PROTEINS: Structure, Function, and Genetics 34:17–28 (1999)

The Particle Concept: Placing Discrete Water Molecules During Protein-Ligand Docking Predictions

Matthias Rarey,* Bernd Kramer, and Thomas Lengauer
German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

ABSTRACT **Water is known to play a significant role in the formation of protein-ligand complexes. In this paper, we focus on the influence of water molecules on the structure of protein-ligand complexes. We present an algorithmic approach, called the *particle concept*, for integrating the placement of single water molecules in the docking algorithm of FlexX. FlexX is an incremental construction approach to ligand docking consisting of three phases: the selection of base fragments, the placement of the base fragments, and the incremental reconstruction of the ligand inside the active site of a protein. The goal of the extension is to find water molecules at favorable places in the protein-ligand interface which may guide the placement of the ligand. In a preprocessing phase, favorable positions of water molecules inside the active site are calculated and stored in a list of possible water positions. During the incremental construction phase, water molecules are placed at the precomputed positions if they can form additional hydrogen bonds to the ligand. Steric constraints resulting from the water molecules as well as the geometry of the hydrogen bonds are used to optimize the ligand orientation in the active site during the reconstruction process. We have tested the particle concept on a series of 200 protein-ligand complexes. Although the average improvement of the prediction results is minor, we were able to predict water molecules between the protein and the ligand correctly in several cases. For instance in the case of HIV-1 protease, where a single water molecule between the protein and the ligand is known to be of importance in complex formation, significant improvements can be achieved. Proteins 1999;34:17–28.**

Key words: molecular docking; flexible docking; protein-ligand interaction; molecular flexibility; conformational analysis; water interactions; drug design

INTRODUCTION

The formation of complexes of protein receptors and ligand molecules takes place in aqueous solution. Water influences this process in many different ways. The attraction of hydrophobic surfaces is based on the presence of water. The ligand as well as the protein must be desolvated before the molecule complex can be formed.[1] Beside these energetic aspects, water also influences the structure of protein-ligand complexes. Several examples are known where single water molecules are lying between the protein and the ligand filling gaps and mediating hydrogen bonds between the two molecules.[2,3] Reviews summarizing the role of water in protein-ligand interactions can be found in references 4, 5, and 6.

In the computer-based prediction of protein-ligand complexes, water is mostly modeled implicitly in the scoring function. In functions calibrated to experimentally determined binding constants (see References 7 and 8 for example), the different energetic contributions reflect the increase or decrease in energy relative to the unbound state of the protein and the ligand in water. For example, the energy contribution of a hydrogen bond is the decrease of energy compared to hydrogen bonds formed with water. In contrast to binding affinity estimation, the influence of water molecules on the structure of protein-ligand complexes is neglected in automated docking algorithms. Water molecules can only be predicted either independently from the individual ligand and then used as a static part of the input or after the prediction of the protein-ligand complex. An algorithm for this task has been developed by Raymer et al.[9]

In this paper we introduce an algorithmic extension of our docking method FlexX,[10] called the *particle concept*, by which we explicitly place single water molecules during protein-ligand docking. The docking algorithm of FlexX consists of three phases and is based on the incremental construction strategy. In the first phase, a set of *base fragments* is automatically selected.[11] Then, the base fragments are placed independently into the active site without considering the remaining parts of the ligand.[12] In the third phase, the ligand is incrementally reconstructed by adding fragments to the placements computed so far.[10] Docking by constructing the ligand inside the active site was first introduced in Reference 13 using a backtracking strategy. Algorithms similar to FlexX are also used in References 14 and 15.

The particle concept places spherical objects between the ligand and the protein during the docking computation. The objects, called *particles*, have the ability to form molecular interactions like hydrogen bonds to the ligand as well as to the protein. In addition, placed particles

*Correspondence to: Matthias Rarey, GMD-SCAI, Schloß Birlinghoven, 53754 Sankt Augustin, Germany. E-mail: Rarey@gmd.de
Received 8 June 1998; Accepted 27 August 1998

interact sterically with the molecules. The particle concept is formalized in a general way and can be used to integrate small molecules, single atoms, or metal ions in the protein-ligand interface. Its main application is the placement of discrete water molecules.

Considering discrete water molecules in an incremental construction docking algorithm has two main advantages. First, during the reconstruction of the ligand, several placements have to be evaluated and the program has to decide which placements are promising and should be investigated further. Here, placed water molecules can improve the scoring scheme. Second, the position of the ligand is determined by the interactions to protein. A water molecule placed between the two molecules can influence the placement such that the interactions of the ligand with the protein as well as with the water molecules are optimized.

In the result section, we report on tests of the docking algorithm using the particle concept on a set of 200 protein-ligand complexes with known structure taken from the Brookhaven Protein Data Bank (PDB).[16] Using the particle concept, we are able to predict water locations which are also seen in the crystal structures. In some cases in which water is known to play a critical role like in HIV-1 protease,[2] we can drastically improve the docking result. We also found cases where the particle concept supports binding modes not observed crystallographically.

METHODS

The physico-chemical interaction and scoring model in FLEXX is of importance for the particle concept and is therefore summarized here. The model is derived from the de novo design tool LUDI.[17,18] A detailed description can be found in Reference 10.

FLEXX molecular interactions are modeled in two different ways. While geometrically less restrictive interactions like the hydrophobic contact surface are considered only in the scoring function, geometrically more restrictive interactions are used to determine the placement of the ligand inside the active site of the protein. A geometry of an interaction used for placement consists of an *interaction surface* which is part of a spherical surface around the *interaction center* (see Figure 3 (top, right) for an example). Interaction surfaces are caps, spherical rectangles, or full spheres. An interaction between two groups takes place if the interaction center of the first group lies approximately on the interaction surface of the second group and vice versa.

In order to score protein-ligand complexes, Böhms function[7] is used with minor modifications:

$$\begin{aligned}\Delta G = \Delta G_0 &+ \Delta G_{rot} \times N_{rot} \\ &+ \Delta G_{hb} \sum_{\text{neutral H–bonds}} f_1(\Delta R, \Delta\alpha) \\ &+ \Delta G_{io} \sum_{\text{ionic int.}} f_1(\Delta R, \Delta\alpha) \\ &+ \Delta G_{aro} \sum_{\text{aro int.}} f_2(\Delta R, \Delta\alpha) \\ &+ \Delta G_{cont} \sum_{\text{cont.}} f_3(\Delta R) \qquad (1)\end{aligned}$$

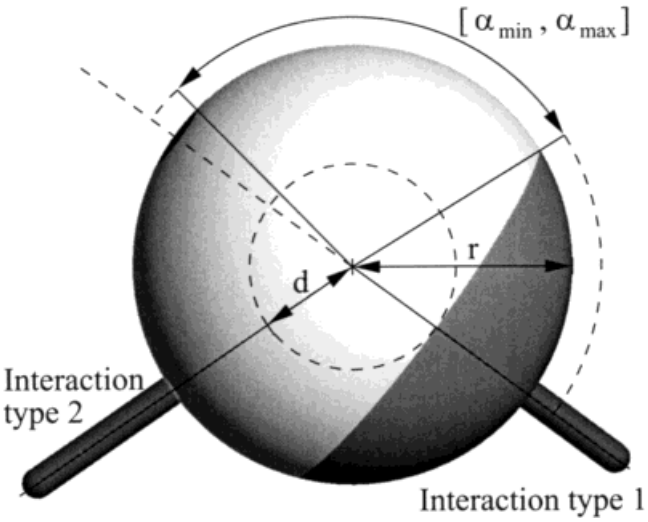

Fig. 1. Geometric parameters of a particle: the radius *r*, the distance offset *d* for each interaction type defining the distance between particle center and interaction center, and the allowed angular range [α_{min}, α_{max}] for each pair of formed interactions. Assuming that an interaction at interaction type 1 is formed, the allowed area for a second interaction with respect to the angular range is shown in light gray, the forbidden area in dark gray.

Here, $f_i(\Delta R, \Delta\alpha)$ is a scaling function penalizing for deviations from the ideal geometry, ΔG_x are constant parameters determined during the calibration of the function.[7] N_{rot} is the number of rotatable bonds in the ligand molecule, ionic interactions are all contacts between charged atoms, aromatic interactions are specific contacts between phenyl rings, methyl groups, and amides. In contrast to Böhms original function, we have added a clash term to the contact score summarized in the contact term (see Reference 10 for details).

The Particle Model

A *particle* is a spherical object which interacts with the protein and the ligand sterically and physico-chemically. Sterically, a particle behaves like a single atom with a user-specified radius.

A particle can form up to four different types of interactions taken from the list of interaction types of FLEXX. The particle model is orientation-invariant, therefore the corresponding interaction geometries must be spherical. In order to approximate the hydrogen bond geometry of water molecules, we added two geometric parameters for the description of interaction geometries at particles shown in Figure 1. The first parameter is the offset *d* of the interaction center from the center of the particle. This is necessary because, in the case of water, the particle center represents the center of the oxygen atom while the interaction center of a hydrogen bond donor is located in the hydrogen atom which is not modeled explicitly.

The second parameter is an angular range [α_{min}, α_{max}] defining the minimal and maximal angle between all pairs of interactions with the same particle. With appropriate

values, an approximate tetrahedral interaction pattern can be enforced for a particle.

The interactions of a particle are classified into up to two groups. For each group, we specify the maximal number of interactions which the particle can form. In the case of the water molecule, we allow the particle to form at most two hydrogen bonds as a donor and two as an acceptor.

A particle contributes to the score of a ligand placement in two different ways. First, interactions formed between the ligand and the particle are handled in nearly the same way as those between the ligand and the protein. The geometric penalty functions normally depend on the angular deviation on the ligand as well as on the protein side. Because the particle has only spherical interactions, the angular deviation cannot be measured directly. Therefore, for each ligand-particle interaction, the minimal angle between the direction to the ligand interaction center and the direction to a protein interaction is determined. The angular deviation of the interaction is then defined to be the difference between this angle and the idealized setting (for example 109.5 degrees for a particle with a tetrahedral interaction pattern). Finally, a penalty is calculated from the angular deviation with a stepwise linear function in the same way as it is done for other geometric deviations in the scoring function.

The second contribution is a penalty $\Delta G_{particle} N_{vac}$ for each placed particle, where N_{vac} is the number of interaction sites of the particle not involved in interactions with the protein or the ligand. Note that we specify the maximal number of interactions for each particle type.

The contact score (last term in the scoring function) accounts for the common contact surface between the molecules. Because a particle does not increase the amount of contact surface it does not contribute to the contact score of the ligand. Nevertheless, the potential score of the ligand-particle contact is needed during the placement optimization, as will be explained later. The particle is therefore categorized either as hydrophobic or hydrophilic. This categorization is also used in the decision whether a particle should be placed or removed again during the docking computation (see below).

Generating Particle Phantoms

In a preprocessing phase, energetically favorable positions of particles inside the active site of the protein are calculated. These positions are called *particle phantoms* as long as they are not used during docking. As docking proceeds, particle phantoms are switched on (from a phantom to a particle) and off (from a particle to a phantom), the positions of the phantoms and therefore of the final particles in the protein-ligand complex are not changed.

Particle phantom locations are computed for each particle type with the following algorithm. For each interaction type of the particle, the corresponding interaction surfaces in the active site of the protein are approximated by discrete point sets. Each point represents an initial phantom position such that an interaction to the protein can be formed.

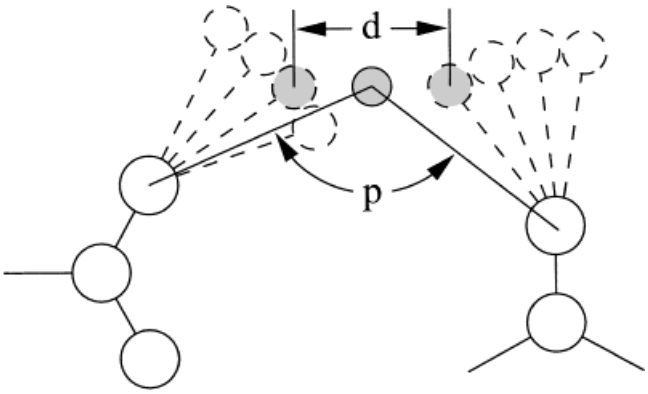

Fig. 2. Clustering particle positions: Dashed circles show initial particle positions forming an interaction to a protein atom (dashed line). Two particle positions (grey dashed circles) can be merged during clustering, if the angle *p* is within the range defined for the particle type. The distance relevant for the clustering algorithm is then the Euclidean distance *d*. After merging, the final position of the particle is in the middle of the initial ones (grey solid circle), the particle forms two interactions to the protein (solid lines).

In the second step, a complete linkage hierarchical clustering algorithm[19] used several times in the docking algorithm[12] is applied to the initial phantom positions. The distance function used for the purpose of clustering is defined as follows. If two phantom positions result from the same interaction surface or if the enclosed angle between the two protein interactions is outside the angular range specified for the particle type, the distance is set to infinity such that these particles cannot be clustered, otherwise the Euclidean distance is used. The distance threshold for the cluster algorithm is set to 1.6 Å. This threshold allows reasonable tolerances for the hydrogen bonds which are about the same as used in docking.

From each cluster, a single particle phantom is generated by merging the set of protein interactions from the phantoms of the cluster and setting the phantom position to the average position of all phantoms in the cluster. The distance function and the merging procedure are illustrated in Figure 2.

Finally, the particle phantoms are filtered using three criteria (parameters for water particles are given in parentheses)

- The number of interactions to the protein must be equal to or greater than a threshold specific for each particle type (2).
- The number of protein interactions must be less than the maximal number of interactions allowed for the particle type (4).
- The overlap volume between the particle phantom and the protein must be less than a fixed threshold (2.5 Å^3).

As an example, the final particle phantoms for water in the particle concept calculated in the active site of HIV-1 protease are shown in Figure 3.

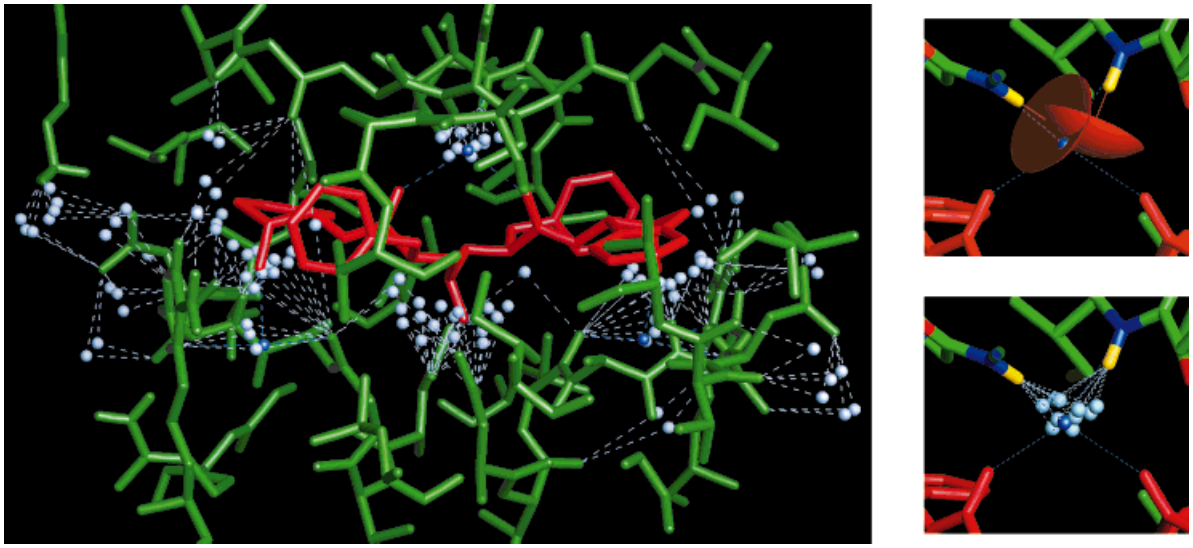

Fig. 3. Particle phantoms (light blue) in the active site of HIV-1 protease. Experimentally observed water molecules are shown in dark blue. **Right:** a detailed view around water molecule HOH-1, with interaction surfaces (top) and with particle phantoms (bottom).

Particles During Ligand Docking

Particles influence the docking computation in several ways. As explained above, particles contribute to the score of the protein-ligand complex. In this section, we describe how particles are selected from the set of precomputed phantoms within the docking algorithm.

Phantoms are implicitly switched to particles whenever an interaction between the ligand and the particle can be formed and the particle does not overlap with the ligand. The search for particle interactions is interleaved with the placement of fragments. First, the base fragments are placed and particle interactions are searched for all resulting placements. Then in the complex construction phase, an iteraction consists of the following steps: First, a new fragment is added in each possible conformation. For each conformation, all particles overlapping with the ligand are switched back to phantoms. Then, interactions from the added fragment to the protein as well as to particles are searched. Finally, the ligand position is optimized to remove clashes and to optimize the formed interactions and the resulting placement is scored. If no clashes to the protein occur, the placement is added to the solution set of the current iteration.

The overall scheme of the incremental construction docking algorithm with particle placement is summarized in Figure 4.

Searching for particle interactions

In contrast to protein-ligand interactions, the interactions to particles are not independent from each other. For example, two phantoms may partially overlap such that only one of them can be switched to a particle. Another dependence occurs if the maximal number of interactions of a specific type at a particle is reached after an interaction from a ligand atom is formed. Then, the particle cannot form another interaction from that group to another ligand atom later in the incremental construction process. In order to avoid a dependence of the interaction lists from the order of the atoms, we apply a two-phase algorithm to the search for new particle interactions.

The goal of the first phase is to generate a candidate list of particle interactions. Each interacting group of the newly added fragment is tested with respect to whether there is a particle type able to form an interaction to this group. In this case, all particles having the right type and approximately the correct distance to the interaction group are added to the candidate list. The candidate list is then sorted by the deviation of the distance between the particle and the interaction group from the optimal distance of the interaction.

In the second phase, the particle interactions of the candidate list are either rejected or accepted in the order of increasing distance deviation. For each particle interaction, the following tests are performed for the corresponding phantom:

1. Does the phantom exceed the maximal allowed number of interactions for one of the interaction types?
2. Is there a pair of interactions with the phantom that violates the angular range defined for the particle?
3. If the phantom has to be switched to a particle (i.e. if this is the first interaction of the ligand to this particle), does it overlap with the ligand?
4. Does the particle interaction violate angular constraints on the ligand side?

If the particle interaction passes these tests, the interaction is added to the list of interactions of the current placement and the corresponding phantom is switched to a particle.

INCREMENTAL CONSTRUCTION WITH PARTICLES

```
%   precomputing
1   compute particle phantoms;
%   base selection phase
2   B ← select a set of base fragments;
%   base placement phase
3   P ← place base fragments B;
4   foreach p ∈ P do
5       search for particle interactions of the ligand with placement p;
6       score placement p; od
%   complex construction phase
7   R ← ∅; P' ← ∅
8   while P ≠ ∅ do
9       foreach placement p ∈ P do
10          extend placement p by an additional fragment;
11          switch particles overlapping with the ligand back to phantoms;
12          search for interactions of the added fragment to the protein;
13          search for new particle interactions of the added fragment;
14          optimize and score placement p;
15          if p places the complete ligand then R ← R ∪ p;
16          else P' ← P' ∪ p; fi; od;
17      P ← select k best of P'; P' ← ∅
18      cluster placements P; od
19  cluster placements R;
20  return R;
```

Fig. 4. The incremental construction algorithm including the particle concept. Line numbers printed in bold indicate statements which implement the particle concept. *B* is the set of base fragments; *P* and *P'* are the sets of partial placements; *R* is the set of resulting placements.

Removing overlapping particles

When a new fragment is added during the incremental construction, the fragment may overlap with a particle placed in a previous iteration. In this case, the particle and its corresponding interactions have to be removed again. Deciding whether a particle should be removed or not is a complicated issue. Removing the particle means that the interactions are also deleted and the particle does not support the current placement anymore. On the other hand there are cases in which the particle occupies the space for a ligand group replacing the particle with its corresponding interactions.

We solve this problem with a hydrophobicity-dependent overlap threshold. If the particle and the overlapping ligand atom are both hydrophilic or hydrophobic, we assume that the ligand atom replaces the particle. Therefore, a small threshold (1.25 Å^3) is used in this case. If the particle and the ligand atom have different hydrophobicity flags, the standard threshold (2.5 Å^3) is used.

Particles during placement optimization

Placement optimization in FLEXX is based on the weighted rigid-body superposition of points adapting the algorithm of Kabsch[20] to this application. The input is a list of triplets (l_i, r_i, w_i) where l_i represents a point of the ligand, r_i a point of the protein, and w_i the weight. Kabsch's algorithm computes a transformation (R, T) of l_i minimizing the weighted sum of squares $\Sigma_i w_i(Rl_i + T - r_i)^2$.

A triplet is generated for each protein-ligand interaction and for each protein-ligand clash. For an interaction, l_i is the interaction center, r_i is the matched point on the interaction surface of the protein and $-w_i$ is the optimal score of the interaction. For a clash between a protein and a ligand atom with an overlap volume exceeding 2.5 Å^3, l_i is the ligand atom center, r_i is the point closest to l_i not overlapping with the protein atom and w_i is the overlap volume.

Concerning particles, only triplets originating from particle-ligand interactions occur since particles overlapping with the protein are removed before placement optimization. While l_i is again the interaction center, r_i and w_i are computed differently. Because the particle interaction is spherical, r_i is simply the point on the line through l_i and the particle center with the optimal interaction distance, w_i is set to a heuristically chosen reduced value (1.0 instead of 4.7 for a hydrogen bond to a water particle) in order to take the inexact position and mobility of the particle into account. While interaction groups of the protein are relatively fixed, particles are more flexible and can optimize the interaction pattern by changing their position. It is therefore more important to optimize the direct interactions to the protein instead of those to the particles.

Placing Water Molecules With the Particle Concept

The main application of the particle concept is handling discrete water molecules during docking computations. In the particle concept, a water molecule is represented as follows. The radius of the water particle is the van-der-Waals radius of oxygen (1.52 Å). The water particle is able to form four interactions to the protein or the ligand, two as a hydrogen bond donor, two as an acceptor. The interaction geometry is spherical with an optimal interaction distance (hydrogen bond length between acceptor and hydrogen atom) of 1.9 Å. If the particle represents the donor, the hydrogen is assumed to be 1 Å apart from the particle center. All angles between the directions of pairs of interactions must be in the range from 70 to 170 degrees.

Obviously, the water molecule is a hydrophilic particle. The scoring contribution for a vacant interaction at a placed particle is set to 1 kJ/mol. A hydrogen bond between a water particle and the ligand contributes a lower score as between the protein and the ligand, namely 2.7 kJ/mol if the hydrogen bond geometry is optimal. The angular scaling parameters for penalizing angular deviations between interactions at the particle (see description of the particle model above) are 40 and 80 degrees, the optimal angle between interactions is set to 110 degrees.

The scoring and geometry parameters are set to reasonable values within the context of the FLEXX model of molecular interactions. It should be noted that they are not derived from empirical data or theoretical (quantum-chemical) calculations.

RESULTS AND DISCUSSION

We have tested the docking algorithm with the particle concept for water on a set of 200 protein-ligand complexes taken from the PDB.[16] The dataset has been assembled without considering the strengths and weaknesses of FLEXX or the importance of water in complex formation. The first part (95 complexes) was chosen due to the availability of experimentally determined binding affinities. This initial set was then merged with the GOLD dataset most of which is published in Reference 21. A list of all PDB entries is given in Table III. A detailed description of the dataset and the docking results achieved with FLEXX will be presented elsewhere.[22] Here we will present some general statistics concerning the usage of the particle concept as well as some selected examples where the advantages and drawbacks of the approach can be demonstrated.

Predicted Phantom Locations for Water

In a first test we compare water molecules explicitly given in the PDB files with water phantom locations computed with the algorithm explained above. Because the phantom locations are not changed during docking, a phantom location near a water molecule found in the PDB file is necessary in order to predict this water molecule during a docking computation.

Water molecules are extracted from the PDB files with the following algorithm. For each water molecule, the number of contacts (hydrophilic atoms within a distance less than or equal to 3.6 Å) to the protein and the ligand are counted. All water molecules forming at least one protein and one ligand contact are then extracted. Note that because hydrogen information and local geometries are not considered, a contact is only a necessary but not sufficient condition for a hydrogen bond.

With this procedure, 335 waters are extracted, 103 of which contain only one protein contact and therefore cannot be predicted. We found that among those there are only 22 waters forming one protein contact and more than one ligand contact. The remaining 81 waters form only two contacts to hydrophilic atoms and are therefore located near the surface of the complex or are energetically unfavorable. This is a justification that our restriction to particles with more than one contact to the protein is reasonable.

FLEXX generates 19,650 phantom locations in total, about 98 per active site, on the average. Normally, a set of 5 to 15 locations describe a water position lying in the intersection of two or three interaction surfaces. As an example, the phantom locations around HOH-1 in a complex of HIV-1 protease with an inhibitor (4phv) is shown in Figure 3.

For each water extracted from the PDB files, we computed the shortest distance to a phantom location. Over the whole set, the average distance is 1.2 Å. Taking only the water molecules with more than one protein contact into account, the average distance reduces to 0.8 Å. Figure 5 gives the average distance with respect to the temperature factor and Figure 6 with respect to the number of

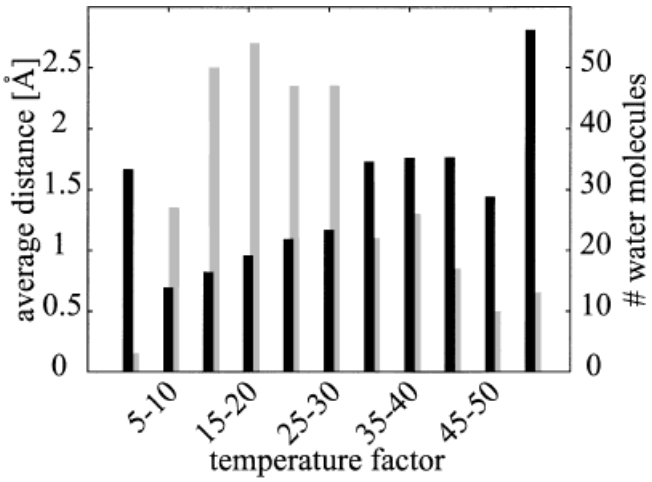

Fig. 5. Average particle distance for water molecules found in the PDB files with respect to the temperature factor (black boxes). The number of PDB water molecules in each temperature factor range is shown with grey boxes. Temperature factors below 1.0 are assumed to be U values and multiplied by $8\pi^2$.

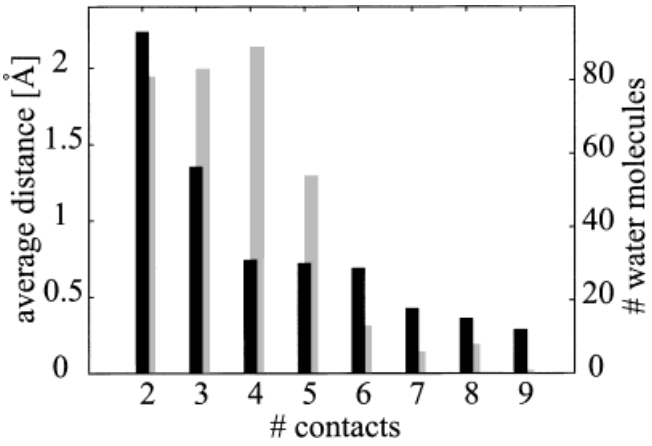

Fig. 6. Average particle distance for water molecules found in the PDB files with respect to the number of contacts (black boxes). The number of PDB water molecules for each number of contacts is shown with grey boxes.

contacts. The temperature factor is related to the mean square displacement of the water molecule. A water molecule with a low temperature factor can be assumed to be located at an energetically favorable location. Temperature factors of 0.0 and 1.0 are assumed to be undefined (19 water molecules) and not taken into account; factors between 0.0 and 1.0 are assumed to be U values instead of B values and are therefore multiplied by $8\pi^2$.

As expected, the average distance increases with an increasing temperature factor and decreases with an increasing number of contacts. This holds except for the lowest temperature factor class from 0 to 5. There are three water molecules in this class, belonging to 1mdr (0.22 Å), lxie (0.44 Å), and 7tim (4.4 Å). The high average distance is caused by the water molecule in 7tim which

TABLE I. RMS Table[†]

RMSD [Å]	1.0	1.5	2.0	2.5	∞
FLEXX run, standard					
Number of examples	95	113	123	133	192
Average rank	23.9	9.5	6.4	8.3	
FLEXX run, particle concept					
Number of examples	88	113	126	135	191
Average rank	14.8	11.5	10.2	8.7	

[†]Row 1: Number of examples which can be predicted with an RMS deviation below a given threshold; row 2: the average of the lowest ranking solution below the threshold over all examples which can be predicted with an RMS deviation below the threshold.

TABLE II. Comparison of Docking Results[†]

Rank difference	1	2	3	4	5
Number of improvements	55	48	42	38	35
Number of deteriorations	46	41	36	34	35
Balance	9	7	6	4	0

[†]Comparison on a case-to-case basis as explained in the text. Rows are: the number of test cases with an improvement, the number of test cases with a deterioration, and the balance for the rank difference values from 1 to 5.

forms only 1 protein contact and can therefore not be predicted.

In summary, 75% of the water molecules having more than one contact to the protein, also have a phantom location within 1.0 Å distance (90% within 1.5 Å distance).

Influence of the Particle Concept on the Docking Results

The evaluation of docking results is a difficult problem. If only one solution is generated, the root mean square deviation (RMSD) to the crystal structure can be computed. But this number does not always reflect the quality of the results. In some cases an RMSD of 2.0 Å can be quite a good result, for example for a large ligand or a ligand with mostly hydrophobic groups, while in other cases it is not. Most docking programs generate a set of suggested structures instead of a single one which makes the evaluation even more difficult.

As a first comparison, we count the number of test cases for which a solution's RMSD does not exceed a given threshold. These numbers are given for a computation without and with the particle concept for water in Table I. While the number of correct test cases for larger RMSD bounds is slightly larger with the particle concept, it decreases for the lowest bound (1.0 Å). The reason for this phenomenon seems to be that if the correct placement is found without applying the particle concept, the waters which are placed only at approximate locations, change the position of the ligand and cause a minor increase in the RMSD. Examples of this kind will be pointed out in the case studies section below.

A more detailed analysis can be achieved by comparing the docking results on a case-by-case basis. For a docking result, we determine the lowest rank of a solution with an RMSD below 1.0, 1.5, 2.0, 2.5 Å respectively. We judge a docking result A with rank sequence (a_0, a_1, a_2, a_3) to be better by d ranks than B with rank sequence $(b_0, \ldots, b_3)$ if there is a k such that $a_i \leq b_i$ for $i < k$ and $a_k \leq b_k - d$.

For 4dfr, for example, we get the rank sequence (6,2,1,1) with and (8,1,1,1) without the particle model. In this case, using the particle model is considered as an improvement for $d \leq 2$ because the highest ranking solution with RMSD less or equal 1.0 Å increases its rank from 8 to 6. A summary of improvements and deteriorations with this measure for $d = 1, \ldots, 5$ is given in Table II.

The largest difference can be seen for $d = 1$ where an improvement occurs in 27.5% of the test cases compared to a deterioration in 23%. The PDB codes of these test cases are listed in Table III.

In order to analyze the number of correctly predicted water molecules, we focus on correctly predicted test cases (RMSD $\leq$ 1.5 Å). These 113 test cases contain 162 water molecules with more than one contact. Considering the solution with RMSD less or equal 1.5 Å and lowest rank, 56 of these water molecules are predicted with a distance error of less or equal 1.5 Å. PDB codes of examples with correctly predicted waters are shown in Table III.

Computation Time

The additional time requirement for the particle concept is minor. The generation of phantom locations take about 1 sec. The average computation time for docking a test case increases by 7% from 72 to 77 sec. All calculations are performed on a SUN Ultra-30 with a single 296 MHz processor and 128 MB main memory.

Case Studies

In this section some of the above mentioned test cases are discussed in more detail.

Improvements

4phv and 1aaq. For these two complexes of HIV-1 protease and inhibitors, using the particle concept for water results in an improvement of the prediction quality. Docking HIV-1 protease inhibitors is a difficult task because often they are large and highly flexible. An additional problem arises from the shape of the pocket forming a kind of a tube instead of a cavity. Our test set currently contains 7 HIV protease complexes. The complexes 1hef, 4hvp, and 9hvp have very large ligands containing 23, 35, and 23 rotatable bonds, respectively, and are not predicted correctly. 1ida (18 rotatable bonds) is also predicted incorrectly, although the ligand is smaller.

1hvr (10 rotatable bonds, 1 flexible ring system) is predicted wrongly under standard conditions. However, a correct prediction is achieved if the central 7-membered ring is rigified (rank 1, 0.7 Å RMSD). In this complex, the central water molecule often seen in HIV-1 protease complexes is replaced by atoms of the central ring, no water molecule is lying between the ligand and protein. As expected, the docking result does not change with the particle concept.

TABLE III. PDB Codes of Test Cases†

Category	No water	Water	Water PDB
Improvements	1cbx 1dwd 1ela 1ghb 1hdc 1lic 1ppl 1srj 1tni 2dbl 2sim 2yhx 4cts 4tln 5p2p 5tmn	1acm 1cde 1ele 1hgh 1ivc 1ive 1nsc 1poc 1rds 1tnk 2ctc 2lgs 2tmn 2ypi 3aah	1aaq 1abe 1abf 1ake 1blh 1byb 1did 1frp 1ivd 1lna 1mld 1pso 1rob 1snc 1tmn 1tph 1trk 1xid 1xie 2xis 4dfr 4phv 6abp 6tim
No change	1ack 1aha *1apt 1cbs 1ctr* 1dbj 1dbk 1dwb *1eed 1elb 1epb 1fen 1fkg* 1fki *1glq 1hdy* 1hsl *1icn 1igj* 1lah *1lpm* 1mbi 1mmq 1mrg *1mrk 1mup 1nco* 1pha 1phd 1phf *1phg 1rbp* 1stp 1tng 1tnj 1tnl 2ada 2cpp 2mth *2plv 3cla* 3ptb *4est 4fab 5cpp 6tmn*	1aco *1aec* 1ase 1avd *1baf 1bbp 1bma 1cdg* 1com 1coy 1cps 1dbb *1eap 1elc 1eld 1etr* 1glp *1hef 1hgg 1hri 1hvr 1ida* 1ldm *1lmo 1mcr 1ppi 1ppm* 1tnh 1ulb *2mcp* 2phh *2r04 2r07* 3gch *3hvt* 4fbp *4hvp 4tmn 5cts 6cpa* 6rsa *8gch*	1lst *1rne* 1wap *2er6* 2gbp 4ts1 5tim *6rnt 7cpa* 8atc *9hvp*
Deteriorations	1azm 1cil 1die 1hgi *1hgj* 1mdr 1pbd *1pph* 1ppk 1rnt *1slt 1tpp* 1tyl 2ak3 *2cgr* 3tpi 5abp	1atl 1dbm *1eta* 1hfc 1hti 1ivb 1ivf 1lcp 1nis 1ppc *1tdb* 1thy 1tka 1tlp 2cht 2pk4 3cpa 4fxn 4hmg	121p 1acj 1dr1 *1dwc* 1hyt 1imb 1ukz 2cmd 4tim 7tim

†The test cases are divided into three classes (improvements, no change, deteriorations) depending on the difference in accuracy between a standard docking computation and a docking computation with particle concept used for water. Columns are: no predicted water in the final docking solution, at least one predicted water molecule in the final docking solution, water molecules predicted in the final docking solution also seen in the PDB file within 1.5 Å distance. The considered solution is the best prediction; i.e., the one with lowest rank and RMSD ≤ 1.0 Å, or the one with lowest RMSD if no solution with RMSD ≤1.0 Å found. PDB-codes in italic are those where no solution with RMSD ≤2.5 Å is found.

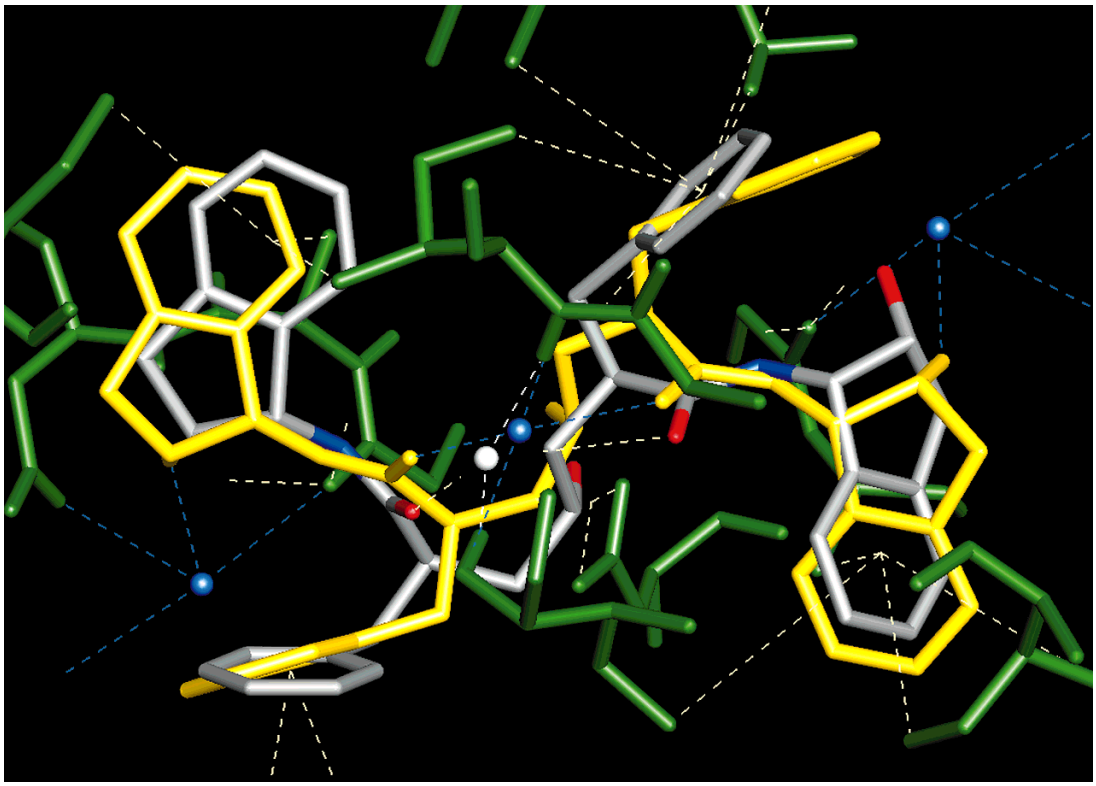

Fig. 7. Docking of the inhibitor L700,417 in HIV-1 protease (4phv), interacting amino acids are shown in green, the ligand in crystal orientation is shown in yellow, the predicted position at rank 2, 1.3 Å RMSD is shown in atom-type colors. Dashed lines visualize interactions between the ligand and the protein. Predicted particles are shown in white, crystallographically observed water molecules in blue.

Here we focus our discussion on 4phv and 1aaq containing HIV-1 protease with two moderately sized inhibitors (17 and 21 rotatable bonds). In both cases, the already mentioned water molecule (HOH-1) shows up in the protein-ligand interface. Using FLEXX without the particle concept results in a completely wrong prediction (12.9 Å RMSD, rank 1) for 1aaq and a partially wrong prediction (2.5 Å RMSD, rank 29) for 4phv. Using the particle concept, both complexes can be predicted correctly including the position of the water molecule (0.9 Å RMSD, rank 1

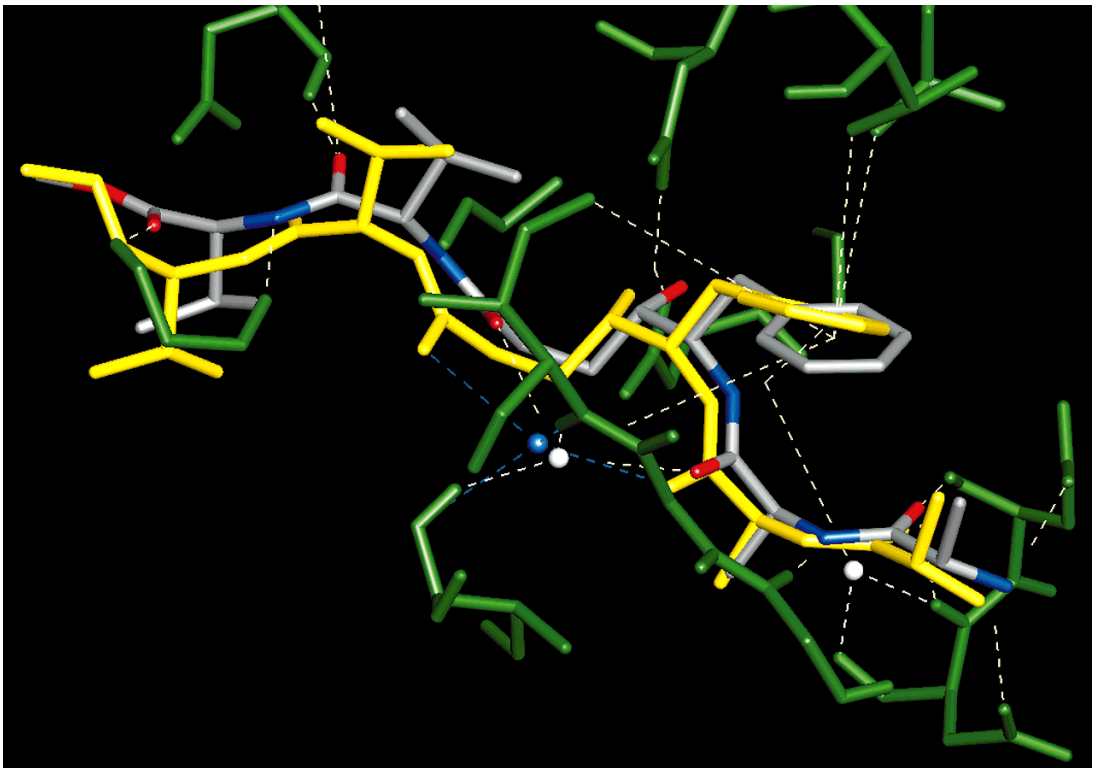

Fig. 8. Docking of a hydroxyethylene isostere in HIV-1 protease (1aaq), placement at rank 1, 0.9 Å RMSD. See Figure 7 for a color legend.

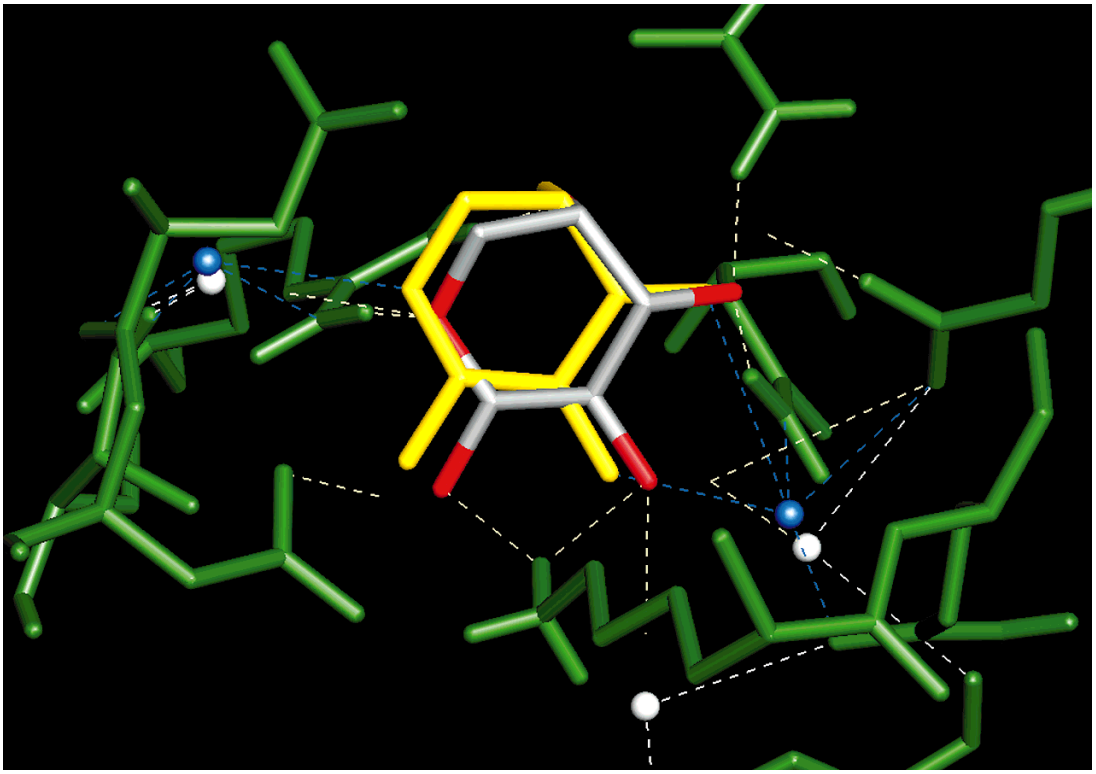

Fig. 9. Docking of L-arabinose to L-arabinose binding protein (1abe), placement at rank 3, 0.5 Å RMSD. See Figure 7 for a color legend.

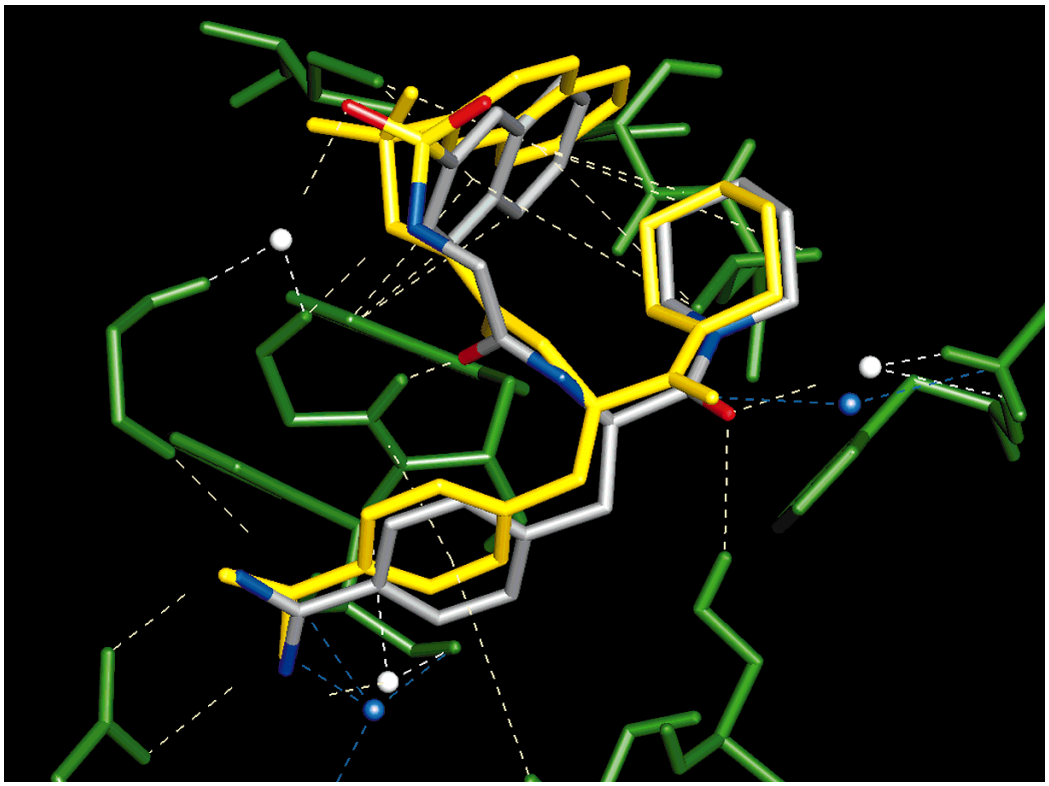

Fig. 10. Docking of NAPAP to α-thrombin (1dwd), placement at rank 1, 1.4 Å RMSD. See Figure 7 for a color legend.

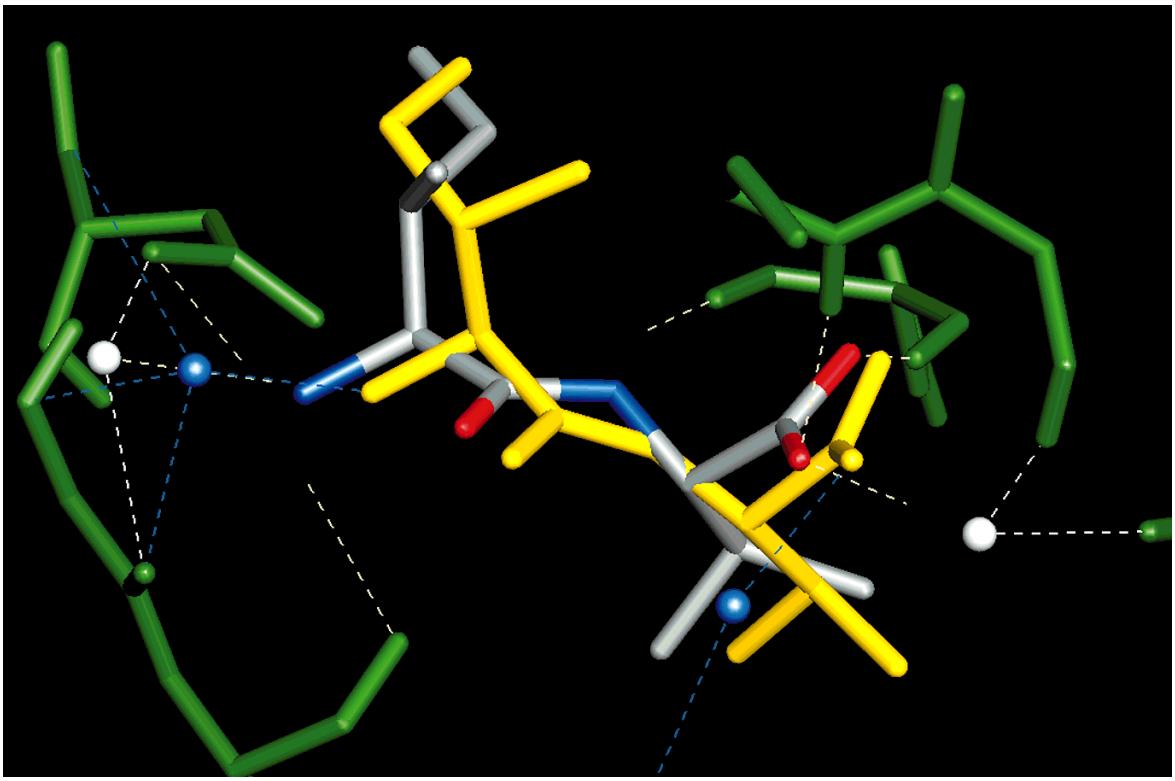

Fig. 11. Docking of the dipeptide ILE-VAL to Trypsinogen (3tpi), placement at rank 1, 1.1 Å RMSD. See Figure 7 for a color legend.

for 1aaq and 1.3 Å RMSD, rank 2 for 4phv). The complexes are shown in Figure 7 and 8.

In principle, a result of similar quality can be achieved if the water molecule HOH-1 is considered as part of the protein and added to the input (as it was done in previous computations with FLEXX[10]). However doing so may be misleading because there are cases where this water molecule is replaced such as in 1hvr. The advantage of the particle concept is that single water molecules do not have to be predefined and can be added on a case-by-case basis.

1abe. The PDB entry 1abe contains a complex of L-arabinose binding protein with its substrate L-arabinose. L-arabinose is a flexible five-membered ring containing four hydroxy groups. The major problem occurring when complexes with sugar molecules have to be predicted is that often there is a variety of placements forming nearly equal numbers of hydrogen bonds to the protein such that the selection of the correct placement is a difficult task.

Docking L-arabinose using FLEXX without particle concept results in a placement with 1.8 Å RMSD at rank 1. The first placement below 1.0 Å RMSD is at rank 3 (0.9 Å), and a solution with 0.5 Å at rank 14.

Using FLEXX with particle concept, the placements vary only slightly, but the ranking is improved. While the previously highest ranking solution (1.8 Å) drops to rank 10, the solutions with low RMSD values decrease to rank 1 (0.8 Å) and 3 (0.5 Å). In both placements, the ligand forms additional hydrogen bonds to water molecules also observed crystallographically (see Figure 9).

1dwd. 1dwd is a well-known test case for docking algorithms containing the complex between human α-thrombin and the inhibitor NAPAP. In the current version of FLEXX, there are several solutions below 1.5 Å RMSD among the lowest ranking, the solution at rank 1 has an RMSD of 1.0 Å.

With the particle concept, a solution with 1.4 Å RMSD is found (see Figure 10) at rank 1. It contains three placed water molecules, two of them are also crystallographically observed. The large RMSD results from a 180 degree rotation of the naphtyl unit. In contrast, the remaining part of the ligand forming the hydrogen bonds to two of the placed water molecules are in better agreement with the crystallographic complex than the solutions obtained with a standard run of FLEXX.

Deteriorations

3tpi. 3tpi contains a complex between trypsinogen and the dipeptide ILE-VAL. With a FLEXX standard run, the complex is predicted with 0.6 Å at rank 1. Using FLEXX with the particle concept, the RMSD of the highest ranking placements increases to 1.1 Å (see Figure 11). This solution contains two additional hydrogen bonds to water particles. Although one of the water particles is near a crystallographically observed water molecule (1.2 Å), these hydrogen bonds cause the increased RMSD.

1dwc. 1dwc is another complex with human α-thrombin, this time with the inhibitor argatroban. Problems with this example arise from geometric deviations in the hydrogen bond network between ligand and protein. Nevertheless, in a standard run of the current version of FLEXX a placement with 1.0 Å RMSD can be found at rank 2 mostly deviating in the orientation of the guanidino group.

1dwc is an example in which FLEXX loses the correct docking solution when the particle model is used. The interactions to particles rearrange the lists of partial docking solutions during the complex construction phase. In this case, solutions with high RMSD are improved in rank resulting in a final solution set in which all solutions have an RMSD above 3.0 Å.

CONCLUSIONS AND FURTHER RESEARCH DIRECTIONS

The particle concept is an extension of our incremental construction docking algorithm realized in the docking tool FLEXX. It features the placement of spherical objects between the ligand and the protein during the complex construction phase. Particles act as single atoms and interact with the ligand sterically and chemically. The main application of the particle concept is the representation of discrete water molecules. A single water molecule is modeled as a particle with the radius of an oxygen atom able to form four hydrogen bonds.

We have applied the extended docking algorithm on a test set of 200 protein-ligand complexes with known structure taken from the PDB. In correctly predicted placements, 35% of the water locations seen in the PDB file are also predicted. Because the overall improvement of the docking results (27.5% of the test cases improve, 23% of them deteriorate) is small, in its current version the particle concept is not a generally applicable feature. However, it has shown to be quite useful in specific test cases like HIV-1 protease where a single water molecule is known to play a critical role.

The focus of this paper is on the methodological integration of discrete water molecules into docking algorithms based on incremental construction. Concerning the integration of discrete water molecules into the scoring scheme of protein-ligand complexes, we have developed an initial model. This integration is crucial for the success of the method and should be investigated further. Looking at wrongly predicted protein-ligand complexes, it becomes clear that cavities formed between the two molecules must be considered more explicitly. The particle concept helps to fill the cavities which are uncritical from the energetic point of view and guides the complex construction such that hydrogen bonds between the ligand and water molecules in these cavities can be formed. Considering energetically unfavorable cavities in the scoring scheme should further improve the docking results.

ACKNOWLEDGMENTS

The authors thank Wolfram Altenhofen (BASF AG, Ludwigshafen), Daniel Hoffmann (GMD SCAI), Gerhard Klebe (University of Marburg), Christian Lemmen (GMD SCAI) for helpful comments on this work.

SUPPLEMENTARY MATERIAL

The FLEXX software package is available for SUN, SGI, and PCs running the Linux operation system. Interested readers should visit our WWW page http://cartan.gmd.de/FlexX or contact the corresponding author.

REFERENCES

1. Andrews PR, Craik DJ, Martin JL. Functional group contributions to drug-receptor interactions. J Med Chem 1984;27:1648–1657.
2. Wlodawer A. Rational drug design: the proteinase inhibitors. Pharmacotherapy 1994;14:9S–20S.
3. Poornima CS, Dean PM. Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. J Comput Aided Mol Des 1995;9:500–512.
4. Levitt M, Park BH. Water: now you see it, now you don't. Structure 1993;15:223–226.
5. Ladbury JE. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. Chemistry and Biology 1996;3:973–980.
6. Israelachvili J, Wennerström H. Role of hydration and water structure in biological and colloidal interactions. Nature 1996;379:219–225.
7. Böhm, H-J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des 1994;8:243–256.
8. Jain AN. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 1996;10:427–440.
9. Raymer ML, Sanschagrin PC, Punch WF, Venkataraman S, Goodman ED, Kuhn LA. Predicting conserved water-mediated and polar ligand interactions in proteins using a k-nearest-neighbors genetic algorithm. J Mol Biol 1997;265:445–464
10. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261:470–489.
11. Rarey M, Kramer B, Lengauer T. Multiple automatic base selection: protein-ligand docking based on incremental construction without manual intervention. J Comput Aided Mol Des 1997;11:369–384.
12. Rarey M, Wefing S, Lengauer T. Placement of medium-sized molecular fragments into active sites of proteins. J Comput Aided Mol Des 1996;10:41–54.
13. Leach AR, Kuntz ID. Conformational analysis of flexible ligands in macromolecular receptor sites. J Comput Chem 1992;13:730–748.
14. Welch W, Ruppert J, Jain AN. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. Chemistry and Biology 1996;3:449–462.
15. Makino S, Kuntz ID. Automated flexible ligand docking method and its application for database search. J Comput Chem 1997;18:1812–1825.
16. Bernstein FC, Koetzle TF, Williams GJB, et al. The protein data bank: A computer based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
17. Böhm H-J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des 1992;6:61–78.
18. Böhm H-J. LUDI: Rule-based automatic design of new substituents for enzyme inhibitor leads. J Comput Aided Mol Des 1992;6:593–606.
19. Duda RO, Hart PE. "Pattern Classification and Scene Analysis." New York: John Wiley & Sons, Inc., 1973. 228 p.
20. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 1976;32:922–923.
21. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
22. Kramer B, Rarey M, Lengauer T. Validation of the FLEXX incremental construction algorithm for protein-ligand docking. Submitted for publication.