Alternative to Consensus Scoring—A New Approach Toward the Qualitative Combination of Docking Algorithms

Antje Wolf,* Marc Zimmermann, and Martin Hofmann-Apitius

Department of Bioinformatics, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Received November 7, 2006

Since the development of the first docking algorithm in the early 1980s a variety of different docking approaches and tools has been created in order to solve the docking problem. Subsequent studies have shown that the docking performance of most tools strongly depends on the considered target. Thus it is hard to choose the best algorithm in the situation at hand. The docking tools FlexX and AutoDock are among the most popular programs for docking flexible ligands into target proteins. Their analysis, comparison, and combination are the topics of this study. In contrast to standard consensus scoring techniques which integrate different scoring algorithms usually only by their rank, we focus on a more general approach. Our new combined docking workflow—AutoxX—unifies the interaction models of AutoDock and FlexX rather than combining the scores afterward which allows interpretability of the results. The performance of FlexX, AutoDock, and the combined algorithm AutoxX was evaluated on the basis of a test set of 204 structures from the Protein Data Bank (PDB). AutoDock and FlexX show a highly diverse redocking accuracy at the different complexes which assures again the usefulness of taking several docking algorithms into account. With the combined docking the number of complexes reproduced below an rmsd of 2.5 Å could be raised by 10. AutoxX had a strong positive effect on several targets. The highest performance increase could be found when redocking 20 protein−ligand complexes of alpha-thrombin, plasmepsin, neuraminidase, and D-xylose isomerase. A decrease was found for gamma-chymotrypsin. The results show that—applied to the right target—AutoxX can improve the docking performance compared to AutoDock and FlexX alone.

INTRODUCTION

Molecular Docking—the positioning of small molecules into the binding site of a target protein—has become a major technique in modern drug design.[1] More and more receptor structures are deposited in the Protein Data Bank (PDB)[2] which makes them available as a target for docking. Improved computing resources potentiated the screening of databases with millions of compounds.[3] This has made docking the method of choice in structure-based virtual screening.[4]

Since the development of the first docking algorithm by Kuntz et al. in the early 1980s[5] a variety of different docking approaches and tools has been created in order to solve the docking problem (for an overview see, e.g., ref 6). Nevertheless, several studies have shown that the docking performance of most tools strongly depends on the considered target and that it is hard to choose the best one in the situation at hand.[7−9] Due to the use of different tools under different conditions for differing targets and the diverse set of criteria in evaluation studies, it is difficult to compare docking tools.[10] This further complicates the decision for the right tool to be used. As Leach et al. state in their recent perspective:[11] 'Clearly, good poses can be produced, but how does one pick the program that will do so reliably for the targets of interest?'. It is extremely difficult to reliably discriminate between binding modes by a single algorithm.[12]

All docking approaches share the problem of the correct ranking of solutions—the scoring step. Most studies show that the tools fail to place the native conformation on the first rank and that there is almost no correlation between predicted and measured binding affinities.[13−15] Consequently, the scoring step is regarded as the main drawback of current docking tools.[12] In spite of these problems docking can serve as a fast virtual screening tool with enrichments often superior to high throughput screening.[3] Moreover, even if not the native binding mode is found, docking can generate interesting hypotheses and alternative binding modes.

As there is no one-solution-to-all-program available, in newer docking experiments more than one docking tool or at least more than one scoring function are used to overcome the weaknesses of single scoring functions. To combine them, consensus scoring schemes have been developed.[16] They integrate the scoring functions by intersection and different rank voting procedures.[17] Although consensus scoring works better in general, it can be worse than the best single function, since the intersection of two nonidentical lists is, by definition, smaller than the individual lists.[16] The better performance of consensus scoring can be attributed to a simple statistical effect: repeated sampling tends to be closer to the true value than any single sampling.[18] The problematic case occurs if the combined scoring functions are correlated, since errors can be amplified rather than diminished in this case.[1] We think the biggest problem consensus scoring poses is that it provides no insights in the charac-

* Corresponding author phone: +49-2241-14-2537; fax: +49-2241-14-2656; e-mail: antje.wolf@scai.fraunhofer.de.

10.1021/ci6004965 CCC: $37.00
Published on Web 05/11/2007

teristics of the scoring functions combined and thus leaves no possibilities for improving them.

In the following we want to show that there are alternatives to consensus scoring. We will present a case study of two of the most widely used docking tools and a new approach toward the combination of these tools in order to improve performance. In contrast to consensus scoring our approach is an integrated system of different modeling schemes and algorithms and considers the particular advantages of the docking tools. It is not based on a statistical effect but provides interpretability of the results and allows further improvement. To integrate interaction schemes or docking algorithms rather than scores is a real challenge because it requires a deeper knowledge of the underlying docking systems.

The commercial docking software FlexX[19] and the freely available tool AutoDock[20] are among the most popular programs for docking flexible ligands into target proteins.[21] FlexX uses incremental construction for building the ligand inside the binding pocket of the protein. Precomputed interaction points of different types are used for the placement of the ligand fragments. AutoDock uses a genetic algorithm with local search for docking and scores the protein−ligand interactions with the help of pairwise affinity potentials from the AMBER[22] force field. Since both tools represent highly diverse docking algorithms, they were chosen for an integrated approach. Our combined docking workflow AutoxX uses the docking algorithm of AutoDock but incorporates parts of the interaction scheme of FlexX. Additionally, a combined visualization helps identifying regions of the protein surface which energetically favor binding.

AutoDock, FlexX, and the combined method AutoxX were evaluated on a test set of 204 diverse protein−ligand complexes from the PDB. Kramer et al.[23] already evaluated their docking tool FlexX by merging the 100 complexes of the GOLD data set[24] and their own test set which resulted in a set of 200 complexes. We enriched these data by five complexes of the plasmepsin family—a protein family we were recently working on in the course of the WISDOM[25,26] project.

MATERIALS AND METHODS

Preparation of the Data Set. For an evaluation of the combined docking method the FlexX test set[27] comprised of 200 complex structures from the PDB plus five complexes of the plasmepsin family were docked with AutoDock, FlexX, and AutoxX. The test set contains 23 identical proteins cocrystallized with different inhibitors. One complex from the FlexX test set had to be excluded because the bound ligand contains a vanadium atom for which no AutoDock parameters were available.

The 200 complexes were obtained from BioSolveIT[28] in a ready-to-dock format for FlexX containing the receptor files and the ligand in the crystal and energy minimized conformation. Since AutoDock expects different input formats, the molecule files had to be prepared again for AutoDock.

Receptor Structures. The preparation of the target structures for AutoDock is composed of the following steps: All water molecules and ligands were removed from the PDB file. Adding of hydrogens, adding of Kollman charges (Gasteiger charges for hetero atoms), merging of nonpolar hydrogens, and adding of solvation parameters was performed using AutoDock Tools (ADT).[29]

Exceptions from the general procedure were made for the following complexes: The test set contained six structures of cytochrome P-450 containing a heme group which is essential for binding. This hetero group needs a special treatment in ADT. We defined a new atom type for Fe with AMBER parameters for van der Waals potential and a Gasteiger charge of +2. To all other atoms of the heme group Gasteiger charges were assigned. To the rest of the protein atoms Kollman charges were added. Fe was substituted by the letter M in the PDBQS file, the input file for receptor structures in AutoDock. It is similar to a PDB file but contains partial charges and solvation parameters for every atom. This way Fe is recognized by the AutoDock affinity grid computing engine *AutoGrid*.

As suggested by Kramer et al.[23] one water molecule was left inside the pocket of the HIV-1 proteases 1aaq and 4phv because it is known to be essential for binding.

Ligands. The ligands of the data set were present in their crystal as well as in their energy minimized conformation in MOL2 format. To prepare them for AutoDock we assigned proper MOL2 atom names. Through this procedure the atom order was changed, and thus it was not possible to use the raw crystal conformations as reference for rmsd calculations in AutoDock. In order to obtain reliable rmsd values we decided to do the following: In cases where the rmsd between the crystal and minimized structure was smaller than 1 Å, the minimized input structure was used as reference—which is the default procedure in AutoDock. In the remaining cases the minimized conformations were docked with the reference placement functionality of FlexX and saved in MOL2 format to obtain a conformation similar to the crystal. The FlexX reference placement algorithm[30] tries to find a placement close to a reference ligand, like the ligand in the X-ray structure in our case. For four ligands FlexX did not find a reference placement. For these ligands the atoms of the minimized and crystal conformation were mapped by hand.

The ligand input PDBQ files were created with ADT. This included adding of Gasteiger charges, merging of nonpolar hydrogens, and the definition of torsions.

Docking Protocols. For all docking runs the latest available versions of the tools were used. These are AutoDock 3.0.5 and FlexX 2.0, respectively. The parameters for docking with FlexX as well as with AutoDock were as recommended in the manuals. In FlexX the options *place_particles* and a maximum overlap volume of 2.5 $Å^3$ was used. The active site was defined by using a cutoff distance of 6.5 Å around the reference ligand in the X-ray structure. For docking with AutoDock the Lamarckian genetic algorithm was applied with the parameters for the genetic algorithm listed in Table 1. Except for the number of energy evaluations these are the default parameters. We decided to choose a higher number of energy evaluations to ensure a sufficient sampling of the conformational space of the ligands. The grid maps were centered on the center of mass coordinates of the ligand in the X-ray structure and were of dimension $61 \times 61 \times 61$ points. The grid spacing was 0.375 Å.

Table 1. Genetic Algorithm Parameters for AutoDock[a]

no. of energy evaluations	$1.5 * 10^6$
no. of generations	27 000
population size	50
no. of runs	10

[a] Except for the number of energy evaluations these are the default parameters.

Algorithm. Before we introduce the algorithm used to combine AutoDock and FlexX, it is necessary to describe how the single tools model protein−ligand interactions.

AutoDock. AutoDock uses precomputed affinity grid maps for each atom type to score the conformations of the ligand found during the genetic algorithm. These three-dimensional grids are placed over the binding site of the protein. At each grid point of the map a probe atom is placed, and the interaction energy between that atom and the protein is precomputed. The resulting values are a linear combination of van der Waals, hydrogen bond, and solvation energies specific for the appropriate atom type. Additionally, an extra electrostatic map is used whose values are then multiplied with the partial charge of the atom. The energetics of a particular ligand configuration is then calculated by trilinear interpolation of affinity values of the eight grid points surrounding each of the atoms in the ligand. AutoDock is a noncommercial tool and therefore has found a large user community, especially in academic institutions. The Lamarckian genetic algorithm works also in cases where the ligand binds via unspecific interactions like hydrophobicity or electrostatics completely since it does not need hydrogen bonds or salt bridges for placement. But compared to FlexX and other docking programs it has a long run-time.[31] In our experiments (with the parameters listed in Table 1) the AutoDock docking took a few minutes to over half an hour for ten runs, whereas the FlexX docking could be done in a few seconds to minutes. Because of its stochastic nature the Lamarckian genetic algorithm has to be run several times to achieve a reliable result. Moreover an appropriate choice of the parameters is necessary. The user has the possibility to adjust a variety of parameters. This makes AutoDock a very flexible tool, but its application requires experienced users.

FlexX. FlexX uses an incremental construction algorithm where the ligand is divided into fragments. A preferably rigid and specific base fragment is chosen and docked in first place. Subsequently, the remaining fragments are added. To each interacting group of the receptor molecule an interaction type and geometry are assigned. The geometry depends on the interaction type and consists of an interaction center, radius, and surface (spheres, spherical caps, or spherical rectangles). There are three different interaction levels: polar interactions (hydrogen bonds and salt bridges) and directed hydrophobic and undirected hydrophobic interactions. For algorithmic reasons the surface is approximated by discrete interaction points. In the FlexX algorithm these points are used for the placement of interaction centers of the ligand to the matching interaction points of the protein. The incremental construction algorithm makes FlexX one of the fastest docking tools available,[31] but the results are strongly dependent on the choice of the base fragment.[32] If it cannot be placed correctly, then the whole ligand will be docked at a wrong site of the protein. The quality of the base fragment depends on the existence of directed interactions like hydrogen bonds and salt bridges. Thus when docking into completely lipophilic binding sites FlexX is likely to fail.[8]

AutoxX. The idea of AutoxX is to make the different approaches of the docking tools AutoDock and FlexX comparable. In order to improve docking results we wanted to combine their specific advantages and introduce an alternative approach to consensus scoring.

FlexX has hard interaction constraints. This is advantageous for hydrogen bonds which have a well-defined geometry. The AutoDock interaction energies are smoother because of the potentials used. In general, these smooth potentials are preferable for the modeling of apolar interactions which are very unspecific. In-depth analysis of the affinity grids computed by AutoDock and FlexX showed that the biggest differences occurred in the modeling of hydrogen bonds. AutoDock uses a 10−12 Lennard Jones potential,[33] whereas FlexX uses known discrete hydrogen-bonding geometries. Therefore we incorporated the FlexX hydrogen bond models into the AutoDock grid maps because we expected it to strengthen the AutoDock docking.

To achieve a unification of the interaction schemes of AutoDock and FlexX, we had to make the different interaction models comparable. In AutoDock the separation of the calculation of the molecular affinity grids from the docking simulation provides a modularity to the procedure. This facilitates the introduction of modified affinity grids. Therefore and because it is algorithmically easier to transform single points into a grid than the other way around we decided to integrate FlexX into AutoDock and not AutoDock into FlexX. Figure 1 shows a schematical depiction of the transformation process.

We transformed the FlexX interaction points into grid maps similar to the ones produced by AutoDock. To increase variability in the FlexX grids Gaussians were applied for the transformation

$$f(x) = h \exp^{-b||x-c||^2} \quad (1)$$

where height h and center c are the FlexX energy contribution of this interaction type and the coordinates of the interaction point, respectively. Each interaction point is represented by one Gaussian. The width b was chosen that way that two adjoining Gaussians of a common interaction surface intersect at $h/2$ (see Figure 1). The interaction points are distributed equidistant on the spherical interaction surface but not in Euclidean space. Since the Euclidean distances of the interaction points vary, average distances were computed for each interaction level for an example protein. For polar interactions the points have an average distance of about 0.7 Å which is twice as much as the spacing of the AutoDock grid maps. The transformation procedure assigns a value $f(x)$ between h and zero to each point x of a FlexX affinity map. If Gaussians are in proximity and interfere with each other, points can also have values smaller than h because the values of the Gaussians are added.

Figure 2 depicts the workflow of the AutoxX approach. In the first step the targets have to be prepared for docking in a tool-specific manner. Second, the AutoDock grid maps and the FlexX interaction points are computed. AutoDock already produces separate map files, and the FlexX interaction points have to be extracted with a special script in the FlexX-own scripting language. Subsequently, the FlexX

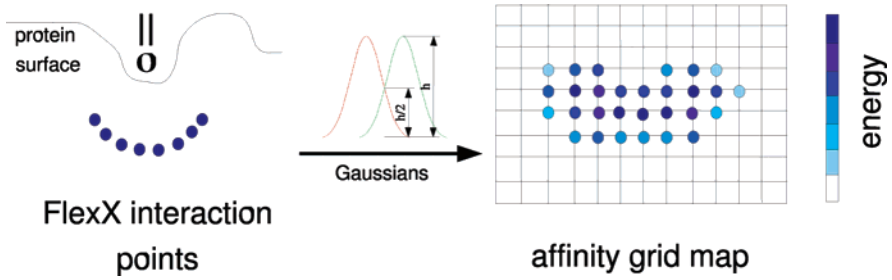

Figure 1. Transformation of FlexX interaction points into an affinity grid. The interaction points of the active site surface of the protein are computed by FlexX and extracted with a script. With Gaussians the single points are transformed into an affinity grid with the same dimension like the ones computed by AutoGrid. Each interaction point is represented by one Gaussian. The height of the Gaussians *h* corresponds to the contributing FlexX interaction energy of the interaction type. The width *b* is chosen that way that the Gaussians intersect at *h*/2.

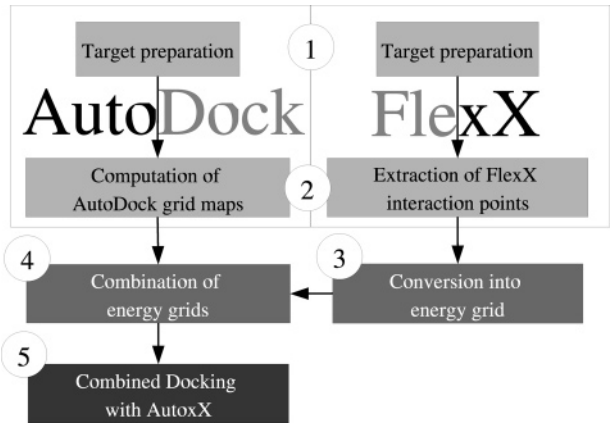

Figure 2. Workflow of AutoxX: 1. target preparation, 2. extraction of FlexX interaction points and computation of AutoDock grid maps, 3. transformation of interaction points, 4. union of maps, and 5. combined docking.

interaction points are transformed into energy grids as described above.

We employed several strategies to unify the AutoDock and FlexX maps. Generally, the FlexX hydrogen bond interactions are incorporated into the maps of those atom types possibly serving as hydrogen acceptor or donor which are oxygen, nitrogen, and polar hydrogen. FlexX donor interactions are mapped to oxygen and nitrogen, FlexX acceptor interactions are mapped to polar hydrogen. In the maps of these atom types the AutoDock energy values are replaced with the FlexX values. The following four replacement strategies were investigated (see also Table 2):

1. The energy values in the AutoDock grid maps are a linear combination of van der Waals, hydrogen bonding, and hydrophobic desolvation energies. When generating these map files, one can switch off any of this contributions. In the first model AutoDock's own hydrogen-bonding model which is based on an angle-dependent 10−12 Lennard-Jones potential is not included when generating the AutoDock maps. The second question is which energy values from the converted FlexX maps to include into the AutoDock maps. Here the generated AutoDock maps are replaced at all grid points where the corresponding FlexX map shows a negative contribution. This means that FlexX has computed interactions of this type in this area. All negative FlexX values are incorporated, independent of the existing AutoDock values at these grid points.

2. Also in the second model the AutoDock hydrogen-bonding potential is not included when generating the AutoDock grid maps. But this time the inclusion of the FlexX energies is dependent on the values computed by AutoDock at these grid points. Only those AutoDock energies are replaced where an "agreement" concerning negative energy contributions between AutoDock and FlexX is found which means that only those FlexX values are integrated where the AutoDock map as well as the FlexX map show a negative contribution at a specific grid point.

Table 2. AutoDock and FlexX Unification Models Tested[a]

	AutoDock hydrogen bond model	AutoDock and FlexX agreement
model 1	no	no
model 2	no	yes
model 3	yes	no
model 4	yes	yes

[a] Variations are the inclusion or omittance of AutoDock's hydrogen-bonding energies and the decision on which FlexX energies to include. Agreement means that both tools calculated a negative energy at a specific grid point.

3. In contrast to the models described above, model 3 includes the AutoDock hydrogen-bonding energies when generating the AutoDock maps. Like in model 1 all negative FlexX grid values are incorporated into the corresponding AutoDock maps, independent of the existing AutoDock values at these points.

4. In the last model the AutoDock hydrogen-bonding contribution is also switched on. Like in model 2 an agreement of AutoDock and FlexX is necessary for the inclusion of FlexX energies which means that both tools calculated a negative energy at a specific grid point.

A problem which occurred in the grid unification models 1 and 3 is the missing of repulsive forces in the AutoxX grid maps because of the introduction of the FlexX interactions. The dominant values in the AutoDock maps stem from van der Waals interactions which are comprised of an attractive and a repulsive part. During the creation of the AutoxX maps some of the values are replaced by FlexX interaction energies. Since FlexX uses another schema to model clashes between the ligand and the protein, sometimes very high repulsive energies in the AutoDock maps are replaced by favorable FlexX energies. During the following AutoxX run, regions which would be excluded by the repulsive term in the original AutoDock maps seem more preferrable. We found that an unification approach is more advantageous where only those FlexX interaction points are introduced which do not contradict with the AutoDock

Table 3. rmsd Values at the First Rank and Minimal Found rmsd as Well as Number of Complexes Docked Below 2 and 3 Å for AutoDock, FlexX, and AutoX[a]

	av first rmsd	av min rmsd	below 2	below 3 [Å]
AutoDock	2.46	1.49	121	147
FlexX	3.90	2.04	85	103
AutoxX	2.17	1.41	126	159

[a] Averaged over all 204 complexes.

repulsive term (models 2 and 4). Since an inclusion of AutoDock's own hydrogen bond model further improved the docking results, we incorporated unification model 4 into our AutoxX approach.

Finally, AutoxX carries out a combined docking with the modified grid maps. During the docking these new modified maps are used instead of the ones created by AutoDock by changing the *docking parameter file* which gives instructions on the docking procedure. With this unification approach the source code of the tools does not have to be modified. The run-time of AutoxX is in the range of AutoDock since the compute-intensive task which are generation of the grid maps and the genetic docking algorithm have to be performed like in the usual AutoDock docking. The additional steps of the AutoxX workflow take just a few seconds and therefore do not carry weight compared to the actual docking.

Visualization. Especially in the field of molecular docking a good visualization is an essential help for result analysis. ADT and the FlexX-own visualization tool FlexV provide a simple view of the docking results. But these built-in visualizations are not able to read the formats of the other tool. For a combined visualization of low-energy regions at the protein surface we used gOpenMol.[34,35]

gOpenMol is a free tool of the Finnish IT center for science[36] for the visualization and analysis of molecular structures and their chemical properties. It is able to display isocontours as well as orbitals and electron densities. Additionally, gOpenMol has an intern command line interpreter based on Tcl/Tk.[37] After transformation into the gOpenMol-readable PLT format AutoDock grid maps can be displayed as isocontours. With our FlexX transformation routine we can create a PLT file out of the interaction points as well which can then be displayed with gOpenMol. Therefore, the FlexX and AutoDock as well as the combined maps can be visualized with the same program which allows a direct comparison of their modeling of the binding site.

RESULTS AND DISCUSSION

Comparison of AutoDock and FlexX. The FlexX results we obtained are very similar to the published ones.[23] There are some slight differences because we used version 2.0 instead of 1.6.5. In our experiments FlexX found no docking solution for two of the complexes: 1baf and 1pha. In Kramer et al. no solution was found for 1hdy and 1pha. They explained a lack of solutions with a rejection of all generated placements because of a significant overlap with the receptor.

As a look at Table 3 shows, the average rmsd between the docked conformations of the 204 ligands found by AutoDock and their conformations in the crystal structure is lower compared to the ones found by FlexX. This is notably striking when considering the solution ranked first where FlexX is worse by more than 1 Å. Additionally, Table 3 contains the minimal rmsd of all docking solutions for one ligand, averaged over all complexes. In contrast to the rmsd at the first rank the minimal rmsd of both tools is quite convincing showing an average deviation from the crystal of 2 Å and 1.5 Å, respectively.

A more detailed analysis shows that the single results of the tools are highly diverse. The correlation between the rmsd found by AutoDock and by FlexX on the first rank is just 0.155. This clearly demonstrates that they perform diverse at different protein−ligand complexes.

The data set used in this study contains proteins cocrystallized with different ligands. In order to investigate possible target effects, we derived a subset of those proteins where at least three complexes were available. These 18 proteins are listed in Table 4. For the proteins alpha-thrombin, L-arabinose-binding protein, and triosephosphate isomerase FlexX is able to reproduce all complexes under 2 Å. Additionally to L-arabinose-binding protein and triosephosphate isomerase AutoDock can do this for FAB fragment and hemagglutinin as well. When redocking the alpha-thrombin complexes, AutoDock fails.

Table 4. Average rmsd Values at the First Rank of AutoDock, AutoxX, and FlexX for Proteins with at Least Three Instances in the Data Set[a]

protein	PDB IDs	AutoDock	AutoxX	FlexX [Å]
alpha-thrombin	*1dwb, 1dwc, 1dwd*	*4.66*	*1.88*	*0.93*
carboxypeptidase a	*1cbx, 1cps, 2ctc, 3cpa, 6cpa, 7cpa*	*1.91*	*1.52*	*3.10*
cytochrome P450-CAM	1pha, 1phd, 1phf, 1phg, 2cpp, 5cpp	1.66	2.05	3.33
D-xylose isomerase	*1did, 1die, 1xid, 1xie, 2xis*	*3.11*	*2.01*	*3.71*
elastase	1bma, 1ela, 1elb, 1elc, 1eld, 1ele, 4est	2.37	2.59	7.67
FAB fragment	1baf, 1dbb, 1dbj, 1dbk, 1dbm, 1igj, 2cgr, 2dbl, 4fab	1.09	1.80	4.27
gamma-chymotrypsin	1ghb, 3gch, 8gch	3.19	4.86	3.31
hemagglutinin	1hgg, 1hgh, 1hgi, 1hgj, 4hmg	1.19	2.12	4.89
HIV-1 protease	*1aaq, 1hef, 1hvr, 4hvp, 4phv, 9hvp*	*3.37*	*2.54*	*9.80*
L-arabinose-binding protein	*1abe, 1abf, 5abp, 6abp*	*1.02*	*0.67*	*1.11*
neuraminidase (sialidase)	*1ivb, 1ivc, 1ivd, 1ive, 1ivf, 1nsc, 2sim*	*4.50*	*3.32*	*3.06*
penicillopepsin	*1apt, 1ppk, 1ppl, 1 ppm*	*3.20*	*2.79*	*5.47*
plasmepsin	*1lee, 1lf2, 1lf3, 1ls5_a, 1ls5_b*	*3.98*	*1.70*	*7.09*
rhinovirus 14	1hri, 2r04, 2r07	2.50	2.70	11.39
ribonuclease T1	1rds, 1rnt, 6rnt	2.28	3.25	4.39
thermolysin	*1hyt, 1lna, 1tlp, 1tmn, 2tmn, 4tln, 4tmn, 5tmn, 6tmn*	*2.20*	*1.64*	*3.60*
triosephosphate isomerase	1hti, 1tph, 2ypi, 4tim, 5tim, 6tim, 7tim	1.21	1.33	1.24
trypsin	*1ppc, 1pph, 1tng, 1tnh, 1tni, 1tnj, 1tnk, 1tnl, 1tpp, 3ptb, 3tpi*	*1.78*	*1.39*	*1.86*

[a] The PDB identifiers of the proteins are given in the second column. The proteins for which AutoxX achieves an improvement of the rmsd over AutoDock are in italics.

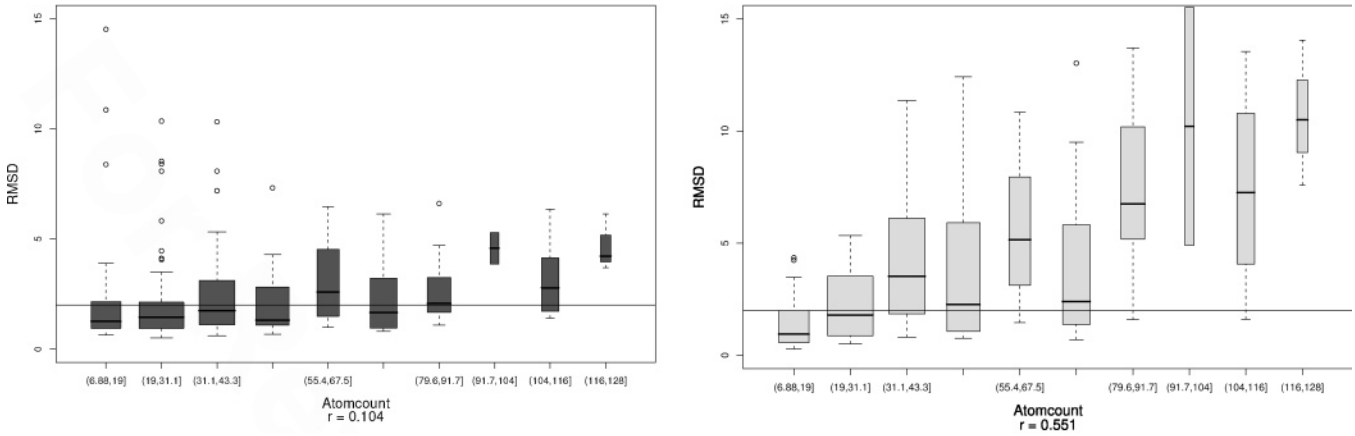

Figure 3. AutoDock (dark gray) and FlexX (light gray) rmsd in dependence of the number of ligand atoms. The boxplot depicts the five-number summary of the data: the smallest nonoutlier observation, the lower quartile, the median, the upper quartile, and the largest nonoutlier observation. The horizontal lines extend to at most 1.5 times the box width. Outliers are represented by small circles. The boxes are drawn with widths proportional to the square roots of the number of observations in the groups. A horizontal line is drawn at 2 Å which marks the barrier for an acceptable rmsd. Below each picture the correlation coefficient between the rmsd and the number of ligand atoms is shown.

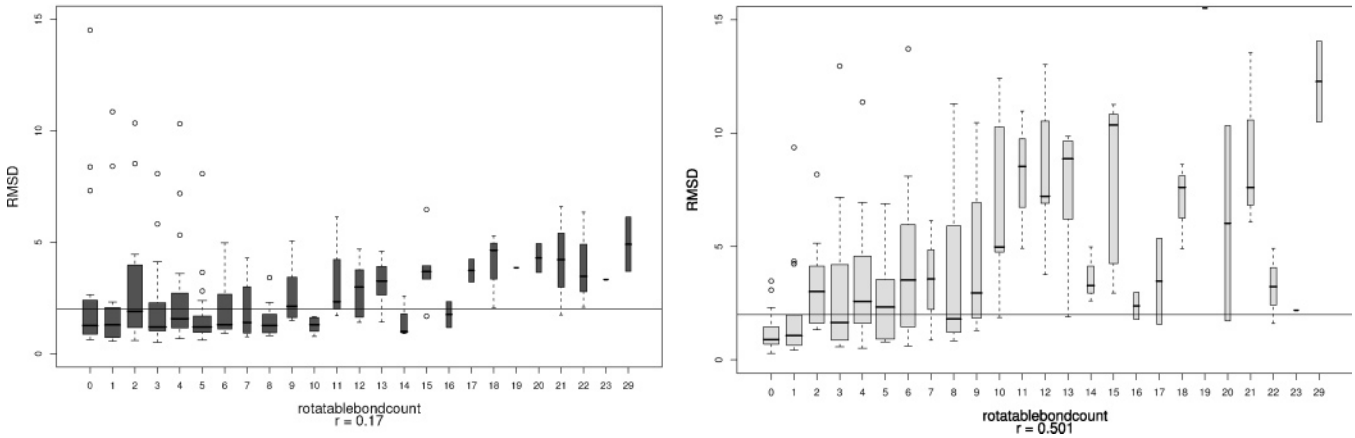

Figure 4. AutoDock (dark gray) and FlexX (light gray) rmsd in dependence of the number of ligand rotatable bonds. For descriptions of the features of a boxplot see Figure 3. Below each picture the correlation coefficient between the rmsd and the number of ligand rotatable bonds is shown.

If one considers the docking accuracy against several chemical properties of the docked ligands, FlexX clearly demonstrates a dependence on the size and flexibility of the ligand (Figures 3 and 4). Similar observations were already noted by others.[23,38] An explanation may be the inadaquate sampling of the conformational space.[38] The AutoDock results seem to be more robust against these properties. Figures 3 and 4 show a slight increase of the median, but the maximal rmsd values are found at smaller, rigid molecules. With adequate parameters the Lamarckian genetic algorithm seems to provide a more exhaustive sampling.

Evaluation of AutoxX. In general, the AutoxX results are more similar to the ones retrieved with AutoDock than with FlexX which is not surprising because AutoxX uses the same algorithm for docking. When looking at the average rmsd values for all 204 complexes (Table 3) one can see an improvement of the minimal rmsd as well as the rmsd at the first rank compared to AutoDock. The minimal found rmsd is very similar with 1.41 Å, but the rmsd at the first rank is decreased by 0.29 Å although the AutoDock value is already at a very low level. This is a decrease by almost 12% of the original AutoDock rmsd value. Even though the FlexX results are not superior to AutoDock, an inclusion of the FlexX hydrogen bond interaction model was able to enhance the AutoDock docking accuracy. For 35 complexes an improvement by more than 1 Å could be seen compared to only 26 complexes which were docked worse by more than 1 Å. With AutoxX the number of complexes docked below 2 Å could be raised by 5, below 3 Å even by 12. These are very encouraging results although an improvement could not be found for all complexes. But this is unrealistic to achieve at such a diverse data set.

As averaging over many observations might not always give a clue, we examined the effect of AutoxX on single targets. Indeed, the performance was very target-specific. AutoxX achieved very positive results for the proteins alpha-thrombin, plasmepsin, neuraminidase, and D-xylose isomerase (see Table 4). Here the average rmsd values compared to AutoDock decrease by 2.78 Å, 2.28 Å, 1.18 Å, and 1.10 Å, respectively. With alpha-thrombin and neuraminidase this resembles more the FlexX results, and one could argue that the FlexX hydrogen bond modeling is more appropriate for these targets. But this does not hold for plasmepsin and D-xylose isomerase. The FlexX rmsd values for the plasmepsin complexes are very high. In spite of that, AutoxX produces good results. The bad rmsd values observed with FlexX are partly caused by symmetry effects which result in a high rmsd although the ligand is well placed inside the

binding pocket. Another reason is that the plasmepsin inhibitors have many rotatable bonds. The interaction model could be good, but the placement algorithm fails. AutoDock's genetic algorithm has a longer run-time but produces more reliable results in this case. A similar trend can be observed for HIV-1 protease. Their ligands are very similar to the ones of plasmepsin, and FlexX fails to find a near-native conformation for them, too. Nevertheless, AutoxX achieves an average improvement of 0.83 Å compared to AutoDock. D-Xylose isomerase catalyzes the interconversion between D-glucose and D-fructose. In our data set the protein is crystallized with different glucose-like ligands. This means these molecules possess many hydroxyl groups able to form many hydrogen bonds. A more thorough hydrogen bond model like the one from FlexX incorporated into AutoDock seems to enhance docking accuracy in this case.

But there are also some targets where the rmsd increases when AutoxX is used for docking. An average increase of more than 1 Å can only be found for gamma-chymotrypsin (1.67 Å). The average rmsd produced by FlexX is in the range of the one computed by AutoDock. Therefore, an inappropriate FlexX model cannot be an explanation why additional FlexX interactions worsened the docking performance compared to AutoDock alone. For a thorough failure analysis the gamma-chymotrypsin complex 1ghb is discussed as an example below.

The success or failing of AutoxX cannot be attributed to one single reason. The explanations are versatile. Therefore we will discuss some positive and negative examples in more detail in the following.

Discussion of Some Single Complexes. 1dwb. This structure contains alpha-thrombin in complex with benzamidine. The complex is docked very well by FlexX with an rmsd of 0.43 Å. AutoDock chooses a totally wrong binding site on the first two ranks which is about 10 Å away from the actual one. AutoxX achieves a very good accuracy of 1.06 Å at the first rank, comparable with FlexX. The ten solutions are clustered into two groups. The second cluster even achieves a better rmsd of 0.65 Å. A visualization of the affinity maps enlightens the reason for the better placement. Figure 5 shows the preferred affinity maps regions of aromatic carbon (A) and polar hydrogen combined with FlexX acceptor interactions (H) at the alpha-thrombin binding site. alpha-Thrombin has a mainly hydrophobic pocket with hydrogen bonds at the catalytic site. In Figure 5 few donor areas can be seen. This is an optimal binding site for FlexX. Because of few polar interactions a very directed docking is possible. This additional information has a positive influence on the AutoDock docking. In the AutoxX-docked conformation the benzene ring of benzamidine is in the preferred region of aromatic carbon, while the amide groups interact via hydrogen bonds with Gly219 and Asp189.

1ls5_a. 1ls5 contains plasmepsin IV chain A complexed with pepstatin A. AutoxX reduces the rmsd at the first rank from 6.35 Å to 2.09 Å. Not only the rmsd at the first rank decreases but also the minimal rmsd found decreases by 0.86 Å, and the average rmsd over all ten runs is reduced by 2.65 Å. This means that not only the ranking of the solutions is better but also every generated solution is much nearer to the crystal structure. The FlexX conformation shows a deviation of 3.23 Å from the crystal. In Figure 6 the atoms which are in preferred FlexX hydrogen bond areas and therefore strongly contributing to the score are labeled. They are in good agreement with the corresponding atoms of the ligand conformation in the crystal where O18 and H12 are forming hydrogen bonds to the catalytic dyad ASP214 and ASP34.

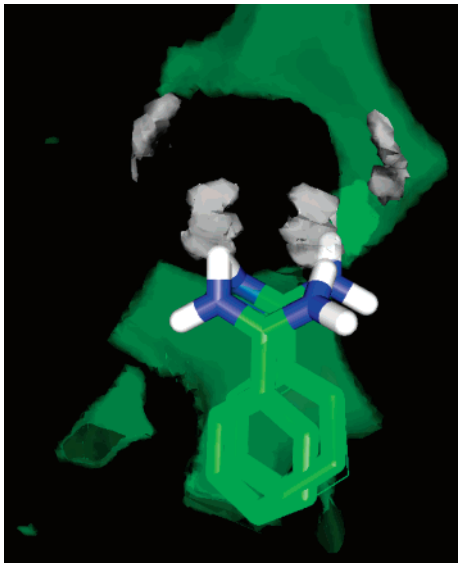

Figure 5. 1dwb. Affinity grid maps of aromatic carbon (green) and polar hydrogen combined with FlexX acceptor interactions (white) for alpha-thrombin with two benzamidine conformations docked by AutoxX. The rmsd values to the crystal conformation (displayed in wireframe) are 1.06 Å and 0.65 Å, respectively.

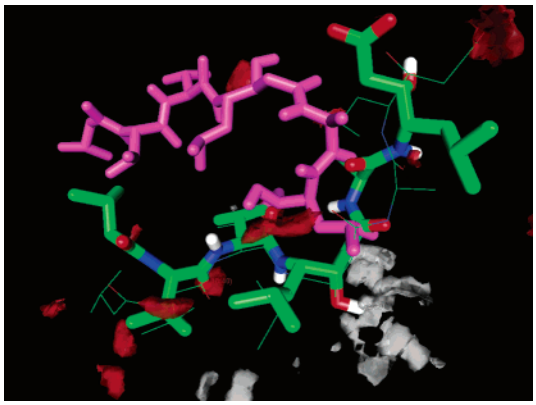

Figure 6. 1ls5_a. Affinity grid maps of polar hydrogen combined with FlexX acceptor interactions (white) and oxygen combined with FlexX donor interactions (red) for plasmepsin. The rmsd values to the crystal conformation (displayed in wireframe) are 2.09 Å for the AutoxX and 6.35 Å for the AutoDock (magenta) conformation. The atoms mainly responsible for the better placement are O18 and H12.

1ghb. 1ghb is the complex for which AutoxX creates the highest loss in docking accuracy compared to AutoDock. It contains N-acetyl-D-tryptophan bound to gamma-chymotrypsin. AutoDock docks the ligand with an rmsd of 1.17 Å at the first rank which is also the minimal rmsd. The FlexX results are moderate with an rmsd of 3.65 Å. AutoxX is not able to dock the complex below 8.26 Å. The AutoxX score mainly consists of the strong contributions of the oxygen of the carboxyl group and the hydrogen of the amide group. In contrast to the AutoDock (magenta) and the crystal conformation (wireframe) one can hardly see hydrophic interactions (green). Here the specific FlexX polar interactions of two groups have distracted the ligand to a wrong binding area.

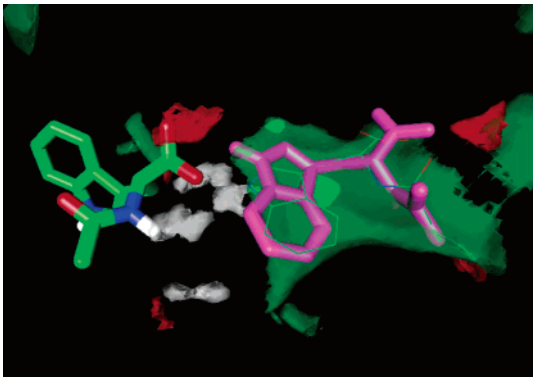

Figure 7. 1ghb. Affinity grid maps of aromatic carbon (green), polar hydrogen combined with FlexX acceptor interactions (white), and oxygen combined with FlexX donor interactions (red) for gamma-chymotrypsin. The ligand conformation docked by AutoxX is on the left side, and the AutoDock conformation (magenta) is on the right. The rmsd values to the crystal conformation (displayed in wireframe) are 8.26 Å and 1.17 Å, respectively.

Failure Analysis of AutoxX. An interesting fact is that the rmsd increase in many cases is caused mainly by an inappropriate ranking of the solutions. Mostly, there is a conformation close to the crystal within the ten solutions, but it is not ranked at the first place. This is a problem of the AutoxX scoring function. But not in all cases the rmsd increase can be explained by a ranking problem. There are a few examples where AutoxX is worse in general (e.g., 1ghb).

An obvious source of failure is an imbalance between H-bonds introduced by FlexX and hydrophobic interactions computed by AutoDock because of the use of the weights of the original scoring functions. Especially, the influence of the H map is sometimes too strong. Thus the ligand hydrophobic parts are not placed into preferrable regions of the A or C map. Our scoring function seems to underestimate hydrophobic interactions which in fact can have a deep impact on the protein−ligand interaction. A regression analysis could put things right but is not possible at the moment because we do not have access to the single terms of the AutoDock scoring function. We are in contact with the AutoDock authors and hope to be able to solve this problem in the near future.

The last explanation we found is an alternative binding mode. In some cases the produced AutoxX conformation does not look incorrect. It is placed well inside the binding pocket, only some parts of the structure are flipped but despite this similar interactions can take place.

Not all of these discussed reasons apply for all complexes. The actual cause of an improvement or impairment of the ligand orientation can only be found when considering the single case.

CONCLUSIONS

In this work we present a new way of making different docking approaches comparable and combining them. We chose the two docking tools AutoDock and FlexX because they are widely used and represent highly diverse algorithms. The idea was to integrate the interaction models in a qualitative manner rather than combining the scores in a quantitative way like in traditional consensus scoring schemes in order to preserve interpretability of the docking results. We developed a method which achieves this without modifying the source code of the tools.

The results of a redocking experiment of 204 diverse complexes from the PDB have shown an improvement of the docking accuracy in general. In a more detailed analysis we have seen that the combined docking only works in about 60% of the cases. We have shown that our combined docking strategy AutoxX works especially well for the proteins alpha-thrombin, plasmepsin, neuraminidase, and D-xylose isomerase. For other targets like gamma-chymotrypsin, where AutoDock alone already reaches a very good docking result, AutoxX can worsen the outcome. Unfortunately, we could not find a general rule determining when AutoxX improves or worsens the docking accuracy. Thus the recommendation we can give is to use a combined docking for the targets mentioned above or when AutoDock alone fails. If one would always choose the best result of both AutoDock and AutoxX one could achieve an average rmsd of 1.89 Å at the first rank, an improvement of 0.57 Å compared to AutoDock alone.

With the visualization of the FlexX, AutoDock, and combined affinity grid maps in one tool it is possible to interpret the influence of the specifities of the single docking tools at a certain binding site on the docking results which cannot be done with results obtained from consensus scoring.

It is obvious that our approach needs refinement. We have included the raw FlexX scoring for hydrogen bonds in the AutoDock scoring function. But all the single terms of the scoring function have weights which were adjusted to experimental binding affinities. We believe that a new calibration of the combined scoring function would lead to better results. At some complexes the hydrogen bonds seem to have too strong an influence on the docking in relation to the other terms.

Like other studies before we have shown that different docking tools perform diverse at different targets and that their performance partly depends on the chemical properties of the considered molecules. A qualitative combination of interaction models and algorithms from different tools like in AutoxX could provide a solution to this problem. By the inclusion of parts of the interaction model of another docking tool we could significantly improve the docking accuracy of several targets AutoDock alone was not able to reproduce correctly. Such an approach could therefore be very useful for virtual screening experiments. The results of this study show that it is not always useful to apply a combined docking, but at the right target it can achieve a great improvement of docking performance in the order of several Å. The challenge is now to predetermine when an application is useful. We are currently working on that topic. By characterizing binding pockets and ligands we want to derive rules for a successful deployment of docking strategies.

AutoxX only considers two of a variety of different docking tools. We selected two diverse docking algorithms to demonstrate that a qualitative combination of them can be an alternative to traditional consensus scoring schemes. But the concept of making interaction models comparable and integrating them is not restricted to AutoDock and FlexX. Potentially, we will include other docking algorithms in our approach.

ACKNOWLEDGMENT

We thank Dr. Frank Cordes for critical reading of the manuscript. We would also like to thank Prof. A. Olson for the AutoDock 3.0 academic license and BioSolveIT for the FlexX 2.0 license and the test set.

REFERENCES AND NOTES

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov*ery **2004**, *3*, 935−949.
(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(3) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.
(4) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.
(5) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Lange, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(6) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151−166.
(7) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242 Evaluation Studies.
(8) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.
(9) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Model.* **2004**, *44*, 793−806.
(10) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, *60*, 325−332.
(11) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of Protein-Ligand Interactions. Docking and Scoring: Successes and Gaps. *J. Med. Chem.* **2006**, *49*, 5851−5855.
(12) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.
(13) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(14) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(15) Warren, G.; Andrews, C.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I.; Woolven, J.; Peishoff, C.; Head, M. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(16) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(17) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discovery Today* **2006**, *11*, 421−428.
(18) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(19) 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.
(20) Morris, G. M.; Goodsell, D. S.; Huey, R.; Hart, W. E.; Halliday, R. S.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(21) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins* **2006**, *65*, 15−26.
(22) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230−252.
(23) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **1999**, *37*, 228−241.
(24) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(25) Jacq, N.; Salzemann, J.; Legré, Y.; Reichstadt, M.; Jacq, F.; Zimmermann, M.; Maass, A.; Sridhar, M.; Kasam, V. K.; Schwichtenberg, H.; Hofmann, M.; Breton, V. Demonstration of In Silico docking at a large scale on grid infrastructure. In *Challenges and Opportunities of HealthGrids*; Hernández, V., Blanquer, I., Solomonides, T., Breton, V., Legré, Y., Eds.; Proceedings of HealthGrid, Valencia, Spain 2006.
(26) WISDOM − Wide In Silico Docking On Malaria. http://scai.fraunhofer.de/WISDOM.html (accessed Jan 3, 2007).
(27) FlexX-200 data set. http://www.biosolveit.de/FlexX/dataset.html (accessed Jan 3, 2007).
(28) BioSolveIT. http://www.biosolveit.de (accessed Jan 3, 2007).
(29) ADT/AutoDockTools. http://autodock.scripps.edu/resources/adt (accessed Jan 3, 2007).
(30) Rarey, M. *FlexX. User Guide, Release 2*; 2005.
(31) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. r. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−763.
(32) 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.
(33) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *AutoDock - Automated Docking of Flexible Ligands to Receptors. User's Guide, Version 3.0.5*; 2001.
(34) Laaksonen, L. A graphics program for the analysis and display of molecular dynamics trajectories. *J. Mol. Graphics Modell.* **1992**, *10*, 33−34.
(35) Bergman, D. L.; Laaksonen, L.; Laaksonen, A. Visualization of solvation structures in liquid mixtures. *J. Mol. Graphics Modell.* **1997**, *15*, 301−306.
(36) CSC − Finnish IT center for science. http://www.csc.fi/english (accessed Jan 3, 2007).
(37) Tcl Developer Site. http://www.tcl.tk (accessed Jan 3, 2007).
(38) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46*, 401−415.

CI6004965