermolysin binding site gave a RMSd for the best docked positions of 0.6 Å and 1.1 Å, respectively. Thus, the correct position was found even for a relatively small inhibitor such as P-*Leu/NH^2 (PDB-id = 2TMN).[44]

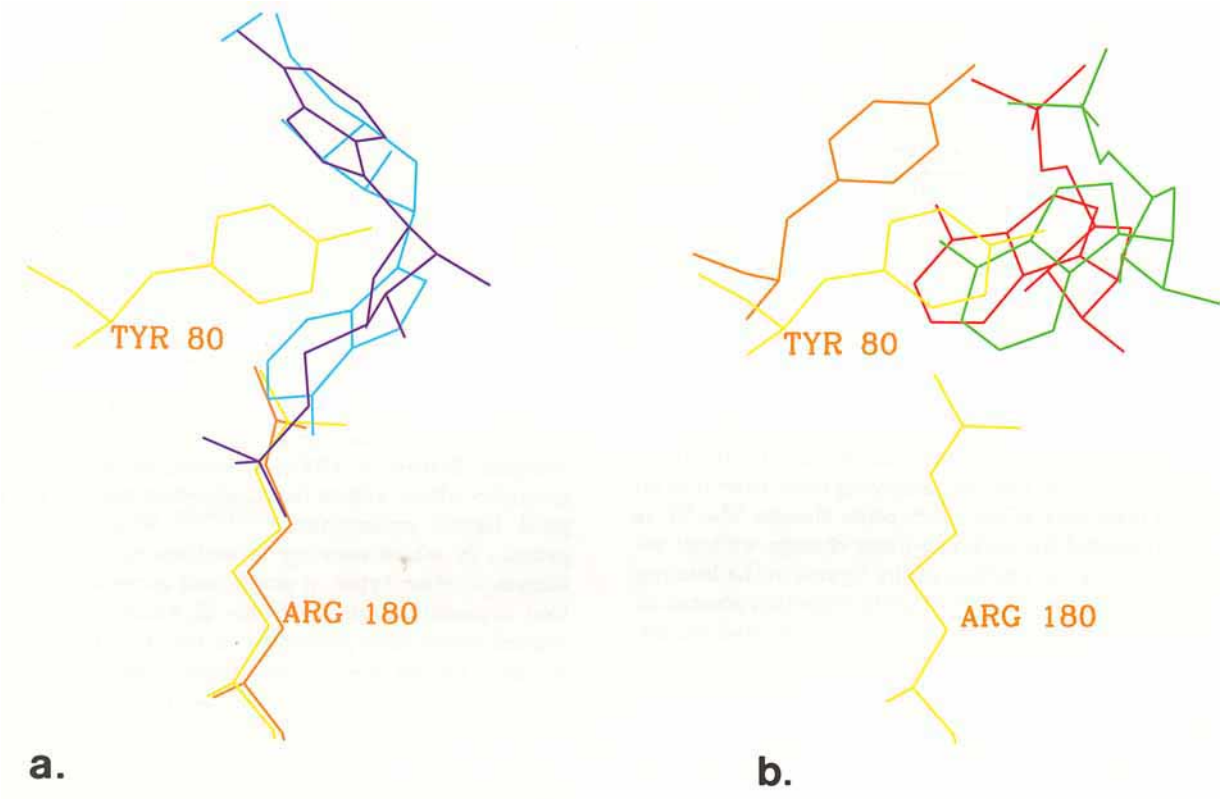


Fig. 4. Docking of formycin-5′-monophosphate in the binding pocket of ricin. It is necessary to allow for the motion of one side chain of ricin to dock the ligand. For the ligand positions with the best (cyan) and second best (blue) complementarity (**a**), the program identified Arg-180 as the residue that is assumed to move, while for the third best position (green, **b**) Tyr-80 was identified. From mutant studies,[54–57] it is known that Arg-180 plays an important role in the stability of the protein and, as shown in a, does not move much on complex formation [compare its position in holo- (orange) and apoprotein (yellow)]. However, as seen in b, Tyr-80 does move extensively on ligand binding [compare its position in holo- (orange) and apoprotein (yellow)].[56] The third best ligand position (green) agrees reasonably well (RMSd = 1.4 Å) with its position in the crystal structure (red).

Hydrogen Bonds and Polar Contacts

The hydrogen bond optimization procedure can improve the accuracy of docking. For example, following this procedure, the RMSd of the peptide-type ligand Cbz-A-A-L(P)-(O)P-OMe in penicillopepsin (PDB-id = 1PPM)[45] improved from 0.7 Å to 0.3 Å (Table II, example 7), and four hydrogen bonds in the complex decreased in length (from 3.1–3.4 Å to 2.9–3.0 Å). This optimization procedure also treats favorable polar interactions that cannot be considered as conventional hydrogen bonds because they are too long. This is demonstrated by the docking of the PTS inhibitor in carbonic anhydrase II (PDB-id = 1CIM).[46] Optimization results in an improved RMSd (from 0.7 Å to 0.3 Å; Table II, example 8) due to reduction (from 4.7 Å to 3.9 Å) in the distance between the two nitrogen atoms shown in Figure 2. Thus, the hydrogen bond optimization step provides a simple way to make small adjustments to docked ligand positions and generally improve favorable polar contacts.

Long favorable polar contacts can also represent hydrogen bonds that would be observed explicitly if adjustment of protein conformation were permitted, or if a water molecule were to mediate the interaction. This is demonstrated by the HIV-1–MVT-101 complex (Table II, example 9; PDB-id = 4HVP[47]; a complete listing of residues in contact with the ligand and of putative H-bonds is available on www page http://swift.embl-heidelberg.de/ligin/.) In the docked complex, Asp-25 in HIV-1 protease is listed as H-bonded to the inhibitor at a contact length of 3.6 Å. This is in agreement with Thompson et al.,[48] who found this residue forming a hydrogen bond of 2.7 Å with a synthetic analog of MVT-101 in a crystal structure. Another example is Ile-50, which is listed in the docked complex as having a polar contact of 5.1Å with the inhibitor. The importance of this long contact in stabilizing the complex is substantiated by three independent crystal structures[47–49] that show a water molecule mediating the contact. The water molecule is considered to be an integral part of the binding pocket.[50]

Protein Side Chain Flexibility

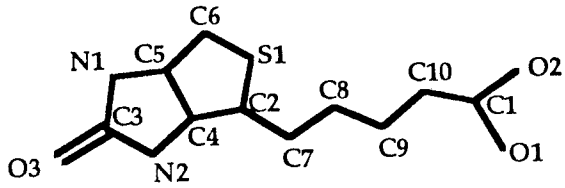
Sometimes binding of a ligand to a receptor involves a local reorientation of one or two side chains in the apoprotein pocket to avoid steric clashes. Our

TABLE III. Normalized Complementarity for Biotin-Streptavidin Complex Following Reclassification of Ligand Atoms*

Biotin atoms†		LIGIN atom class‡							
PDB atomic label	LIGIN class	I. Hydrophilic	II. Acceptor	III. Donor	IV. Hydrophobic	V. Aromatic	VI. Neutral	VII. Neutral-donor	VIII. Neutral-acceptor
C7	IV	0.76	0.72	**0.75**	0.68	0.76	0.75	0.75	0.73
C8	IV	0.71	0.71	0.71	0.68	0.71	0.71	**0.71**	0.71
O3	II	0.68	0.68	**0.56**	**0.48**	0.68	0.68	0.57	0.68

*Underscored value is the NC for the best docked position of the ligand before atom substitution. Values shown in bold are discussed in text.

†Atomic labels are taken from the 1STP PDB file:



‡LIGIN atom classification is defined in Table I.

method takes this phenomenon into account by simply omitting clashing side chains from calculation of the CF. Examples 10–13 of Table II summarize data for several ligands that were docked into aporeceptors by this method.

The first case is stearic acid in the binding site of the adipocyte lipid-binding protein (PDB-id = 4LIF).[51] In the holoreceptor (Table II, example 10a), docking is straightforward, and the RMSd between the crystallized and docked coordinates is 0.7 Å. In the aporeceptor (Table II, example 10b), the ligand could not be fitted into the active site of the protein (NC = 0.27) unless overlap with any one receptor side chain was allowed. The program chose to neglect Phe-57 (the residue that clashed the most with the ligand on docking), yielding an NC of 0.60 and RMSd of 1.0 Å. Indeed, different positions for Phe-57 have been observed in the crystal structures of the apo- and holoreceptor.[51] In Figure 3, the overall good fit for the stearic acid ligand docked by our method (blue), compared with the experimental position from the crystal structure (red), is shown, as well as the markedly different orientation of the Phe-57 residue in the crystalline holo (orange) and apo receptor (orange dotted) structures.

A further example of this type is given by docking of N-ethyl retinamide into retinol binding apoprotein (PDB-id = 1ERB and 1HBQ[52]; Table II, example 11). Docking is possible only following motion of Phe-36 (accompanied by an increase in NC from 0.22 to 0.79).

However, not all cases requiring local reorientation involve low NC values for the best docked positions in the apoprotein. For example, formicin-5′-monophosphate could not be docked correctly into the apo form of ricin[53] when overlap with receptor side chains was not allowed (NC = 0.66 but RMSd = 5.1). This receptor-ligand complex thus exemplifies a case in which specific prior knowledge of protein flexibility would be required for our method to be predictive. The correct ligand position was readily found in the holoreceptor (Table II, example 12a), but in the aporeceptor, even when overlap with one residue was allowed, only the third hit was close to the experimental one (Table II, example 12b). In the best two cases, the ligand overlaps with the side chain of Arg-180 (Fig. 4A), while in the third, it overlaps with the side chain of Tyr-80 (Fig. 4B). This would suggest that either Tyr-80 or Arg-180 adopts another conformation when the ligand binds. Since mutagenesis studies indicate that Arg-180 is vital for the stability of the protein,[54–57] one is led to consider reorientation of Tyr-80.

Another example requiring prior knowledge of protein flexibility is the Met repressor DNA-binding regulatory protein (PDB-id = 1CMB and 1CMC).[58] Docking of the corepressor ligand (s-adenosylmethionine) into the apoform of this protein requires motion of both the Phe-65 and Glu-2 side chains (Table II, example 13). We caution, however, that neglecting two side chains will only be useful if, as in this case,[58] experimental evidence shows that the ligand actually binds the receptor and satisfactory docking solutions cannot be found otherwise.

Ligand Design

The LIGIN program provides lists of: residues in contact with the ligand; putative hydrogen bonds between receptor and ligand; and NC values for the best docked position following reclassification of atoms in the ligand. These lists clarify factors governing complex formation and assist in the design of improved ligands. For example, in the streptavidin–biotin complex (PDB-id = 1STP),[59] the best docked position for biotin has an NC of 0.68 (Table II, example 2). However, as shown in Table III, replacement of hydrophobic atom C7 by any of the other 7 atom classes would lead to better complementarity

(0.72–0.76). Examination of the crystal structure and the list of atomic contacts of the docked ligand suggests that replacement by a H-bond donor atom would result in better interactions with nearby backbone and side chain oxygen atoms, yielding NC = 0.75. In this case, hydrophobic atom C8 necessarily becomes a neutral donor, which would also yield an improvement in complementarity (NC = 0.71). In contrast, replacement of the ureido oxygen atom (O3), a hydrogen bond acceptor, with a hydrophobic or H-bond donor atom would reduce the NC for this atom from 0.68 to 0.48 or 0.57, respectively. These predicted changes in complementarity are in full agreement with experimental binding data and free energy calculations[60] for biotin, thiobiotin, and iminobiotin. In the last two compounds, the O3 ureido oxygen is, respectively, replaced by a more hydrophobic sulfur atom and by a hydrogen bond donor (NH group) and, indeed, they have a weaker affinity for streptavidin. The full list of CF changes following atom substitution for biotin is available on www page http://swift.embl-heidelberg.de/ligin/.

CONCLUSIONS

Surface complementarity between ligand and receptor is the guiding principle for ligand docking in our approach. Surface complementarity incorporates information about the shape and chemical nature of the atoms of the interacting molecules. Here we have refined the definition of surface complementarity and shown how it can be used to dock ligands (taking some account of receptor flexibility), analyze binding modes, and suggest new ligands. We have tested the program for a wide range of ligand–protein complexes for which the coordinates of the holoreceptor and, in some cases, aporeceptor, are available. The advantages of using surface complementarity are particularly apparent when ligands are docked into spacious receptor pockets. In such cases, our method performs well because the definition of contact surface allows loose contacts (up to a solvent-separated distance) to be considered and it optimizes favorable polar contacts, both loose and tight.

In its present form, the LIGIN program does not take into account flexibility of the ligand, but there are no intrinsic limitations other than computer time to its incorporation. Local flexibility in protein receptors is rarely observed in the examples available from the Brookhaven Protein Data Bank. In the few cases in which local side chain flexibility is required for ligand docking, the program is helpful in determining a ligand position close to that experimentally observed and identifies which residue(s) reorient on ligand binding.

In our schema, atom types are distributed into eight classes. While it would be possible to introduce further classes, we find that if the chemical characteristics of an atom are unclear, it is best to treat this atom as neutral if it is in the ligand, and to remove it entirely from the calculation if it is part of the target receptor. Surface complementarity can be calculated with a different atom class substituted at each atom position in the docked ligand. This permits an analysis of all contributions to the complementarity function. This, in turn, is useful for understanding the atomic intricacies of a ligand–receptor complex and in designing improved ligands.

The WHAT IF and LIGIN programs are available from G.V. and V.S., respectively.

ACKNOWLEDGMENTS

The authors thank Drs. Chris Sander, Rob Hooft, and Yehudit Weisinger-Lewin for many helpful discussions and Michael Degenhardt for running tests of the LIGIN program. This work was partly supported by the BMFT. V.S. thanks the EMBL for providing financial support, and M.E. acknowledges the support of the Forschheimer and Wilstätter Centers at the Weizmann Institute of Science.

REFERENCES

1. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. A geometric approach to macromolecule–ligand interactions J. Mol. Biol. 161:269–288, 1982.
2. Connolly, M.L. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. Biopolymers 25:1229–1247, 1986.
3. Jiang, F., Kim, S.-H. "Soft docking": Matching of molecular surface cubes. J. Mol. Biol. 219:79–102, 1991.
4. Masek, B.B., Merchant, A., Matthew, J.B. Molecular skins: A new concept for quantitative shape matching of a protein with its small molecule mimics. Proteins 17:193–202, 1993.
5. Yue, S.-Y. Distance-constrained molecular docking by simulated annealing. Protein Eng. 4:177–184, 1990.
6. Boehm, H.-J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. J. Comput. Aided Mol. Design 6:61–78, 1992.
7. Bohacek, R.S., McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a *de novo* design method incorporating combinatorial growth. J. Am. Chem. Soc. 116:5560–5571, 1994.
8. Goodford, P.J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem. 28:849–857, 1985.
9. Wodak, S.J., DeCrombrugghe, M., Janin, J. Computer studies of interactions between macromolecules. Prog. Biophys. Mol. Biol. 49:29–63, 1987.
10. Miranker, A., Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. Proteins 11:29–34, 1991.
11. Caflisch, A., Niederer, P., Anliker, M. Monte Carlo docking of oligopeptides to proteins. Proteins 13:223–230, 1992.
12. Abagyan, R., Totrov, M., Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. J. Comput. Chem. 15:488–506, 1994.
13. Mizutani, M.Y., Tomioka, N., Itai, A. Rational automatic search method for stable docking models of protein and ligand. J. Mol. Biol. 243:310–326, 1994.
14. Knegtel, R.M.A., Antoon, J., Rullmann, C., Boelens, R., Kaptein, R. MONTY: A Monte Carlo approach to protein–DNA recognition. J. Mol. Biol. 235:318–324, 1994.
15. Leach, A.R. Ligand docking to proteins with discrete side-chain flexibility J. Mol. Biol. 235:345–356, 1994.
16. Goodsell, D.S., Olson, A.J. Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202, 1990.

17. Hart, T.N., Read, R.J. A multiple-start Monte Carlo docking method. Proteins 13:206–222, 1992.
18. Di Nola, A., Roccatano, D., Berendsen, H.J.C. Molecular dynamics simulation of the docking of substrates to proteins. Proteins 19:174–182, 1994.
19. Jones, G., Willett, P., Glen, R.C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. 245:43–53, 1995.
20. Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M., McCammon, J.A. A molecular mechanics/grid method for evaluation of ligand–receptor interactions. J. Comput. Chem. 16:454–464, 1995.
21. Meng, E.C., Shoichet, B.K., Kuntz, I.D. Automated docking with grid-based energy evaluation. J. Comput. Chem. 13:505–524, 1992.
22. Bacon, D.J., Moult, J. Docking by least-squares fitting of molecular surface patterns. J. Mol. Biol. 225:849–858, 1992.
23. Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R. A geometry-based suite of molecular docking processes. J. Mol. Biol. 248:459–477, 1995.
24. Cherfils, J., Janin, J. Protein docking algorithms: Simulating molecular recognition. Curr. Opin. Struct. Biol. 3:265–269, 1993.
25. Kuntz, I.D., Meng, E.C., Shoichet, B.K. Structure-based molecular design. Acc. Chem. Res. 27:117–123, 1994.
26. Cherfils, J., Duquerroy, S., Janin, J. Protein–protein recognition analyzed by docking simulation. Proteins 11:271–280, 1991.
27. Shoichet, B.K., Kuntz, I.D. Protein docking and complementarity. J. Mol. Biol. 221:327–346, 1991.
28. Wang, H. Grid-search molecular accessible surface algorithm for solving the protein docking problem. J. Comput. Chem. 12:746–750, 1991.
29. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., Vakser, I.A., Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. U.S.A. 89:2195–2199, 1992.
30. Helmer-Citterich, M., Tramontano, A. PUZZLE: A new method for automated protein docking based on surface shape complementarity. J. Mol. Biol. 235:1021–1031, 1994.
31. Vakser, I.A., Aflalo, C. Hydrophobic docking: A proposed enhancement to molecular recognition techniques. Proteins 20:320–329, 1994.
32. Meng, C.E., Kuntz, I.D., Abraham, D.J., Kellogg G.E. Evaluating docked complexes with the HINT exponential function and empirical atomic hydrophobicities. J. Comput. Aided. Mol. Design 8:299–306, 1994.
33. Sobolev, V., Edelman, M. Modeling the quinone-B binding site of the photosystem-II reaction center using notions of complementarity and contact surface between atoms. Proteins 21:214–225, 1995.
34. Bernstein, F.C., Koetzle, T.F., Williams G.J.B., Meyer, E.F., Rodgers, J.R., Kennand, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 112: 535–542, 1977.
35. Bondi, A. van der Waals volumes and radii. Phys. Chem. 68:441–451, 1964.
36. Lee, B, Richards, F.M. The interpretation of protein structure: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 197.
37. Kasinos, N., Lilley, G.A., Subbarao, N., Haneef, I. A robust and efficient automated docking algorithm for molecular recognition. Protein Eng. 5:69–75, 1992.
38. Levitt, M., Perutz, M.F. Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol. 201:751–754, 1988.
39. Basharov, M.A., Vol'kenstein, M.V., Golovanov, I.B., Nauchitel', V.V., Sobolev, V.M. The fragment-fragment interaction method. Part 1. Estimating molecule-molecule interactions. J. Gen. Chem., U.S.S.R. 59:435–447, 1989.
40. Himmelblau, D.M. "Applied Nonlinear Programming." New York: McGraw-Hill, 1972.
41. Vriend, G. WHAT IF: A molecular modelling and drug design program. J. Mol. Graph. 8:52–56, 1990.
42. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C., Kraut, J. Crystal structure of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 A resolution. J. Biol. Chem. 25:13650–13662, 1982.
43. Lauble, H., Kennedy, M.C., Beinert, H., Stout, C.D. Crystal structures of aconitase with isocitrate and nitroisocitrate bound. Biochemistry 31:2735–2748, 1992.
44. Tronrud, D.E., Monzingo, A.F., Matthews, B.W. Crystallographic structural analysis of phosphoramidates as inhibitors and transition-state analogs of thermolysin. Eur. J. Biochem. 157:261–268, 1986.
45. Fraser, M.E., Strynadka, N.C.J., Bartlett, P.A., Hanson, J.E., James, M.N.G. Crystallographic analysis of transition-state mimics bound to penicillopepsin: Phosphorus-containing peptide analogues. Biochemistry 31:5201–5214, 1992.
46. Smith, G.M., Alexander, R.S., Christianson, D.W., McKeever, B.M., Ponticello, G.S., Springer, J.P., Randal, W.C., Baldwin, J.J., Habecker, C.N. Positions of His-64 and a bound water in human carbonic anhydrase II upon binding three structurally related inhibitors. Protein Sci. 3:118–125, 1994.
47. Miller, M., Schneider, J., Sathyanarayana, B.K., Toth, M.V., Marshall, G.R., Clawson, L., Selk, L., Kent, S.B.H., Wlodawer, A. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. Science 246:1149–1152, 1989.
48. Thompson, S.K., Murthy, K.H.M., Zhao, B., Winborne, E., Green, D.W., Fisher, S.M., DesJarlais, R.L., Tomaszek, T.A., Meek, T.D., Gleason, J.G., Abdel-Meguid, S.S. Rational design, synthesis, and crystallographic analysis of a hydroxyethylene-based HIV-1 protease inhibitor containing a heterocyclic P_1'-P_2' amide bond isostere. J. Med. Chem. 37:3100–3107, 1994.
49. Jhoti, H., Singh, O.M.P., Weir, M.P., Cooke, R., Murray-Rust, P., Wonacott, A. X-ray crystallographic studies of a series of penicillin-derived asymmetric inhibitors of HIV-1 protease. Biochemistry 33:8417–8427, 1994.
50. Gustchina, A., Sansom, C., Prevost, M., Richelle, J., Wodak, S.Y., Wlodawer, A., Weber, I.T. Energy calculations and analysis of HIV-1 protease-inhibitor crystal structures. Protein Eng. 7:309–317, 1994.
51. Xu, Z., Bernlohr, D.A., Banaszak, L.J. The adipocyte lipid-binding protein at 1.6-Å resolution. J. Biol. Chem. 268: 7874–7884, 1993.
52. Zanotti, G., Malpeli, G., Berni, R. The interaction of N-ethyl retinamide with plasma retinol-binding protein (RBP) and the crystal structure of the retinoid-RBP complex at 1.9-Å resolution. J. Biol. Chem. 268:24873–24879, 1993.
53. Monzingo, A.F., Robertus, J.D. X-ray analysis of substrate analogs in the ricin A-chain active site. J. Mol. Biol. 227: 1136–1145, 1992.
54. Frankel, A., Welsh, P., Richardson, J., Robertus, J.D. Role of arginine 180 and glutamic acid 177 of ricin toxin A chain in enzymatic inactivation of ribosomes. Mol. Cell. Biol. 10:6257–6263, 1990.
55. Ready, M.P., Kim, Y., Robertus, J.D. Site-directed mutagenesis of ricin A-chain and implications for the mechanism of action. Proteins 10:270–278, 1991.
56. Kim, Y., Robertus, J.D. Analysis of several key active site residues of ricin A chain by mutagenesis and X-ray crystallography. Protein Eng. 5:775–779, 1992.
57. Weston, S.A., Tucker, A.D., Thatcher, D.R., Derbyshire, D.J., Pauptit, R.A. X-ray structure of recombinant ricin A-chain at 1.8 Å resolution. J. Mol. Biol. 244:410–422, 1994.
58. Rafferty, J.B., Somers, W.S., Saint-Girons, I., Phillips, S.E.V. Three-dimensional crystal structure of *Escherichia coli* Met repressor with and without corepressor. Nature 341:705–710, 1989.
59. Weber, P.C., Ohlendorf, D.H., Wendoloski, J.J., Salemme, F.R. Structural origins of high-affinity biotin binding to streptavidin. Science 243:85–88, 1989.
60. Miyamoto, S., Kollman, P.A. Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. Proteins 16:226–245, 1993.