J. Mol. Biol. (1996) **261**, 470–489

JMB

A Fast Flexible Docking Method using an Incremental Construction Algorithm

Matthias Rarey[1]*, Bernd Kramer[1], Thomas Lengauer[1] and Gerhard Klebe[2]

[1]*German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI), Schloß Birlinghoven, 53754 Sankt Augustin, Germany*

[2]*BASF AG, Main Laboratory Carl-Bosch Strasse, 67056 Ludwigshafen, Germany*

**Corresponding author*

We present an automatic method for docking organic ligands into protein binding sites. The method can be used in the design process of specific protein ligands. It combines an appropriate model of the physico-chemical properties of the docked molecules with efficient methods for sampling the conformational space of the ligand. If the ligand is flexible, it can adopt a large variety of different conformations. Each such minimum in conformational space presents a potential candidate for the conformation of the ligand in the complexed state. Our docking method samples the conformation space of the ligand on the basis of a discrete model and uses a tree-search technique for placing the ligand incrementally into the active site. For placing the first fragment of the ligand into the protein, we use hashing techniques adapted from computer vision. The incremental construction algorithm is based on a greedy strategy combined with efficient methods for overlap detection and for the search of new interactions. We present results on 19 complexes of which the binding geometry has been crystallographically determined. All considered ligands are docked in at most three minutes on a current workstation. The experimentally observed binding mode of the ligand is reproduced with 0.5 to 1.2 Å rms deviation. It is almost always found among the highest-ranking conformations computed.

Keywords: molecular docking; flexible docking; protein ligand interaction; molecular flexibility; drug design

Introduction

Ligands with low molecular weight can bind specifically to biological macromolecules. As a consequence, they interfere with biochemical pathways, e.g. through enzyme inhibition or the modulation of signal transduction and can be used as drugs. 3D structures of new therapeutically relevant target proteins are becoming available at a dramatically increasing rate either through structure determination by X-ray crystallography and NMR spectroscopy or, with lower accuracy, through homology modeling.

Due to this growing structural knowledge the docking problem is becoming essential in rational drug design: How can a particular ligand be docked into the binding pocket of a given protein? Increasing interest is attributed to the automatic screening of ligand databases by computational methods in the lead discovery process. Once a ligand has been successfully docked into a protein its affinity has to be estimated.

To the best of our knowledge, the first approach to this problem was published by Platzer *et al*. (1972a,b) who performed conformational energy calculations for a set of substrates binding to chymotrypsin. Although the method is described only for this specific application, it contains the two major parts required for a docking tool: the prediction of the geometry and an estimation of the free energy of protein–ligand complexes.

Kuntz *et al*. (1982) presented the DOCK program. DOCK is based on a sphere-matching procedure, in which the conformation of the ligand is kept fixed during the computation in order to keep the problem computationally tractable on the computers then available. Since then, many docking methods have been published (see Kuntz, 1992; Blaney & Dixon, 1993; Kuntz *et al*., 1994, for reviews). Only a few of them consider ligand

Abbreviations used: CSD, Cambridge Structural Database; PDB, Protein Data Base; rms, root-mean-square.

0022–2836/96/330470–20 $18.00/0

flexibility directly. Here, we will summarize only such methods.

Partial ligand flexibility has been integrated into the DOCK approach by dividing the ligand into a few fragments, which are docked independently and subsequently fused (DesJarlais *et al*., 1986). The hinge-bending algorithm (Sandak *et al*., 1995) is based on the same strategy.

In principle, the docking problem can be tackled by applying energy minimization or simulation techniques (Di Nola *et al*., 1994; Luty *et al*., 1995). The main disadvantages of these algorithms are that the results depend on the initial placement of the ligand and that the algorithms do not explore the solution space exhaustively. Goodsell & Olson (1990) apply simulated annealing to the docking problem. Although this global optimization technique in principle avoids getting trapped in local minima, differences in results obtained using simulated annealing on different starting solutions show that similar problems as with greedy minimization techniques occur in practice.

Distance geometry affords a closed-form representation of the conformational space of a molecule (Crippen & Havel, 1988) in the form of a matrix containing upper and lower bounds for all atom-atom distances. Some approaches for flexible ligand docking are based on this concept (Ghose & Crippen, 1985; Smellie *et al*., 1991; Billeter *et al*., 1987). Unfortunately, the distance matrix is a highly redundant description of the conformational space of the ligand and thus, only a small portion of all distance matrices which are consistent with the initially computed upper and lower bounds represent meaningful molecular conformations. Transferring a ligand from distance space to Euclidean space is difficult. Because meaningful energy functions including conformational energy of the ligand and terms restricting the volume of protein–ligand overlap cannot be applied to a ligand represented in distance space, rejecting out bad solutions in an early stage of computation is impossible.

Mizutani presents an algorithm which is based on a complete combinatorial search over all possible matches between hydrogen bond patterns in the ligand and receptor, respectively (Mizutani *et al*., 1994). Distance filters are applied in order to limit the search time. Nevertheless, the algorithm's asymptotic run time behavior is exponential in the number of hydrogen-bonding partners.

Recently, two flexible docking methods based on a genetic algorithm have been published (Oshiro *et al*., 1995; Jones *et al*., 1995). Both approaches produce predictions with high quality for the test cases. Unfortunately, the run time for these predictions is relatively high, because one has to perform several runs of the genetic algorithm in order to get a reliable prediction.

The most promising approach to docking from our point of view is the algorithm of Leach & Kuntz (1992). The algorithm is based on the incremental construction method, which can be described as follows. In a first phase, a pre-selected base or anchor fragment of the ligand is placed into the active site independently of the rest of the ligand. Leach considers the base fragment as rigid and uses a variant of the DOCK algorithm, which is based on hydrogen bonding, for this purpose. In the second phase, the remaining fragments of the ligand are added to the initially placed anchor fragment. For this incremental construction of the complex, Leach uses a backtracking algorithm combined with force-field calculations.

Incremental construction is mainly used in the area of *de-novo* ligand design (see Lewis & Leach, 1994; Colman, 1994 for overviews). Tools following this principle are the peptide design tool GROW (Moon & Howe, 1991), the *de-novo* design tools LUDI (one possible mode; Böhm, 1992a,b), LEGEND (Nishibata & Itai, 1991), GenStar (Rotstein & Murcko, 1993), and SPROUT (Gillet *et al*., 1990, 1994).

In this paper, we present a new docking method, and report on experiences with a corresponding docking tool called FLEXX in reproducing protein–ligand complexes determined by X-ray crystallography. FLEXX incorporates the concepts from four different software tools applied in structure-based drug design. First of all, the overall strategy for docking of flexible ligands is incremental, analogous to the algorithm of Leach & Kuntz (1992). The selection of the base fragment is the only interactive step during a complex prediction run with FLEXX.

The principles of modeling the protein–ligand interactions were adapted from the *de-novo* design tool LUDI (Böhm, 1992). Specifically, placements of the ligand are computed on the basis of pairwise assignments of interaction geometries. In our opinion, physico-chemical properties provide the most useful information for guiding the placement of the ligand. If the ligand is small or relatively flexible, shape alone is only a weak descriptor. For estimating the binding energy and for ranking the (partial) placements obtained, we apply the scoring function of Böhm (1994) with minor changes (see Theory and Algorithm for details).

The concepts of handling the conformational flexibility of the ligand are taken from the conformational search program MIMUMBA, developed by Klebe & Mietzner (1994). Conformations are generated according to a discrete raster. Bond lengths and angles are kept invariant as given in the input structure. A set of up to 12 discrete torsion angle values is assigned to each acyclic single bond by matching representative torsional fragments onto the ligand. Conformational preferences are determined for each fragment by a statistical evaluation of the Cambridge Structural Database (CSD; Allen *et al*., 1979). The receptor is regarded as rigid in our approach.

For placing the base fragment, we use a new algorithmic approach based on a pattern recognition technique called pose clustering (Linnainmaa *et al*., 1988). A complete description of how to apply

the pose clustering idea to docking can be found in Rarey *et al*. (1996). The incremental construction process is based on a simple greedy strategy: After adding a fragment in all possible conformations to all placements found in the previous iteration, only the *k* best placements are taken into the next iteration. This strategy was first used in the peptide design tool GROW, developed by Moon & Howe (1991). The elementary steps in each iteration of the construction process are described in Theory and Algorithm.

FLEXX is continuously being tested on an increasing set of known protein–ligand complexes taken from the PDB (Bernstein *et al*., 1977). Initial results were reported by Rarey *et al*. (1995). We are currently working with a set of 19 complexes. Four target proteins appear twice in the test set: thermolysin, thrombin, ribonuclease and carboxypeptidase A. In all test cases, the run time is below three minutes on a workstation (SUN SPARCstation 20). Thus, our approach is fast enough for screening larger sets of ligands for their binding affinity to a given receptor. For most complexes, a binding mode closely approaching the experimental geometry is predicted among the few highest ranking placements.

Theory and Algorithm

Modeling chemical phenomena (issues, features) in FLEXX

The chemical model underlying FLEXX is based on the work of Böhm (Böhm, 1992a, 1994) and Klebe & Mietzner (1994). The model can be divided into three areas: conformational flexibility, protein–ligand interactions, and the scoring function used for ranking the solutions generated.

Ligand conformational flexibility

The model describing the conformational flexibility of the ligand was developed by Klebe & Mietzner (1994) for conformation generation in MIMUMBA. The goal of the computer program MIMUMBA is to enumerate a large number of low-energy conformations of a small organic molecule. Conformational flexibility is modeled discretely. Bond lengths and angles are used as given in the input structure (except inside ring systems, see below). Therefore, reasonably minimized geometries should be used.

A set of preferred torsion angles is assigned to each acyclic single bond. Exceptions are bonds adjacent to a methyl group or to a planar amino group. These bonds remain unchanged as given in the initial conformation. Records from a database containing about 900 molecular fragments with a central single bond are assigned automatically to all other bonds using a subgraph matching algorithm. In addition to the identification of the fragment definition, each entry contains a histogram for the occurrence of torsion angles for this fragment in the Cambridge Structural Database (CSD; Allen *et al*., 1979). Torsion angles which are highly populated in the histogram are selected for the generation of low energy conformations. If several fragments can be matched to a single bond, the intersection of their sets of torsion angles is assigned to the bond. Details on the database and its generation can be found in Klebe & Mietzner (1994). With this approach, up to 12 torsion angles are assigned to each single bond.

Multiple conformations for rings are computed with the program SCA, developed by De Clercq (1984a,b) (the program is available by QCPE (Hoflack *et al*., 1989)). The program computes alternative conformations for polycyclic systems containing five- to seven-membered rings. For practical purposes in drug design, this is sufficient in most cases. SCA normally takes only a few seconds for a typical drug. Thus the generation of ring conformations is performed online after loading the ligand.

Protein–ligand interactions

In this section, we discuss the model of protein–ligand interactions used in FLEXX. The model is derived from the work of Böhm (1992a,b) and Klebe (1994).

The intermolecular interactions can be classified by the strength of their geometric constraints. Interactions are geometrically restrictive, if they are observed with only small deviations in geometric parameters such as distances or angles. Geometrically restrictive interactions are used in FLEXX for placing the ligand into the active site. Such interactions are mainly hydrogen bonds, but some hydrophobic interactions, for example those between phenyl rings and methyl groups, can be used for this purpose, as well.

Because most of the geometrically restrictive interactions have a favored interaction distance, we use the following model to describe molecular interactions based on spherical surfaces. We assign an interaction type and an interaction geometry to each interacting group of the molecule. The interaction geometry contains an interaction center *c*, an interaction radius *r*, and an interaction surface, which is a part of the spherical surface with radius *r* around *c*. Possible types of interaction surfaces are complete spheres, cones, capped cones and spherical rectangles (see Figure 1). For algorithmic reasons, the interaction surfaces on the receptor side are approximated by finite sets of points, called the interaction points. An interaction between two groups *A* and *B* is formed if

(1) The interaction types of *A* and *B* are compatible.

(2) The interaction center of *A* lies approximately on the interaction surface of *B* and *vice versa* (see Figure 2).

The interaction types and compatibilities are

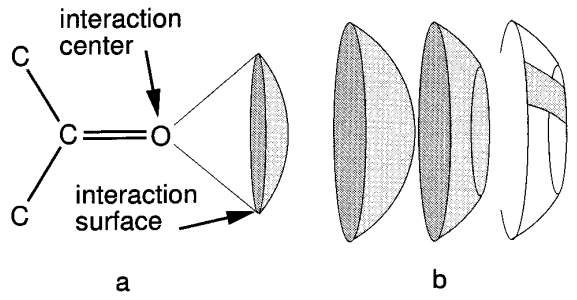

Figure 1. Interaction geometries. a, Interaction center and surface pertained to a carbonyl group. b, Three of the four different types of interaction surfaces: cones, capped cones and spherical rectangles.

shown in Table 1; geometries currently used in FLEXX are summarized in Figure 3.

Estimating the free energy of binding

The ranking of the generated solutions is performed using a scoring function similar to that developed by Böhm (1994) which estimates the free binding energy ΔG of the protein–ligand complex.

$$\Delta G = \Delta G_0 + \Delta G_{rot} \times N_{rot} + \Delta G_{hb} \sum_{\text{neutral H-bonds}} f(\Delta R,\Delta\alpha) + \Delta G_{io} \sum_{\text{ionic int.}} f(\Delta R,\Delta\alpha) + \Delta G_{aro} \sum_{\text{aro int.}} f(\Delta R,\Delta\alpha) + \Delta G_{lipo} \sum_{\text{lipo. cont.}} f^*(\Delta R) \quad (1)$$

Here, $f(\Delta R, \Delta\alpha)$ is a scaling function penalizing deviations from the ideal geometry (see below) and N_{rot} is the number of free rotatable bonds that are immobilized in the complex. The terms ΔG_{hb}, ΔG_{io}, ΔG_{rot}, and ΔG_0 are adjustable parameters. These values and the function f are taken as developed by

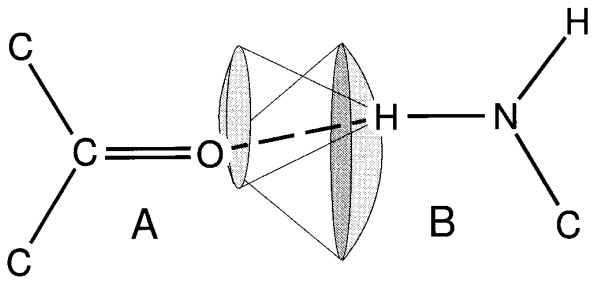

Figure 2. Condition for the formation of interactions: a hydrogen bond between the carbonyl oxygen and the nitrogen. The interaction centers are the oxygen and the hydrogen atom forming the hydrogen bond. They have to fall mutually on the surrounding interaction surfaces.

Table 1. Interaction types of FLEXX

H-acceptor	H-donor
Metal acceptor	Metal
Aromatic-ring-atom, methyl, amide	Aromatic-ring-center

In each row, the interaction types in the left and the right column can be matched. The interaction types aromatic-ring-center / aromatic-ring-atom are used to generate the preferred t-shaped arrangement of neighboring aromatic ring systems.

Böhm. In addition, we take into account the interactions of aromatic groups with the new parameter $\Delta G_{aro} = -0.7$ kJ/mol. The last term (ΔG_{lipo}) is a modification of Böhm's lipophilic contact energy. In the original function, this energetic contribution is intended to be proportional to the lipophilic contact area estimated with a grid method. However, our preliminary experiments with this definition of contact energy have generated placements that deviate markedly from the crystal structure. We therefore decided to calculate this term as a sum over all pairwise atom-atom contacts. It is essential that the function $f^*(\Delta R)$ in (1) account for contacts with a more or less ideal distance and penalize forbiddingly close contacts. For this reason, we choose:

$$f^*(\Delta R) = \begin{cases} 0 & \Delta R > 0.6 \text{ Å} \\ 1 - \dfrac{\Delta R - 0.2}{0.4} & 0.2 \text{ Å} < \Delta R \leqslant 0.6 \text{ Å} \\ 1 & -0.2 \text{ Å} < \Delta R \leqslant 0.2 \text{ Å} \\ 1 - \dfrac{-\Delta R - 0.2}{0.4} & -0.6 \text{ Å} < \Delta R \leqslant -0.2 \text{ Å} \\ \dfrac{\Delta R + 0.6}{0.2} & \Delta R \leqslant -0.6 \text{ Å} \end{cases}$$

with $\Delta R = R - R_0$. Here, R is the distance between the atom centers and R_0 is its ideal value assumed to be the sum of both van-der-Waals radii, each increased by 0.3 Å.

The overall docking algorithm of FLEXX

The docking algorithm in FLEXX is based on an incremental construction strategy, which consists of three phases:

(1) Base selection. The first phase of the docking algorithm is the selection of a connected part of the ligand, the base fragment.

(2) Base placement. In the second phase, the base fragment is placed into the active site independently of the rest of the ligand.

(3) Complex construction. In the last phase, called the construction phase, the ligand is constructed in an incremental way, starting with the different placements of the base fragment.

In the current version of FLEXX, the base selection is performed interactively. The docking algorithm is quite sensitive to the selection of the base fragment. If a part of a ligand is selected that has

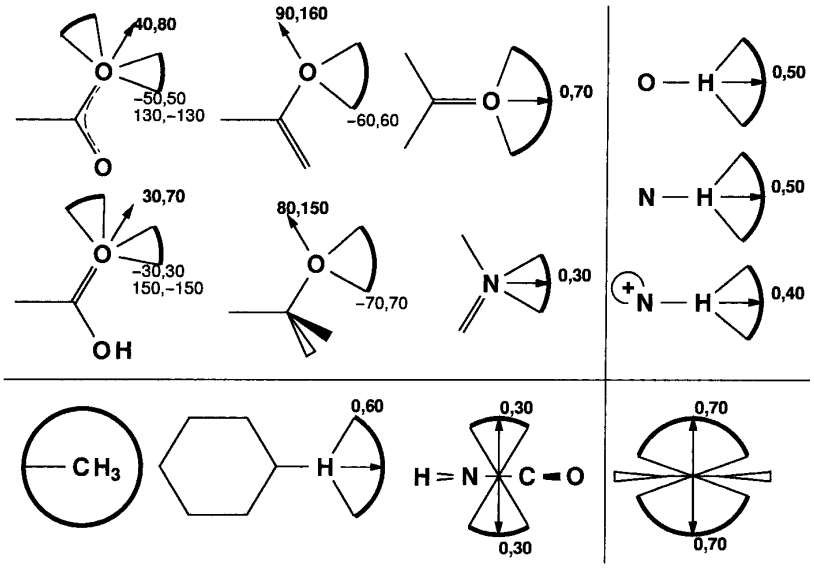

Figure 3. Interaction geometries as used in FLEXX: H-bond acceptors (upper left), H-bond donors (upper right), H-atoms at aromatic rings, methyl, and amide groups (bottom left), aromatic ring center (bottom right). Metal interaction geometries are always spheres and are not shown here. Interaction distances are 1.9 Å (hydrogen bonding), 2.0 Å (metal), and 4.5 Å (hydrophobic interactions). Angular ranges given by bold numbers are measured relative to the indicated vector in the drawing plane; all other angles are measured perpendicular to the drawing plane. The aromatic ring is drawn perpendicular to the drawing plane (bottom right). The sketched geometries are representative for several functional groups, e.g. the interaction geometry for the carbonyl oxygen is also used for sp^2 oxygens on sulfur and phosphorus.

no clearly predominant directional interactions with the receptor, the docking algorithm obviously has problems in predicting the correct binding mode.

As the size of the base fragment and the number of putative interaction groups increases, so does the probability of predicting the correct binding mode of the base fragment, but the number of conformations of such a base fragment unfortunately increases, as well. This causes longer run times, because substantial internal flexibility of the base fragment cannot be handled efficiently in our base placement algorithm (otherwise we would not need the incremental construction). Thus, the number of potential interaction groups should be maximized while the number of alternative conformations of the base fragment should be minimized.

Once the base fragment is selected, the remaining part of the ligand is automatically divided into fragments. We obtain the best results if the fragments are small. Thus, we cut the remaining part of the ligand at each rotatable, acyclic single bond.

The FLEXX base placement algorithm

The algorithm used for placing the base fragment is described in detail by Rarey *et al*. (1996). Here, we only give a short summary of the method.

The goal of the base placement algorithm is to find positions of the base fragment in the active site such that a sufficient number of favorable interactions between the fragment and the protein can occur simultaneously. This problem is related to problems in the area of computer vision and pattern recognition. One problem in this area is to detect an object in a photographic scene. Here, we have to identify a ''position'' of the object in the scene, such that most points of the object can be mapped onto points in the scene. Fisher *et al*. (1995) have already successfully applied a computer vision technique, called geometric hashing, to the geometric docking approach DOCK of Kuntz *et al*. (1982).

The algorithm from computer vision that we have adapted to the docking problem is called pose clustering (Linnainmaa *et al*., 1988; Olson, 1994). Here, we describe the adapted version of the algorithm.

Assume, first, that the base fragment is a rigid object. As mentioned above in the modeling section, the interaction surfaces on the receptor side are approximated by finite sets of interaction points. A transformation of the base fragment into the active site is uniquely defined by mapping three interaction centers of the fragment onto three interaction points of the receptor by simply superposing the three point pairs onto each other (assuming that the point sets are not collinear) as illustrated in Figure 4. We call two triangles δ-compatible, if the corresponding edge lengths

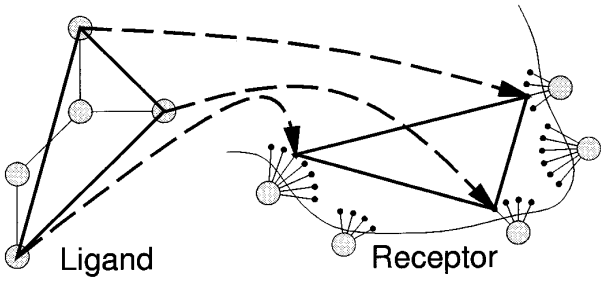

Figure 4. The fragment placing algorithm: mapping three interaction centers (grey spheres) of the ligand onto three discrete interaction points in the active site (black dots) defines a unique transformation of the ligand into the active site.

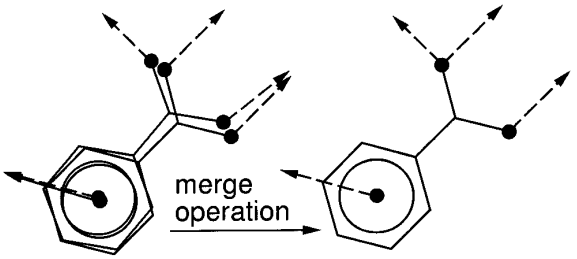

Figure 5. Merging initial placements. Two initial placements with different sets of interactions but similar transformations (left) will be transformed into a single placement (right) by merging the sets of interactions and recomputing the transformation.

differ by at most δ and the corresponding corners have compatible interaction types.

Therefore, the first step of the base placement algorithm has to solve the following problem: For each triangle of interaction centers of the base fragment find all δ-compatible triangles of interaction points in the active site of the receptor.

We have developed a new data structure for this problem based on line segment hashing and an efficient query algorithm for constructing triangles out of line segments (Rarey *et al*., 1996). Compared to hashing triangles directly as in geometric hashing (Fischer *et al*., 1995), the storage requirement is reduced from cubic to quadratic in the number of points.

A transformation is derived from each match between a triangle of interaction centers of the base fragment onto a triangle of interaction points in the receptor. Now, two filters are applied to the placement: First the angular constraints of the interaction geometries of the fragment are checked (i.e. does the interaction center on the receptor side approximately coincide with the interaction surface on the ligand side). Second, the placed fragment is checked for overlap with the receptor.

Normally the list of placements obtained from the first step contains many similar transformations, for two reasons. First, changing a point in a triangle to a nearby point on the same interaction surface changes the transformation only slightly. Second, often a fragment can form more than three interactions simultaneously.

Therefore, the second step in the base placement algorithm clusters the placements according to an appropriate distance function. For this purpose, we use the rms deviation between two placements.

The method we apply for clustering, is a complete-linkage hierarchical cluster algorithm (Duda & Hart, 1973). The main strategy is that the two clusters with minimal distance are merged into one cluster iteratively as long as the minimal distance between two clusters is less than a predefined threshold. We have developed a time-efficient algorithm for this step. The details of the cluster algorithm can be found in Rarey *et al*. (1996).

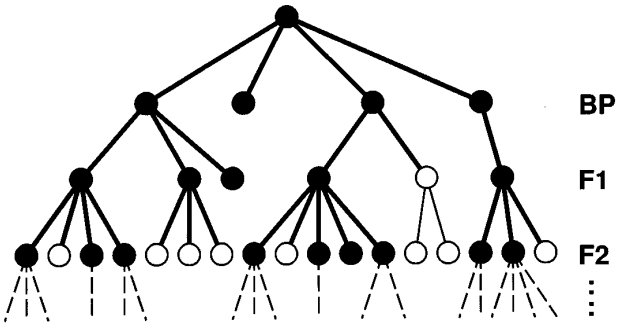

Figure 6. Search tree during the complex construction algorithm. The first level contains different placements of the base fragment and levels $i \geqslant 2$ placements of the partial ligand up to fragment $i - 1$. Black nodes represent placements which will be considered in the next iteration.

All placements inside the same cluster are combined to one solution by merging the lists of interactions and recomputing a superposition of all interaction centers of the ligand onto the interaction points of the receptor (see Figure 5) (Kabsch, 1976).

In a third step, a final overlap test is performed and for the non-overlapping placements energies are computed with Böhm's function (Böhm, 1994).

In the case of small base fragments it may happen that no match between triangles can be found. Then we can apply a variation of the described algorithm which matches pairs of interaction centers onto pairs of interaction points in the active site (instead of triangles). The remaining degree of freedom (the rotation around the axis defined by the pair of interaction points) can be fixed by rotating the ligand such that the interaction centers of the two interaction groups on the receptor side lie on the interaction surfaces of the interaction groups on the ligand side.

The FLEXX complex construction algorithm

Once a set of favorable placements for the base fragment has been computed, we can start the incremental construction process that adds the remaining fragments to the alternative placements of the base fragment.

We formulate the incremental construction as a tree search problem. A node in the tree represents a placement of a connected part of the ligand. On the first level of the tree, the different placements of the base fragment are arranged. Each of the following levels contains the alternatives for adding the next fragment to the placement represented by the parent node (see Figure 6). The order in which the fragments are added is kept constant during the tree search. If there are alternatives, we first add fragments which can form hydrogen bonds or salt bridges because these interactions are more directional and thus geometrically more precisely defined. The geometric conditions for attaching a fragment depends on the partial placement. Therefore, the number of successors in the tree changes with the tree node (see Figure 6).

The goal of the tree search is to find the leaves which contain placements with favorable binding energies as estimated by the scoring function. Because the search tree grows exponentially in the number of degrees of freedom, a complete search of the tree is infeasible. Instead we can use the binding energy of the partially placed ligand represented by the interior nodes of the tree to guide the search.

Two alternative tree search techniques have already been applied in other tools for structure-based drug design. Leach & Kuntz (1992) use a backtracking algorithm in their docking approach. We follow a simple greedy heuristic which has been used before in the *de-novo* peptide design tool GROW, developed by Moon & Howe (1991). It works as follows: after adding a fragment, the energetically k best partial placements found are considered in the next iteration. This heuristic is very fast and is robust with respect to extensions of the model as well as modifications in the different steps of complex construction. With an appropriate setting of the parameter k, the algorithm can be prevented from getting stuck in local minima. Because the run time and space requirements increase with k, efficient algorithms and data structures are required. The results presented in this paper are all obtained by setting k to 500.

In the following, we explain the elementary steps during one iteration of the complex construction algorithm. Each iteration starts with a set of different placements of the partial ligand, called the solution set. In the first iteration, this set is initialized with the placements found in the base placement phase.

Adding a fragment

At the beginning, the new fragment is added to all placements in the current solution set in all possible conformations. All extended placements having internal overlap in the ligand or a strong overlap with the receptor (maximal receptor atom–ligand atom overlap volume $\geqslant 4.5\,\text{Å}^3$) are rejected.

In order to reduce space, the ligand's conformations and placements are stored in a tree structure. Each node in the tree represents a set of conformations and a fixed position in a global coordinate system. New sets of conformations can be generated using three different operations: expansion (adding a new fragment), reduction (reducing a set to a subset), and transformation (applying the same motion to all conformations in the set). A ligand conformation and placement can be extracted from the structure by traversing from the node containing the ligand conformation and placement to the root of the tree. We call this data structure the confset tree.

In order to keep the overlap tests efficient, we use a box hashing technique. A cubic grid covering three-dimensional space is aligned to the Cartesian coordinate axis. Each cube has an integer address computed from its coordinates in space. An atom is stored in the cube containing the center of the atom. To check a ligand atom for overlap with the receptor, we inspect the receptor atoms whose centers are located in all cubes intersecting a sphere centered in the ligand atom whose radius is the sum of the Van-der-Waals radius of the ligand atom and the largest atom of the receptor (see Figure 7). The same data structure is used for all three-dimensional neighborhood queries in the docking algorithm.

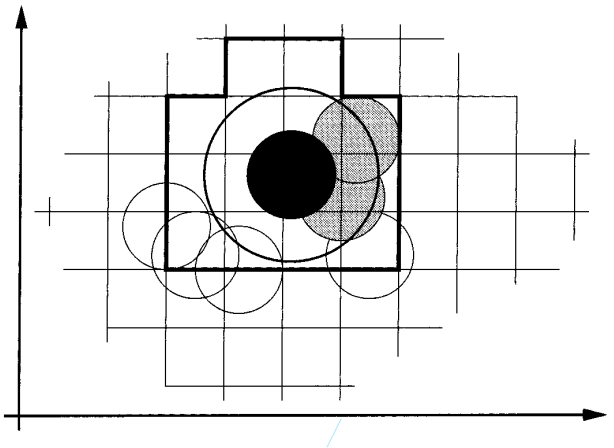

Figure 7. Box hashing. The black circle represents the ligand atom on which the query is based. The ligand atom radius is enlarged by the radius of the largest atom in the receptor (large surrounding circle). The cubes checked for receptor atom intersection with the ligand atom are surrounding by a rectagon outlined in bold.

Searching for new interactions

For each extended placement, the algorithm searches for matching interaction groups. This is done by scanning through the interaction groups of the fragment to be added and, for each such group, looking for chemically complementary interaction groups in the receptor. For this search, the box hashing technique is used again. We allow large distance and angle tolerances for the interaction geometries (distance tolerance of 2.0 Å, angle tolerance of 20°). In most cases, the subsequent optimization step can improve the inaccurate geometries of new interactions.

The space requirements for handling the lists of interactions for all placements can be reduced by arranging the interactions in a tree-like structure. Because extended placements are derived from placements of a smaller portion of the ligand, interactions can also be arranged in a tree structure. In principle, a singly-linked list of interactions can be assigned to each placement. All placements, which are derived from the same parent placement in the previous iteration have the same list of interactions except for the interactions involving the new fragment. If the lists of interactions are sorted by the number of the ligand fragment to which they belong, the lists differ only in the head

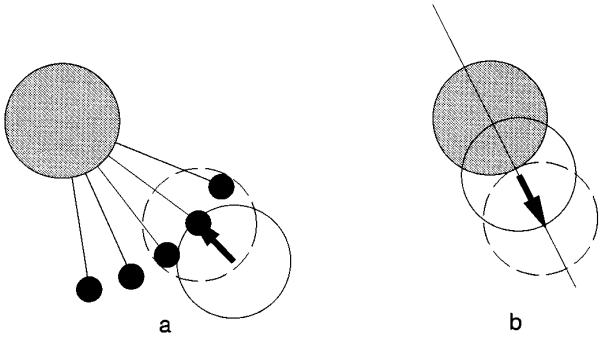

Figure 8. Point generation during the optimization of a placement. For each protein–ligand interaction (a) and each protein–ligand overlap (b), a pair of matching points is generated. The arrow points from the matching point of the ligand to the matching point of the receptor. Receptor atoms are represented by grey spheres, ligand atoms by white spheres. The ideal position of the ligand atom is shown by the broken drawn circle.

elements. Thus, all placements that have the same parent placement can share the tail of the list.

Optimizing the placement

The location of an extended placement in the active site must be optimized if (a) new interactions are found or (b) the placement contains slightly overlapping atoms between the receptor and the ligand (maximal receptor atom–ligand atom overlap volume $\geqslant 2.5$ Å^3).

The respective optimization is based on a weighted superposition of points (Kabsch, 1976). The input of the optimization is a set of triples (l_i, r_i, w_i), $i = 1, \ldots, n$, where l_i is a point of the ligand, r_i a point of the receptor, and w_i is a weight. In the superposition routine, l_i are fitted onto r_i minimizing the sum of distance squares $\Sigma_{i=1}^{n} w_i (l_i - r_i)^2$.

For each interaction and each overlap the point pairs and their weights are generated as follows: For an interaction, l_i is the interaction center of the interacting group of the ligand, r_i is one of the discrete interaction points on the surface of the interaction group of the receptor, and $-w_i$ is the energy contribution of the interaction, given optimal geometry. For an overlap between a ligand atom *a* and a receptor atom *b*, l_i is the center of atom *a*, r_i is the point on the line through the centers *a* and *b* onto which the center of *a* has to be placed such that no overlap occurs (see Figure 8). The weight is set to the overlap volume between *a* and *b*.

After superposition, another overlap test is performed. If the ligand still overlaps, all new interactions with substantial distance violations are removed and the superposition is done a second time. If there are still overlapping atoms between the molecules, the solution is rejected.

Even with the box hashing technique, most of the time is spent in performing the overlap tests between the molecules. We have developed an extension of our data structure to speed up the overlap test even more. We call this feature the atom cache. The idea behind a cache is to store answers from recent queries to a data structure in order to use this information again when a similar query is posed in the near future.

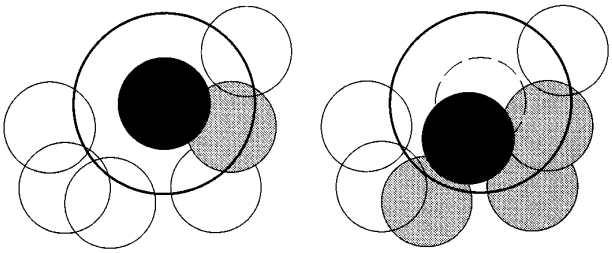

Figure 9. Atom cache. The ligand atom is shown in black. The surrounding circle represents the query with the enlarged radius. Receptor atoms are shown in white (non-overlapping) and grey (overlapping).

All queries to the box hashing data structure have the form ''List all receptor atoms overlapping with the ligand atom *X* at position *p*''. During the incremental construction process, in many cases, the ligand atom does not move very far between two queries.

We can take advantage of this phenomenon. We attach the answer of a box hashing query for ligand atom *X* together with its actual location *p* directly to the atom. In the query, the radius of the ligand atom is extended by a constant value δ (0.4 Å). In a subsequent query, we first check whether the ligand atom has moved farther than δ since the last query. If this is the case, we have to perform a new search in the box hashing data structure. Otherwise, we can re-use the answer from the previous query (see Figure 9). The same technique is applied for computing the hydrophobic contacts required in the evaluation of the scoring function.

Selection and clustering of the solution set

All placements produced are ranked by energy and the *k* best are kept for further analysis. A simple way of performing this selection is to store all placements, sort them, and delete those with a rank greater than *k*. A more time and space-efficient way of selecting the *k* best solutions is to use a priority queue of fixed length *k*. The queue is sorted by descending energy values. Thus, the head of the queue always contains the *k* best solution. Inserting an element into the queue and deleting the head of the queue costs logarithmic time in *k* and during the computation, we have to store at most $k + 1$ solutions simultaneously.

After selecting the *k* best solutions, we perform a clustering in order to remove similar placements. Here, the complete-linkage hierarchical cluster algorithm from the base placement phase is applied again. The distance between two placements is the rms deviation between the coordinates of the ligand (rms threshold 0.7 Å). In addition, placements that differ in the position or direction of the vector at

which fragments will be added in subsequent iterations are not allowed to belong to the same cluster. From each cluster, only the solution with highest rank is used in the next iteration.

General remarks on the input data

Before we describe the results obtained, we give some general remarks on the preparation of the input data.

The ligand

The following steps preprocess the ligand for docking. First, the ligand is extracted from the PDB file and transformed into SYBYL mol2 file format. Correct atom types including hybridization states (mol2 notation; TRI, 1994) as well as correct bond types are defined. Hydrogen atoms are added to all atoms with reasonable geometries, formal charges are assigned to each atom. The ligand is centered and an energy minimization is performed with the TRIPOS force field (TRI, 1994).

FLEXX analyzes the structure of the ligand and detects local topological symmetries at single bonds whose torsion angle can be changed by less than 360° such that the same conformation results, e.g. a C_2-symmetrical para-substituted phenyl ring or a carboxylate group. The computation of rms deviations also considers this local symmetry.

The receptor

For the receptor, a so-called receptor description file (rdf) is provided. The file contains information on the chains and hetero groups to consider, on how to resolve ambiguities in the PDB file (alternate location indicators, atom names containing an *A*), on how to add hydrogens to polar atoms in the active site of the receptor, etc. A template that contains physico-chemical information about the amino acid is assigned to each amino acid. All assignments are made according to default rules except for the definition of the torsion angle at the hydroxyl group of tyrosine. Here, a torsion angle of 0° or 180° is selected by visual inspection of the protein.

FLEXX requires a definition of the active site of the protein. If the protein is given without a ligand, the active site can be defined either manually or with automatic tools (e.g. Peters *et al*., 1996). Optionally, in cases of a cocrystallized ligand, these coordinates can be used to define the neighboring active site. In the test cases, all atoms of the protein, with a distance of up to 6.5 Å to 8 Å from an atom of a known ligand at its crystalline position are selected to define the active site. This distance is set by visual inspection such that the active site is completely enclosed.

In addition FLEXX, requires information about the accessibility of receptor atoms by water molecules, i.e. are on the Connolly surface (Connolly, 1983) of the protein. This information is computed by FLEXX only once for each protein. The run time for this computation is between 30 and 60 seconds depending on the size of the protein and is not considered as part of the presented run times for the docking.

Results and Discussion

Evaluating the quality of the conformational model

To analyze the quality of the MIMUMBA conformational model, we performed the following test. For a test suite of 19 protein–ligand complexes (Figures 10 and 11) with known crystal structure, we compute ligand conformations that closely approximate the ligand's crystal conformation using the same strategy as in the docking algorithm. For the examples, the rms deviations of the generated conformation with respect to the ligand conformation in the crystal are given in Table 2, accuracy column. Assuming that the computed conformation is the one with the lowest possible rms, this value is a lower bound for the accuracy of our docking algorithm.

While the accuracy of the model is important for the quality of the results, the size of the conformational set is important for the time-complexity of the algorithm. We compute a combinatorial number of conformations by multiplying the number of alternatives for each degree of freedom (see column complexity in Table 2). Because internal clashes discard a part of these conformations, this combinatorial number is an upper bound for the number of possible low-energy conformations in the model. For a better estimate of this number, we have enumerated the number of conformations without clashes for methotrexate which is approximately 10% of the combinatorial number of all conformations.

All 19 protein–ligand complexes were used as test cases for our docking tool. In the following we discuss in some detail the special features of several docking experiments. A complete overview of the results is given in Table 3.

Complexes with small ligands

The base placement algorithm has been tested for the complexes 5tim, 1ldm, 2phh, 3ptb and 1ulb, since no incremental construction phase is required.

5tim

The complex of triose-phosphate isomerase and a sulfate ion (5tim) has been solved by Wierenga *et al*. (1991) to 1.83 Å. The binding site consists of 49 accessible atoms potentially involved in ligand binding. The docking algorithm generates six different placements; the one ranked third according to FLEXX's energy estimate has an rms deviation of 0.87 Å with respect to the X-ray structure. The calculated binding energy of −16.3 kJ/mol is in

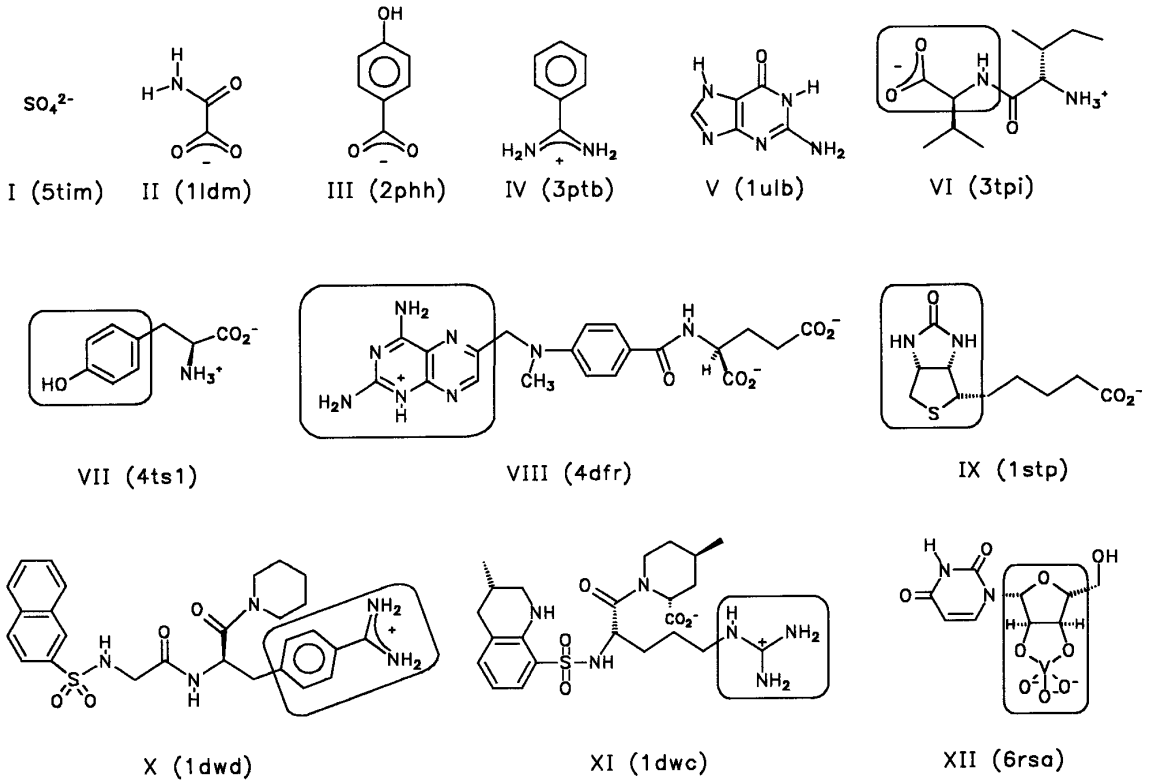

Figure 10. Chemical formulae of the ligands in the test suite (1 to 12). Base fragments used are surrounded by boxes.

good agreement with the experimental value of –13.1 kJ/mol (Verlinde *et al*., 1991). Hydrogen bonds to Gly173 and Gly235 have been found. In the crystal structure five additional interactions to water molecules are present, but we did not include these water molecules in the binding site. In this situation, the algorithm searches for alternative hydrogen bonds. In fact, nearly all solutions show additional hydrogen bonds to Lys13 or Ser213. As a consequence, the calculated and the observed positions of the sulfate ion differ slightly.

1ldm

The complex 1ldm of lactate dehydrogenase and oxamate (II) has been solved (Abad-Zapatero *et al*., 1987) to a resolution of 2.1 Å. The ligand binds (with $\Delta G = -30.8$ kJ/mol) with eight hydrogen bonds in a polar pocket composed of Gln100, Arg106, Arg169, His193 and Thr245. The docking algorithm generates 28 placements. The highest-ranking solution exhibits six hydrogen bonds, all of which are also observed experimentally. Convincing agreements in structure (rms = 0.62 Å) and binding affinity (–32.8 kJ/mol) are found.

2phh

The complex 2phh between *p*-hydroxybenzoate hydroxylase and *p*-hydroxybenzoate (III) has been solved by van der Laan *et al*. (1989). In the crystal structure the ligand is bound through its carboxylate group to Tyr222, Ser212 and Arg214 and its hydroxyl group to Tyr201. The docking algorithm generates 179 different solutions. The predicted binding energies of –32.1 kJ/mol and –30.0 kJ/mol, respectively, for the two highest-ranking solutions are very close to the experimental binding energy of –26.7 kJ/mol (Entsch *et al*., 1976). They are separated from the predicted energies of the next solutions by more than 3 kJ/mol. The according rms deviations of 0.58 Å and 0.29 Å, respectively, show that nearly the same binding geometry as in the crystal structure has been generated.

3ptb

Benzamidine (IV) is a ligand of trypsin. The structure of the complex (3ptb) has been resolved (Marquart *et al*., 1983). Presumably the ligand binds protonated to the enzyme. The docking algorithm produces 14 different placements. The energy of the highest-ranking solution (–28.3 kJ/mol) is about 14 kJ/mol lower than that of the proceeding solution, and is in very good agreement with the measured value of –27.2 kJ/mol (Mares-Guia & Shaw, 1965). Although one water molecule has been omitted from the active site, a convincing structural agreement is found: the rms deviation is 0.48 Å. The hydrogen bonds to Ser190, Gly219 and Asp189 have been reproduced.

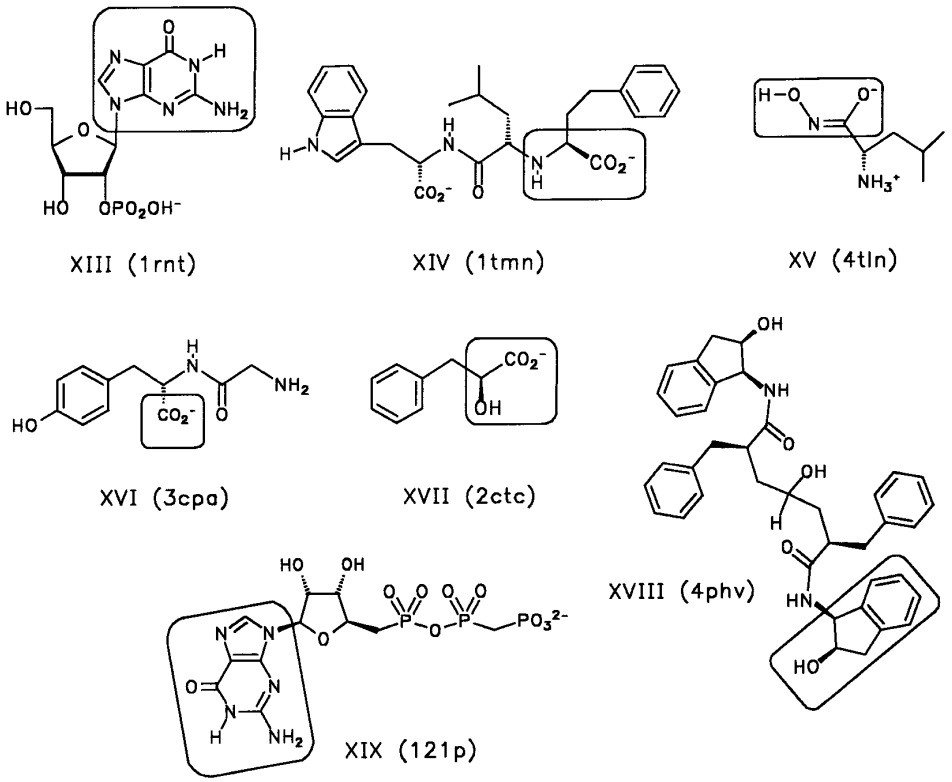

Figure 11. Chemical formulae of the ligands in the test suite (13 to 19). Base fragments used are surrounded by boxes.

1ulb

The complex 1ulb (Ealick *et al*., 1991) of purine-nucleoside phosphorylase with guanine (V) has a free binding energy of –30.2 kJ/mol (Zollner, 1993). It binds into a pocket composed of Glu201, Asn243 and Lys244. These residues form strong and specific hydrogen bonds with the ligand. The docking algorithm produces 134 solutions. The highest-ranking solution has a binding energy of –17.9 kJ/mol and an rms value of 0.65 Å. All but one of the experimentally observed hydrogen bonds have been reproduced in this solution. The one bond not realized is the interaction between the

Table 2. Conformational model, performance parameters

1 Ligand no.	2 PDB	3 Acc.	4 Com.
I	5tim (Wierenga *et al*., 1991)	0.04	1
II	1ldm (Abad-Zapatero *et al*., 1987)	0.22	2
III	2phh (van der Laan *et al*., 1989)	0.05	3
IV	3ptb (Marquart *et al*., 1983)	0.09	2
V	1ulb (Ealick *et al*., 1991)	0.18	1
VI	3tpi (Marquart *et al*., 1983)	0.22	7.0×10^5
VII	4ts1 (Brick & Blow, 1987)	0.09	1.2×10^2
VIII	4dfr (Bolin *et al*., 1982)	0.40	3.9×10^7
IX	1stp (Weber *et al*., 1989)	0.29	6.7×10^5
X	1dwd (Banner & Hadvary, 1991)	0.34	2.5×10^8
XI	1dwc (Banner & Hadvary, 1991)	0.47	7.2×10^9
XII[a]	6rsa (Borah *et al*., 1985)	0.06	1.2×10^2
XIII	1rnt (Arni *et al*., 1987)	0.53	8.1×10^3
XIV	1tmn (Monzingo & Matthews, 1984)	0.44	7.8×10^{10}
XV	4tln (Holmes & Matthews, 1981)	0.54	2.4×10^3
XVI	3cpa (Rees & Lipscomb, 1983)	0.17	1.7×10^5
XVII	2ctc (Teplyakov *et al*., 1993)	0.23	4.5×10^2
XVIII	4phv (Bone *et al*., 1991)	0.69	1.1×10^{13}
XIX	121p (Krengel, 1991)	0.54	2.6×10^6

Columns from left to right: 1, ligand number; 2, PDB-entry with reference; 3, accuracy (smallest rms deviation from crystal conformation in Å); 4, complexity (theoretical number of conformations).

[a] The ring system of uridine vandate was kept invariant.

Table 3. Summary of docking results

1	2	3	4	5	6	7	8	9	10	11	
	No.	Low-energy		Best prediction			Run time		ΔG	ΔG	
PDB	Sol.	ΔG	rms	ΔG	rms	rank	BP	CC	Cryst.	Exp.	
5tim	6	–16.3	1.99	–10.4	0.87	3	3.12	0.00	–7.8	–13.1	(Verlinde *et al*., 1991)
1ldm	28	–32.8	0.62	–32.8	0.62	1	11.38	0.00	–29.1	–30.8	(White *et al*., 1976)
2phh	179	–32.1	0.58	–32.1	0.58	1	36.37	0.00	–30.3	–26.7	(Entsch *et al*., 1976)
3ptb	28	–28.3	0.48	–28.3	0.48	1	18.86	0.00	–30.2	–27.2	(Mares-Guia & Shaw, 1965)
1ulb	134	–17.9	0.65	–17.9	0.65	1	46.56	0.00	–23.2	–30.2	(Zollner, 1993)
3tpi	215	–24.0	0.58	–24.0	0.58	1	6.65	27.36	–28.1	–24.5	(Bode, 1979)
4ts1	156	–16.1	2.01	–11.8	0.71	15	6.30	7.11	–21.9	–32.0	(Wells & Fersht, 1986)
4dfr	80	–62.9	1.34	–62.2	0.90	2	38.00	55.80	–53.7	–55.3	(Bolin *et al*., 1982)
1stp	163	–31.2	0.81	–31.2	0.81	1	10.73	14.06	–33.0	–76.4	(Weber *et al*., 1992)
1dwd	332	–42.6	2.12	–38.6	0.63	23	24.02	1:21.94	–37.9	–48.6	(Bode *et al*., 1990)
1dwc	181	–32.0	2.66	–29.1	1.20	13	27.91	1:16.18	–22.7	–44.0	(Hilpert *et al*., 1994)
6rsa	131	–35.9	0.85	–35.9	0.85	1	34.64	11.38	–37.8	–28.5	(Lindquist *et al*., 1973)
1rnt	123	–42.0	1.48	–31.4	0.96	47	39.54	33.93	–30.2		
1tmn	237	–39.1	0.87	–39.1	0.87	1	21.47	1:25.86	–35.5	–41.6	(Matthews, 1988)
4tln	138	–15.8	4.50	–8.6	0.93	106	1.45	1.66	–9.4	–21.2	(Matthews, 1988)
3cpa	120	–35.3	3.08	–32.9	1.06	3	54.70	36.00	–24.1	–22.1	(Bunting & Myers, 1975)
2ctc	144	–30.3	–0.65	–30.3	0.65	1	1:01.98	5.50	–24.8	–22.2	(Teplyakov *et al*., 1993)
4phv	9	–38.2	1.04	–38.2	1.04	1	0.65	1:30.45	–42.8	–52.2	(Bone *et al*., 1991)
121p	44	–50.0	2.00	–30.5	1.14	43	31.91	49.49	–73.4		

Columns from left to right: 1, PDB-entry; 2, number of solutions produced; 3, solution with lowest energy, ΔG predicted and 4, rms deviation; 5, best prediction, ΔG predicted; 6, rms deviation; 7, rank by predicted energy; 8, elapsed run time of base placement algorithm; and 9, complex construction algorithm; 10, predicted ΔG values for the crystal structure; 11, experimentally observed ΔG with reference.

oxygen of the ligand and the amino group of Lys244. Whereas some lower-ranking solutions reveal an interaction to Lys244 and none to Asn243, no placement has been generated showing H-bonds either to Lys244 and Asn243 at the same time. Such a situation would involve unfavorable binding angles for both hydrogen bonds, according to the applied interaction model.

Short peptides

The complexes 3tpi and 4ts1 of our test suite contain short peptidic ligands. These ligands have a relatively high flexibility together with a low number (four) of specific interaction centers distributed over the entire molecule. This complicates the identification and placement of a base fragment.

3tpi

The complex of trypsinogen with the dipeptide Ile-Val (VI) has been structurally solved by Marquart *et al*. (1983). Its free binding energy is –24.5 kJ / mol (Bode, 1979). We selected the carboxylate group, the C^{α} atom and the nitrogen of the valine residue as a base fragment. The base placement and the subsequent incremental construction procedure reproduced convincingly the experimental structure with the same hydrogen-bonding pattern. The rms deviation amounts to only 0.58 Å and the calculated binding energy –24.0 kJ / mol is very close to the experimental value.

4ts1

For the complex 4ts1 of a mutant of tyrosyl-transfer-RNA synthetase with the ligand tyrosine (VII) we first selected the same type of base fragment ($NH—CH—CO_2^-$), however with disappointing results. Supposedly, this is due to the binding mode of the carboxylate group: in the crystal structure this group only binds to three water molecules (HOH353, HOH387 and HOH402) which mediate interactions to the protein. However, in the docking experiment, these water molecules have been removed.

Using the phenol subunit as a base fragment, an acceptable base placement is found. The complete docking algorithm produced 156 placements. The highest-ranking solution showed a completely different H-bond pattern. The first reasonable solution (with rank 14) has an rms deviation of 0.71 Å. Although the experimentally observed hydrogen bonds have been detected by this solution, the placement appears to be so unfavorable that only a low binding energy of –11.8 kJ / mol is found. This is substantially less than the experimental value of –32.0 kJ / mol (Wells & Fersht, 1986). One explanation might be the negligence of the energy contribution of the three water-mediated interactions.

Dihydrofolate reductase

4dfr

Dihydrofolate reductase plays an important role in the process of DNA replication and is therefore a target for cytostatic drugs as well as for

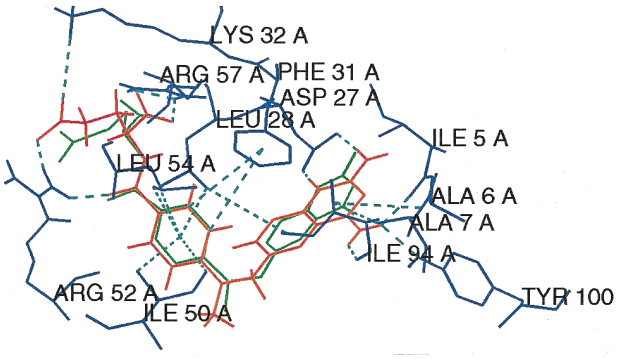

Figure 12. Docking results for methotrexate (VIII). Ligand geometry in the crystal structure, green; solution with the second best ranking, red. For the calculated geometry hydrogen atoms are also displayed. Broken lines represent the protein–ligand interactions.

anti-bacterial drugs. The structure of the protein in the complex with the inhibitor methotrexate (VIII) has been solved with 1.7 Å resolution (Bolin *et al*., 1982).

Methotrexate can be divided into three parts. The first is a pteridine ring, which binds with high complementarity in the deeply buried part of the active site of dihydrofolate reductase. In the center a hydrophobic moiety is composed by a methylen-amino and a phenyl group. The third portion comprises a flexible chain bearing two terminal carboxylate groups. Experimentally determined water molecules in the active site have been omitted from consideration. Including them in the docking experiment has only minor influence on the predicted binding modes.

The ligand contains two rigid ring systems and ten rotatable bonds, including the amide bond next to the phenyl ring and omitting the single bond to the terminal methyl group. In our discrete conformation model 3.9×10^7 putative ligand conformations are obtained; the most similar conformation has an rms distance of 0.4 Å with respect to the crystal conformation.

Methotrexate is a favorable test case for an incremental construction algorithm. Its pteridine ring represents an ideal base fragment; the algorithm finds 107 different solutions with rms deviations of 0.29 to 15.0 Å compared to the crystal structure. The highest-ranking placement has an rms deviation of 0.48 Å from the observed crystal structure and a predicted free energy of –23.6 kJ/mol. The placement with 0.29 Å rms deviation from the crystal structure ranks at position nine with a predicted energy of –20.7 kJ/mol. The placements with rms values beyond 1.9 Å are energetically clearly separated from those with low rms deviations. They start from rank 12 and possess energies of –14.4 kJ/mol and more.

After adding the remaining 10 fragments in the complex construction phase, 80 placements are generated. The highest-ranking complex has an rms deviation of 1.33 Å from the crystal structure and a predicted binding energy of –62.9 kJ/mol. In this solution the ligand differs only in the terminal chain which forms alternative hydrogen bonds to Arg57 and Arg52 compared to the X-ray structure. The second best solution (ΔG_{calc} = – 62.2 kJ/mol) has an rms of 0.90 Å. In this placement, the carboxylate group forms hydrogen bonds to Arg52 and Lys32.

Streptavidin

1stp

The complex streptavidin-biotin is an interesting test case for a docking algorithm, because of its very high binding affinity. The measured free binding energy is –76.6 kJ/mol (Weber *et al*., 1992). The structure has been solved by Weber *et al*. with a resolution of 1.55 Å (Weber *et al*., 1992).

Biotin (IX) consists of a bicycle and a carboxylate group linked *via* a four-membered alkyl chain. In the complex, the carboxylic group is assumed to be deprotonated. The fused ring system composed of two five-membered rings, one with a ureido group and a second with a sulfur atom, turns out to be a useful base fragment.

Whereas the predicted binding energy (–31.2 kJ/mol) of the highest-ranking solution is not in agreement with experiment the binding geometry is reproduced with an rms deviation of 0.81 Å.

In this solution the alkyl chain adopts another conformation, but it is embedded into the same hydrophobic pocket as observed in the X-ray structure. The bicycle and the carboxylate group are correctly placed. All seven hydrogen bonds to amino acids of the receptor are therefore detected. The ureido group binds to Asn23, Ser27 and Tyr43 *via* its carbonyl oxygen and to Ser45 and Asp128 through its nitrogen atoms. For the carboxylate group, we find hydrogen bonds to Asn49 and Ser88. In addition, experimental data exhibit three hydrogen bonds between this group and HOH315 and HOH445. We were unable to reproduce these interactions since we did not include water molecules in our calculations. Despite these missing bonds, the carboxylate was fixed in its correct position through the remaining hydrogen bonds.

The absence of these interactions is the major cause of the deviation between the observed and the calculated ΔG. Another reason has been given by Weber *et al*., who pointed out that the ureido group of the ligand can be easily polarized in the protein. This leads to salt-bridge-like hydrogen bonds, thus contributing much more strongly to binding affinity.

Thrombin

The serine protease thrombin catalyzes important reaction steps in the blood clotting cascade. In recent years there have been extensive efforts to find selective thrombin inhibitors, which are potential antithrombotic drugs. We tested FLEXX with two thrombin complexes (PDB codes 1dwd

and 1dwc). NAPAP (X) and argatroban (XI) are both potent lead structures, structurally investigated by Banner & Hadvary (1991) to a resolution of 3.0 Å.

1dwd

Thrombin usually cleaves peptides behind an arginine. This position is preferred because it is specifically recognized through a strong bidentate salt bridge between the guanidinium group of arginine and Asp189. All thrombin inhibitors have one fragment in common that is capable of interacting with Asp189. This fragment should be a hydrogen donor with positive charge. In the case of NAPAP (X), it is a benzamidino group. In our docking experiment we used this moiety as the base fragment. It forms several interactions with the receptor and is rigid by comparison with the remaining fragments of NAPAP. The highest-scoring placement deviates by an rms value of only 0.54 Å from the experimental structure and exhibits the above-mentioned interaction with Asp189. Furthermore the interaction with Gly219 could be observed.

After the complete docking procedure we obtain 332 different solutions. The highest-ranking placement ($\Delta G_{calc} = -42.6$ kJ/mol) differs mainly in the position of the naphthyl and the sulfonyl group of the ligand. Although there is one H-bond missing (between the sulfonamide nitrogen and Gly216) in comparison with the X-ray structure, this solution is stabilized by a large number of hydrophobic contacts of the naphthyl group in this alternative position. The total rms deviation is 2.12 Å. The solution at rank 5 approximates the experimental situation with an rms deviation of 1.01 Å. The geometrically closest solution (rms = 0.63 Å) has rank 23 but differs in energy by only 4 kJ/mol from the best solution and shows the same interaction pattern as in the crystal structure. There is only a small difference concerning the interactions of the carbonyl oxygen of the phenylalanyl part of NAPAP. In the crystal this group forms a hydrogen bond to a water molecule which has not been included in the protein structure in the docking experiment. The algorithm suggests an alternative: a bond to Ser195. This example again underlines the importance of considering interactions mediated *via* water molecules.

1dwc

Similarly to NAPAP, argatroban (XI) consists of two hydrophobic ring systems (tetrahydroquinoline and piperidine group). The guanidinium group is an appropriate base fragment. All successful base placements fill the specificity pocket. However, they all show slightly different interaction patterns compared to the crystal. Thus, the algorithm manages to find placements that allow an H-bridge to Gly219, which does not occur in the experimental structure.

Nevertheless, due to the high flexibility of the alkyl chain connecting the guanidinium group with the remaining part of the molecule, the additional fragments capable of forming hydrogen bonds are placed correctly. Thus, the docking solution with the smallest observed rms deviation (1.20 Å) still shows good agreement with respect to the positions of the two lipophilic groups and the guanidinium group. This solution has rank 13 and its energy is less than 3 kJ/mol above the energy of the highest-ranking solution ($\Delta G_{calc} = -32.0$ kJ/mol) which has an rms deviation of 2.66 Å from the crystal structure.

On a relative scale, the binding affinities of both ligands (NAPAP and argatroban; Hilpert *et al*., 1994) are correctly predicted.

Ribonuclease

Ribonucleases play an important role in RNA transcription. We have chosen two resolved complexes of ribonuclease A (6rsa; Borah *et al*., 1985) and ribonuclease T_1 (1rnt; Arni *et al*., 1987) to further test our docking approach.

6rsa

The structure of the complex of ribonuclease A with the ligand uridine vanadate (XII) has been solved by Borah *et al*. (1985).

The ligand consists of two large fragments: a uracil group and a bicycle with two fused five-membered rings, one of them containing a vanadate group. Because of the large number of potential hydrogen acceptors, the bicycle is selected as the base fragment. Since the ring generation program SCA is not able to handle vanadium, we used the ring conformation as given in the PDB, although there might be alternative conformations.

The complete docking algorithm revealed 128 different solutions. The first three solutions fall in an interval of less than 0.1 kJ/mol and are separated by 2 kJ/mol from the subsequent ones. These first three structures differ from each other only in the conformation of the hydroxy methylene side-chain. FLEXX was unable to find any interaction for the hydroxy group. As in the crystal structure, this group extends into the solvent. The bicycle and the pyrimidine moiety are placed analogously to the experimental structure. The binding affinity is overestimated by about 10 kJ/mol, presumably due to lacking parameters for the vanadate group.

1rnt

The ligand in this complex with ribonuclease T_1 is the nucleotide guanosine 2'-monophosphate (XIII). We selected the guanine group as base fragment. Due to a pK_a value of about 7 for the second ionisation step, the sugar phosphate ester exists under physiological conditions as a mixture of mono and dianions. For our docking experiments

we tested both forms. The calculations using the monoions produced the lower ΔG values and geometries closer to the crystal structure. Accordingly, only these results are presented.

As expected, the base fragment could be placed very accurately, due to specific interactions with Asn44, Tyr45, Glu46 and Asn98. The position of the adjacent ribose ring is difficult to predict correctly since it extends beyond the receptor surface into solvent. Nevertheless, the O3'-phosphate group at the ribose ring interacts with the protein. The protonated oxygen of the phosphate group is hydrogen bonded to the carboxylate group of Glu58. In principle, the highest-ranking docking solution shows the correct binding mode although it has an rms deviation of 1.48 Å. The solution closest to the experimental structure (with an rms of 0.96 Å) is found at rank 47.

Thermolysin

Thermolysin is a zinc protease. The following two complexes have been solved by Matthews (1988).

1tmn

The ligand *N*-(1-carboxylato-3-phenylprop-1-yl)-L-leucyl-L-tryptophan (XIV) has two carboxylate groups. We selected the N—CH—CO_2-fragment in the leucyl part as the base fragment. Considering 24 different conformations, FLEXX finds 148 placements. Within the first ten solutions there are at least two structures with rms deviations below 1 Å with respect to the crystal structure.

The complete docking algorithm generates 237 solutions. The highest-ranking solution (rms = 0.87 Å) has a binding energy of –39.1 kJ/mol, which is close to the experimentally observed value of –41.6 kJ/mol. The solution most similar to the crystal structure (rms = 0.69 Å) has rank 7 and differs only by 1.3 kJ/mol from the best one.

4tln

The ligand in this complex which has been solved with a resolution of 2.3 Å (Matthews, 1988) is much smaller than the previous inhibitor. L-leucyl-hydroxylamine (XV) is a hydroxamic acid that can exist in a number of different tautomers. In agreement with Judson *et al*. (1995), we selected the one shown in Figure 11. The hydroxamic acid was chosen as base fragment for the docking experiment.

In agreement with the experimentally observed structure, we reveal four base placements with bidented zinc coordination. The third-ranking solution has an rms deviation of only 0.59 Å.

After the subsequent addition of the remaining three fragments the algorithm predicts binding modes that differ from the experimental structure. The highest-ranking solution of FLEXX has an rms deviation of 4.5 Å. The experimentally observed geometry shows a synperiplanar arrangement of the ammonium and the *i*-propyl group. In our model, such a torsional angle (0°) is extremely unfavorable. By visual inspection of the experimental structure we were unable find any obvious steric or electronic reason for such a surprising conformation.

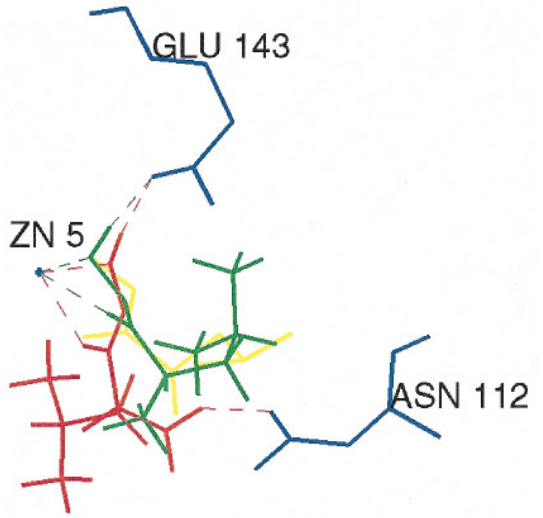

Figure 13. Docking results for XV. Geometry of the ligand in the crystal, yellow; highest ranking solution, red; first solution with an rms below 1 Å, green. For the calculated geometries, hydrogen atoms are also displayed. Broken lines represent the protein–ligand interactions.

The first solution with an rms below 1 Å has rank 118. It differs by more than 7 kJ/mol in the binding energy presumably because it shows one H-bond less than the experimental structure (Figure 13).

Carboxypeptidase A

Carboxypeptidase A specifically cleaves polypeptide chains. Like thermolysin, it is a zinc protease.

3cpa

The ligand in the complex 3cpa is the dipeptide glycyl-L-tyrosine (XVI). The experimental structure solved by Rees & Lipscomb (1983) to 2.0 Å resolution shows complexation of the zinc through the carbonyl group of the glycine residue.

Since the carboxylate group with only two interaction centers is selected as the base fragment, the alternative base placement algorithm (forming two instead of three interactions) is applied automatically.

After the complete docking procedure, FLEXX finds two distinct binding modes among the highest ranking solutions. In comparison to the crystal structure, the first and second solution show a completely different orientation of the base fragment. The carboxylate group forms hydrogen bonds not to Arg145 but to Arg127. As a consequence, the complex construction ends up with high rms deviations (3.08 and 2.62 Å). The third solution with a 2.5 kJ/mol less favorable

binding energy shows good agreement with the X-ray structure. The rms deviation is 1.06 Å.

2ctc

The ligand L-phenyllactate (XVII) can be formally divided into five fragments, three of which form the base fragment HO—CH—CO_2. Because of its internal flexibility 15 different conformations have to be considered. In the complex construction phase, only hydrophobic interactions can be used. However, a favorable position of the phenyl ring can be determined. FLEXX generates 144 different solutions. The highest-ranking solution approximates the experimentally observed binding pattern (rms deviation = 0.65 Å), only the hydrogen bond to Tyr248 is missing.

HIV protease

4phv

Inhibitors of HIV-1 protease are potential drugs in AIDS treatment. The ligand VAC (XVIII) in the complex 4phv, which has been solved by Bone *et al*. (1991), is another example with a high binding affinity (–52.2 kJ/mol). It is the largest inhibitor in our test set, with 17 rotatable bonds. This ligand reflects the pseudo C_2 symmetry of HIV-1 protease.

In our docking experiment we included the highly conserved water molecule, HOH1, at Ile50 and Ile50' with explicitly modeled hydrogen atoms. Altogether the active site consists of 69 potential interaction partners. The selected base fragment is the indanyl moiety.

The base placement algorithm finds 20 placements with the highest ranking solution showing an rms deviation of 0.85 Å. Also the pseudo C_2 symmetrical placement of the ligand is generated, but receives only rank 6, indicating that slightly less favorable interactions with the protein have been generated.

In the course of complex construction the amide group is subsequently added. Because of the interaction of its nitrogen with Gly27 and its oxygen with one of the hydrogens of HOH1, the docking algorithm establishes a solution which is in good agreement with the crystal ligand structure. Without the included water molecule the algorithm is not able to reproduce the crystallographic binding mode.

Analyzing the selected construction pathways dependent on the presence of HOH1 shows that this water molecule guides all nine complete solutions to virtually the same result, which closely approaches the crystal structure. The highest-ranking solution (see Figure 14) has an rms deviation of 1.04 Å and a calculated binding energy of –38.2 kJ/mol. The difference of about 14 kJ/mol compared to the experimental value indicates that, although the calculated structure fits the X-ray structure perfectly, not all possible interactions have been established.

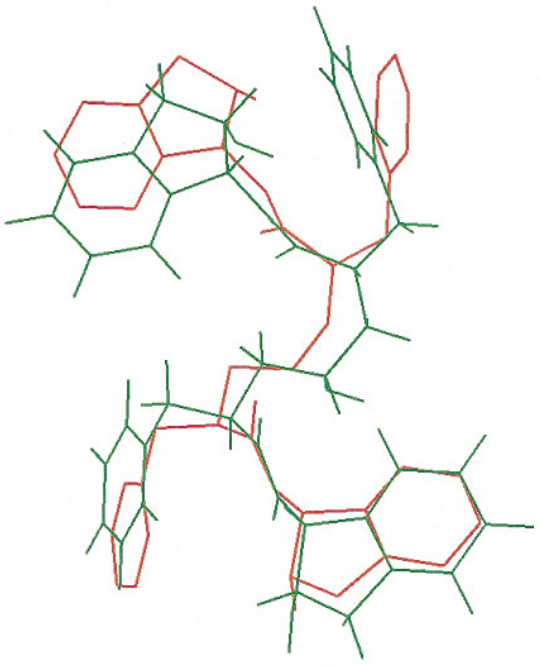

Figure 14. Docking results for XVIII. Ligand geometry in the crystal structure, red; best solution, green. For the calculated geometry hydrogen atoms are also displayed.

RAS protein

121p

The complex 121p between the oncogene protein H-RAS P21 and its inhibitor guanosine 5'-B,G-methylene-triphosphate (XIX) has been solved by Krengel (1991) with a resolution of 1.54 Å. Here a Mg^{2+} is part of the active site.

The ligand's guanine group displays an ideal base fragment. The base placement results in 102 different solutions, with the highest-ranking solution having 0.37 Å rms deviation. It exhibits the same hydrogen bonds as the crystal structure.

In the phase of complex construction, the nine remaining fragments are added. The algorithm reveals 44 solutions. The structure with the smallest rms deviation (1.14 Å) has rank 43. This solution has the correct base placement and the ribose ring and the phosphate chain fall close to the experimentally observed positions. The conformation of the triphosphate differs slightly and the terminal phosphate group does not coordinate the Mg^{2+}. Accordingly, this solution coordinates the metal ion only through one oxygen of the central phosphate unit. Most likely this missing interaction is responsible for a difference of about 20 kJ/mol in binding energy compared to the highest-ranking solution.

The highest-ranking solution has an rms deviation of 2 Å with respect to the crystal structure, but it exhibits a correct orientation of the base fragment and the terminal two phosphate units. All important interactions observed experimentally are also exhibited in this solution. Only the ribose and the first phosphate group deviate in

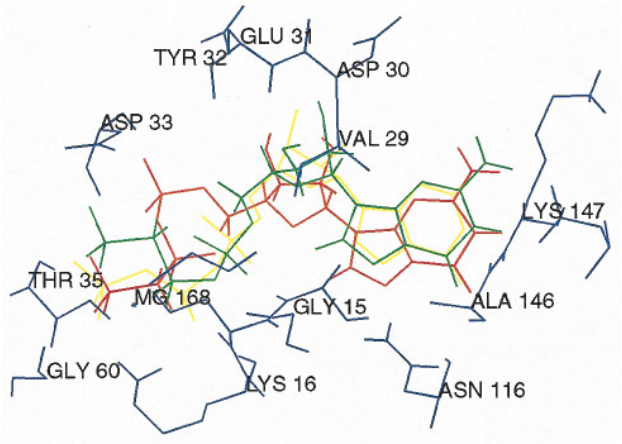

Figure 15. Docking results for XIX. Ligand geometry in the crystal structure, yellow; highest ranking solution, red; solution with rank 43, green. For the calculated geometries, hydrogen atoms are also displayed.

position. This part of the ligand extends beyond the binding site into the solvent (see Figure 15). However, we believe that this solution is a realistic docking.

Summary of results

All solutions obtained by FLEXX are summarized in Table 3. The first solution given corresponds to the lowest predicted free energy of binding (columns 3 to 4). The second one is the energetically most favorable prediction (columns 5 to 7) that approximates the experiment by less than 1 Å rms deviation. In cases, where no predicted placement satisfies this criterion, we have given the solution with the lowest rms. For comparison we list the binding energy obtained by applying the scoring function to the crystal structure (column 10) as well as the experimental binding energies (column 11). Note that the crystal configuration as well as the predicted configuration are not located in local minima with respect to the scoring function.

In columns 8 and 9, the run time of the base placement (column 8) and complex construction phase (column 9) are shown as elapsed run time on a SUN SPARCstation 20 (125 MHz HyperSPARC processor). Data preparation times (loading and assigning of chemical properties such as interaction geometries, torsion angles, etc.) are not given in the Table. The preparation of a ligand takes approximately two seconds and the computation of the ring conformations with SCA (depending on the complexity of the ring systems) takes approximately three seconds. The assignments concerning the protein take approximately ten seconds including the time for generating the hash table for the base placement algorithm (see Theory and Algorithm).

Conclusions

In this paper, we have presented a new approach to flexible ligand docking based on incremental construction. The approach contains efficient data structures for all steps that may present performance bottlenecks. Thus, a large number of alternative placements can be considered during the construction process within an acceptable amount of run time. Docking a typical drug molecule considering ligand flexibility (all rotatable bonds and flexible ring systems) can be performed in two to three minutes on a standard workstation. To the best of our knowledge this is ten times faster than any alternative approach for flexible ligand docking. The algorithm is fast enough for interactive work or for searching for potential inhibitors in a larger set of ligands. With appropriate pre-screening steps, a database search is possible.

Using a set of 19 protein–ligand complexes whose three-dimensional structure is known we have shown that FLEXX is able to reproduce the experimentally observed complexes. Docking algorithms are always based on simplifications which are necessary in order to keep the docking process computationally feasible. In summary, these simplifications are receptor rigidity, a discrete model to describe conformational flexibility, the model to predict the geometry of protein–ligand interactions, Böhm's empirical scoring function, and the heuristics used inside the algorithmic engines, especially the incremental construction method itself and the greedy strategy during complex construction. The results we obtain with FLEXX demonstrate that the simplifications in our model are reasonable. An exception is receptor rigidity, a problem that we have evaded by using the input structure as given in the bound state.

The major drawback of a docking algorithm based on incremental construction is the need for defining a base fragment. But we have shown that FLEXX is able to place even very small or flexible parts of the ligand correctly. Thus, it should be possible to find an appropriate base fragment in practically all kinds of ligands. The selection of the base fragment can be automated by evaluating a scoring function considering the number of possible interactions and the number of conformations of the fragments. A few alternatives can also be tried in order to avoid that a fragment with a limited number of putative contacts to the receptor is selected.

In several earlier works on structure-based drug design, a two-phase approach has been proposed (Kuntz *et al*., 1982; DesJarlais *et al*., 1988). In a first phase, protein–ligand complexes are generated solely on the basis of geometric conditions, and, in a second phase, the generated complexes are ranked by energy. This approach is based on the hypothesis that steric complementarity is a good descriptor for protein–ligand complexes.

Steric complementarity is of course necessary for low-energy protein–ligand complexes because the important interactions are weak and short-range. But most ligands show inherent conformational flexibility and can adopt a shape which is complementary to the receptor. Thus, steric

complementarity is a useful descriptor for protein–ligand complexes only if the ligand conformation is known, which is quite an unrealistic assumption from the practical point of view.

FLEXX is only one step towards a reliable docking tool. But the method provides a robust basis for further improvements on the models as well as on the algorithms used. We summarize the problems which, in our opinion, are most limiting and should be tackled in future.

As we have seen from the complex of HIV-1 protease, water plays an important role in complex formation. An appropriate strategy for handling discrete water molecules mediating hydrogen bonds between the ligand and the protein is necessary. Considering the receptor as a rigid object is of course the most severe restriction in FLEXX. Besides methods based on molecular dynamics, only one approach handling receptor flexibility has so far been published (Leach, 1994). Leach's algorithm models side-chain flexibility and is based on tree search techniques. Unfortunately, tests on the method have been reported on only two small examples.

In some of our examples the predicted free energy of binding deviates substantially from the experimentally observed energy. In these cases, important energetic contributions have not yet been modeled and an improvement of the scoring function is required. Here, work is in progress (Böhm, unpublished). Of course, using a simple function based on a single configuration of a receptor-ligand complex for scoring can never be more than a rough approximation of the free energy of binding. But Böhm has shown that his scoring function is able to achieve predictions in good agreement with experimentally observed energies for a wide set of examples.

Finally, FLEXX is limited to medium-size ligands such as those used in the test set. This is not a problem of limited time or space. Rather, it is a limitation resulting from the heuristic greedy strategy employed during complex construction. The assumption underlying the greedy strategy is that the partial solutions during the construction of the optimal solution are always nearly optimal. As the number of fragments increases, this assumption becomes more and more unrealistic. We have shown that ligands with 17 rotatable bonds are manageable, which is sufficient for most practical purposes in drug design. If one wants to predict complexes containing ligands with more than 20 fragments, other search strategies seem to be required.

Acknowledgements

This work was performed as part of the RELIWE-Project, which is partially funded by the German Federal Ministry for Education, Science, Research, and Technology (BMBF) under grant no. 01 IB 302 A. The cooperation partners in the Project are BASF AG, Central Research Ludwigshafen, E.Merck Preclinical Research Darmstadt, EMBL Heidelberg and GMD institutes IPSI (Institute for Integrated Publication and Information Systems) Darmstadt and SCAI (Institute for Algorithms and Scientific Computing) St. Augustin. We thank all our partners in the project for an excellent cooperation, especially Hans-Joachim Böhm, Thomas Mietzner, and Hugo Kubinyi from BASF AG, and our colleagues in St. Augustin Stephan Wefing and Christian Lemmen.

References

Abad-Zapatero, C., Griffith, J. P., Sussman, J. & Rossmann, M. G. (1987). Refined crystal structure of dogfish M_4 apo-lactate dehydrogenase. *J. Mol. Biol.* **198**, 445–467.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink-Peters, T., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallog. sect. B*, **35**, 2331–2339.

Arni, R., Heinemann, U., Maslowska, M., Tokuoka, R. & Saenger, W. (1987). Restrained least-squares refinement of the crystal structure of the ribonuclease T_1*2' guanylic acid complex at 1.9 Å resolution. *Acta Crystallog. sect. B*, **43**, 548–554.

Banner, D. W. & Hadvary, P. (1991). Crystallographic analysis at 3.0-Å resolution of the binding to human thrombin of four active site-directed inhibitors. *J. Biol. Chem.* **266**, 20085–20093.

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). The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Billeter, M., Havel, T. F. & Kuntz, I. D. (1987). A new approach to the problem of docking two molecules: the ellipsoid algorithm. *Biopolymers*, **26**, 777–793.

Blaney, J. M. & Dixon, J. S. (1993). A good ligand is hard to find: automated docking methods. *Perspect. Drug Disc. Design*, **1**, 301–319.

Bode, W. (1979). The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. *J. Mol. Biol*. **127**, 357–374.

Bode, W., Turk, D. & Stürzebecher, J. (1990). Geometry of binding of the benzamidine- and arginine-based inhibitors *N*-α-(2-naphthyl-sulphonyl-glycyl)-DL-*p*-amidinophenyl-alanyl-piperidine (NAPAP) and (2*R*,4*R*)-4-methyl-1-[*N*-alpha-(3-methyl-1,2,3,4-tetrahydro-8-quinolinesulphonyl)-L-arginyl]-2-piperidine carboxylic acid (MQPA) to human alpha-thrombin—X-ray crystallographic determination of the NAPAP-trypsin complex and modeling of NAPAP-thrombin and MQPA-thrombin. *Eur. J. Biochem.* **193**, 175–182.

Böhm, H.-J. (1992a). The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Design*, **6**, 61–78.

Böhm, H.-J. (1992b). LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Design*, **6,** 593–606.

Böhm, H.-J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Design*, **8**, 243–256.

Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. & Kraut, J. (1982). Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase

refined at 1.7 Å resolution. I. General features and binding of methotrexate. *J. Biol. Chem.* **257**, 13650–13662.

Bone, R., Vacca, J. P., Anderson, P. S. & Holloway, M. K. (1991). X-ray crystal structure of the HIV protease complex with L-700,417, an inhibitor with pseudo C_2 symmetry. *J. Am. Chem. Soc.* **113**, 9382–9384.

Borah, B., Chen, C.-W., Egan, W., Miller, M., Wlodawer, A. & Cohen, J. S. (1985). Nuclear magnetic resonance and neutron diffraction studies of the complex of ribonuclease A with uridine vanadate, a transition-state analogue. *Biochemistry*, **24**, 2058–2067.

Brick, P. & Blow, D. M. (1987). Crystal structure of a deletion mutant of a tyrosyl-tRNA synthetase complexed with tyrosine. *J. Mol. Biol.* **194**, 287–297.

Bunting, J. W. & Myers, C. D. (1975). Reversible inhibition of carboxypeptidase A. IV. Inhibition of specific esterase activity by hippuric acid and related species and other amino acid derivates and a comparison with substrate inhibition. *Can. J. Chem.* **53**, 1993–2004.

Colman, P. M. (1994). Structure-based drug design. *Curr. Opin. Struct. Biol.* **4**, 868–874.

Connolly, M. L. (1983). Analytical molecular surface calculation. *J. Appl. Crystallog.* **16**, 548–558.

Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation*. Research Studies Press Ltd, Taunton, England.

DeClercq, P. J. (1984a). Systematic conformational analysis. A microcomputer method for the semiquantitative evaluation of polycyclic systems containing five-, six-, and seven-membered rings. 1. Program characteristics. *Tetrahedron*, **40**(19), 3717–3727.

DeClercq, P. J. (1984b). Systematic conformational analysis. A microcomputer method for the semiquantitative evaluation of polycyclic systems containing five-, six-, and seven-membered rings. 2. Scope and limitations. *Tetrahedron*, **40**(19), 3729–3738.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D. & Venkataraghavan, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **29**, 2149–2153.

DesJarlais, R. L., Sheridan, R. P., Seiberl, G. L., Dixon, J. S., Kuntz, I. D. & Venkataraghavan, R. (1988). Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **31**, 722–729.

DiNola, A., Roccatano, D. & Berendsen, H. J. C. (1994). Molecular dynamics simulation of the docking of substrates to proteins. *Proteins: Struct. Funct. Genet.* **19**, 174–182.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., New York.

Ealick, S. E., Babu, Y. S., Bugg, C. E., Erion, M. D., Guida, W. C., Montgomery, J. A. & Secrist J. A., III (1991). Application of crystallographic and modeling methods in the design of purine nucleoside phosphorylase inhibitors. *Proc. Natl Acad. Sci. USA*, **88**, 11540–11544.

Entsch, B., Ballou, D. P. & Massey, V. J. (1976). Flavin-oxygen derivatives involved in hydroxylation by p-hydroxybenzoate hydroxylase. *J. Biol. Chem.* **251**, 2550–2563.

Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **248**, 459–477.

Ghose, A. K. & Crippen, G. M. (1985). Geometrically feasible binding modes of a flexible ligand molecule at the receptor site. *J. Comput. Chem.* **6**(5), 350–359.

Gillet, V. J., Johnson, P. & Sike, S. (1990). Automated structure design in 3D. *Tetrahedron*, **3**(6), 681–696.

Gillet, V. J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z. & Johnson, A. P. (1994). SPROUT: Recent developments in the *de novo* design of molecules. *J. Chem. Inf. Comput. Chem.* **34**, 207–217.

Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Protein: Struct. Funct. Genet.* **8**, 195–202.

Hilpert, K., Ackermann, J., Banner, D. W., Gast, A., Gubernator, K., Hadvary, P., Labler, L., Müller, K., Schmid, G., Tschopp, T. B. & van de Waterbeemd, H. (1994). Design and synthesis of potent and highly selective thrombin inhibitors. *J. Med. Chem.* **37**, 3889–3901.

Hoflack, J., DeClercq, P. J. & Cauberghs, S. (1989). SCA: Systematic conformational analysis. *Quantum Chemical Program Exchange* QCPE program QCMP079, Indiana University, Bloomington, IN, USA.

Holmes, M. A. & Matthews, B. W. (1981). Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis. *Biochemistry*, **20**, 6912–6920.

Jones, G., Willet, P. & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53.

Judson, R. S., Tan, Y. T., Mori, E., Melius, C., Jaeger, E. P., Treasurywala, A. M. & Mathiowetz, A. (1995). Docking flexible molecules: A case study of three proteins. *J. Comput. Chem.* **16**(11), 1405–1419.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A*, **32**, 922–923.

Klebe, G. (1994). The use of composite crystal-field environments in molecular recognition and the *de-novo* design of protein ligands. *J. Mol. Biol.* **237**, 221–235.

Klebe, G. & Mietzner, T. (1994). A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Design*, **8**, 583–606.

Krengel, U. (1991). Struktur und Guanosintriphosphat-Hydrolysemechanismus des C-terminal verkürzten menschlichen Krebsproteins P21-h-RAS. Thesis, Heidelberg.

Kuntz, I. D. (1992). Structure-based strategies for drug design and discovery. *Science*, **257**, 1078–1082.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. L. & Ferrin, T. E. (1982). A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161**, 269–288.

Kuntz, I. D., Meng, E. C. & Shoichet, B. K. (1994). Structure-based molecular design. *Acc. Chem. Res.* **27**(5), 117–123.

Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **235**, 345–356.

Leach, A. R. & Kuntz, I. D. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. *J. Comput. Chem.* **13**, 730–748.

Lewis, R. A. & Leach, A. (1994). Current methods for site-directed structure generation. *J. Comput.-Aided Mol. Design*, **8**, 467–475.

Lindquist, R. N., Lynn, J. L. Jr & Lienhard, G. E. (1973). Possible transition-state analogues for ribonuclease. The complexes of uridine with oxovanadium (IV) ion and vanadium (V) ion. *J. Am. Chem. Soc.* **95**, 8762–8768.

Linnainmaa, S., Harwood, D. & Davis, L. S. (1988). Pose determination of a three-dimensional object using

triangle pairs. *IEEE Trans. Pattern Anal. Machine Intelligence,* **10**(5), 634–646.

Luty, B. A., Wasserman, Z. R., Stouten, P. F. W., Hodge, C. N., Zacharias, M. & McCammon, J. A. (1995). A molecular mechanics / grid method for evaluation of ligand-receptor interactions. *J. Comput. Chem.* **16**(4), 454–464.

Mares-Guia, M. & Shaw, E. (1965). Studies on the active center of trypsin. The binding of amidines and guanidines as models of the substrate side chains. *J. Biol. Chem.* **240**, 1579–1585.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallog. sect. B,* **39**, 480–490.

Matthews, B. W. (1988). Structural basis of the action of thermolysin and related zinc peptidases. *Acc. Chem. Res.* **21**, 333–340.

Mizutani, M. Y., Tomioka, N. & Itai, A. (1994). Rational automatic search method for stable docking models of protein and ligand. *J. Mol. Biol.* **243**, 310–326.

Monzingo, A. F. & Matthews, B. W. (1984). Binding of *N*-carboxymethyl dipepetide inhibitors to thermolysin determined by X-ray crystallography. A novel class of transition-state analogues for zinc peptidases. *Biochemistry,* **23**, 5724–5729.

Moon, J. B. & Howe, W. J. (1991). Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins: Struct. Funct. Genet.* **11**, 314–328.

Nishibata, Y. & Itai, A. (1991). Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron,* **47**(43), 8985–8990.

Olson, C. F. (1994). Time and space efficient pose clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 251–258, IEEE Computer Society Press, Seattle, Washington.

Oshiro, C. M., Kuntz, I. D. & Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Design*, **9**, 113–130.

Peters, K. P., Fauck, J. & Frömmel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201–213.

Platzer, J. E. B., Momany, F. A. & Scheraga, H. A. (1972a). Conformational energy calculations of enzyme-substrate interactions. *Int. J. Peptide Protein Res.* **4**, 187–200.

Platzer, J. E. B., Momany, F. A. & Scheraga, H. A. (1972b). Conformational energy calculations of enzyme-substrate interactions. *Int. J. Peptide Protein Res.* **4**, 201–219.

Rarey, M., Kramer, B. & Lengauer, T. (1995). Time-efficient docking of flexible ligands into active sites of proteins. In *Proceedings of the Third International Conference on Intelligent Systems in Molecular Biology* (Rawlings, C. *et al*., eds), pp. 300–308, AAAI Press, Menlo Park, California.

Rarey, M., Wefing, S. & Lengauer, T. (1996). Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Design,* **10**, 41–54.

Rees, D. C. & Lipscomb, W. N. (1983). Crystallographic studies on apocarboxypeptidase A and the complex with glycyl-l-tyrosine. *Proc. Natl Acad. Sci. USA,* **80**, 7151–7154.

Rotstein, S. H. & Murcko, M. A. (1993). GenStar: a method for *de novo* drug design. *J. Comput.-Aided Mol. Design,* **7**, 23–43.

Sandak, B., Nussinov, R. & Wolfson, H. J. (1995). An automated computer vision and robotics-based technique for 3-D flexible biomolecular docking and matching. *Comput. Appl. Biosci.* **11**(1), 87–99.

Smellie, A. S., Crippen, G. M. & Richards, W. G. (1991). Fast drug-receptor mapping by site-directed distances: a novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Chem.* **31**, 386–392.

Teplyakov, A., Wilson, K. S., Orioli, P. & Mangani, S. (1993). High resolution crystal structure of the complex between carboxypeptidase A and L-phenyl lactate. *Acta Crystallog. sect. D,* **49**, 534–540.

TRIPOS Associates, Inc., St. Louis, Missouri, USA. (1994). *SYBYL Molecular Modeling Software Version 6.x*.

van der Laan, J. M., Schreuder, H. A., Swarte, M. B. A., Wierenga, R. K., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1989). The coenzyme analogue adenosine 5-diphosphoribose displaces fad in the active site of *p*-hydroxybenzoate hydroxylase. An X-ray crystallographic investigation. *Biochemistry,* **28**, 7199–7205.

Verlinde, C. L. M. J., Noble, M. E. M., Kalk, K. H., Groendijk, H., Wierenga, R. K. & Hol, W. G. J. (1991). Anion binding at the active site of trypanosomal triosephosphate isomerase. Monohydrogen phosphate does not mimic sulphate. *Eur. J. Biochem.* **198**, 53–57.

Weber, P. C., Ohlendorf, D. H., Wendoloski, J. J. & Salemme, F. R. (1989). Structural origins of high-affinity biotin binding to streptavidin. *Science,* **243**, 85–88.

Weber, P. C., Wendoloski, J. J., Pantoliano, M. W. & Salemme, F. R. (1992). Crystallographic and thermodynamic comparison of natural and synthetic ligands bound to streptavidin. *J. Am. Chem. Soc.* **114**, 3197–3200.

Wells, T. N. C. & Fersht, A. R. (1986). Use of binding energy in catalysis analyzed by mutagenesis of the tyrosyl-tRNA synthetase. *Biochemistry,* **25**, 1881–1886.

White, J. L., Hackert, M. L., Buehner, M., Adams, M. J., Ford, G. C., Lentz, P. J. Jr, Smiley, I. E., Steindel, S. J. & Rossmann, M. G. (1976). Comparison of the structures of apo dogfish M_4 lactate dehydrogenase and its ternary complexes. *J. Mol. Biol.* **102**, 759–779.

Wierenga, R. K., Noble, M. E. M., Vriend, G., Nauche, S. & Hol, W. G. J. (1991). Refined 1.83 Å structure of trypanosomal triosephosphate isomerase, crystallized in the presence of 2.4 M-ammonium sulphate. A comparison with the structure of the trypanosomal triosephosphate isomerase-glycerol-3-phosphate complex. *J. Mol. Biol.* **220**, 995–1015.

Zollner, H. (1993). *Handbook of Enzyme Inhibitors*. VCH Publishers, Weinheim.

Edited by F. E. Cohen

(Received 11 March 1996; received in revised form 6 June 1996; accepted 10 June 1996)