Time-Efficient Docking of Flexible Ligands into Active Sites of Proteins

Matthias Rarey and **Bernd Kramer** and **Thomas Lengauer**
German National Research Center for Computer Science (GMD)
Institute for Algorithms and Scientific Computing (SCAI)
Schloß Birlinghoven
53754 Sankt Augustin, Germany
rarey@gmd.de bernd.kramer@gmd.de lengauer@gmd.de

Abstract

We present an algorithm for placing flexible molecules in active sites of proteins. The two major goals in the development of our docking program, called FLEXX, are the explicit exploitation of molecular flexibility of the ligand and the development of a model of the docking process that includes the physico-chemical properties of the molecules. The algorithm consists of three phases: The selection of a *base fragment*, the placement of the base fragment in the active site, and the incremental construction of the ligand inside the active site. Except for the selection of the base fragment, the algorithm runs without manual intervention. The algorithm is tested by reproducing 11 receptor-ligand complexes known from X-ray crystallography. In all cases, the algorithm predicts a placement of the ligand which is similar to the crystal structure (about 1.5 Å RMS deviation or less) in a few minutes on a workstation, assuming that the receptor is given in the bound conformation.

Keywords: molecular docking, flexible docking, receptor ligand interaction, molecular flexibility, conformational analysis, drug design

1 Introduction

Most biochemical processes in living systems are based on the specific binding of small organic molecules, the *ligands*, to the active sites of proteins, called *receptors* in this context. A major goal of pharmaceutical research is to control such processes by designing molecules with high binding affinity to a given receptor molecule. The first step in the drug design process is the search for a *lead structure*, i. e. a molecule with high binding affinity to a given receptor. If the three-dimensional structure of the receptor is known, rational drug-design techniques are applicable. The *docking problem*, which is the key problem in rational drug design, can be stated as follows:

> Given a three-dimensional structure of a receptor and the structure of a ligand molecule, predict the binding affinity of the ligand to the receptor and the geometry of the receptor-ligand complex.

The prediction of the binding affinity of a ligand molecule allows for the extraction of potential drugs from a large set of molecules, while knowledge of the geometry of the complex gives important insights into the binding mode and makes a focussed optimization of the potential drug molecule possible.

Since the historically first docking algorithm was published (Kuntz *et al.* 1982), a wide variety of different methods for this problem have been developed. Overviews of docking and structure generation algorithms can be found in (Lewis & Leach 1994; Kuntz, Meng, & Shoichet 1994). Most of the docking algorithms handle the ligand molecule as a rigid object (Kuhl, Crippen, & Friesen 1984; DesJarlais *et al.* 1988; Jiang & Kim 1991; Smellie, Crippen, & Richards 1991; Lawrence & Davis 1992; Meng, Shoichet, & Kuntz 1992; Bacon & Moult 1992; Kasinos *et al.* 1992; Shoichet & Kuntz 1993). Thus, the run time grows linearly with the number of energetically favorable conformations, and this number can be large under realistic conditions. Therefore, more recent approaches try to integrate molecular flexibility directly into their concepts (DesJarlais *et al.* 1986; Goodsell & Olson 1990; Leach & Kuntz 1992; Sandak, Nussinov, & Wolfson 1994; Kearsley *et al.* 1994; Mizutani, Tomioka, & Itai 1994). One possibility of doing so is to simulate the docking process in a more or less abstract way (Goodsell & Olson 1990; Yue 1990; Di Nola, Roccatano, & Berendsen 1994). So far, these approaches have turned out to be quite time-consuming and the quality of the results depends on the initial placement of the ligand. More frequently, flexible docking is based on partitioning the ligand into small connected pieces, called *fragments*. There are two different strategies for recombining the fragments. First, fragments may be placed in the active site independently and reconnected by using the yet unplaced parts of the molecule (DesJarlais *et al.* 1986; Sandak, Nussinov, & Wolfson 1994). Alternatively, one selected fragment, often called *base* or *anchor fragment*, may be placed in the active site and the ligand reconstructed inside the active site incrementally starting with the base fragment (Leach & Kuntz 1992).

The overall goal of our research is to develop a software tool for docking that is fast and reliable.

We want the tool to be fast enough to allow for docking large sets of ligands into a given receptor pocket within a matter of hours (as part of a search over a database of ligands) or suggesting alternative conformations for docking a single ligand within a matter of minutes (for interactive dialogs at a workstation). The reliability of the output is a little harder to quantify. We do not believe that, with the methods and computers available today, it is possible to give accurate estimates of free energy of a representative set preferred binding modes within a matter of minutes on a workstation. However, we require that the tool compute a representative set of low-energy binding modes. Furthermore these binding modes should be ranked approximately accurately with respect to free energy. Finally, the 3d-coordinates of the binding modes should be accurate to within the limits of the error involved in experimental measurements (roughly up to 1A rms deviation).

In order to achieve these goals, we employ the *incremental construction* strategy that was originally developed for de-novo design of ligands (Moon & Howe 1991). With the receptor held rigid, we model the more essential flexibility of the ligand explicitly. Geometric and physico-chemical properties of the molecules are modeled explicitly, as well, and only chemically meaningful placement of the ligand are generated.

For the sake of speed, we developed a new algorithm for placing the base fragment, and a number of time- and space-efficient techniques for the incremental construction process. Instead of using time-consuming force-field calculations, we base the scoring of the (partial) placements on a variation of H.-J. Böhm's empirical energy function (Böhm 1994) designed for the structure generation tool LUDI (Böhm 1992a; 1992b). Therefore, we are able to give a rough estimate of the binding affinity, in addition to computing the geometry of the complex.

Apart from the selection of the base fragment, our tool works without manual intervention. Up to date, we have successfully tested our tool on 11 docking examples. The run times are within a few minutes on a SUN SPARCstation 20. We are constantly increasing our test set, in order to make the tool more reliable.

The following section describes the chemical modeling in FLEXX. Section 3 contains a brief description of the algorithms and data structures that FLEXX uses. The last section gives a summary of the results obtained with our docking tool.

2 The chemical modeling in FLEXX

Receptor structures The first part of the input for the docking problem is the three-dimensional conformation of the receptor, especially, of the active site. FLEXX gets this information from a standard PDB file (Bernstein *et al.* 1977) and a user-defined receptor-description file. The active site must be identified by the user. This only weakly diminishes the power of our approach, because we use FLEXX as an element of a software-package that includes tools and databases for predicting and storing geometries of active sites. The receptor-description file contains the assignment of amino acid templates to the receptor. The templates contain atom and bond types, possible interacting groups and the numbers and relative positions of polar hydrogen atoms.

Ligand structures We model the flexibility of the ligand by allowing for a finite set of torsion angles for each freely rotatable single bond and finite sets of low-energy conformations for each ring system inside the ligand. All remaining torsion angles inside the ligand as well as its bond angles and bond lengths are taken from an energy-minimized conformation of the ligand. In order to assign finite sets of torsion angles to each single bond, FLEXX uses data of a statistical analysis of the CSDB (Allen *et al.* 1979) generated by G. Klebe for the conformational search program MIMUMBA (Klebe & Mietzner 1994). The calculation of the structure and the strain energy of the possible conformations of ring systems is carried out with the QCPE program SCA (Hoflack & De Clercq 1988). Only conformations with energies lying below the threshold of 20 kJ/mol (relative to the observed minimum) are chosen.

Molecular interactions The receptor-ligand interaction is modeled by a few special types of interactions. These are hydrogen bonds, metal-acceptor bonds and a few types of hydrophobic contacts (see Table 1). An interaction is modeled by an *interaction center* and an *interaction surface* located on a sphere around the center (fig. 1). An interaction between two groups A and B occurs, if the interaction center of group B lies on the interaction surface of group A and vice versa (fig. 2). In FLEXX the interaction surfaces in the receptor are modeled by sets of discrete *interaction points.*

Furthermore, a repulsive interaction between receptor and ligand is implicitly taken into account by the clash tests which FLEXX performs in order to eliminate placements with large overlap volumes. We use van-der-Waals radii from Pauling (Pauling 1939) in combination with a united atom model, and set the thresholds for individual atom-atom overlaps to $2.5Å^3$ and for the average atom-atom overlaps to $1.0Å^3$.

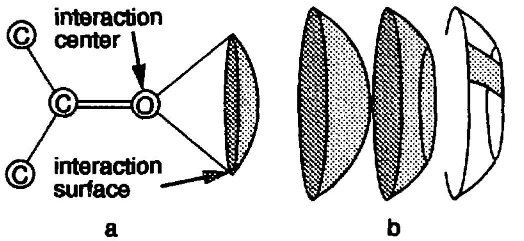

Figure 1: Interaction surface of a carbonyl group (a), and the three different types of interaction surfaces that are sections of a sphere: cone, cone section, spherical rectangle (b).

As mentioned above, FLEXX uses a function similar to that developed by H.-J. Böhm (Böhm 1994) in order to score the placements and to estimate the free binding energy ΔG of the protein-ligand complex.

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

Here, $f(\Delta R, \Delta\alpha)$ is a suitable function and N_{rot} counts the free rotatable bonds that are hindered in the complex. The terms ΔG_{hb}, ΔG_{io}, ΔG_{rot}, and ΔG_0 are gaugeable parameters. These values and the function f are taken from Böhm. However the lipophilic contact term has been changed and the last term is a new and takes into account the interactions of aromatic groups (with the new parameters $\Delta G_{lipo} = -0.34$kJ/mol and $\Delta G_{aro} = -0.7$kJ/mol).

The lipophilic contact energy term of Böhm is intended to be proportional to the lipophilic contact area. Böhm estimates this area with a grid method. However, our preliminary investigations with this notion of contact energy have generated placements that deviate markedly from the crystal structure. Therefore, we chose to calculate this term over a sum of pairwise atom-atom contacts, instead. It is important that the function $f^*(\Delta R)$ in (1) penalize forbiddingly close contacts. Because of that, we choose:

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

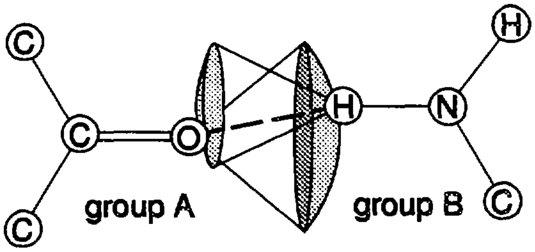

Figure 2: Hydrogen bond between a carbonyl oxygen and a nitrogen. The interaction occurs, if the interaction centers (oxygen and hydrogen atom) lie on the interaction surface of the counter group.

Here, ΔR is the difference of the distance between the atoms centers and the sum of both van-der-Waals radii, each increased by 0.3Å.

Finally it should be mentioned that the energy function does not take into account any explicit intramolecular interaction. We implicitly account for such a term by dropping all ligand conformations with unfavorable torsion angles or ring conformations or internal clashes.

3 The FLEXX algorithm

FLEXX follows an *incremental construction* strategy, which consists of three phases:

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

Base placement The base fragment is the part of the ligand, which is placed into the active site first and independently of the rest of the ligand. This is done in the second phase, the so-called *base placement* phase.

Construction In the last phase, the ligand is constructed in an incremental way, starting with the different placements of the base fragment. This phase is called the *construction phase*.

In the current version of FLEXX, the base selection is not automated; it must be done interactively by the user of the docking tool. The base fragment selection involves a tradeoff between two properties of the fragments. On the one hand, the base fragment should have a large set of interacting groups, such that a placement of the fragment into the active site can be found. On the other hand, the number of low-energy conformations of the base fragment should be small, because each conformation must be considered separately in the base placement phase.

3.1 Placing the base fragment

The fragment placing algorithm is based on a technique from the area of pattern recognition, called

Type of interaction group	Example	Geometry	Center of sphere (section)
H-bond donors	O–H	cone	H
	N–H	cone	H
metal ions	Zn^{2+}	sphere	Zn
H-bond or metal-ion acceptors	C=O in COO^-	two spher. rect.	O
	C=O, S=O, V=O	cone	O
	R–O–H	two spher. rect.	O
	R_2O, R_3N	cone	O, N
aromatic groups		two cones above and below plane	center of the ring
groups interacting with aromatic groups	aromatic groups	cone	H
	methyl group	sphere	C
	amide group	two cones above and below plane	middle of C–N bond

Table 1: A few examples of the different interaction type groups and their geometric requirements (see fig. 1). (For the complete list and the quantitative data of the surfaces please contact the authors)

pose clustering (Linnainmaa, Harwood, & Davis 1988; Olson 1994). The basic idea of the algorithm is the following: A transformation of a fragment in a given conformation into the active site of the receptor is determined by three simultaneously occurring interactions between the fragment and the receptor (see fig. 3). Thus, the algorithm considers each triple of interaction centers (q_0, q_1, q_2) in the fragment and searches for triples of interaction points (p_0, p_1, p_2) in the active site with the following three properties:

1. The interaction types of q_i and p_i are complementary, i.e., an interaction between q_i and p_i is chemically meaningful, for $i = 0, 1, 2$.
2. The pairwise distances between the interaction centers and the interaction points are nearly equal, i.e. $|\,\|q_i - q_{i+1\bmod 3}\| - \|p_i - p_{i+1\bmod 3}\|\,| \leq \delta$, for $i = 0, 1, 2$. Here, δ is a parameter describing the tolerance in the matching step.
3. After the fragment is placed on the interaction points, the additional angular constraints at the ligand's interaction groups are valid.

A pair of triangles, which fulfill the first two properties is called *δ-compatible.*

In principle, a fragment may form more than three interactions. Thus, a subsequent clustering step is needed, which merges similar transformations of the fragment to a single resulting placement (see fig. 4). Finally, each remaining placement computed is tested for overlap with the receptor. The non-overlapping placements are ranked by their decreasing free energy of binding, as estimated by the algorithm.

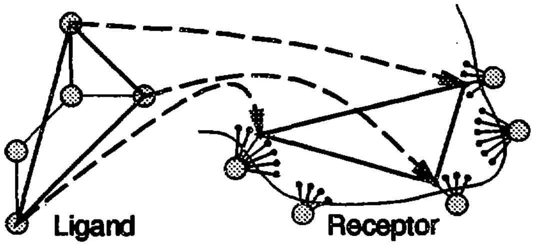

Figure 3: The fragment placing algorithm: mapping three interaction centers (grey spheres) of the ligand onto three discrete interaction points in the active site (black dots) defines a unique transformation of the ligand into the active site.

The approach contains two algorithmic parts, which we consider in more detail: the search for similar triangles in three-dimensional space and the clustering of transformations for the fragment.

Matching: searching for similar triangles During the first part of the fragment placing algorithm, a search for all δ-compatible triangles (p_0, p_1, p_2) of interaction points must be performed for each triangle (q_0, q_1, q_2) of interaction centers of the fragment. In a naive approach, the search time is $O(m^3)$ for each triple of interaction centers, where m is the number of interaction points in the active site. This time can be reduced by precomputing all possible triangles and using the interaction types and edge lengths in order to address the triangle in a hash table. Under the practical assumption that the access of the triangles via a hash function needs $O(1)$ time, the run time of the

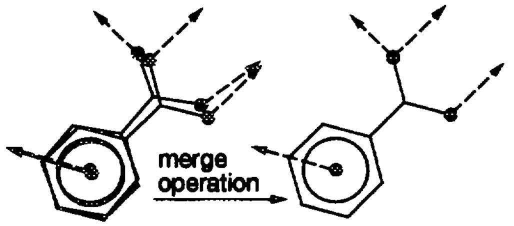

Figure 4: Merging initial placements: Two initial placements with different sets of interactions but similar transformations (left) will be combined to one placement (right) by merging the sets of interactions and recomputing a transformation.

search is proportional to the number of δ-compatible triangles retrieved and is asymptotically optimal. Nevertheless, the preprocessing time and storage requirement increases to $O(m^3)$, which is too large for practical cases.

We have developed a data structure for the triangle searching problem which needs $O(m^2)$ storage and is as fast as the triangle hashing method, in practice. Instead of storing each triangle, each line segment is precomputed and stored in a table. Let (p_0, p_1) be a line segment, consisting of two interaction points with interaction types $t(p_i), i = 0, 1$. The interaction types determine the table, where the line segment is stored. Each table divides the possible range of line segment lengths equally into buckets of width 2δ. The line segment (p_0, p_1) of length $d = ||p_0 - p_1||$ is stored in bucket k, with $2\delta k \leq d < 2\delta(k+1)$.

For retrieving all line segments, which are δ-compatible to a given line segment (q_0, q_1), we proceed analogously. Note that two buckets must be considered during retrieval: the bucket containing the length of the line segment $||q_0 - q_1||$ and one adjacent bucket. Each line segment (p_0, p_1) contained in one of the two buckets must be tested with respect to whether it fulfills the distance property, i. e., $|\,||q_0 - q_1|| - ||p_0 - p_1||\,| \leq \delta$. In each bucket, the line segments will be sorted by the *index of the first interaction point* p_0 . This is an important feature of the triangle query algorithm, which will now be explained.

Assume we are searching for all triangles which are δ-compatible with a given triangle (q_0, q_1, q_2) out of a set P of interaction points. The algorithm starts by searching for all line segments being δ-compatible with (q_0, q_1) and (q_0, q_2). We obtain two sorted lists L_i of line segments from the set P, whose elements are δ-compatible with (q_0, q_i), for $i = 1, 2$, respectively. All δ-compatible triangles are then constructed from the lists L_i by a list merging procedure. Let (p_i, p_j) be the current head of list L_0, and let (p_k, p_l) be the head of L_1. If $i = k$ and $|\,||p_j - p_l|| - ||q_1 - q_2||\,| \leq \delta$, then we have a new δ-compatible triangle (p_i, p_j, p_l). It is straightforward to prove that the algorithm detects exactly all δ-compatible triangles out of the set of interaction points P.

Clustering: combining placements Each matching of a triple of interaction centers of the ligand onto a triple of interaction points of the receptor yields an initial placement P. P consists of a set M of three interactions between the two molecules and a transformation (t, R) of the fragment into the active site, where t is an translation vector and R is a rotation matrix. Obviously, there can be initial placements with similar transformations (see fig. 4). Thus, a subsequent clustering step is necessary. We apply a complete-linkage hierarchical clustering algorithm (Duda & Hart 1973), which is not be explained in detail, here (Rarey, Wefing, & Lengauer 1995). A survey of cluster algorithms can be found in (Duda & Hart 1973; Murtagh 1983; Dorndorf & Pesch 1994).

The distance between two placements $P_i = (t_i, R_i, M_i)$, $i = 0, 1$ is defined to be the root-mean-square deviation between the atoms of the fragment placed at the two positions, i. e. $d(P_0, P_1) = \sqrt{(}1/n \sum_{i=1}^{n}((R_1 x_i + t_1) - (R_2 x_i + t_2))^2)$. Here, the fragment consists of n atoms with initial coordinates $x_1, \ldots, x_n$. A new placement out of a cluster of initial placements is computed by merging the sets of matches from each initial placement. The transformation of the new placement is recomputed by superposing (Ferro & Hermans 1977) the interaction centers of the matches to the corresponding interaction points (see fig. 4).

3.2 Constructing the ligand within the active site

Once the base fragment is placed into the active site, we can start an incremental construction process, beginning at the base fragment. We use a simple greedy strategy: In each iteration, a new fragment is joined to all placements found so far in all possible conformations. Placements having overlap with the receptor are eliminated. For the resulting placements, the binding energy is estimated and the best k placements are taken into the next iteration. A similar technique is already used in the peptide design tool GROW (Moon & Howe 1991).

In contrast to the original GROW approach, FLEXX must (and can) handle arbitrary organic molecules. We gain speed by basing our estimates and optimizations on simple discrete models instead of time-consuming force-field calculations. Our approach contains error-correcting mechanisms, in order to minimize the sensitivity of the results to small misplacements of the base fragment. Also, we added efficient algorithms and data structures, which are briefly summarized below.

We now describe the process of attaching a new fragment in more detail.

Let L_i be the set of placements after the i-th iteration. Assume that these placements place fragments $f_0, \ldots, f_i$ into the active site. In iteration $i+1$ we do the following:

1. For each placement P in the set of solution L_i: join fragment f_{i+1} to the part of the ligand already placed in all possible conformations, yielding a set of possible placements L'_{i+1} for the expanded part of the ligand.
2. Remove all placements from L'_{i+1} which exhibit significant overlap with the receptor.
3. For each placement P in L'_{i+1} do

3.1. Search for interactions occurring between the new fragment f_{i+1} and the receptor and append them to the list of matches.

3.2. Optimize the transformation of the placement such that the interactions have small deviations from their ideal geometries.

3.3. Test, whether the optimized transformation produces significant overlap with the receptor. If so, take the non-optimized transformation and remove matches with strong deviations from their ideal interaction geometries.

3.4. Estimate the free energy of binding.

4. Select the k best placements from the modified set L'_{i+1}.
5. Cluster the k best placements and select only the best placement of each cluster for the next iteration. This step is necessary to ensure higher diversity in the set of placements. The result is the set L_{i+1} of placements considered in the next iteration.

Conformations of the partially constructed ligand and matches of interacting groups are stored in hierarchical data structures, such that placements that start differing in fragment f_i share the same data for the fragments f_0 to f_{i-1}. All parts of the energy function which are additive over the fragments of the ligand are computed incrementally (Step 3.4). In order to speed up overlap tests, the surface atoms of the active site of the receptor are stored in three-dimensional hashing-tables. This concept is extended by cashing receptor atoms in the neighborhood of a ligand atom. This enables a very fast overlap test for only slightly moved ligands (Step 3.3). The optimization of the interaction geometries is done by a superposition algorithm (Ferro & Hermans 1977). Storing and sorting of *all* generated solutions in order to select the k energetically most favorable ones can be avoided by using a priority queue with k elements (Step 4). The elements in the priority queue are sorted by decreasing energy values. Thus, only k elements must be stored simultaneously and the selection time is reduced to $O(n \log k)$. The clustering is done by the hierarchical complete-linkage clustering algorithm, mentioned in section 3.1 (Step 5). The distance between two placements is taken to be their root-mean-square deviation. In addition, two placements can only be clustered, if the connection vectors (the vectors, which determine the location and direction in which subsequent fragments are attached to the fragments already placed) point roughly to the same location and in the same direction.

4 Results

Eleven known receptor-ligand complexes from the Brookhaven Protein Data Bank (Bernstein *et al.* 1977) have been used to test the docking algorithm and the scoring function described above. The names of the complexes are given in table 2. Beside flexible ligands with up to 10 free rotatable bonds also some rigid ligands are included in the test set. The active site of the receptor is taken to include all atoms with distances less than 6.5Å to the ligand in the crystal structure. All ligand structures have been minimized with the TRIPOS force field of the molecular modeling program SYBYL (TRI 1994). One exception is uridine vanadate, which cannot be handled by standard force fields because of the vanadium atom. Here we used the structure given in the PDB file. For the same reason we could not perform a conformational analysis by SCA for this example and were forced to held the ring systems rigid.

In all cases the algorithm reproduces the experimental structures within acceptable precision and with total run times on a SUN SPARCstation 20 of less than 90 seconds. This time does not include the preparation time (e.g., for the generation of receptor interaction points and the hashing tables), which arises only once for each receptor. In general, the times for the base placement or the total placement increase linearly with the number of interaction points in the active site and the numbers of interaction centers and conformations of the base fragment or the whole ligand respectively. Exceptions may occur in the case of symmetric patterns of interaction centers in the receptor or ligand.

The next important aspect of the quality of a docking algorithm is the quality of the predicted ligand placement and the accuracy of the estimate for free binding energy or, at least, of the ranking by energy values. Table 3 shows our calculated ΔG values and RMS deviations from the crystal structure of the energetically highest ranking solution and the highest ranking solution with an RMSD less than or close to 1Å. In addition, the table contains the rank of this solution, the total number of generated solutions and the experimental free binding energy. In all test cases,

Protein-ligand complex	PDB	Rec.	Lig.	Lig.	Run Time		
	entry	IP	IC	Conf.	Prep.	BP	CB
DHFR–methotrexate	4DFR	616	30	$3.14 \cdot 10^8$	10	22	55
Streptavidin–biotin	1STP	681	6	$6.67 \cdot 10^5$	8	16	16
Carboxypeptidase A–phenyllactate	2CTC	585	11	$9.00 \cdot 10^2$	8	68	1
α-Thrombin–NAPAP	1DWD	867	32	$9.80 \cdot 10^8$	15	13	71
Trypsinogen–ile-val	3TPI	723	12	$7.00 \cdot 10^5$	15	6	11
Ribonuclease A–uridine vanadate	6RSA	473	18	$1.20 \cdot 10^2$	6	73	9
Trypsin–benzamidine	3PTB	453	11	2	6	10	–
PHBH–*p*-hydroxybenzoic acid	2PHH	409	11	6	7	54	–
Lactate dehydrogenase–oxamate	1LDM	395	6	2	6	12	–
Triosephosphate isomerase–sulfate	5TIM	738	4	1	12	3	–
PNP–guanine	1ULB	787	20	1	17	64	–

Table 2: Protein-ligand complexes used as test cases for the flexible docking algorithm. Number of interaction points in the receptor (Rec. IP), number of interaction centers of the ligand (Lig. IC), number of ligand conformations (Lig. Conf.), and run times for preparation, base placement (BP) and complex build up (CB) phase. Times are given in seconds CPU time on a SUN SPARCstation 20. Complexes with no complex build up time have ligands with only one fragment.

the predicted binding modi are similiar to the crystal structure. In five cases the RMS deviation of the highest ranking solution is below 1Å, in the remaining cases a placement with RMSD < 1.1Å is found among the twelve highest ranking. Even solutions with RMS deviations above 1Å are, in fact, good binding positions. For instance, those fragments of methotrexate, which lie deep in the receptor, deviate much less than the mainly hydrophilic parts of the ligand that make contact with the water surrounding the protein (see figure 5). In the 1DWD test case the situation is similiar (figure 6). Here a naphthyl group has been turned around by an angle of 180 degrees, whereas the rest of the NAPAP molecule fits very well. It may be supposed that the predicted naphthyl position binds with comparable strength.

In summary, we find that FlexX is able to reproduce the geometries of receptor-ligand complexes with a precision that lies within the accuracy of the experimental data. In addition to the binding mode observed in nature, FlexX finds other low-energy binding modes that can be of interest for drug design. FlexX runs an order of magnitude faster than currently available docking methods regarding ligand flexibility (Leach & Kuntz 1992; Goodsell & Olson 1990; Mizutani, Tomioka, & Itai 1994) without a loss of accuracy.

However, there are discrepancies between observed and predicted free-binding energies. In cooperation with our project partners, we will tackle this problem by further improving the energy function and extending our test set. We also will re-calibrate the energy function as a consequence of the change in the computation of the lipophilic contact area.

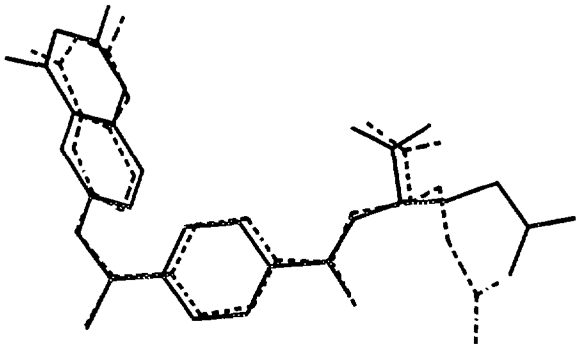

Figure 5: Energetically highest ranking solution of the test case DHFR–methotrexate. The predicted placement is shown in solid lines, the experimentally observed placement is shown in dashed lines (picture created with WHATIF (Vriend 1990)).

From our point of view, the main limitations of FlexX are the required existence of a suitable base fragment in the ligand and the rigidity of the receptor. Of course, ligands whose binding is based only on hydrophobic interactions cannot be docked by FlexX. But these examples are of less importance because drugs normally containing hydrophilic groups for pharmacological reasons. Thus, one future goal is to reduce the requirements for a *placeable* base fragment. Furthermore, we will focus on receptor flexibility. The speed-up achieved with FlexX is not based on methods that rule out receptor flexibility, in principle such as grid-based energy evaluations. Even if accounting for larger conformational changes of the receptor such as relative movements of domains, changes

PDB entry	Solut.	best ΔG ΔG	RMSD	RMSD < 1 ΔG	RMSD	rank	ΔG_{exp}
4DFR	60	-69.9	1.35	-66.7	0.64	4	-55.3
1STP	128	-32.3	1.15	-28.0	1.05	7	-76.4
2CTC	22	-29.8	1.65	-29.8	1.03	2	-22.2
1DWD	342	-46.8	1.56	-41.8	0.98	12	-48.6
3TPI	208	-22.0	0.99	-22.0	0.99	1	-24.5
6RSA	121	-34.5	0.81	-34.5	0.81	1	-28.5
3PTB	24	-29.4	0.45	-29.4	0.45	1	-27.2
2PHH	61	-24.6	1.82	-23.7	0.52	5	-26.7
1LDM	10	-35.3	1.54	-33.6	0.66	2	-30.8
5TIM	43	-16.5	1.23	-10.6	0.70	11	-13.1
1ULB	54	-19.3	0.66	-19.3	0.66	1	-30.2

Table 3: Protein-ligand complexes used as test cases for the flexible docking algorithm. Number of solutions, ΔG and RMS deviations from the crystal structure of the energetically highest ranking solution and the highest ranking solution with an RMSD less than 1Å (or solution with best RMSD for 1STP and 2CTC), and experimental free binding energy. (ΔG is given in kJ/mol, RMSD in Å.)

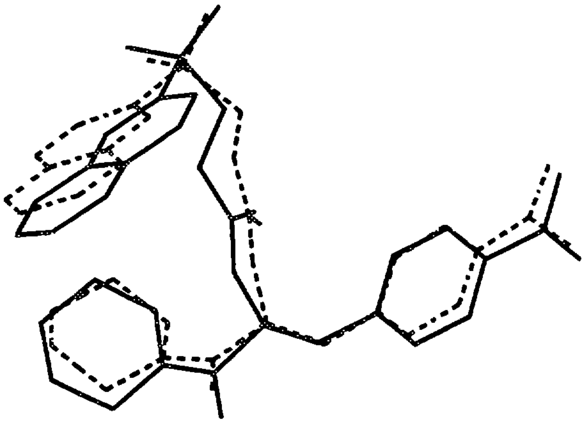

Figure 6: Energetically highest ranking solution of the test case α-Thrombin–NAPAP. The predicted placements is shown in solid lines, the experimentally observed placement is shown in dashed lines (pictures created with WHATIF (Vriend 1990)).

in loop conformations, or a complete rearrangement of side chains, seems unfeasable it should be possible to integrate small but important conformational changes such as the rotation of endstanding groups in our algorithmic framework.

Acknowledgements

We thank Christoph Bernd and Claus Hiller, who have done parts of the implementation of FLEXX. We would also like to thank our partners in the RELIWE-Project, especially Gerhard Klebe, Thomas Mietzner, and Joachim Böhm, from BASF main laboratory for preparing parts of the input data, making MIMUMBA available to us, and introducing us to Molecular Modeling. We thank Gert Vriend for making the WHATIF modeling package (Vriend 1990) available to us. The RELIWE-Project is partially funded by the German Federal Ministry for Education, Science, Research, and Technology (BMBF) under grant 01 IB 302 A.

References

Allen, F.; Bellard, S.; Brice, M.; Cartwright, B.; Doubleday, A.; Higgs, H.; Hummelink-Peters, T.; Kennard, O.; Motherwell, W.; Rodgers, J.; and Watson, D. 1979. The cambridge crystallographic data centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallographica* B35:2331–2339.

Bacon, D. J., and Moult, J. 1992. Docking by least-squares fitting of molecular surface patterns. *Journal of Molecular Biology* 225:849–858.

Bernstein, F.; Koetzle, T.; Williams, G.; Meyer, E. J.; Brice, M.; Rodgers, J.; Kennard, O.; Shimanouchi, T.; and Tasumi, M. 1977. The protein data bank: a computer based archival file for macromolecular structures. *Journal of Molecular Biology* 112:535–542.

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

Böhm, H.-J. 1992b. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *Journal of Computer-Aided Molecular 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. *Journal of Computer-Aided Molecular Design* 8:243–256.

DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; and Venkataraghavan, R. 1986. Docking flexible ligands to macromolecular receptors by molecular shape. *Journal of Medical Chemistry* 29:2149–2153.

DesJarlais, R. L.; Sheridan, R. P.; Seiberl, G. L.; Dixon, J. S.; Kuntz, I. D.; and Venkataraghavan, R. 1988. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *Journal of Medical Chemistry* 31:722–729.

Di Nola, A.; Roccatano, D.; and Berendsen, H. 1994. Molecular dynamics simulation of the docking of substrates to proteins. *PROTEINS: Structure, Function and Genetics* 19:174–182.

Dorndorf, U., and Pesch, E. 1994. Fast clustering algorithms. *ORSA Journal on Computing* 6(2):141–153.

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

Ferro, D. R., and Hermans, J. 1977. A different best rigid-body molecular fit routine. *Acta Crystallographica* A33:345–347.

Goodsell, D. S., and Olson, A. J. 1990. Automated docking of substrates to proteins by simulated annealing. *PROTEINS: Structure, Function and Genetics* 8:195–202.

Hoflack, J., and De Clercq, P. 1988. The SCA program: An easy way for the conformational evaluation of polycyclic molecules. *Tetrahedron Computer Methodology* 44:6667.

Jiang, F., and Kim, S. 1991. Soft docking: Matching of molecular surface cubes. *Journal of Molecular Biology* 219:79–102.

Kasinos, N.; Lilley, G.; Subbarao, N.; and Haneef, I. 1992. A robust and efficient automated docking algorithm for molecular recognition. *Protein Engineering* 5(1):69–75.

Kearsley, S.; Underwood, D.; Sheridan, R.; and Miller, M. 1994. Flexibases: A way to enhance the use of molecular docking methods. *Journal of Computer-Aided Molecular Design* 8:565–582.

Klebe, G., and Mietzner, T. 1994. A fast and efficient method to generate biologically relevant conformations. *Journal of Computer-Aided Molecular Design* 8:583–606.

Kuhl, F. S.; Crippen, G. M.; and Friesen, D. K. 1984. A combinatorial algorithm for calculating ligand binding. *Journal of Computational Chemistry* 5(1):24–34.

Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R. L.; and Ferrin, T. E. 1982. A geometric approach to macromolecule–ligand interactions. *Journal of Molecular Biology* 161:269–288.

Kuntz, I.; Meng, E.; and Shoichet, B. 1994. Structure-based molecular design. *Accounts in Chemical Research* 27(5):117–123.

Lawrence, M. C., and Davis, P. C. 1992. CLIX: A search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. *PROTEINS: Structure, Function and Genetics* 12:31–41.

Leach, A. R., and Kuntz, I. D. 1992. Conformational analysis of flexible ligands in macromolecular receptor sites. *Journal of Computational Chemistry* 13:730–748.

Lewis, R., and Leach, A. 1994. Current methods for site-directed structure generation. *Journal of Computer-Aided Molecular Design* 8:467–475.

Linnainmaa, S.; Harwood, D.; and Davis, L. 1988. Pose determination of a three-dimensional object using triangle pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(5).

Meng, E. C.; Shoichet, B. K.; and Kuntz, I. D. 1992. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry* 13(4):505–524.

Mizutani, M.; Tomioka, N.; and Itai, A. 1994. Rational automatic search method for stable docking models of protein and ligand. *Journal of Molecular Biology* 243:310–326.

Moon, J., and Howe, W. 1991. Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *PROTEINS: Structure, Function and Genetics* 11:314–328.

Murtagh, F. 1983. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal* 26(4):354–359.

Olson, C. 1994. Time and space efficient pose clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 251–258.

Pauling, L. 1939. *The nature of chemical bond.* Ithaca, NY: Cornell Univ Press. The united atom radii have been calculated by increasing the van der Waals radii by 0.1Å for every bound H-atom.

Rarey, M.; Wefing, S.; and Lengauer, T. 1995. Placement of medium sized molecular fragments into the active site of proteins. Arbeitspapier 897, German National Research Center for Computer Science, GMD.

Sandak, B.; Nussinov, R.; and Wolfson, H. 1994. 3-D flexible docking of molecules. *IEEE Workshop on Shape and Pattern Matching in Computational Biology* 41–54.

Shoichet, B. K., and Kuntz, I. D. 1993. Matching chemistry and shape in molecular docking. *Protein Engineering* 6(7):723–732.

Smellie, A.; Crippen, G.; and Richards, W. 1991. Fast drug-receptor mapping by site-directed distances: A novel method of predicting new pharmacological leads. *Journal of Chemical Information and Computer Science* 31:386–392.

TRIPOS Associates, Inc., St. Louis, Missouri, USA. 1994. *SYBYL Molecular Modeling Software Version 6.0.*

Vriend, G. 1990. WHATIF: A molecular modeling and drug design program. *Journal of Molecular Graphics* 8:52–56.

Yue, S. 1990. Distance-constrained molecular docking by simulated annealing. *Protein Engineering* 4(2):177–184.