J Mol Model (2003) 9:47–57
DOI 10.1007/s00894-002-0112-y

ORIGINAL PAPER

Tanja Schulz-Gasch · Martin Stahl

Binding site characteristics in structure-based virtual screening: evaluation of current docking tools

Received: 30 September 2002 / Accepted: 11 November 2002 / Published online: 14 January 2003

Abstract Two new docking programs FRED (OpenEye Scientific Software) and Glide (Schrödinger, Inc.) in combination with various scoring functions implemented in these programs have been evaluated against a variety of seven protein targets (cyclooxygenase-2, estrogen receptor, p38 MAP kinase, gyrase B, thrombin, gelatinase A, neuraminidase) in order to assess their accuracy in virtual screening. Sets of known inhibitors were added to and ranked relative to a random library of drug-like compounds. Performance was compared in terms of enrichment factors and CPU time consumption. Results and specific features of the two new tools are discussed and compared to previously published results using FlexX (Tripos, Inc.) as a docking engine. In addition, general criteria for the selection of docking algorithms and scoring functions based on binding-site characteristics of specific protein targets are proposed.

Keywords Docking · Enrichment factor · FlexX · FRED · Glide · Scoring function · Virtual screening

Introduction

Modern approaches for finding new leads for therapeutic targets are increasingly based on three-dimensional information about receptors. As more and more pharmaceutically relevant target 3D structures become available, efficient techniques for exploiting the information contained in these structures gain importance over random experimental screening. [1] Structure-based virtual screening methods [2, 3, 4] now play a major role in lead finding. In particular, software tools for small-molecule docking [5, 6, 7] have recently been applied successfully to large compound libraries. [8, 9, 10, 11]

T. Schulz-Gasch (✉) · M. Stahl
Pharmaceuticals Division, Molecular Design,
F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland
e-mail: Tanja.Schulz-Gasch@roche.com
Tel.: +41-61-6888309, Fax: +41-61-6886459

Database docking is an approach to solving the problem of identifying those compounds in a database of small organic compounds that display favorable steric as well as electrostatic interactions to the target binding site. Docking programs consist of two essential parts: an algorithm that searches the conformational, rotational and translational space available to a candidate molecule within the binding site, and an objective function to be minimized during this process. This function calculates a crude measure of binding affinity or receptor–ligand complementarity and is usually referred to as a *scoring function*. [12, 13, 14, 15] In order to be successful as a virtual screening tool, a docking program must be able to find docking solutions (called *poses*) for active molecules in accordance with experiment, it should be able to separate active compounds from inactive ones, and it should use as little CPU time as possible per compound to be applicable to large libraries.

There are three principal algorithmic approaches to docking small molecules into macromolecular binding sites. [7] A first class of algorithms aims at simultaneously optimizing the conformation and orientation of the molecule in the binding site. Because of the tremendous complexity of this combined optimization problem, systematic solutions are out of reach, and stochastic algorithms such as genetic algorithms or Monte Carlo simulations are usually employed. Docking programs based on such stochastic algorithms, in particular, can give very accurate docking solutions even for very large and flexible ligands. [16, 17] For practical virtual screening, however, this class of algorithms falls behind the other two for its lack of speed, especially because docking runs have to be repeated several times for confident structure prediction. [18, 19] It will therefore not be regarded further here. A second class of algorithms separates the conformational search of the small molecule from its placement in the binding site. A conformational analysis is carried out first, and all relevant low-energy conformations are then rigidly placed in the binding site, whereby only the remaining six rotational and translational degrees of freedom of the rigid conformer must be

considered. We will refer to this approach as "multiconformer docking". Finally, the third class of docking algorithms exploits the fact that most molecules contain at least one small, rigid fragment that is able to form specific, directed interactions with a receptor. Such so-called base fragments are docked rigidly at various favorable positions of the binding site. Docking solutions are then built starting from these various initial base fragment positions in an *incremental construction* process, thereby exploring the (torsional) conformational space of the newly added fragments.

Scoring functions have a two-fold task. First, they serve as an objective function to differentiate between diverse poses of a single ligand in the receptor binding site. Second, after docking a compound database, a scoring function is needed to estimate binding affinities of different receptor–ligand complexes and to rank order the compounds. Due to the crucial role of scoring, a large number of functions have been developed. They can be classified in three categories. The most important class, both in terms of usage and number of available functions, is *empirical* scoring functions. [14] They approximate the free energy of binding as a weighted sum of terms, each term being a function of the ligand and protein coordinates and describing a different type of interaction such as lipophilic contacts and hydrogen bonds between receptor and ligand. The second class of scoring functions is based on molecular mechanics force fields. The binding affinity is estimated by summing up the electrostatic and van der Waals interaction energies between receptor and ligand. Finally, so-called *knowledge-based* scoring functions [15] are derived from statistical analyses of experimentally determined protein–ligand X-ray structures. The underlying assumption is that interatomic contacts occurring more frequently than average are energetically favorable. Knowledge-based functions are sums of many atom-pair contact contributions for protein and ligand atom type combinations. We have omitted this latter class of scoring functions from the present analysis for two reasons: first, their performance has proven to be rather unpredictable in a previous virtual screening study, [20] and second, failure cases cannot be explained easily and functions cannot be improved in a straightforward way. This is because the many individual atom-pair functions are difficult to interpret apart from donor–acceptor or lipophilic–lipophilic contact terms, which usually display the well-known distance dependence of these particular interactions.

Scoring functions that contain no directional (angular) terms and that have large distance cutoffs can be regarded as *soft* functions, because their values do not change abruptly with slight changes of ligand orientation and emphasize lipophilic contacts and general steric fit. Soft scoring functions are knowledge-based ones like PMF [21] and DrugScore, [22] but also the "piecewise linear potential" (PLP) [23] and the Gaussian shape fitting procedure by OpenEye. [24] The empirical functions ChemScore [25, 26] and the closely related FlexX scoring function [27] are "hard", because they contain angular terms for hydrogen bond interactions and emphasize these directed interactions more strongly. The Screen Score function was developed as a compromise between the hard, directed terms of the FlexX scoring function and the softer PLP potential. [20] Force fields also belong to the category of hard functions, because they naturally include not only attractive, but also repulsive interactions that lead to steeper potential surfaces. The distinction between hard and soft will become more apparent as applications of each of the functions mentioned are discussed.

The rough overview in the preceding paragraphs makes it clear that many options for combinations of docking algorithms and scoring functions are available. However, it is by no means clear under what circumstances a particular combination will fail or give good results. In the past, newly developed tools have rarely been tested on large and consistent test sets. Furthermore, even for the same algorithm, implementation details vary greatly and include many heuristic parameters [28] that are difficult to optimize. First comparative studies for virtual screening have given some insight into comparative performance. [29, 30] We have compared various scoring functions in combination with the docking program FlexX. [20] In the present contribution, we build on this study and evaluate two very recently developed docking programs Glide (Schrödinger, Inc.) and FRED (OpenEye Scientific Software) that are two different implementations of multiconformer docking. Results are compared to those previously obtained with the incremental construction program FlexX. [27, 31, 32, 33] We show that multiconformer docking and incremental construction algorithms are complementary to each other, and that either approach is especially powerful in combination with specific scoring functions and receptor characteristics. Thus, we consider this study as an important step towards general guidelines for setting up efficient virtual screening runs.

Materials and methods

Preparation of compound libraries

Inhibitors for the seven targets in Table 1 as well as the subset of the WDI database were taken from a previous publication. [20] For Glide docking studies, the inhibitors and the selected WDI subset were converted to mae format (Maestro, Schrödinger Inc.) and optimized by means of the MMFF94 force field. For docking studies with FRED as a docking engine, multiconformer libraries of the known inhibitors and the WDI subset in a binary format were produced by OMEGA (OpenEye Scientific Software). Modifications applied to the default settings of OMEGA were (i) rejection of conformers with an energy difference to the global minimum of >5.0 kcal mol^{-1} (GP_ENERGY_WINDOW), (ii) maximum number of output conformers 400 (GP_NUM_OUTPUT_CONFS) and (iii) low energy selection (no random selection) of conformers from the final ensemble (GP_SELECT_RANDOM false).

Preparation of protein target structures

Protein target structures were used as described in [20] and further modified to be used in FRED or Glide docking calculations. For

Table 1 Number and origin of active compounds used in this docking study [20]

Number of compounds	Target	Origin
128	Cyclooxygenase 2	[38, 39, 40]
55	Estrogen receptor	[41, 42, 43]
72	p38 MAP kinase	Roche, [44]
36	Gyrase B	Roche
67	Thrombin	[45, 46]
43	Gelatinase A and general MMP	WDI, PDB, [47]
51	Neuraminidase	PDB, Roche

Glide calculations, proteins and co-crystallized ligands were combined and hydrogen atoms were added within Maestro. The complexes were stored in the MacroModel dat format. The pprep script shipped by Schrödinger was used to check protonation states and tautomeric forms. All receptor–ligand complexes were minimized within MacroModel with the OPLS-AA force field by application of the autoref.pl script. Progressively weaker restraints (tethering force constants 3, 1, 0.3, 0.1) were applied to nonhydrogen atoms only. This refinement procedure is recommended by Schrödinger (technical notes for version 1.8), because Glide uses the full OPLS-AA force field at an intermediate docking stage and is claimed to be more sensitive towards geometric details than other docking tools. Minimizations were performed until the average root mean square deviation of the nonhydrogen atoms reached 0.3 Å. For FRED calculations, polar hydrogens were added by means of the interactive modelling program MOLOC [34] and were used structurally unchanged in further docking studies.

FlexX calculations were performed as described previously. [20] We include a short introduction to the FRED and Glide docking algorithms, because details about these tools have not been published yet.

FRED docking

FRED docking calculations were performed with FRED version 1.1 (OpenEye Scientific Software). For efficient handling of large compound databases, FRED distributes jobs via PVM [35, 36] over multiple processors. The first stage in docking is a shape fitting process, which takes a set of ligand conformers as input and tests them against a "bump map" (a Boolean grid with true values where ligand atoms can potentially be placed). Orientations that clash with the protein or are distant from the active site are rejected. The crude docking solutions are further tested against a pharmacophore feature if specified, and any poses that do not satisfy the pharmacophore are rejected. Poses surviving the shape fitting routine can then be passed through up to three scoring function filters in the screening process. Various options are available for optimization with respect to the built-in scoring functions: optimization of hydroxyl group rotamers, rigid body optimization, torsion optimization, and reduction of the number of poses that are passed on to the next scoring function. Available scoring functions in FRED are ChemScore, PLP, ScreenScore, and Gaussian shape fitting. The latter is a proprietary function of OpenEye. Qualitatively, the Gaussian scoring function has favorable values when the ligand and protein have a high surface contact and little volume overlap.

Glide docking

Glide calculations were performed with Impact version v18007 (Schrödinger, Inc.). The para_glide facility offers a way to break up databases into several segments that can be run on multiple processors. Schrödinger recommends the performance of test calculations with different scaling factors for the receptor and ligand atom van der Waals radii, because steric repulsive interactions might otherwise be overemphasized, leading to rejection of overall correct binding modes of active compounds. Grid calculations with Glide represent a time-consuming process and require on average 30–60 min CPU time on SGI R10k processors, depending on the box size to specify the active site and the scaling factors applied. We found that full van der Waals radii can be used for most targets, except COX-2, where optimal results were obtained with a scaling factor of 0.9 for receptor atoms and 0.8 for ligand atoms, p38 MAP kinase with 0.9 for receptor atoms, and gelatinase A with 0.9 for ligand atoms. In contrast to FRED, Glide generates conformations internally and passes these through a series of filters. The first places the ligand center at various grid positions of a 1 Å grid and rotates it around the three Euler angles. At this stage, crude score values and geometric filters weed out unlikely binding modes. The next filter stage involves a grid-based force field evaluation and refinement of docking solutions including torsional and rigid-body movements of the ligand. The OPLS-AA force field is used for this purpose. A small number of surviving docking solutions can then be subjected to a Monte Carlo procedure to try to minimize the energy score. The final energy evaluation is done with GlideScore, Schrödinger's implementation of the ChemScore function. This differs from ChemScore in slightly different weighting factors for each term and an additional steric repulsive term.

Schrödinger states that there is an effect of the ligand input geometry on poses generated by Glide. We have found that this is indeed the case to a great extent. Trial runs to reproduce the X-ray structures of the thrombin–NAPAP complex or the HIV protease–Saquinavir complex resulted in significantly different (rmsd values >2) rank 1 solutions with different low-energy conformers as starting conformations. This points to a lack of coverage of conformational space in the internal conformer generator. For our test cases, most of the known inhibitors were extracted from crystal structures or built by modeling ligands into binding sites, and therefore are already close to the native ligand. Thus, the GLIDE enrichment factors presented in this study might be overestimated and the enrichment factors given should be interpreted with caution.

Scoring functions

The principal scoring function used in Glide is called GlideScore, a modified version of ChemScore with slightly different weighting factors for each term and an additional steric repulsive term. ChemScore does not penalize mismatches between lipophilic and hydrophilic groups. To overcome this limitation, a penalty term was included in GlideScore that adds 3 kcal mol^{-1} to the score for each polar group buried in a lipophilic environment. A so-called composite scoring function (GlideComp) is also available. It is a weighted sum of the GlideScore (default weighting factor 0.6) and the OPLS-AA nonbonded energy (0.08). Schrödinger claims (Technical Notes for version 1.8), that this composite score in general yields better database enrichment factors than either GlideScore or Coulomb–vdW by itself. In addition, threshold values for individual score components can be used to filter out compounds with low receptor–ligand complementarity. The two built-in filters are (i) a "strength of interaction" filter, i.e. the OPLS-AA nonbonded energies (van der Waals and Coulomb terms) and (ii) a "specificity" filter, for which the hydrogen bonding and metal ligation energies from GlideScore are used. For re-scoring Glide docking solutions with ScreenScore, we used Glide rank 1 solutions and rescored with the FlexX implementation of ScreenScore.

Scoring functions implemented in FRED are Gaussian shape scoring, ChemScore, PLP and ScreenScore. The Gaussian shape function describes the shape of individual atoms as spherical Gaussian functions and returns favorable score values for a large surface contact between ligand and receptor and low volume overlap. ScreenScore was derived through a combination of PLP and FlexX terms. [20] The ScreenScore implementation in FRED does not include an angular term for metal contacts, and features an additional clash term that penalizes heavy atom clashes with less than 0.5 Å overlap by 1 kJ mol^{-1}, and more severe clashes by 10 kJ mol^{-1}. The overlap is calculated as the difference between the atom distances and the sum of the atom van der Waals radii.

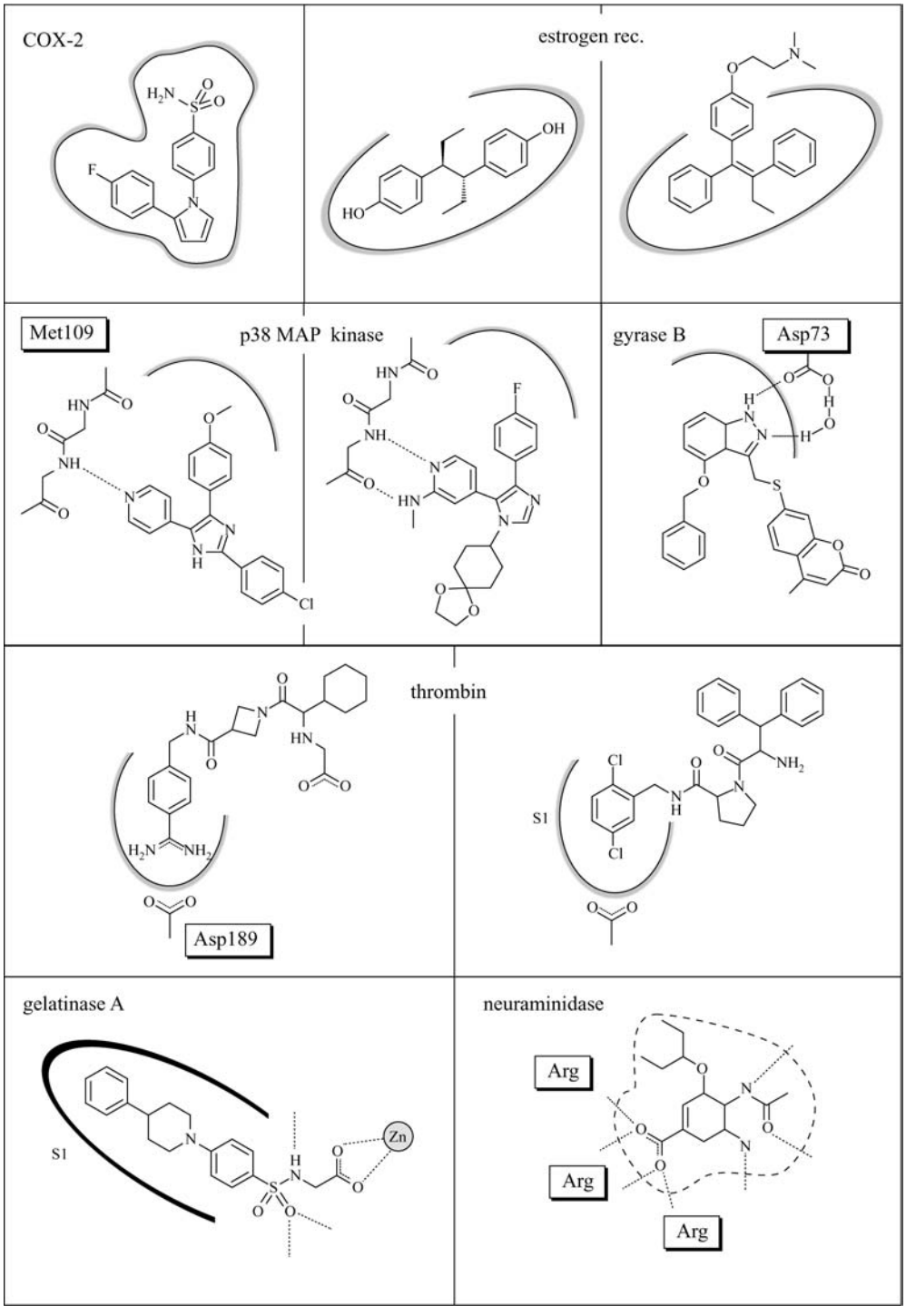

Fig. 1 Binding site characteristics and representative inhibitors for the different protein targets. The targets are shown in increasing order of binding site polarity. Explicitly drawn pocket outlines represent solvent-inaccessible parts of the binding cavities

Hardware and average run times

Preparation of compound libraries, receptor setups, grid calculations (Glide) and docking runs were performed on multiprocessor SGIs, either SGI R12k 400 MHz or SGI R10k 195 MHz. OMEGA is able to convert approximately 100,000 ligands per processor and per day to multiconformation libraries in binary format to be used in FRED docking. Further preprocessing of protein targets used in FRED docking is reduced to addition of polar hydrogen atoms. Grid calculations with Glide require on average 30–60 min CPU time on SGI R10k processors. Average CPU time consumptions on SGI R10k for docking calculations are about 2 min per ligand for FlexX, 6–7 min for Glide and 13.5 s for FRED (with all optimization flags in FRED activated).

Results and discussion

Figure 1 shows schematic binding site outlines and representative inhibitors of the protein targets investigated. For the sake of discussion, it is useful to group the seven targets into three classes according to the nature of their

binding sites. COX-2 and the estrogen receptor can be classified as buried, lipophilic binding sites, whereas the binding sites of the remaining five targets are more polar and also more solvent exposed. Neuraminidase and gelatinase A are the most polar of these. It will become apparent that this classification is reflected in the performance of particular combinations of docking tools and scoring functions.

The main body of virtual screening results are presented in Fig. 2. This figure gives an overview of the enrichment of known inhibitors obtained with various docking/scoring combinations for all seven test targets. For a detailed discussion, we need the distinction between the term "objective function" estimating the receptor–ligand interaction energy, which is used and minimized during the docking process, and the term "scoring function", denoting a function that is used for rank ordering of ligands relative to each other. The data in Fig. 2 were obtained with combinations of objective and scoring functions that are specific to each tool: in the case of FlexX, the objective function during the docking phase was the native FlexX function, for FRED calculations the Gaussian shape function was used in the docking phase, and for Glide GlideScore and the OPLS-AA force field serve as objective functions (for details see below and in the methods section). It should be noted that the FlexX scoring function is very closely related to ChemScore. For the seven targets, FlexX/ChemScore results (not shown) are only slightly worse than the FlexX/FlexXScore results depicted in column 1 of the seven panels in Fig. 2. Here we chose the native FlexX scoring function rather than ChemScore because its terms are adjusted to the details of the geometric interaction scheme used in FlexX.

For didactic reasons, the following analysis of these results deals with various aspects of the docking and scoring problem consecutively. Eventually, of course, it is the interplay of all factors – the choice of the docking algorithm, the definition of the active site, the objective and scoring functions employed, and the use of additional constraints – that influences the outcome of virtual screening runs. To illustrate individual factors, we will exemplify important findings with reference to further calculations.

Shape fitting versus incremental construction

For COX-2 and the estrogen receptor, the binding mode of ligands is determined by the overall shape of the binding pocket rather than by directed hydrogen bonding interactions. In these cases, the FRED multiconformer docking approach in conjunction with the Gaussian shape fitting as objective scoring function alone can lead to satisfactory enrichment of known inhibitors, superior to that obtained with FlexX (Fig. 3a). However, as soon as hydrogen bonding plays a role, pure shape fitting gives rather poor results (Fig. 3b). In contrast to multiconformer approaches, the incremental construction algorithms often fail for completely lipophilic binding sites, since there are no clear criteria for placing the initial ligand fragments and the search easily gets trapped in irrelevant regions of the binding pocket. For instance, irrespective of the choice of the scoring function, FlexX performs worse for COX-2 than FRED or Glide protocols.

The binding site of thrombin is an interesting intermediate case. It is the only other target, apart from COX-2, for which Gaussian shape fitting alone gives at least moderate enrichment. This is due to the narrow S1 pocket, for which the Gaussian shape function can effectively detect compounds with high complementarity. Furthermore, the thrombin test set of known inhibitors contains thrombin inhibitors with purely lipophilic S1 moieties rather than the usual positively charged donor groups (Fig. 1). On the other hand, the majority of the inhibitors do form hydrogen bonds with the Asp 189 side chain at the bottom of the S1 pocket. This residue is an ideal anchor point for incremental construction algorithms. The final scoring function, however, should be relatively soft, because otherwise nonhydrogen bonding inhibitors receive much lower ranks than the polar ones. This can be demonstrated for the example shown in Fig. 1, where ChemScore ranks this ligand with a purely lipophilic moiety in the S1 pocket on one of the last positions amongst known inhibitors and for the softer function ScreenScore the same ligand can be found among the top third of known inhibitors.

In contrast to COX-2 and the estrogen receptor, the binding sites of p38 MAP kinase, gyrase, thrombin, gelatinase A and neuraminidase are solvent exposed and hydrogen bonding is an essential element of inhibitor binding. In these cases, pure shape fitting often leads to the selection of the wrong binding mode, because hydrogen-bonded complexes often display less surface contact than nonhydrogen-bonded alternative poses and thus often do not receive favorable scores. Therefore, in the case of FRED, good performance is only obtained if not just one but all poses generated in the shape fitting routine are passed on to a screening process with a different scoring function. Enrichment drops considerably if only a fraction of the poses is passed on to the scoring function for all seven test targets (results not shown).

On the other hand, functional groups in the receptor site capable of forming hydrogen bonding interactions can serve as good anchor points for incremental construction algorithms. Once the initial fragment is correctly placed, the incremental construction process is very efficient, since the confines of the binding site force the algorithm to focus on relevant conformations only. Optimum performance of FlexX is observed for those binding sites that have very obvious anchor points (e.g. Asp 189 at the bottom of the S1 pocket of thrombin). The presence of too many polar groups in the binding site reduces performance again, due to the difficulty of discriminating between many alternative placements of initial fragments with equally good hydrogen bonding geometry.

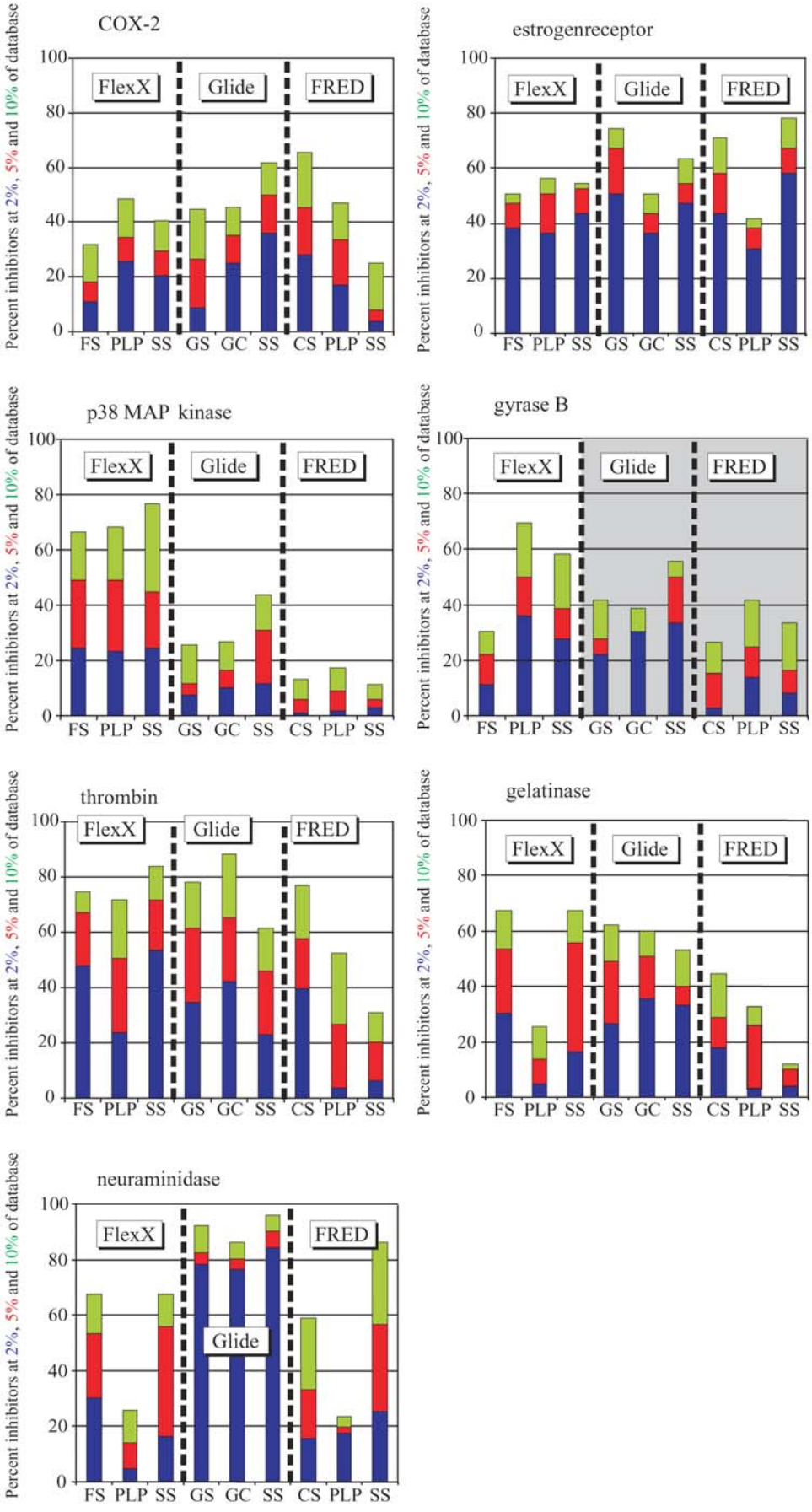

Fig. 2 Results obtained with FlexX, Glide and FRED in combination with different scoring functions for the seven investigated protein targets. Abbreviations for scoring functions are: FS, FlexXScore; SS, ScreenScore; GS, GlideScore; GC, GlideComp; CS, ChemScore; PLP, pairwise ligand–protein potentials. In the case of FlexX, the objective function during the docking phase was the native FlexX scoring function, for FRED calculations the Gaussian shape function was used in the docking phase, and for Glide GlideScore and OPLS-AA serve as objective functions (details see methods section). FRED and Glide results for gyrase B are shaded in *gray* since poses seem to be unreasonably predicted

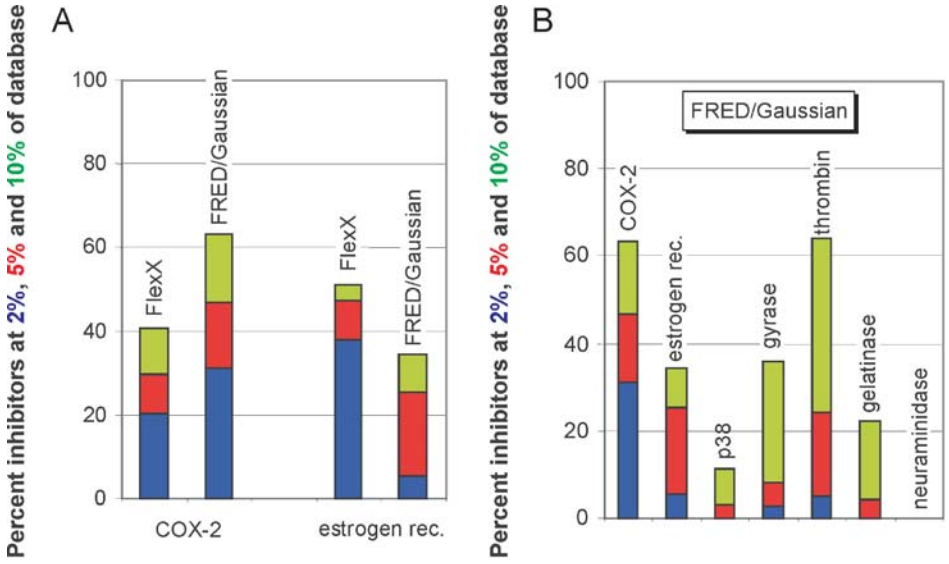

Fig. 3 **a** Results for COX-2 and the estrogen receptor gained by a pure shape fitting routine (FRED) in comparison to FlexX results. **b** Shape fitting results (FRED) for the seven investigated protein targets

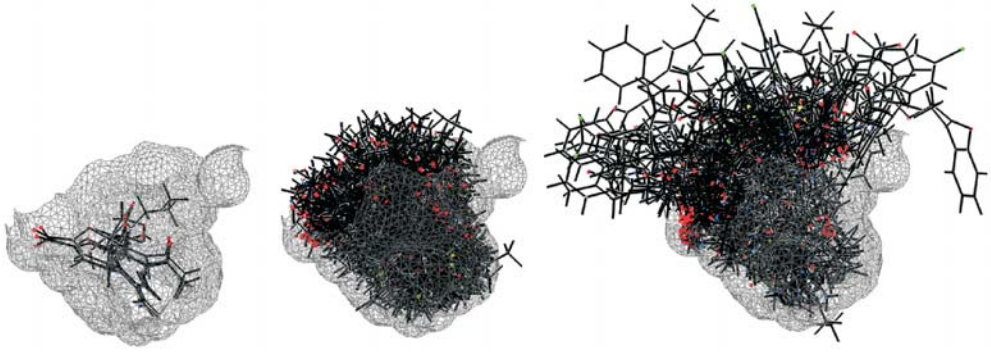

Fig. 4 The effect of different binding site definition as they are applied in FRED and FlexX. Active site surface of neuraminidase with correctly docked structures of two known ligands (left), 150 top ranking structures from the neuraminidase library docking run with FRED (center), 150 top ranking structures from the corresponding FlexX calculation

Binding site definition

The success of docking calculations depends crucially on the specification of a binding site to constrain the search. The smaller the binding site, the faster and the more reliable are calculations, because fewer alternative binding orientations can be generated that compete with one another. For both Glide and FRED calculations, the binding site is defined by a rectangular box oriented parallel to the main axes of the coordinate system, whereas for FlexX calculations one has to define a set of protein atoms that should be considered as being part of the binding site or "active". Curiously enough, much of the performance difference between FRED or Glide and FlexX can be attributed to this difference in binding site definition. The effectiveness of the two approaches in reducing the search space to relevant regions depends on the nature of the binding site. The following two examples illustrate this point.

The neuraminidase binding site is relatively shallow and solvent exposed. Due to the presence of many polar groups, FlexX finds many initial placements of base fragments with a good hydrogen bonding pattern for all those compounds that are capable of forming hydrogen bonds. The incremental construction process then allows "growing" many solutions out of the binding pocket into the solvent-exposed region (Fig. 4), since there are no conformational restrictions in this area. As a result, many solutions only partially occupy the binding site but still obtain high scores, because they form a number of geometrically correct hydrogen bonds. In contrast, for this solvent-exposed binding site, bounding boxes as employed by FRED or Glide act as efficient filters for removing poses whose center of mass (or some other definition of the molecule center) is located in the solvent region.

The ATP binding site of p38 MAP kinase is a good example for the opposite extreme. The binding site is a relatively narrow lipophilic cleft, which means that a bounding box cannot trim down the solution space significantly, and there is no easy way to orient the box such that it encloses only the relevant region of the cavity. Thus, it can happen that these docking tools explore irrelevant side pockets, especially if the box boundaries can only be adjusted in increments of 2 Å, as in Glide. In contrast, the "active atom" approach of FlexX has the advantage that few selected residues (in this case the hinge region with Met 109 at the center) can be chosen as the binding site, such that base fragments will be placed only there. This selection resembles the definition

of a 3D pharmacophore, which will be discussed further below.

Scoring functions and binding site characteristics

A target-specific selection of the docking algorithm can increase the performance in virtual screening. However, the choice of the docking algorithm cannot be made independently from the choice of the objective function. Furthermore, none of the currently available scoring functions would perform equally well in virtual screening for all types of receptor sites, even when only correctly docked poses of active compounds would have to be ranked relative to a random library. Thus, it is even more important to derive general guidelines for scoring than for docking algorithms.

We have seen that a multiconformer docking algorithm can function properly in conjunction with a very soft objective function during the docking phase (FRED/Gaussian shape fitting, Fig. 3a). The more narrow and lipophilic the target active site is, the more effective is the initial use of a soft scoring function. Using Glide, we have observed that scaling van der Waals radii of the ligand and receptor to 0.9 or 0.8, which essentially leads to a softer van der Waals potential, increases the enrichment rates for COX-2, whereas for all other six targets the full radii gave the best results.

In contrast, incremental construction algorithms rely on the presence of specific directed interactions, which must therefore also be taken into account in the objective function. Thus, for polar active sites that form a significant number of hydrogen bonds to the inhibitors, FlexX results are generally worse with PLP as the objective function than when PLP is used to re-score poses that have been generated with the harder FlexX or ChemScore functions. For highly lipophilic active sites, however, FlexX performs well with PLP as the objective function. This is exemplified in Fig. 5 for the targets COX-2 and p38 MAP kinase. For COX-2, the performance difference between using PLP as the objective function and using PLP as the scoring function after docking with FlexXScore is not very pronounced, but both procedures lead to significantly better results than docking and scoring with FlexXScore. For the p38 ATP binding site, where the formation of specific hydrogen bonds plays a major role in receptor–ligand binding, using PLP as the objective function drastically decreases inhibitor enrichment. The ATP binding site of gyrase B is even more lipophilic, and also more shallow and solvent exposed than the p38 MAP kinase ATP binding site. Again, PLP is the best function for scoring poses generated by FlexX. It should be noted that, for the gyrase B target, Glide and FRED suffer from the active site definition problem discussed above for p38 MAP kinase, although the enrichment, especially with Glide, seems to be better than for p38 MAP kinase. Visual inspection of the highest ranked poses generated by FRED and Glide

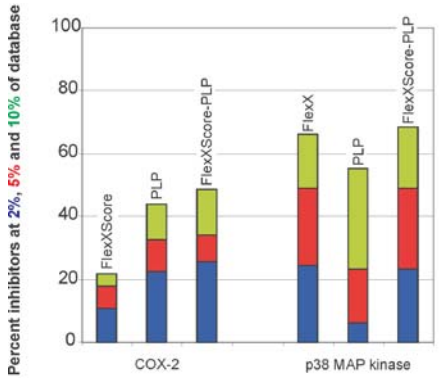

Fig. 5 Results for COX-2 and p38 MAP kinase to demonstrate that when using PLP for re-scoring rather than as the objective function results are improved

for gyrase B, however, showed that most of the inhibitor binding modes generated were not in accordance with experiment. This is the only case in our test suite where considerable enrichment was obtained with obviously wrong binding modes. The Glide and FRED results for gyrase B should therefore not be compared to the FlexX results, where correct inhibitor binding modes were generated.

Using a soft objective function, followed by optimization with restrictive repulsive and angular terms seems to be a good general strategy for multiconformer docking. In this context, one particular feature of Glide should be highlighted. In Glide, intermediate force field filtering and minimization are always executed after the initial placement of the conformers and before the final scoring with GlideScore. The OPLS-AA nonbonded terms are used to describe the protein–ligand interactions. Because the nonbonded interactions naturally include electrostatic as well as steric (van der Waals) repulsive interactions, this procedure can help to weed out many mismatching docking solutions that other scoring functions could not detect, because they do not contain the necessary penalty terms. The force field filtering step is especially efficient for targets with many polar functional groups. For this reason, Glide performs particularly well for the two most polar targets neuraminidase and gelatinase A. Especially impressive are the enrichment rates for neuraminidase, where the majority of active compounds are contained within the top 2 percentiles of the ranked database. In Glide, poses can optionally be re-scored with the so-called GlideComp function, a weighted sum of the GlideScore and OPLS nonbonded terms. This function should be more sensitive towards electrostatic mismatches than GlideScore itself, but otherwise have similar properties to GlideScore, its main component. Interestingly, GlideComp performs slightly better than GlideScore for all targets (except the estrogen receptor), especially in enriching inhibitors in the top 2 percentiles. This points to the fact that penalty terms can play an important role in recognizing false positives that form many favorable interactions but in addition display re-

pulsive interactions that render the computed binding mode unlikely to occur.

We have generally observed that FRED virtual screening runs are more effective if all poses are fully optimized with each scoring function and not simply re-scored. This optimization includes rotation around all torsional angles including those of terminal OH groups as well as solid body movements. However, not all scoring functions are equally suited for flexible optimization. ScreenScore was derived without flexible optimization in mind, and so far has been used together with only the PLP repulsive term in FlexX. The FRED ScreenScore implementation adds the repulsive part of the FlexX contact ("lipo" and "ambig") term to this. Obviously, this does not lead to a sufficient balance between attractive and repulsive terms. Optimization with PLP leads to clash-free poses and generally better enrichment than rigid re-scoring, whereas poses optimized with ScreenScore often display steric clashes between receptor and ligand, which are obviously outweighed by attractive polar interactions. Especially for narrow binding sites, this leads to serious deficiencies in virtual screening. Therefore, FRED/ScreenScore results are particularly poor for COX-2 and thrombin (in the case of Gelatinase A, enrichment is extremely low because of the missing angular dependence of the metal interaction term in the current FRED ScreenScore implementation). However, if ScreenScore is used to score poses generated by FlexX or Glide, results are generally very satisfactory. In both cases, no optimization is performed and the poses are already relatively clash free (especially in the case of Glide due to the force field optimization step).

Score components as filters

Schrödinger recommends using both the nonbonded OPLS-AA energies and the hydrogen bonding energies from GlideScore separately as computational filters to weed out poses with low "strength of interaction" and "specificity", respectively. This is done by requiring that each of these components should be higher than a user-defined threshold value for each pose. The use of such filters is of course a highly subjective step in a virtual screening procedure, and it is useful only if one is aware of its consequences. We have not used these filters here, since for most of our targets' meaningful threshold values would have led to the exclusion of a significant fraction of the active compounds. It is our impression that setting threshold values for hydrogen bonding contributions is a less powerful tool for focusing a virtual screen than using direct pharmacophore constraints, as discussed in the next section. The nonbonded energies depend strongly on the scaling of van der Waals radii for nonpolar atoms and are therefore also difficult to use with a specific threshold for filtering. The score component filters will thus be useful only for well-established virtual screening protocols where each of the heuristic search parameters has been refined.

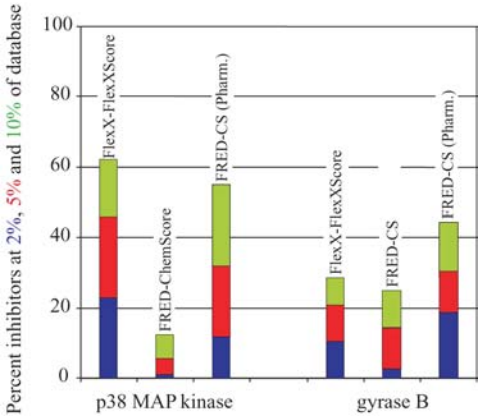

Fig. 6 Enrichment of inhibitors for p38 MAP kinase and gyrase B. Objective scoring functions are FlexXScore for FlexX, and Gaussian for FRED. Inclusion of a pharmacophore constraint (definition see text) in the FRED calculations led to similar results to those obtained with FlexX. In Glide, pharmacophore constraints cannot be defined

Docking under pharmacophore constraints

With the programs FRED and FlexX, docking can optionally be performed under pharmacophore constraints. This feature is especially useful for the inclusion of hydrogen bonding constraints in multiconformer docking. We have seen above that it is difficult to focus the calculations of multiconformer docking runs on the essential regions of the p38 MAP kinase and gyrase B ATP binding sites. For p38 MAP kinase, the pharmacophore was defined as a sphere with 0.5 Å radius centered at the position of the acceptor nitrogen atom of an inhibitor complexed with p38 MAP kinase (PDB entry 1bmk). A nitrogen or oxygen acceptor atom was required to be placed within this sphere in hydrogen bonding distance to the NH of Met109 in p38 MAP kinase. Enrichment factors gained in this way are close to those obtained with FlexX (Fig. 6). For gyrase B, a single pharmacophore sphere was used to enforce hydrogen bonding to a conserved water molecule (Fig. 1), which again led to much higher enrichment. Even though pharmacophore constraints can be defined at arbitrary points in space and with arbitrary radii, they are only very roughly observed during the calculation. The actual spatial extension of the volume segment occupied by ligand atoms passing the pharmacophore depends on the resolution of the grid used for shape fitting as well as on the proximity of the pharmacophore point to the nearest grid point. Nevertheless, the results we have obtained with single point pharmacophore constraints in FRED are quite encouraging.

CPU time consumption

Table 2 gives an overview on CPU time consumption for docking with FRED and Glide. For FlexX calculations on parallel 16 SGI R12k 400-MHz processors an average

Table 2 Overview on CPU time consumption in FRED and Glide docking procedures. These benchmark calculations were performed on SGI R10K processors. FRED calculations include a single screen using ChemScore as the scoring function and full optimization

Receptor	Average no. of rotors	Average no. of heavy atoms	FRED average docking time (s)	Glide average docking time (s)
COX-2	4.1	24.7	5	134
Estrogen rec.	4.2	26.7	15	290
p38 MAP kinase	4.7	26.6	9	133
Gyrase B	5.6	27.5	13	144
Thrombin	9.7	32.2	15	562
gel-A	9.9	30.6	13	513
Neuraminidase	6.5	21.1	8	207
WDI subset	5.6	24.2	13	400

wall clock run time of 52 s/molecule can be assumed [19]. The benchmark FRED results were run with a single re-scoring process with ChemScore including full optimization (ligand hydroxyl groups, solid body and torsion angles). FRED run times of the docking process mainly depend on two factors: (i) the shape properties of the active site, which significantly affect the performance of the shape fitting routine, and (ii) the number of conformers generated by OMEGA. For instance, the shape fitting routine is very efficient for the very narrow and confined COX-2 binding site, and less so for the larger estrogen binding site. The majority of the CPU time is spent in re-scoring and optimization processes. However, even with several subsequent re-scoring steps including full optimization, FRED is about an order of magnitude faster than FlexX. With respect to run times, Glide is competitive neither to FRED nor to FlexX. As can be seen from Table 2, the average docking time per ligand is several minutes. Glide docking time significantly increases if the average number of rotors per ligand increases. Schrödinger claim an average docking time of a library with compounds containing 0–20 rotatable bonds of approximately 360 s (SGI R10K) per ligand. This is in rough accordance to the approximately 400 s (SGI R10K) we have measured per ligand for the WDI subset.

Summary and conclusion

We have described structure-based virtual screening experiments for seven different targets that differ significantly in the characteristics of their binding sites. The binding sites can roughly be grouped into three different classes: lipophilic buried cavities (COX-2, estrogen receptor), targets of intermediate polarity with hydrogen bonding motifs common to the majority of inhibitors (p38 MAP kinase, gyrase B, thrombin) and targets with very polar, solvent-exposed binding sites (neuraminidase, gelatinase A). The calculations were performed with a variety of different objective and scoring functions in combination with three fast, state-of-the-art docking programs. The results show clearly that the performance of docking algorithms, objective functions and scoring functions strongly depends on characteristics of the target structure. We have arrived at a number of general guidelines that should help to direct the selection of the best combination for a particular virtual screening problem:

- For lipophilic binding sites where general steric fit of ligands outweighs the importance of hydrogen bonding, the method of choice is multiconformer docking in combination with a soft objective function and a harder scoring function.
- Binding sites that are predominantly lipophilic but feature polar groups that must form specific interactions to ligands are best dealt with by incremental construction algorithms.
- For polar binding sites whose ligands have to form networks of directed interactions, one has the choice between incremental construction docking with hard scoring functions or multiconformer docking with a hard scoring function at least as an intermediate filtering step.
- Incremental construction algorithms require a hard objective function. If lipophilic interactions are of particularly great importance, a softer function can be used for scoring.
- Overall, ChemScore (or the related GlideScore) seems to be the most generally applicable and robust scoring functions to be used in combination with multiconformer docking.
- Virtual screening runs are the more successful the more narrowly and focused the search constraints are defined. Consequently, such details as the definition of the binding site boundaries are critical. In the present study, this is done either by means of a rectangular box (FRED, Glide) or by a set of receptor atoms (FlexX). It would be desirable to have both options available in all tools. The inclusion of additional pharmacophore constraints is possible in FRED or FlexX. [37] This is helpful to incorporate previously gained knowledge on ligand binding into virtual screening runs.

We have found that FRED as a docking engine in combination with ChemScore as a scoring function is a good general method for structure-based virtual screening. Considering its high speed, FRED is certainly an especially attractive tool. Where the performance of FRED is suboptimal, FlexX has specific strengths, so that these two tools complement each other well. Glide performs especially well where the intermediate force field optimization step is necessary to filter out electrostatically

mismatching poses, but is relatively slow compared to the two other tools. For Glide, one should also keep the enormous influence of the ligand input geometry on the poses generated in mind, and that enrichment factors given in this manuscript should be interpreted really carefully.

To sum up, we believe that current docking tools are mature enough for routine applications in the pharmaceutical industry. As long as they are not used as black boxes, but with knowledge about the underlying algorithms and heuristic assumptions, they provide a good basis for rational compound selection.

Acknowledgements The authors thank Tom Halgren from Schrödinger Inc. for providing Glide evaluation licenses and many discussions about this software, Jörg Weiser and Gerd Räther from Anterio Consult & Research for support, Anthony Nicholls, Matt Stahl and Mark McGann from OpenEye Software for providing FRED and OMEGA evaluation licenses and the implementation of ScreenScore into FRED. Matthias Rarey is thanked for his continuing support of FlexX and many fruitful discussions. We thank our colleagues at Roche Basel and in the Roche biostructure community for supporting our work.

References

1. Bailey D, Brown D (2001) Drug Discovery Today 6:57–59
2. Walters WP, Stahl MT, Murcko MA (1998) Drug Discovery Today 3:160–178
3. Gane PJ, Dean PM (2000) Curr Opin Struct Biol 10:401–404
4. Good A (2001) Curr Opin Drug Disc Dev 4:301–307
5. Dixon JS, Blaney JM (1998) Docking—predicting the structure and binding affinity of ligand–receptor complexes in designing bioactive molecules. In: Martin YC, Willet P (eds) American Chemical Society, Washington, D.C., pp 175–197
6. Stahl M (2000) Structure-based library design in virtual screening for bioactive molecules. In: Schneider G, Boehm H-J (eds) Virtual screening for bioactive molecules. VCH, Weinheim, pp 229–259
7. Muegge I, Rarey M (2001) Small molecule docking and scoring. In Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry. VCH, New York, p 1
8. Burkhard P, Hommel U, Sanner M, Walkinshaw MD (1999) J Mol Biol 287:853–858
9. Filikov AV, Monan V, Vickers TA, Griffey RH, Cook PD, Abagyan RA, James TL (2000) J Comput-Aided Mol Design 14:593–610
10. Jordan DB, Basarab GS, Liao D-I, Johnson WMP, Winzenberg KN, Winkler DA (2001) J Mol Graphics Mod 19:434–447
11. Liebeschuetz JW, Jones SD, Morgan PJ, Murray CW, Rimmer AD, Roscoe JME, Waszkowycz B, Welsh PM, Wylie WA, Young SC, Martin H, Mahler J, Brady L, Wilkinson K (2002) J Med Chem 45:1221–1232
12. Ajay Murcko MA (1995) J Med Chem 38:4953–4967
13. Tame JRH (1999) J Comput–Aided Mol Design 13:99–108
14. Boehm H-J, Stahl M (1999) Med Chem Res 9:445–462
15. Gohlke H, Klebe G (2001) Curr Opin Struct Biol 11:231–235
16. Vieth M, Hirst JD, Dominy BN, Daigler H, Brooks III CL (1998) J Comput Chem 14:1623–1631
17. Abagyan R, Trotov M, Kuznetsov DJ (1994) Comput Chem 15:488–506
18. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) J Comput–Aided Mol Design 10:293–304
19. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
20. Stahl M, Rarey M (2001) J Med Chem 44:1035–1042
21. Muegge I, Martin YC (1999) J Med Chem 42:791–804
22. Gohlke H, Hendlich M, Klebe G (2000) J Mol Biol 295:337–356
23. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB, Fogel LJ, Freer ST (1995) Chem Biol 2:317–324
24. F. Brown, A. Nicholls, H. Almond, M. McGann (manuscript submitted)
25. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) J Comput-Aided Mol Design 11:425–445
26. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD (1998) Proteins 33:367–382
27. Rarey M, Wefing S, Lengauer T (1996) J Comput-Aided Mol Design 10:41–54
28. Salo J–P, Yliniemelä A, Taskinen J (1998) J Chem Inf Comput Sci 38:832–839
29. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) J Med Chem 42:5100–5109
30. Bissantz C, Folkers G, Rognan D (2000) J Med Chem 43:4759–4767
31. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261:470–489
32. Rarey M, Kramer B, Lengauer T (1997) J Comput-Aided Mol Design 11:369–384
33. Rarey M, Kramer B, Lengauer T (1999) Bioinformatics 15:243–250
34. Moloc (2002) F Hoffmann-La Roche/Gerber Molecular Design
35. Sunderarm V, Dongarra J, Geist A, Manchek R (1994) Parallel Computing 20:531–547
36. http://www.epm.ornl.gov/pvm/pvm_home.html
37. Hindle SA, Rarey M, Buning C, Lengauer T (2002) J Comput-Aided Mol Design (in press)
38. Carter JS (1997) Exp Opin Ther Patents 8:21–29
39. Friesen RW, Brideau C, Chan CC, Charleson S, Deschenes D, Dubé D, Ethier D, Fortin R, Gauthier JY, Girard Y, Gordon R, Greig GM, Riendau D, Savoie C, Wang Z, Wong E, Visco D, Xu LJ, Young RN (1998) Bioorg Med Chem Lett 8:2777–2782
40. Kalgutkar AS (1999) Exp Opin Ther Patents 9:831–849
41. http://www.protherics.com/
42. Magarian RA, Overacre LB, Singh S, Meyer KL (1994) Curr Med Chem 1:61–104
43. Fink BE, Mortensen DS, Stauffer SR, Zachary DA, Katzenellenbogen JA (1999) Chem Biol 6:205–219
44. Hanson GJ (1997) Exp Opin Ther Patents 7:729–733
45. Wiley MR, Fisher MJ (1997) Exp Opin Ther Patents 7:1265–1282
46. Sanderson PEJ, Nayor–Olsen AM (1998) Curr Med Chem 5:289–304
47. Beckett RP, Whittaker M (1998) Exp Opin Ther Patents 8:259–282