**FLEXE: Efficient Molecular Docking Considering Protein Structure Variations**

Holger Claußen*, Christian Buning, Matthias Rarey and Thomas Lengauer

Side-chain or even backbone adjustments upon docking of different ligands to the same protein structure, a phenomenon known as induced fit, are frequently observed. Sometimes point mutations within the active site influence the ligand binding of proteins. Furthermore, for homology derived protein structures there are often ambiguities in side-chain placement and uncertainties in loop modeling which may be critical for docking applications. Nevertheless, only very few molecular docking approaches have taken into account such variations in protein structures.

We present the new software tool FLEXE which addresses the problem of protein structure variations during docking calculations. FLEXE can dock flexible ligands into an ensemble of protein structures which represents the flexibility, point mutations, or alternative models of a protein. The FLEXE approach is based on a united protein description generated from the superimposed structures of the ensemble. For varying parts of the protein, discrete alternative conformations are explicitly taken into account, which can be combinatorially joined to create new valid protein structures.

FLEXE was evaluated using ten protein structure ensembles containing 105 crystal structures from the PDB and one modeled structure with 60 ligands in total. For 50 ligands (83%) FLEXE finds a placement with an RMSD to the crystal structure below 2.0 Å. In all cases our results are of similar quality to the best solution obtained by sequentially docking the ligands into all protein structures (cross docking). In most cases the computing time is significantly lower than the accumulated run times for the single structures. FLEXE takes about five and a half minutes on average for placing one ligand into the united protein description on a common workstation.

The example of the aldose reductase demonstrates the necessity of considering protein structure variations for docking calculations. We docked three potent inhibitors into four protein structures with substantial conformational changes within the active site. Using only one rigid protein structure for screening would have missed potential inhibitors whereas all inhibitors can be docked taking all protein structures into account.

**Keywords:** structure-based drug design, flexible protein-ligand molecular docking, protein flexibility, protein ensembles, aldose reductase

---

*Corresponding author

Abbreviations used: PDB, Protein Data Bank; RMSD, root-mean-square deviation; CPU, central processing unit; CSD, Cambridge Structural Database.

E-mail address of the corresponding author: Holger.Claussen@gmd.de

---

**Introduction**

In recent years the search for novel drugs has evolved from a process of trial and error into a sophisticated procedure including several computer-based approaches. In structure-based design the structures of known target proteins are used to discover new compounds of therapeutical relevance. The approaches can be classified roughly...
into two categories: de novo design and docking. The former method designs new ligands to fit the protein target, whereas the latter is used to decide whether existing compounds possess a good steric and chemical complementarity to the given protein. For reviews on computer-aided structure-based design methods, see Kuntz,1 Rosenfeld et al.,2 Lengauer & Rarey,3 Kubinyi4 and Rarey.5

In order to use computational methods for structure-based design, several assumptions have to be made. For example early-docking tools treated both the ligand and the protein as rigid structures for efficiency reasons. At present, standard applications of current docking tools like DOCK,6 GOLD,7,8 and FLEXX9 use flexible ligands, but keep the protein structure essentially rigid, except for a few terminal H-bond donors and acceptors, and assume one single protein conformation even for complexes with different ligands. Therefore, ligands requiring larger conformational changes within the protein upon binding cannot be placed correctly by these methods.

Using experimentally determined structure ensembles of proteins with different ligands, several types of flexibility are observed.10 The scale ranges from simple side-chain rotations, over smaller adjustments of single loops, and up to large movements of complete domains. In some cases there are mutations of single amino acid residues (point mutations) within the active site which may influence the ligand binding. Furthermore, there are a lot of sequences for which no 3D structure is so far known. Although the overall structure of some of these proteins may be derived from homology modeling, there are often ambiguities in side-chain placement and uncertainties in loop modeling which may be critical for docking applications.

Even though in most cases only a few side-chains vary within the active site and the conformational changes tend not to be very large,11 flexibility can be quite critical for docking applications. For example, if a rotatable side-chain fills a sub-pocket of an active site into which a ligand should be docked, then the correct binding mode cannot be predicted without rotating the side-chain. We refer to this as the problem of protein structure variation during docking calculations. There are only a few approaches which focus on different aspects of this problem.

Leach11,12 explores the conformational space of protein side-chains keeping the backbone rigid in order to find the global minimum energy combination of amino acid side-chains and ligand conformations within a specified energy cutoff for a given orientation of the ligand. His approach is based on dead-end elimination and the A*-algorithm to search the large combinatorial space. The examination of conformations near to the global energy minimum shows that most differences arise from changes in side-chain conformations that are essentially independent from each other and only very few result from a concerted change in conformations of two or more residues.

Sandak et al.14-17 model flexibility by the concept of hinge-bending. This is a method adopted from computer vision techniques used for recognizing flexible objects assembled from rigid parts. The hinge-bending approach was originally used to model ligand flexibility, but the roles of the ligand and the protein can be interchanged since the mathematical problem is symmetrical. However, the flexibility can only be considered for one docking partner at a time. The hinges have to be defined manually and they are limited to a small number, therefore, side-chain flexibility is out of the scope of this approach.

Knecht et al.18 use weighted averages with respect to energy and geometry over ensembles of protein structures to describe the flexibility of proteins. Their ensembles are sets of crystal and NMR structures from the PDB. Ligand flexibility is not taken into account in this approach. The force-field terms for the particular structures are combined and stored on a single scoring grid that is used in DOCK 3.5,19-22 Repulsive potentials are only included if the potentials from all structures are repulsive. Although several protein structures are used during the calculation, it should be noted that the combinatorial nature of the problem resulting from explicitly distinguished alternative conformations is neglected. The interactions of the particular conformations are blurred by averaging, which may cause maxima in the potential related to unrealistic conformations. In addition, the active site is enlarged by neglecting the repulsive terms.

Schnecke et al.23-26 focus on docking as a screening method. The aim of their screening tools Spec-titope and Slide is the very fast and rough docking of many ligands from a large database. They use multi-level hashing to match an anchor fragment of the ligand to the template points describing the binding site. The anchor fragment is the part of the ligand which can be matched onto the template points using the ligand conformations given in the database. All flexible bonds within this anchor fragment are rigifed and so this approach depends on the initial conformation of the ligand in the database and the protein from which the template points are derived. Ligand and protein flexibility are modeled as a post optimization process by resolving collisions between the placed ligand and the protein by directed rotations of single bonds of both the flexible parts of the ligand or the side-chains of the protein. Mean-field theory is used to decide which rotations to use to resolve collisions and improve shape complementarity. In order to evaluate their approach they created several different binding-site templates from three different HIV-1 protease structures and screened three different ligand databases for putative ligands with low energy of binding. Their results show that, in spite of the handling of protein flexibility, the choice of the structure used as the target is still very important for the screening results.25
Our new software tool FLEXE is able to take into account several side-chain conformations and, to some extent, even loop movements while placing a flexible ligand into the active site. The flexibility of the protein is considered directly during ligand placement and not as a post optimization. The approach is based on the assumption that the overall structure of the protein and the general shape of the active site have to be conserved due to the bonding specificity. The occurrence of large main-chain variations, such as domain movements are out of the scope of our model.

The main idea is to describe the protein structure variations with a set of protein structures representing the flexibility, mutations or alternative models of a protein. FLEXE selects the combination of partial structures which suits best the given ligand with respect to the scoring function. Therefore, the variability considered by FLEXE is defined by the differences within the given input structures.

According to Knegtel et al.\cite{18} we refer to the set of protein structures representing the variation as the ensemble, although it is not an ensemble in strict terms of statistical thermodynamics. However, the set of protein structures contains mostly low energy conformations of the same protein, which also dominate a thermodynamical ensemble because they are highly populated. In this sense the two terms are related.

FLEXE is based on the so-called united protein description created from the superimposed structures of the ensemble. Similar parts of the structures are merged whereas dissimilar areas are treated as separate alternatives. The different structures of the ensemble are not only considered simultaneously, but can also be combined to form new overall structures during the docking process. This concept can easily be extended to rotamer libraries. Since FLEXE is derived from FLEXX, all substantial concepts like the interaction scheme, the incremental construction algorithm, and the scoring function are adapted to the ensemble approach.

Due to the recombination of protein structures, dependencies between different alternatives within the united protein description occur. These are caused by logical and geometric exclusion resulting in the concept of incompatibility between structure elements of the protein. Incompatibility is internally represented as a graph. Valid protein structures are independent sets of alternatives within this graph that fulfill certain constraints. During the incremental construction of the ligand optimal independent sets of alternatives with regard to the binding energy are determined for each partial ligand placement.

Here, using FLEXE, the ensemble of structures only models conformers present in the input structure (i.e. between four and 16 structures) and the possible recombinations of them. This may be a limitation in some cases. This is not a limitation in methods based on the Mean-field theory or the A*-algorithm. However, in principle the ensembles used with FLEXE are not limited to experimentally determined protein structures. For example, structures picked from a molecular dynamics simulation, generated by using rotamer libraries, or ambiguous homology models can be used as well. However, in order to evaluate FLEXE we need to know the correct binding mode. For this reason we use mainly experimentally determined protein structures for validations purpose.

FLEXE has been evaluated with ten protein structure ensembles containing 105 crystal structures from the PDB\cite{27} in total. In addition we used one homology model of human aldose reductase. The structures within the ensembles have a highly similar backbone trace, but different conformations for several side-chains, point mutations, and slight variations within loops. Sixty structures contain ligands which are used as a reference for the correct binding mode.

All ligands of an ensemble were docked into the united protein description of the ensemble and the root mean square deviation (RMSD) to the ligand's position in the crystal structure was determined. For comparison, all ligands were docked into all structures in the ensemble separately with FLEXX (cross docking) and the solutions merged into one ranking list. The RMSD of the ligands is of the same order of magnitude for both tools. Taking into account the top ten solutions, FLEXE finds a ligand position with an RMSD below 2.0 Å to the reference structure in 67% and FLEXX in 63% of the cases. The CPU time for FLEXE for the base placement and the complex construction is about five and a half minutes on average on a Sun Ultra SPARC 10 workstation and in most cases is lower than the sum of the corresponding run times for the cross-docking experiment with FLEXX.

The cross docking experiment with aldose reductase which has a highly flexible active site shows the necessity of taking into account various protein structures for docking calculations. We docked the three potent inhibitors sorbinil, tolrestat, and zopolrestat causing substantial conformational changes within the active site of the enzyme. For all ligands the ensemble approach of FLEXE and the cross docking with FLEXX find good solutions. However, FLEXX can only correctly place tolrestat and zopolrestat in the protein structures which were crystallized together with these ligands. Thus, using FLEXX with just one rigid protein structure for screening would have missed potent inhibitors.

Results

Test data set

The ensembles used with FLEXE are not restricted to experimentally determined protein structures. However, in order to evaluate FLEXE we need a reference for the correct binding mode. Therefore,
FLEXE has been evaluated with ten protein ensembles containing 105 crystal structures from the PDB and one homology model (see Table 1). Ideal ensembles for this test should meet the following criteria. (i) The structures should have very similar sequences and backbone traces. (ii) There should be different side-chain conformations available originating from complexes with different ligands. (iii) There should be several different ligands. (iv) Point mutations are allowed. They are especially interesting if they influence the binding of different ligands. (v) Flexible loops should be limited to short segments (containing a few amino acid residues).

Not all ensembles here satisfy all criteria. All members of an ensemble have highly similar backbone traces, different conformations for some side-chains or point mutations, and slight variations within some loops. Due to the limitations of available PDB structures, the ligands are often quite similar and sometimes even identical. For the aldose reductase case we combine three PDB structures of the porcine aldose reductase enzyme with a homology model of human aldose reductase (Podjarny, A., van Zandt, M., Krämer, O. & Klebe, G., personal communication). In this case, the overall sequence identity is about 86% but the active site is highly conserved. The size of the ensembles varies from seven to 16 members.

All proteins are either apo structures or complexed with a small molecule. Ligands that are covalently bound or too small (e.g. an azide ion) were disregarded. The remaining 60 ligands are marked in Table 1.

Preparation of the ligands

The preparation of the ligands for FLEXE is performed as for FLEXX using SYBYL. First, the ligand coordinates of the non-hydrogen atoms were extracted from the original PDB. They are used as reference for the calculation of the RMSD values later on. In the case in which there are identical ligands for different PDB entries, we used all ligands as separate references since they differ sometimes slightly. We obtained the ligand input files by defining correct atom types (including hybridization states) and correct bond types, adding hydrogen atoms, assigning formal charges to each atom, and finally energy-minimizing the reference structure. The energy minimization guarantees a low-energy conformation with suitable bond length and angles. This new geometry and the fact that the minimized structure is not translated according to its ensemble structure guarantees that there is no implicit docking information about the protein-ligand complex of the PDB structure in the ligand input file. In general, all carboxylic-acid and phosphoric-acid groups are ionized while all amino, amidino, and guanidino groups, but no amide groups, are protonated.

Preparation of the ensemble structures

For each structure, the description of an ensemble contains the definition of the protein atoms (via chain identifiers and hetero groups), the resolution of ambiguities in the PDB file (alternate location indicators, etc.), the location of hydrogen atoms at hetero atoms, and the definition of the active site. The assignment of hydrogen positions is made on the basis of default rules except for the definition of the torsion angles at the hydroxyl groups of the amino acid residues serine, threonine, tyrosine, and the hydrogen position inside the histidine side-chain. Here, torsion angles (either 0°, 180° for tyrosine; 60°, 180°, 300° for serine, threonine) and the optimal tautomeric histidine state are selected by visual inspection of the protein. The side-chains of lysine and arginine residue are protonated and the carboxylate groups of aspartic and glutamic acid are ionized. Water molecules contained in the PDB file have been removed.

In order to define the active sites of the proteins, all members of an ensemble are superimposed together with their reference ligand structure. All atoms including metal ions are selected that are located less than 6.5 Å apart from an atom of any ligand of the ensemble. Therefore, the active site is defined by the union of all ligands of the ensemble. In addition, the complete amino acid is selected if at least one of its atoms is picked.

The superimposed protein structures and reference ligand positions are stored for later cross docking experiments with FLEXX using the same definitions of active sites.

Evaluation of FLEXE

We docked all ligands of each ensemble into the united protein structure with FLEXE. The results are shown in Table 2. For each ligand the RMSD of the first solution, the best solution within the first ten, and the best solution within all solutions is given. In addition, the rank of the best and the first prediction with an RMSD below 2.0 Å and 2.5 Å, respectively, are shown. In the literature, an RMSD threshold of 2.0 Å is usually used for acceptable docking solutions. The increased cutoff of 2.5 Å takes into account the lack that the geometry of the united protein description may be slightly distorted. This can occur because the instances are clustered while building the united protein description using a maximum distance of 1.0 Å.

The RMSD of the first solution is often used to rate the performance of a docking tool because in practical screening applications one cannot inspect several placements of a large set of ligands. However, this placement depends not only on the docking algorithm but also highly on the scoring function. Here, we present a new approach for handling protein structure variations using one of the various scoring functions which are proposed for docking applications.
ing functions performs well under all circumstances. Therefore, one cannot expect to find the best solution ranked first. Since improving the scoring function is out of the scope of this work, we therefore discuss the quality of the docking on the basis of the best RMSD within the first ten solutions. We do this in order to reduce the influence of the scoring function on the one hand, while on the other hand keeping the number of solutions manageable.

The minimal RMSD found is also given in order to estimate the optimal RMSD that could be achieved with an ideal scoring function. The rank of this best solution is an indicator for the quality of the scoring function. However, one should keep in mind that this placement is not itself independent of the scoring function, because the scoring function is used during the docking calculation to choose the optimal independent set of instances and to rank the partial solutions.

The rank of the first solution with an RMSD below 2.0 Å shows how many solutions would have to be scanned to find a good placement. For practical applications this rank is therefore much more important than the rank of the best solution.

Table 3A summarizes the number of solutions with minimal RMSD and below 1.0, 1.5, 2.0 and 2.5 Å, respectively. This table reveals that for 83% of the ligands, a placement with an RMSD below 2.0 Å is found with FLEXE considering all solutions. This first hit is on average ranked 25th with a standard deviation of 60 ranks, whereas the best prediction...
<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Ligand</th>
<th>(Min.) RMSD (Å)</th>
<th>Best solution</th>
<th>Rank of first sol.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Rank 1</td>
<td>Rank 10</td>
<td>RMSD (Å)</td>
</tr>
<tr>
<td><strong>Aldose reductase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Alpha-momorcharin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Carboanhydrase II</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(satisfactory)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Carboxypeptidase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(satisfactory)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dihydrofolate reductase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Isocitrate dehydrogenase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mandelate racemase</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ricin</strong> (satisfactory)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trypsin</strong> (good)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For each Ligand, the RMSD of the first solution and the best RMSD within the first ten solutions are listed together with the RMSD and the rank of the best prediction and the rank of the first solution with an RMSD below 2.5 Å.
Table 3. Statistics on results

<table>
<thead>
<tr>
<th>RMSD</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. FLEXE first solution on any rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># hits</td>
<td>28</td>
<td>39</td>
<td>50</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>Per cent</td>
<td>46.7</td>
<td>65.0</td>
<td>83.3</td>
<td>90.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Avg. rank</td>
<td>24.0</td>
<td>25.2</td>
<td>24.7</td>
<td>11.9</td>
<td>73.3</td>
</tr>
<tr>
<td>SD</td>
<td>37.8</td>
<td>60.6</td>
<td>60.2</td>
<td>25.0</td>
<td>80.8</td>
</tr>
<tr>
<td>B. FLEXE first solution on top ten ranks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># hits</td>
<td>20</td>
<td>29</td>
<td>40</td>
<td>43</td>
<td>60</td>
</tr>
<tr>
<td>Per cent</td>
<td>33.3</td>
<td>48.3</td>
<td>66.7</td>
<td>71.7</td>
<td>100.0</td>
</tr>
<tr>
<td>SD</td>
<td>5.4</td>
<td>11.8</td>
<td>12.0</td>
<td>4.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Avg. rank</td>
<td>11.9</td>
<td>12.0</td>
<td>12.0</td>
<td>4.0</td>
<td>12.0</td>
</tr>
<tr>
<td>C. FLEXE first solution on top ten merged ranks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># hits</td>
<td>25</td>
<td>35</td>
<td>38</td>
<td>46</td>
<td>60</td>
</tr>
<tr>
<td>Per cent</td>
<td>41.7</td>
<td>58.3</td>
<td>63.3</td>
<td>76.7</td>
<td>100.0</td>
</tr>
</tbody>
</table>

This table summarizes the number of ligands for which a solution below 1.0, 1.5, 2.0, 2.5 Å is found in absolute numbers and as percentage considering all solutions of FLEXE (A), the first ten solution of FLEXE (B) and the first ten solution of the merged ranking list of FLEXE (C). In addition the average rank and the standard deviation (SD) of the rank is given for (A).

Docking Considering Protein Structure Variations

For all the ensembles of aldose reductase, alphamorphorcharin, dihydrofolate reductase, isocitrate dehydrogenase, mandelate racemase and trypsin FLEXE works very well. The first solution with an RMSD below 2.0 Å is found for all proteins investigated. This is a ranking problem which can be seen in the RMSD distribution of the first hit for each protein (see Figure 1). Such large ligands lead to many solutions and are also problematic for most other docking tools including FLEXE. Finally, the ligand A.G.1apg is not completely bound to the protein in the crystal structure, about one half of the ligand lies outside the active site.

is on average ranked 73rd with a standard deviation of 81 ranks. This low rank and the great standard deviation indicates that there are problems with the scoring function. The first hit (RMSD below 2.0 Å) is sometimes not within the first ten solutions, therefore Table 3B shows the same statistics taking into account only the first ten solutions predicted by FLEXE: For 67% of the ligands, a placement with an RMSD below 2.0 Å is found. Since all predictions are within the first ten solutions, no statistics on the rank is given.

We do not discuss all ensembles in detail, but classify the results into two groups: good solutions which are mostly correct and satisfactory solutions which partially failed.

**Good solutions**

For the ensembles of aldose reductase, alphamorphorcharin, dihydrofolate reductase, isocitrate dehydrogenase, mandelate racemase and trypsin FLEXE works very well. The first solution with an RMSD below 2.0 Å occurs frequently on rank 1 and the best RMSD within the first ten solutions is often around 1.5 Å or less. There is one ligand (FES.1jol) for dihydrofolate reductase for which no solution below 2.0 Å is found and two others where the first solution with an RMSD below 2.0 Å is ranked low down in the list: ZST.model (aldose reductase) and TOS.1pph (trypsin).

For the dihydrofolate reductase there are two types of ligands. The first type contains the methotrexates (MTX) of different PDB structures which differ slightly from each other and the second type consists of the five folate derivatives. (See Figure 1) The heterocycles of the latter ligands are all placed in a binding mode similar to the heterocycle of the methotrexate, which lies rotated by 180° in comparison to the folate crystal structure (see Figure 2). Therefore, the RMSD of the folate placements are about one angström higher than the RMSDs of the methotrexates. The reason for this rotation lies in the missing water molecules, which as yet cannot be handled in FLEXE.

**Satisfactory solutions**

These include carboanhydrase II, carboxypeptidase, ricin and seryl tRNA synthetase. For several ligands there is no solution with an RMSD below 2.0 Å within the first ten solutions and the better ones are ranked lowly. For some ligands even the best prediction has an RMSD of more than 2.0 Å and for five ligands no placement with less than 2.5 Å RMSD is found at all.

Several effects lead to these results. The first is a ranking problem which can be seen with the ligands AZM.1zsb, GLY.4cpa, or AMP.1obt where the first solution below 2.5 Å as well as the best one are ranked low down the list. Secondly, there are ligands which are quite large and have greater than ten rotatable bonds: EG1.1cnw, (14) ZAF.6cpa, (14) or FVF.7cpa (17) (see Figure 1). Such large ligands lead to many solutions and are also problematic for most other docking tools including FLEXE. Finally, the ligand A.G.1apg is not completely bound to the protein in the crystal structure, about one half of the ligand lies outside the active site.

**Comparison with FlexX**

In order to take into account protein structure variations with a docking tool in which the protein structure is kept rigid, a ligand has to be docked sequentially into all different protein structures (cross docking) and the solutions of the particular docking runs have to be combined in an automatically reproducible way. Picking the protein structures leading to the solution with minimal RMSD is not possible because the reference is usually unknown. Therefore, we merge the particular solution lists predicted by FLEXX and resort the predictions according to the scores.

FLEXE is derived from FLEXX as an integrated approach for efficient handling of protein structure variations using the same scoring function as FLEXX. Therefore, the quality of the predictions by FLEXE in terms of RMSD should be comparable with the quality of solutions selected from such a sequential cross docking.

We docked all ligands of all ensembles with FLEXX separately into each of the superimposed structures of the ensemble (cross docking) using the same reference structure and definition of the active site as in FLEXE. This yielded 727 complexes in total. For each ligand, we merged the solutions predicted for all ensemble structures and reranked them according to their score. Table 3C summarizes the results taking into account the first ten solutions of this merged ranking list.

For 38 ligands (63.3%) there is a solution with an RMSD below 2.0 Å. For FLEXE, this number is slightly higher (40 lig., 66.7%), but FLEXX finds sig-
Figure 1. Chemical formulae of the ligands mentioned in the text. The ligands within a box belong to the same ensemble (see also Table 1).
nificantly more solutions below 1.5 Å than FLEXE. We see the reason for this in the distortion of the ensemble structures by clustering. Therefore, the results are quite similar for the larger thresholds.

Again, we do not discuss all ensembles in detail, but take some cross docking experiments as examples. The results are represented in color coded matrices (Figures 3-5). The ligands correspond to the rows, the protein structures to the columns. The first column contains the results for the ensemble used in FLEXE, the second column the outcome of the merging approach, and the other columns the values for the particular docking run with FLEXX. Note that the merging approach does not necessarily find the minimal RMSD solution of the particular docking runs due to the different scores of the ligands within the various protein structures.

Dihydrofolate reductase

The dihydrofolate reductase is a nice example for which the FLEXE approach works well (see Figure 3). With one exception (FFO.jol), the RMSDs of the ligands docked with FLEXE are comparable or better than the best predictions in the merged ranking list of FLEXX. For the ligands DDF.1dyj, FOL.ra2, and MTX.ra3, FLEXE finds significantly better solutions than the predictions from the merged ranking list of FLEXX, although there are comparable predictions among the individual docking runs from FLEXX, which obviously do not score well enough to reach the top-ten ranks in the merged solution list.

Again, the two types of ligands namely, the methotrexates (MTX) and the five folate derivates, can be easily distinguished due to the color coding in the matrix. The higher RMSD of the folate derivates is caused by a misplacement of their heterocycles, which are placed by both tools in a binding mode similar to the heterocycles of the methotrexate. As explained before, this misplacement is due to missing water molecules. Since they cannot yet be handled by FLEXE water was also excluded for the FLEXX runs to allow for comparability.

The cross-docking experiment also shows that some protein structures yield better results for most ligands than others (1dhj or 1drb for example). Since all structures have a similar resolution it would be difficult to choose the right protein structure in advance.

In the test case, described above, FLEXX finds a solution for at least one protein structure. In order to analyze whether FLEXE still produces a good prediction if this is not the case, we created an additional ensemble for dihydrofolate reductase containing only the three PDB structures 1dhj, 1drb and 1jol which performed worst with FLEXX and repeated the cross docking experiment. Figure 4 compares the results of this experiment with the results produced with the complete ensemble.

The predictions of FLEXE with the reduced ensemble are slightly worse than for the complete ensemble of dihydrofolate reductase structures, whereas the predictions of the merging approach with FLEXX are significantly worse compared to that of the whole ensemble, especially for the methotrexates (MTX). Therefore, all methotrexates binding modes would only have been found by FLEXE using the reduced ensemble. This experiment shows that FLEXE is able to recombine different structures to new conformations that are more suitable for docking a particular ligand. The conformational space taken into account by FLEXE is much greater than covered by the separate ensemble structures. FLEXE is therefore more robust against local unfitness than the merging approach using FLEXX because local unfitness can be compensated for by alternatives during ligand placement. This is impossible when merging the solutions afterwards.

Aldose reductase

Aldose reductase catalyzes the reduction of glucose to sorbitol. As this reaction is believed to be linked to the pathogenesis of diabetic complications affecting the nervous, renal, and visual systems, the development of therapeutic agents has attracted intense effort.

Two areas of the active site of the aldose reductase are involved in ligand binding: (i) a recognition region for hydrogen-bonds and (ii) a hydrophobic contact zone. Upon binding to different ligands, the aldose reductase opens a specificity pocket in the hydrophobic area which can alter its shape by adopting different conformations. This flexibility can explain the large variety of possible substrates of aldose reductase.

We used an ensemble of four protein structures for the cross docking experiment. Three of these were the crystal structures of porcine aldose reductase (1ah0, 1ah3, 1ah4), all crystallized by
The structures 1ah0 and 1ah3 are complexed with the potent inhibitors sorbinil (SBI.1ah0) and tolrestat (TOL.1ah3), respectively, whereas the PDB entry 1ah4 contains the native holoenzyme. We combined these structures with a homology model of the complex of the inhibitor zopolrestat (ZST.model) with the human aldose reductase. (Podjarny, A., van Zandt, Krämer, O. & Klebe, G., personal communication). For this complex only the coordinates of the Cα-atoms are currently published (1mar). The human aldose reductase has an overall sequence identity of about 86% to the porcine sequence, but has a very conserved active site.

Figure 6(a) shows the active site of the united protein description containing the three inhibitors as given in the reference structure. There are two different binding regions: the hydrophobic contact zone mainly between the tryptophan residues Trp20, Trp79, Trp111, Trp219 and the amino acid residues Thr48, His110 and Trp111 forming hydrogen bonds to the ligands. This part of the active site is highly conserved over the four ensemble structures and is merged to one structure by clustering in the united protein description.

The specificity pocket is the flexible region of the active site where even backbone movements do occur (loop Ala299-Cys303). The pocket is closed by Phe122 and Leu300 when binding sorbinil (conf. 2) and opens in two different ways to assimilate tolrestat (conf. 1) or zopolrestat (conf. 3). Since the specificity pocket is not involved in binding sorbinil, FLEXX is able to place this ligand into all ensemble structures with an RMSD less than 1.0 Å. The original complex can even be reproduced with an RMSD of 0.43 Å (Figure 5).

Figure 3. Cross-docking dihydrofolate reductase. The color-coded Table shows the RMSD of the best prediction within the top-ten solutions for each ligand (row) predicted by FLEXX (1.col.), the merged ranking list (2.col.) and by FLEXX for the particular ensemble structures.
Due to the different conformations of the specificity pocket when binding tolrestat and zopolrestat, FLEXX can correctly place these ligands only into their own structures. The best solutions in foreign structures have RMSDs of more than 3.5 Å for the tolrestat and 6.0 Å for the zopolrestat. This means that using only one of the ensemble structures with FLEXX would definitely have missed inhibitors. Figure 5 shows the RMSD of the best solutions predicted by FLEXE on any rank, as well as the analogous results of the cross docking of all inhibitors into all the ensemble structures separately with FLEXX. We do not use the best solution within the top ten ranks to be sure that there is no better solution on any lower rank. This example shows the necessity of considering several different protein structures to decide whether a ligand can bind to a protein.

The united protein description of FLEXE as well as merging the cross-docking ranking lists predicted by FLEXX are different ways to take into account these protein structures variations. Both approaches select suitable conformations and find good placements with about or even less than 1.0 Å RMSD for all three inhibitors.

The best solutions predicted by FLEXE together with the reference structures for each ligand are shown in Figure 6(b)-(d). In these images, only the amino acid residues forming interactions with the ligand are shown and the alternative instances not used for the particular solution have been faded out.

Sorbinil (Figure 6(b)), which consists of three connected rings, is rigid. Three hydrogen bonds to Thr48, His110 and Trp111 are found and the aromatic ring is placed in the hydrophobic area. A good solution with 0.58 Å is found on rank one and the best one with 0.54 Å on rank two (Table 3).

In the case of tolrestat (Figure 6(c)), two hydrogen bonds between its carboxylate group and the amino acid residues His110 and Trp111 are found, while the naphthyl ring forms hydrophobic contacts with Trp20, Trp79, Trp111, Trp219 and Phe115, Phe122 and Leu300. Conformation one of Phe122 and conformation one of Leu300 are used by FLEXE, such that the trifluormethyl group and the methoxy group of the inhibitor can be placed into the upper part of the specificity pocket in agreement with the experimental structure. Again, there is a good solution (1.09 Å) ranked first and a slightly better solution (1.05 Å) on rank five. The best solution in the merged ranking list of FLEXE is ranked second.

Like tolrestat, zopolrestat (Figure 6(d)) forms two hydrogen bonds between its carboxylate group and His110 and Trp111. The phthalazinone ring lies within the conserved hydrophobic zone and the benzothiazole ring fills the deeper part of
the specificity pocket opened by Leu300. FLEXE takes the protein conformation stemming from the tolrestat complex (conf. 1) instead of conformation three. Hence, there is a larger overlap between the nitrogen of the benzothiazole ring and Leu300 tolerated by FLEXE. For Phe122, FLEXE uses conformation three which is the orientation stemming from the modelled complex between zopolrestat and aldose reductase. Therefore, Phe122 can form hydrophobic interactions with the benzothiazole ring as well as with the phthalazinone ring of the zopolrestat. The best solutions are ranked 133rd by FLEXE and first in the merged ranking list of FLEXX.

Run time

Table 4 summarizes the run time needed for preparing the input and for docking. The table shows for each ensemble the average run time for docking a ligand into a single-protein structure with FlexX, the accumulated average run time for sequentially placing the ligand into all protein structures with FlexX, and the average run time for simultaneously considering the whole ensemble with FlexE. The CPU times are measured for both tools on a Sun Ultra 10 machine having an Ultra SPARC 2e processor with 440 MHz and 512 MB RAM.

The preparation of the protein structure is not a critical factor with respect to screening applications because these computations have to be done only once for an ensemble (FlexE) and each protein structure (FlexX). For FlexE this phase contains building the united protein description as well as pre-computing and partitioning the incompatibility graph, while for both tools it includes indexing of possible triplets of interactions points. The preparation phase accumulates on average up to a few minutes for the whole ensemble with FlexX and takes less than 30 minutes with FlexE with one exception: carboxypeptidase. This case needs about three hours due to a very unfavorable combination

Figure 6. Aldose reductase. The Figure shows the united protein description containing (a) the reference structures of the three inhibitors sorbinil (magenta), tolrestat (yellow) and zopolrestat (cyan). In addition the best prediction and the reference structures of (b) sorbinil, (c) tolrestat and (d) zopolrestat are given separately. In the latter images only the amino acid residues forming interactions with the ligand are shown and the alternative instances not used for the particular solution are faded out.
of a large ensemble which leads to complex dependencies and a large active site.

Docking the ligand into the protein is actually the runtime-critical phase with respect to larger screening applications. A single docking run with FLEXX takes on average about a minute. However, a separate docking run is necessary for each ensemble structure in order to take into account all conformations given in the ensemble, resulting in an accumulated average run time of ten minutes for a whole ensemble.

For eight of the ten ensembles FLEXE is about a factor of two faster than FLEXX regarding the accumulated run time. FLEXE needs on average about 5.5 minutes for placing one ligand into the united protein structure taking into account not only all ensemble structures, but also the combinatorial combination of the conformations contained in the ensemble. This recombination of structures is not covered by the sequential approach with FLEXX.

Conclusions

Here, we describe a new docking tool FLEXE which is able to take into account protein structure variations. The idea of our approach is to represent protein flexibility, point mutations, or alternative models of a protein by an ensemble of feasible structures and combine them to form new valid protein structures during the docking process using an independent set search algorithm on the so-called incompatibility graph. FLEXE is a tool for efficiently handling and recombining alternative conformations. The variations considered by FLEXE are defined by the differences within the ensemble structures, and are therefore not limited to experimental data.

In contrast to the merging approach using FLEXX and the discrete docking approaches mentioned in the introduction, FLEXE is able to take into account ligand and protein flexibility simultaneously and directly while placing the ligand into the active site and not as a post optimization as in methods such as merging the ranking lists or resolving collisions. Therefore, FLEXE is independent from initial conformations of the ligand or the protein.

FLEXE treats the instances separately and recombines them in a discrete combinatorial way. Hence, the interactions are not blurred by averaging over distinct alternative instances, which may correspond to unrealistic protein conformations.

The example of aldose reductase demonstrates the necessity of considering protein structure variations for docking calculations. Using a single rigid conformation would have missed inhibitors which could be placed into another ensemble structure.

We applied our method to ten ensembles of protein structures and compared the results with the merged ranking lists that arose from cross-docking runs with FLEXX. The results show that our ensemble approach is able to cope with several side-chain conformations and even movements of loops. Motions of larger backbone segments or even domain movements are not covered by this approach.

For 67% of the test cases we obtained docking solutions with an RMSD below 2.0 Å within the top ten solutions predicted by FLEXE. This is comparable with 63% found in the merged ranking list of FLEXX. However, the run time of FLEXE is on average lower than the accumulated run time needed by FLEXX to dock the ligand sequentially into all members of the ensemble covering only a fraction of the whole conformational space considered by FLEXE, because FLEXE is able to create novel combinations of the ensemble conformations. The cross docking into the reduced ensemble of dihydrofolate reductase shows the advantage of combining structural elements of different ensemble members during ligand placement. FLEXE can dock a ligand correctly into an ensemble of structures into which the ligand cannot be docked by
merging the ranking lists of FlexX afterwards. Docking into all possible combinations of ensemble parts with FlexX would further increase the accumulated run time.

The gap in run time further increases if we take into account rotamer libraries to enrich the number of possible alternative conformations of a protein when for example only a few structures are available. Then the merging approach using FlexX would become combinatorially too complex, whereas FlexE even has the potential to efficiently handle such a big conformational space.

There are still some problems in ranking the solutions for both FlexE and the merging approach using FlexX. However, improving the scoring function was not aim of this work. This will be the focus of further investigation.

Up to now, intramolecular interactions within the protein are only considered in so far that two instances are incompatible if they overlap with each other. But there are non-overlapping interactions as well, which are partly favorable and partly unfavorable. Taking this into account could further improve the results.

**Materials and Methods**

FlexE is based on FlexX whose methods and recent developments have been described in detail elsewhere. Results on the evaluation of FlexX and a screening application with FlexX have also been reported. Therefore, we just summarize the main concepts of FlexX which are adapted for FlexE and describe the new approaches of FlexE in more detail.

**Ligand conformational flexibility**

The conformational flexibility of the ligand is modeled by a discrete set of preferred torsion angles at acyclic single bonds, and multiple conformations for ring systems. Torsion angles at multiple bonds, bond lengths and bond angles are used as given in the input structure. The torsion angles are taken from a database containing about 900 molecular fragments with a central single bond which has been derived from the Cambridge Structure Database (CSD) by Klebe & Mietzner. By this method up to 12 low-energy torsion angles can be assigned to each single bond.

Multiple conformations for rings are computed with the program CORINA. The number of ring atoms for each elementary ring is limited to eight. Larger rings are considered rigid and the input structure is used.

RMS deviations due to the described model of ligand flexibility are typically less than 1.0 Å and in most cases are even less than 0.5 Å (e.g. methotrexate: 0.4 Å).

**Interaction scheme**

The model of molecular interactions used in FlexX and FlexE has been adopted from Bohm and Klebe. For each group forming interactions an interaction geometry is assigned consisting of the position of a center and the shape of a spherical interaction surface. Two groups interact if the interaction center of each group is lying approximately on the interaction surface of the counter group (Figure 7). For algorithmic reasons, the interaction surfaces on the protein side are approximated by a finite set of so-called interaction points.

The interactions are divided into three different types, from level 3 for highly-directional interactions such as H-bonds down to level 1 for directionally unspecific such as hydrophobic interactions. The higher-level types are preferred in the selection and placement of base fragments. Only if there are not enough high-level interactions will the algorithm descend to lower-level interaction types.

**Scoring function**

The ranking of the docking results is performed with a modification of the scoring function developed by Bohm.

\[
\Delta G = \Delta G_0 + \Delta G_{rot} \times N_{rot} \quad (1)
\]

\[
+\Delta G_{\text{hb}} \sum f(\Delta R, \Delta \alpha) \quad (2)
\]

\[
+\Delta G_{\text{int}} \sum f(\Delta R, \Delta \alpha) \quad (3)
\]

\[
+\Delta G_{\text{aro}} \sum f(\Delta R, \Delta \alpha) \quad (4)
\]

\[
+\Delta G_{\text{lipo}} \sum f^*(\Delta R) \quad (5)
\]

The first two terms (equation (1)) of the function are a fixed ground term (\(\Delta G_0 = 5.4 \text{ kJ/mol}\)) and a term taking into account the loss of entropy during ligand binding due to the hindrance of rotatable bonds in the ligand.
(\(\Delta G_{\text{rot}} = 1.4\ \text{kJ/mol}\)). The following terms (equations (2)-(4)) are sums over all pairwise interactions. The last part (equation (5)) of the scoring function evaluates the atom-atom contacts between protein and ligand, i.e. hydrophobic contacts and forbiddingly close contacts (clashes). The functions \(f, f^*\) are heuristic distance and angle-dependent penalties (see \(^9,30\) for details).

**Incremental construction algorithm**

The docking algorithm is divided into three parts: (i) the selection of the base fragments; (ii) the placement of the base fragments; and (iii) the incremental construction of the whole ligand within the active site. First, the ligand is fragmented into components by severing all acyclic single bonds. FLEXE/X automatically forms a set of alternative base fragments by selecting single components or combinations of them.\(^36\)

The base fragments are placed into the active site using one of two different algorithms. The first triangle matching algorithm superposes triples of interaction centers of a base fragment with triples of appropriate interaction points in the active site. If a base fragment has fewer than three interaction centers or if the number of placements is less than 100, then the second line matching algorithm is started. It matches pairs of interaction centers with pairs of interaction points. Because of geometric ambiguity, multiple placements are generated by rotation around the axis defined by the interaction points and centers. Both base placement algorithms typically generate a large number of solutions. A reduction by clash tests and clustering follows.

Starting with the different base placements the complete ligand is constructed by linking the remaining components step by step in compliance with the torsional database. After adding one component, new interactions are searched for and the scoring function is used to select the best partial solutions which are used for the next extension step. The maximum number of solutions taken into account in the next iteration is \(400 + 100n_b\), where \(n_b\) is the number of different base fragments.

**New concept of FLEXE**

The FLEXE approach is based on the united protein description which handles the similarities and differences of the protein structures of the ensemble. All side-chains and backbone parts are treated separately and the dependencies between them are represented in this so-called incompatibility graph. An algorithm for finding sets of pairwise compatible instances is applied to this graph to ensure that valid protein structures are used as a basis for the docking algorithm. These steps are described in more detail in the following sections.

**United protein description**

The basis of FLEXE is the united protein description which administers protein structure variations. The description is generated from an ensemble of protein structures. Each member of the ensemble must be a valid protein structure which shows one possible conformation of the protein. The particular structures are superimposed and combined to a united protein description (Figure 8). The maximum number of allowed structures per ensemble is currently set to 30.

Since we assume highly similar backbone traces for the members of an ensemble, the most straightforward way to superimpose the structures is to fit the backbone atoms of the particular structures. We implemented two simple methods to superimpose the structures of an ensemble onto a reference structure, which we take to be the first structure of the ensemble. The first method applies the Kabsch algorithm\(^48\) to fit two sets of atoms e.g. the backbone atoms given in the ensemble description. The second method iterates the first method, such that all pairs of atoms with a distance greater than a user defined threshold are ignored for the next step. This procedure emphasizes the differences and improves the fitting in conserved regions of the structures. Alternatively, the superposition can be performed externally with any other tool.

**Figure 8.** Ensemble approach. Active site of alpha-momorcharin. The united protein description (right) is created from the superimposed structures of the ensemble (left). Similar parts of the structures are merged whereas dissimilar areas are treated as separate alternatives.
The superimposed structures are combined to create the united protein structure by clustering the alternative side-chain conformations and backbone parts, which we call instances (see Figure 9). The clustered instances can be recombined to form new valid protein structures regardless of the structure from which they originally stem. Therefore, the structures we dock into are not limited to the original ensemble structures.

The method we apply for clustering is a complete-linkage hierarchical cluster algorithm. The strategy of hierarchical clustering is such that 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. Complete-linkage means that the distance between two clusters is defined as the maximum distance between the elements of the clusters.

We cluster the instances of each part (Figure 9) separately. The instances of the particular parts are the elements of the clusters and the distance between two elements is the mean distance between the atoms of the instances if they are of the same type (i.e. they are both backbone parts or side-chains of the same amino acid residue), otherwise the two instances are not clustered. We use a threshold of 1.0 Å as a trade off between distorting and clustering the instances.

Incompatibility graph

Two instances of the united protein structure are incompatible if they cannot be realized simultaneously. The incompatibility between the instances is represented as a graph by using all instances as nodes and connecting pairs of incompatible nodes with edges (see Figure 10).

We distinguish between three kinds of incompatibility: (i) logical: two instances are alternatives of each other; (ii) geometric: two logically compatible instances overlap; (iii) structural: two instances of the same chain are unconnected. Logical incompatibility is implied by the construction of the united protein description: If two instances belong to the same part they are incompatible.

An overlap test is performed to test for geometric incompatibility, tolerating an overlap volume of 5.5 Å³ between two instances. This threshold was chosen such that instances of experimentally determined protein structures do not clash with each other. Covalent bonds between adjacent backbone instances, adjacent backbone and side-chain instances, close contacts between two cysteine residues (disulfide bridge) and those with a proline neighbor (ring closure) are exempt from this test.

In addition, compatible instances have to be linked to avoid combinations of instances of different loops which are too far away from each other because such combinations would create absurd protein structures. We assume two instances to be directly linked if the bond lengths of the bonds between them agree with the expected length up to a tolerance of ±1.0 Å, which corresponds to the threshold for clustering the instances. Two adjacent components are linked if there is at least one pair of linked backbone instances of the two components. Otherwise a chain break is supposed. Two instances are always compatible if there is a chain break in between them. In order to determine whether two instances are linked via a sequence of instances we use a dynamic programming algorithm which will not be described in detail here.

A valid protein structure corresponds to a completely disconnected subgraph (independent set) in the incompatibility graph, containing exactly one node per part. Therefore, finding a valid protein structure is tantamount to searching for an independent set in the incompatibility graph. In order to speed up this search the incompatibility graph is divided into maximum connected subgraphs, which can be treated independently from each other (see Figure 10). Maximum connected subgraphs are minimal subsets of nodes connected by at least one edge such that there are no edges between nodes of different subgraphs.

In some rare cases there is a conflicting instance which is incompatible with a single instance of a part (see Figure 10). It occurs if one of the ensemble structure itself is invalid or may be produced by clustering if the threshold for clustering is too large and the tolerance for volume overlap between two instances is too small. Such a conflicting instance is excluded from further computations.

Surface and interaction geometries

FLEXE uses the Connolly molecular surface to decide whether an atom of the protein is solvent-accessible. Only atoms at the surface can form an interaction with the ligand. An analogous definition of a surface in the united protein description of FLEXE is difficult because the surface of an instance cannot be defined independently from proximate compatible instances. Therefore, FLEXE does not compute a surface for the united protein description, but allows all atoms that are part of the active site to form interactions.

To each instance that can form interactions we assign an interaction geometry consisting of a set of interaction points. Each interaction point is individually tested for overlap with the protein taking into account the compatibility. See for example Figure 11. The interaction points of instance O1 cannot clash with the instance O2 and vice versa, even though the points are placed within the van de Waals radius of the other instance, because the instances O1 and O2 are incompatible. In addition, interaction points of instance O1 interfere with the instance
S2 which is compatible to O1. However, there is an alternative instance S1 which also is compatible to O1 and does not overlap with any interaction point of instance O1. All interaction points of both instances O1 and O2 can therefore be used.

Selection of instances

During the incremental construction algorithm the ligand is placed fragment by fragment into the active site of the united protein description. After each construction step, all possible interactions between the (partially) placed ligand and all instances of the united protein structure are determined. For each particular instance, the scoring function is applied in order to estimate the score of this instance. The score of an independent set of instances in the incompatibility graph can then be calculated as the sum over the scores of its nodes. The independent set of instances with the highest score represents the protein structure which best suits to the (partially) placed ligand with respect to the scoring function. The score of this optimal independent set therefore determines the final score of the (partial) solution.

The optimal independent set can be assembled from independent sets of the maximum connected subgraphs. Thereby, only those subgraphs have to be considered which contain at least one node with a score that is not zero because the other subgraphs can not contribute to the total score.

We use a modified version of the Bron-Kerbosch algorithm for finding high-scoring independent sets within the maximum connected subgraphs with more than three nodes. The original algorithm was intended for finding all cliques of an undirected graph, which is complementary to the independent set problem. The Bron-Kerbosch algorithm enumerates all independent sets by augmenting an initial independent set in a recursive fashion. At each step, the partial independent set P is extended by a node selected from a set C of so-called candidate nodes. This selection is crucial for the performance of the algorithm. We refer the reader to the original paper for a detailed description of the algorithm.

The time consuming search for optimal independent sets has to be done after each construction step and for each maximum connected subgraph. Therefore, it is the most time-critical step of FLEXE. In order to speed up the search, we check after each selection if the union P∪C of the partial independent set P and the set of candidates C still contain instances of all parts. If this is not the case the algorithm can backtrack the recursion early, since a independent set must contain exactly one instance per part in order to describe a valid protein structure. In addition, we stop the algorithm if the first independent set is found because we only need the score of the optimal independent set. Although it cannot be guaranteed that the first-found independent set is the optimal one, this “greedy” strategy finds best scoring independent set frequently, because we sort the candidate set C by descending score before starting the enumeration.

World Wide Web resource

More detailed results of the presented test for all ensembles and all cross docking experiments as well as the input files of the presented data set will be made available on the World Wide Web. The FLEX software package is available for SUN, SGI, and PCs running the Linux operation system. FLEXE will be made available in the near future. Interested readers should visit our website at http://cartan.gmd.de/FlexX and http://cartan.gmd.de/FlexE, or contact the corresponding author for details.
Acknowledgments

The authors thank Alberto D. Podjarny (UPR de Biologie Structurale IGBMC, Illkirch, France), Michael van Zandt (Institute for Diabetes Discovery (IDD) Branford, Connecticut, USA), Oliver Krämer, and Gerhard Klebe (University of Marburg, Germany) for providing the homology model of the human aldose reductase.

We also thank Martin Stahl (F. Hoffmann-La Roche Ltd., Basel, Switzerland) for fruitful discussions on this topic, our cooperation partners for various helpful comments during the method development, and Sally Hindle for many suggestions on earlier versions of the manuscript.

This work is part of the Relimo project, funded by the bmb+f (Bundesministerium für Bildung und Forschung) and the participating industrial partners Boehringer Ingelheim Pharma KG and Merck KGaA, Darmstadt under grant 0311 620.

References


Edited by J. Thornton
(Received 3 July 2000; received in revised form 13 February 2001; accepted 16 February 2001)