Structure-Based Drug Design: Docking and Scoring

Romano T. Kroemer[1,2,*]

[1]*Computational Sciences, Department of Chemistry, Nerviano Medical Sciences, Viale Pasteur 10, 20014 Nerviano (MI), Italy;* [2]*New address: Chemical Sciences, Sanofi-Aventis, 13 Quai Jules Guesde, 94403 Vitry-sur-Seine, France*

Abstract: This review gives an introduction into ligand – receptor docking and illustrates the basic underlying concepts. An overview of different approaches and algorithms is provided. Although the application of docking and scoring has led to some remarkable successes, there are still some major challenges ahead, which are outlined here as well. Approaches to address some of these challenges and the latest developments in the area are presented. Some aspects of the assessment of docking program performance are discussed. A number of successful applications of structure-based virtual screening are described.

INTRODUCTION

The need for a rapid search for small molecules that may bind to targets of biological interest is of crucial importance in the drug discovery process. One way of achieving this is the *in silico* or virtual screening (VS) of large compound collections to identify a subset of compounds that contains relatively many hits against the target, compared to a random selection from the collection. The compounds that are virtually screened can stem from corporate or commercial compound collections, or from virtual compound libraries. If a three-dimensional (3D) structure or model of the target is available, a commonly used technique is structure-based virtual screening (SBVS) [1]. Here a so-called ‘docking program’ is used to place computer-generated representations of a small molecule into a target structure (or in a user-defined part thereof, e.g., the active site of an enzyme) in a variety of positions, conformations and orientations. Each such docking mode is called a ‘pose’, Fig. (**1**). In order to identify the energetically most favorable pose (also referred to as ‘pose prediction’), each pose is evaluated (‘scored’) based on its complementarity to the target in terms of shape and properties such as electrostatics. A good score for a given molecule indicates that it is potentially a good binder. This process is repeated for all molecules in the collection, which are subsequently rank-ordered by their scores (i.e., their predicted affinities). This rank-ordered list is then used to select for purchase, synthesis, or biological investigation only those compounds that are predicted to be most active. Assuming that both the poses and the associated affinity scores have been predicted with reasonable accuracy, this selection will contain a relatively large proportion of active molecules, i.e., it will be ‘enriched’ with actives compared to a random selection.

High-throughput docking has become increasingly important in the context of drug discovery [2–4]. Despite the technical challenges in reliably predicting the binding mode of a molecule [5] and its binding affinity relative to other compounds [6], in many cases docking campaigns have yielded significant hit rate improvements compared to random screening [7-10]. For a number of reasons even in the age of “real” high-throughput screening (HTS) there is a place for VS: External compound collections can easily be virtually screened and only compounds that are predicted to inhibit the target will then be acquired. Virtual libraries can be screened *in silico* and the results can be used to select scaffolds and to help design the final library to be synthesized. When no HTS is envisioned or has been carried out yet, a discovery effort can be jump-started by VS of corporate and/or public compound collections. When no biochemical or other functional assay is available, VS may be the only way of identifying inhibitors of a specific target. Compared to HTS, VS is fast and inexpensive. Also, it is conceivable that VS is complementary to HTS, i.e. compounds that are falsely not detected as active (false negatives) in first-round screening of an HTS campaign can be highlighted by their docking scores and therefore be re-tested in the confirmation rounds of a given screening campaign.

BASIC REQUIREMENT FOR DOCKING: 3D ATOMISTIC REPRESENTATION OF THE RECEPTOR

In order to perform a structure-based virtual screening exercise it is necessary to have the receptor structure(s) of interest at hand. Most commonly the structure of the receptor has been determined by experimental techniques such as X-ray crystallography or NMR. For proteins, if the structure is not available, one can resort to the techniques of protein-structure prediction [11]. Most commonly applied are ‘threading’ and homology modeling [12, 13]. During threading – or fold recognition – an assessment is made whether a given amino acid sequence is compatible with one of the structures in a database of known folds. Homology – or comparative – modeling relies on a clear relationship or homology between the sequence of the target protein and at least one known structure.

Be it from experiment or prediction, once the three-dimensional structure of the protein of interest has been obtained, it can be analyzed using a variety of computational

*Address correspondence to this author at the Chemical Sciences, Sanofi-Aventis, 13 Quai Jules Guesde, 94403 Vitry-sur-Seine, France; Tel: +33 (0)1 5893 2028; Fax : +33 (0)1 5893 8063;
E-mail: romano.kroemer@sanofi-aventis.com,

1389-2037/07 $50.00+.00

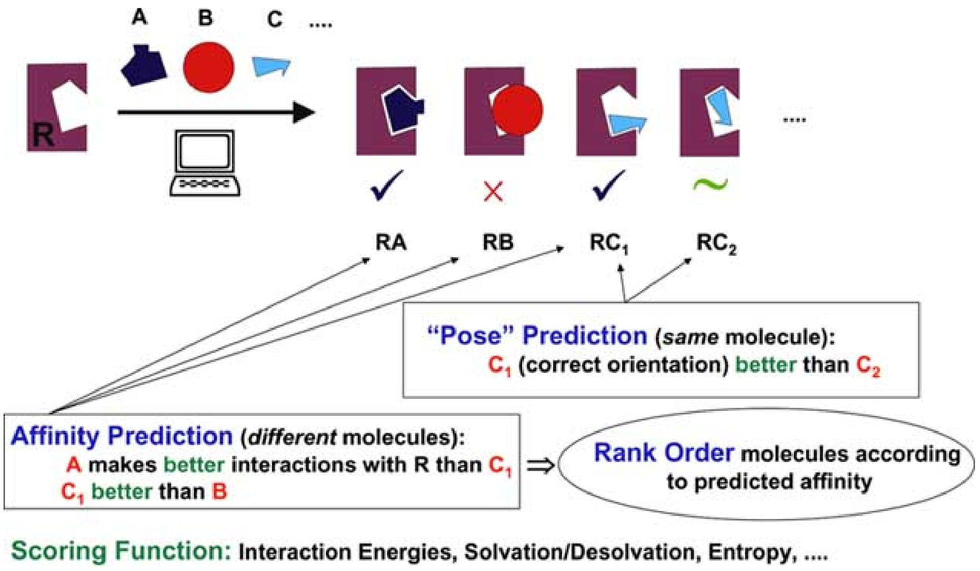

A
B
C ....
R
....
RA
RB
RC1
RC2
"Pose" Prediction (same molecule):
C1 (correct orientation) better than C2
Affinity Prediction (different molecules):
A makes better interactions with R than C1
C1 better than B
⇒
Rank Order molecules according to predicted affinity
Scoring Function: Interaction Energies, Solvation/Desolvation, Entropy, ...

Fig. (1). Illustration of docking and scoring. R symbolizes a receptor structure, A, B and C represent small molecules to be docked into the receptor.

techniques. If the function of the protein is unknown, it may be important to search its structure for putative binding sites [14]. These binding sites can subsequently be explored for the binding of selected molecules, or they can be compared with other, known, binding sites. In many cases the binding site (and the function of the protein) is known by reference (e.g. the protein can be assigned to a protein family with known function), or the protein has been co-crystallized with a ligand. An analysis of the binding-site characteristics and/or the interactions with a given ligand can lead to important insights for the design of novel ligands or the docking of putative ligand molecules [15, 16].

SEARCHING AND POSE PREDICTION

Searching for the correct binding mode (pose prediction) of a molecule is typically carried out by performing a number of trials and keeping those poses that are energetically best. It involves finding the correct orientation and, as most ligand molecules are flexible, the correct conformation of the docked molecule. This implies that the degrees of freedom to be searched include translational and rotational degrees of freedom of the ligand as a whole, as well as its internal degrees of freedom, i.e., predominantly the rotatable bonds. The search stops once a certain number of trials have been carried out and/or a sufficient number of poses have been found for a molecule. In order to explore a large search space, algorithms have been developed that keep track of previously discovered minima and guide the search into new regions [17-19]. The decision to keep a trial pose is based on the computed ligand–receptor interaction energy (score) of that pose. To identify and rank-order many different poses of a molecule during the search in a reasonable time, several programs calculate a 'dock score' (a crude score based on a simple energy function such as a force field with an electrostatic term and repulsive and attractive Van-der-Waals terms), which can be evaluated very rapidly during the docking process, while a more sophisticated function is used to calculate the final 'affinity score' for that molecule.

SCORING OR AFFINITY PREDICTION

Affinity scoring functions are then applied to the energetically best pose or *n* best poses found for each molecule, and comparing the affinity scores for different molecules gives their relative rank-ordering. The implicit assumption is that for a given molecule the best pose according to the affinity score is among the *n* saved poses identified with the dock score. For comprehensive overviews of various scoring schemes used to predict binding affinity, see references [20] and [21].

Many of the scoring functions fall into one of two main groups. One main group comprises **knowledge-based scoring functions** that are derived using statistics for the observed interatomic contact frequencies and/or distances in a large database of crystal structures of protein–ligand complexes. It can be assumed that only those molecular interactions that are close to the frequency maxima of the interactions in the database favor the binding event and therefore increase the overall binding affinity, whereas interactions that have been found to occur with low frequency in the database are likely to destabilize binding and decrease the affinity. The observed frequency distributions are converted to what is usually referred to as potentials of mean force or knowledge-based potentials. Several such potentials to predict binding affinity have been developed (e.g., PMF [22], DrugScore [23], SmoG [24], Bleep [25]). All these approaches differ mainly in the size of the training database that was employed and in the types of molecular interaction that were considered.

The other main group contains scoring schemes based on physical interaction terms [26]. These so-called **energy component methods** are based on the assumption that the change in free energy upon binding of a ligand to its target can be decomposed into a sum of individual contributions:

$$\Delta G_{bind} = \Delta G_{int} + \Delta G_{solv} + \Delta G_{conf} + \Delta G_{motion}$$

The individual terms in this equation account for the main energetic contributions to the binding event, as follows: specific ligand–receptor interactions (ΔG_{int}), the interactions of ligand and receptor with solvent (ΔG_{solv}), the conformational changes in the ligand and the receptor (ΔG_{conf}) and the motions in the protein and the ligand during the complex formation (ΔG_{motion}). In principle, a separation into individual terms is only possible if the system of interest is divided into mutually independent variables [27]. However, many of the individual terms are highly correlated with each other and they can affect the binding affinity in more than one way (i.e., positive or negative contribution) [28]. Moreover, the free-energy contributions are not calculated as ensemble mean values, but are usually computed from a single structure. Also, the assumption of additivity in biochemical processes is not strictly valid [29]. Despite these approximations, energy component methods are very appealing as the simplifications result in functions that can be evaluated very rapidly, which is important in a high-throughput docking setting. More importantly, they have also been successfully applied to the prediction of protein–ligand affinity [30-32]. Two classes of function can be defined within this group of energy-component methods: in first-principle scoring functions the terms are directly derived from physicochemical (statistical mechanics) theory and are not fitted to experimental data [33, 34]. The other class comprises empirical or regression-based methods, which assume an often linear statistical relationship between the total free-energy change upon binding (i.e., the binding affinity) and a number of terms that characterize the binding event. Most frequently, these terms include descriptors for hydrogen bonds and ion pairs, the amount of buried and contact surface, and molecular flexibility of the ligands. A training set of crystal structures of ligand-protein complexes and associated binding affinity data is used to optimize the coefficients of the regression equation. Many popular scoring functions have been derived this way (e.g., LUDI [35], ChemScore [36], Validate [37], GOLD score [38, 39], PLP [40, 41], FlexX score [42], ScreenScore [43], Autodock3 [44, 45]). Various approaches to derive optimal coefficients for regression-based scoring functions exist. Most of them aim to reproduce experimental binding affinities. However, if the only goal is to classify molecules (i.e., to distinguish binders from nonbinders) and one does not mind sacrificing the correct rank-ordering within the group of binding molecules, optimizing the scoring function by maximizing the score gap between binders and nonbinders is appealing [46].

DOCKING PROGRAMS AND THEIR UNDERLYING ALGORITHMS

A large number of docking programs and search algorithms have been published. One criterion for classifying the underlying algorithms is the way the ligands are treated during docking. In some of these algorithms the ligand is built up incrementally, starting from a docked 'base fragment'. Programs that follow this approach include Hammerhead [47], DOCK [48-50], and FlexX [51]. In other programs, such as AutoDock [44, 52], Genetic Optimization for Ligand Docking (GOLD) [53], ICM-Dock [54, 55] and QXP [56], the ligand is treated in its entirety.

In addition to ligand flexibility, it may be desirable to keep at least part of the receptor flexible in order to allow for conformational changes that are necessary to accommodate the ligand, a phenomenon referred to as 'induced fit.' Because it is computationally expensive, few docking programs allow protein flexibility. Notable exceptions are the latest versions of AutoDock [52], FlexE [57], QXP [56], Affinity [58] and the latest version of ICM-Dock [54, 55]. The way flexibility is handled differs from program to program. For example, FlexE uses multiple receptor conformations, Affinity allows any selection of atoms to be mobile with a user-defined tethered buffer region between the fixed and mobile regions, and QXP allows user-defined parts of the protein, e.g., selected side chains or a particular loop, to move.

Another criterion to classify docking programs would be according to the search strategy employed. Roughly speaking one could distinguish between programs trying to maximize shape complementarity – often based on geometric criteria – and programs incorporating an energy-driven or stochastic algorithm. Well known representatives of the former group are DOCK [48-50], FlexX [51] and FRED [59, 60]. Among the latter group programs such as AutoDock [44, 52], ICM-Dock [54, 55], QXP [56] and GOLD [53] can be found.

A list of popular docking programs is given in Table 1. To illustrate some of the radically different approaches incorporated in some docking programs, the algorithms and scoring functions implemented in three of them (FlexX, GOLD, and QXP) are described in some detail below. It should be noted that this does not imply that they are better than other docking programs that are available.

FlexX [51] uses an incremental buildup algorithm where ligands are docked starting with a base fragment. Base fragments are generated by severing all noncyclic bonds in a given ligand. All base fragments identified for a given ligand serve as starting points for the docking. After placement of a base fragment (in different positions) the complete ligand is constructed by adding the remaining components back on. Each component is added in accordance with a set of predefined, allowed torsion angles, thus allowing for ligand flexibility. At each step the interactions are evaluated and the best solution is selected according to the docking score. The docking score uses the model of molecular interactions developed by Böhm [69, 70] and Klebe [71]. For each moiety that can make an interaction, interaction centers and surfaces (usually spheres) are defined. Two moieties interact if the interaction center of one of them is situated at or near the interaction surface of the other one. Different levels of interaction are defined and the program attempts to satisfy high-level interactions (such as hydrogen bonds) first. Subsequently, the docked results are scored using a modified version [72] of the Böhm scoring function [73]. This function takes into account the loss of ligand entropy upon binding (counting the number of rotatable bonds in the ligand), hydrogen bonds, ionic interactions, aromatic interactions, and lipophilic contacts. For more details, see references [51] and [72].

GOLD is based on a genetic algorithm (GA) [53, 74], that mimics the process of evolution by applying genetic operators to a collection of putative poses for a given ligand

Table 1. Selection of Popular Docking Programs

Program	ALGORITHM	REFERENCES
AutoDock	Lamarckian GA	[44, 52]
DOCK	Shape matching (sphere images)	[48-50]
DOCK (NWU version)	Shape matching (sphere images)	[61, 62]
FlexX	Incremental construction	[51]
FRED	Shape matching (gaussian functions)	[59, 60]
Glide	Descriptor matching/MC	[63-65]
GOLD	GA	[53]
Hammerhead	Incremental construction	[47]
ICM	MC minimization	[54, 55]
LigandFit	Shape matching (moments of inertia)	[66]
QXP	MC minimization, tree searching and pruning	[56]
SLIDE	Descriptor matching	[67]
Surflex Dock	Surface-based molecular similarity	[68]

(in GA terms, a population of chromosomes). GOLD chromosomes contain conformational information of the flexible parts of the protein (OH of Ser, Thr and Tyr as well as lysine NH_3^+) and of the ligand, as well as hydrogen bonds and lipophilic interactions. Chromosome decoding yields the corresponding 3D pose for the ligand, which is followed by a least-square (LS) fitting procedure [75], with the objective to maximize the overlap between ligand and receptor features. The energy of the resulting pose (fitness) consists of three terms: (1) hydrogen-bonding energy, (2) internal energy of the ligand, and (3) steric interaction energy.

The population of chromosomes evolves through sequential application of genetic operations. Newly generated chromosomes are decoded, the fitness of the corresponding pose evaluated, and the chromosome is kept if it is fitter than the least-fit chromosome in the pool. After the application of a predefined number of genetic operations the algorithm terminates, and the poses with the highest scores are saved.

QXP (Quick Explore [56]) is part of the Flo+ program and contains two conceptually different docking algorithms: MCDock and ZipDock. The MCDock algorithm goes through repeated cycles of Monte Carlo followed by energy minimization in order to generate and refine an ensemble of low-energy ligand poses. New poses are added to the ensemble if they are low in energy and dissimilar (as defined by an RMSD threshold) to the poses already present. As the number of cycles increases the number and size of the perturbations are reduced, which makes the MCDock procedure very efficient in finding low-energy solutions.

The ZipDock algorithm was designed to carry out a near-systematic search. First, a representative basis set of (200 by default) ligand conformers is aligned and stored in a conformer tree. The conformer tree is docked rigidly and as a single entity, using a set of 2000 rotations and translating each of these 2000 instances to the centers of small spheres that fill the binding site. For each combination of rotation and translation, the interaction energy of each atom of each conformer in the tree is evaluated using a potential-energy grid. By combining parts of various conformers from the tree that exhibit favorable energies, one generates new conformers with overall favorable energies. As each low-energy conformer is discovered, it is evaluated using a rapid, approximate scoring method and added to the ensemble of best poses subject to the same energy and dissimilarity criteria that MCDock uses.

In terms of pose prediction accuracy, ZipDock and MCDock perform about equally well, but the ligands that fail to dock correctly with MCDock are often different from those that fail with ZipDock. Therefore, a combination of the ZipDock and MCDock algorithms – called FullDock – is available in QXP as well.

An interesting aspect of QXP is that it allows the user to define parts of the binding site as flexible. These sections can move during the energy minimization steps of the process and are fully mobile or tethered to their initial positions. For the next cycle of searching, the program uses the binding site coordinates from the previous cycle (which may be modified as a result of allowing protein flexibility) with a random probability of 90%. Resetting occasionally to the starting coordinates prevents unrealistic protein motion caused by indefinite propagation through subsequent cycles.

Originally the program used a molecular mechanics force field as scoring function. The improved 'plus' incarnations (e.g. MCDock+) employ an empirical potency score to score each pose that is generated and to rank different poses of the same ligand. The main terms in this empirical score account for receptor-ligand atom-atom contacts, hydrogen bonds, steric repulsion, desolvation, internal ligand strain, and ligand and receptor entropy. Interestingly, this scoring function has been optimized to predict both the relative potencies of inhibitors in experimental structure–activity relationship series and the crystallographic binding modes of those inhibitors for which complex structures are available.

TARGET-BASED OR "TUNED" SCORING FUNCTIONS

All docking programs come with built-in, "general", scoring functions, but there are several reasons why these may not be the best choice for all purposes and why one may develop or optimize a scoring function. The built-in scoring functions of docking programs were developed to work across a large set of target proteins (i.e. to be of general use), but that does not necessarily mean that they are the best functions for a particular target. If one has experimental binding affinities or inhibition constants against a specific target for a set of compounds and associated binding mode information (from crystal structures or even from prior docking experiments), one can 'tune' the scoring function to that target (hence the expression 'target-based' or 'tuned' scoring function). This can be achieved by building a regression equation with various scoring function terms and adjusting the regression coefficients to maximize the correlation between observed and calculated affinities. The generation of target-based scoring function may also involve the addition of novel terms. Subsequently the tuned scoring function is applied in docking experiments against that same target. In the following a number of examples are given.

Most programs use a single scoring function for initial pose evaluation and acceptance on the one hand and for affinity prediction on the other, but it has been observed that scoring functions optimized for affinity prediction do not necessarily reproduce binding modes best and vice versa. Recently, Giordanetto et al. modified the scoring function in the program QXP by adding additional terms and refitting the resulting functions to experimental data [76]. They observed a significant improvement in affinity prediction for their test set. However, when the pose prediction performance of these novel scoring functions was examined, the original (unmodified) QXP scoring function was found to perform much better. The original QXP function had been derived by also taking deviations from crystal structure poses into account in the fit procedure. By only using affinity data and descriptors derived from crystallographic complexes, Giordanetto et al. improved affinity prediction, but at the expense of pose prediction performance. It may therefore be advisable to generate separate functions for pose and affinity prediction.

Also, one needs a fast scoring function for the initial pose evaluation as one quickly needs to accept or reject many generated poses, while a slower scoring function may be acceptable for a more precise assessment of the binding affinity of the surviving poses. For these reasons the ICM-Dock program [54, 55] has two different built-in functions: one for pose evaluation (the docking function) and one for affinity prediction (the scoring function). Alternatively, one can take the poses from a docking program that were scored with its built-in scoring function and rescore them in a separate step outside the docking programs.

One example of target-based tuning is provided by the COMBINE approach [77]. Here one carries out a PLS analysis of interaction energies in protein–ligand complexes to identify those interactions that contribute most to the variance in ligand affinity. This information can then be used to predict the affinities of other ligands or to guide structure-based design efforts [78]. Interestingly, in a recent publication it was found that a variation of this approach works better in lead optimization than in hit identification [79].

Another method, the adaptation of fields for molecular comparison (AFMoC) approach, is based on interaction fields for known ligands inside a given receptor-binding site [80]. These fields are correlated to experimental affinities using PLS. In order to prevent the AFMoC scoring function from becoming too biased towards the training set, it can be mixed with a more general scoring function. A mixing parameter then determines how much both scoring functions contribute to the final score. In a recent study on inhibitors of DOXP-reductoisomerase, this mixing parameter has been found to have a major influence on the results [81]. Setting the parameter to 0.5 (i.e., 50% contribution from the DrugScore scoring function and 50% contribution from the AFMoC fields) allowed for good affinity prediction of related ligands as well as structurally different inhibitors.

Yet another approach to tuning a scoring function is followed in the latest version of the docking program QXP [56]. Using docking results from an initial run with compounds whose experimental binding affinity is known, the user has the option to refit the scoring function to the experimental numbers. This refitted scoring function can subsequently be applied in another docking run, where it serves not only as a scoring function but also as a docking function during the pose generation process.

In general it should be said that tuning of scoring functions is statistical in nature, and it is possible that the resulting scoring function contains physically unrealistic terms and coefficients. The risk of straying from a physically meaningful model is that the tuned function may work well only with the same target and very similar compounds as those in the training set. There are several ways of reducing this risk. One can tune scoring functions to a family of related targets (e.g., serine proteases, kinases). Or one can use only scoring function terms that have physical meaning and that are known to work well in validated scoring functions, and make sure that the coefficients remain realistic.

RESCORING

It has been observed that the scoring functions that come with docking programs do not always yield the best affinity predictions. One way to address this is rescoring. Here one takes the poses generated by a docking program and applies one or more alternative scoring functions to those poses. Rescoring and tuning are conceptually similar and the distinction is fuzzy. A key difference is that tuning uses scoring function terms as variables to derive an improved scoring function, while rescoring uses the scoring functions themselves. Other, less distinct differences are that one generally refers to tuning when the scoring function is optimized for a target or target family, but is still supposed to be transferable. The tuned function is also often the result of a significant development and intended for use both as a docking function (for pose evaluation) and a scoring function (for affinity prediction), while rescoring is a strictly postprocessing effort to improve affinity prediction. As we will see below, rescoring may actually also involve fitting to experimental data, which makes it more similar to tuning. Several approaches have been taken to rescoring. One is built on the physics of the binding process, while others apply a set of scoring functions and combine the results to generate a final score. Combining the results can be done simply by giving all scoring functions equal weights (consensus scoring) or, if one has experimental binding or inhibition data, by developing statistical models with optimized weights for the various scoring functions.

Rescoring – Solvation-Based Scoring

Solvation plays an important role in molecular recognition, but appropriate treatment of solvent effects in scoring functions still remains a major challenge. In many scoring functions these effects are considered only partially, neglected altogether, or included indirectly, as in some knowledge-based scoring schemes. A more rigorous way of treating solvation effects in the estimation of binding affinities has become known as MM-PBSA or MM-GBSA scoring, where MM stands for molecular mechanics, PB and GB for Poisson–Boltzmann and Generalized Born, respectively, and SA for solvent-accessible surface area. The MM-PBSA approach has been pioneered by Kollman et al., and its basis is a thermodynamic cycle for complex formation in aqueous solution [82, 83]. The key element is that the electrostatics of (de)solvation and ligand-receptor interactions are treated in a more sophisticated manner using PB or GB instead of simple Coulomb-based terms. The (de)solvation process can be divided into polar and apolar contributions. The associated energies, the polar free energy of solvation and the apolar free energy of solvation, are calculated with the PB or GB approach and using an expression containing a surface area term, respectively [84].

Recently, first applications of MM-PBSA as a more sophisticated scoring function in the context of SBVS have become known. In contrast to earlier applications, where it was combined with molecular dynamics (MD) simulations, the recent examples demonstrate its value also for 'snapshot scoring,' i.e., the evaluation of the MM-PB(GB)SA expression for one or a few poses per ligand. These poses had been generated using a conventional docking program and not by means of a lengthy MD simulation. Researchers at Wyeth [85] and SGX Pharmaceuticals [86] presented evidence that MM-PBSA scoring can lead to an improvement compared to conventional scoring. It was shown that, given a number of precomputed poses per ligand, re-ranking of the poses with MM-PBSA leads to a better separation between correct and incorrect poses. This improvement was due to a reduction of both false negatives and false positives. Also, it was illustrated that enrichment was significantly higher when MM-PBSA was used to rescore larger databases of docked ligands. Treatment of a substantial number of compounds was computationally feasible, as the compute-intensive part of the MD simulations including explicit water had been replaced by pose generation with a fast docking program.

Kuhn and coworkers recently demonstrated the application of MM-PBSA scoring to several different data sets [87]. Docking was performed using the programs FlexX [51] (ScreenScore function [6]) and FRED [59, 60] (ChemScore function [36]). Prior to rescoring, the complex structures were minimized in the presence of explicit water and counterions. For one data set (neuraminidase) it was shown that the correct inhibitor pose is normally identified by docking, but that the ChemScore function is not able to differentiate between true and false positives. In this case MM-PBSA scoring led to a significant improvement. Also, it was shown that MM-PBSA scoring improved pose prediction and consequently enrichment for p38 MAP kinase inhibitors. Analysis of these cases indicated that a major deficiency of conventional scoring functions is the lack of an energy penalty for the desolvation of mismatched, i.e., polar–apolar – protein–ligand interactions, which MM-PBSA can improve upon. The authors concluded that the application of MM-PBSA to a single structure is generally valuable for rescoring after docking and for distinguishing between strong and weak binders.

In another study MM-PBSA scoring was incorporated as the last step in a hierarchical database screening process [83]. In this case 20 poses per ligand were generated using short MD simulations (20 ps after equilibration), starting from the previously docked ligand orientations. Although only a limited number of ligands was considered, an encouraging correlation between experimental and MM-PBSA binding free energies was found, and it was noted that the overall strategy achieved not only high efficiency but also high reliability.

In an earlier study the effects of different solvent models were compared using a data set of 189 protein–ligand complexes [88]. In contrast to the afore-mentioned examples, comparing PB solvation with simpler solvent models – such as GB, constant or distance-dependent dielectric function, in conjunction with the CHARMm force field – did not indicate an advantage of more sophisticated solvation models in terms of the rank correlation between predicted and observed binding energies. Interestingly, the authors also noted that for pose prediction steric complementarity between the ligand and the receptor appears to be the most important factor.

Rescoring – Consensus Scoring

All scoring functions may exhibit pathological behavior with certain compound classes or functional groups. To minimize the impact of this problem and to reduce statistical noise, composite scoring methods have been introduced [89-91]. Rather than using a single scoring function, several scoring functions are combined such that in order to be classified as a potential binder, a molecule has to be scored well by a number of different scoring functions. Such composite scoring functions come in two flavors. Consensus methods combine the results of various single scoring functions in a predefined, unbiased manner without any training. Statistical composite methods, on the other hand, attempt to optimize the affinity prediction by developing a model that is based on a training set, which is relevant for the target being studied. Both approaches will be described below. For examples of the successful application of composite scoring methods, see references, [89, 92 and 93].

The premise of consensus scoring is that the more scoring functions agree that a compound is active (or inactive), the more reliable the prediction is. So, a compound that receives a high score from multiple scoring functions is more likely to be a good inhibitor in an actual assay than a compound that receives a high score from only a single function. Wang and Wang [90] simulated a docking and scoring experiment by taking known binding affinities for a set of compounds to which they added a random error to mimic the behavior of a scoring function. They repeated this process for several scoring functions and subsequently carried out consensus scoring. They found that ''consensus scoring outperforms any single scoring for a simple statistical reason: the mean value of repeated samplings tends to be closer to the true value.'' In other words, consensus scoring will work better than a single scoring function if the scoring functions are largely independent of each other and if the individual scoring functions themselves are equally predictive. Several ways of combining scoring functions exist.

In the **rank-by-vote** procedure, one lets several (N) scoring functions 'vote' on all compounds. For each scoring function, all compounds are rank-ordered by their score and the highest scoring compounds (e.g. the highest scoring 2%) receive a vote. Subsequently, for each compound the votes from all N scoring functions are added together. All compounds with N votes are predicted to be active. One can also allow one dissenting vote and regard all compounds with at least N-1 votes as active, but that does not necessarily increase enrichment [93].

A potential problem with the latter procedure is that, depending on the agreement between different scoring functions, more or less compounds may pass the test of receiving N (or N-1) votes. In order to pre-define the number of compounds that are selected by a consensus scoring scheme, a novel, "non-reducing", consensus scoring scheme was recently developed [93]. This variant involves iteratively increasing the number of compounds that receive a vote by descending the rankordered list of each scoring function, one compound at a time, until the number of compounds with N votes equals that pre-defined number. This procedure starts by giving a vote only to the single best-scoring compound per scoring function and ends by giving votes to all compounds considered. One can also call this variant a worst-rank consensus scheme because the worst rank a molecule has according to all scoring functions determines its final rank.

Another consensus approach uses the **mean rank** of each compound, i.e., the average value of the ranks of that compound according to each of the scoring functions that are allowed to vote. Unlike rank-by-vote, the mean rank procedure is by definition nonreducing.

Neither approach is consistently better than the other. The results depend on the docking program and on the number and the nature of the scoring functions that are chosen to vote. Although the appeal of these methods is that they are unbiased and that their application does not require any experimental binding data, in practice it is important to run some pilot experiments to determine how many and which scoring functions work best (i.e., yield the best enrichments).

In the preceding paragraph it was described how the results of rescoring docking poses with various scoring functions can be combined in an unbiased way. If one has experimental binding affinity data, however, one can optimize the affinity prediction by developing a model that is based on this data (the training set). Recently, Bayesian statistics (BS) has made inroads in drug discovery and developing composite scoring functions is one possible application of BS. Cotesta et al. [93] found that in the majority of cases BS performs better than the individual scoring functions and than the unbiased consensus approaches in terms of enrichment (i.e., hit rate increases). It also works well in distinguishing between moderately active and very active molecules, making the approach also suitable for lead optimization.

CHALLENGES AND IMPROVEMENTS

Much work has been invested in the generation of better docking programs and scoring functions over the past years and, although much progress has been made, improvement is still necessary. In this section some of the fundamental challenges in docking and scoring and the ways that researchers have started to address them are outlined. Some general approaches to improve the performance of docking and scoring are presented, too.

Challenge 1 - Docking into Flexible Receptors

One of the most challenging problems in docking and scoring is the treatment of flexible receptors. Numerous examples have become known where the same protein adopts different conformations depending on which ligand it binds to [94, 95]. As a consequence, docking using a rigid receptor representation corresponding to a single receptor conformation will fail for those ligands that require a different protein conformation in order to bind.

An example where relatively small conformational changes can have already a large effect is given by the following cross-docking example: in an in-house evaluation of the pose prediction performance of several docking programs at Pharmacia, a number of publicly known CDK2 inhibitors was docked into a receptor conformation corresponding to CDK2 in complex with adenosine triphosphate (ATP) (Protein Data Bank (PDB) code 1QMZ), after the natural ligand had been removed. One of the ligands that was especially difficult to be docked correctly was hymenialdisine. Examination of the crystal structure of this ligand in complex with CDK2 (PDB code 1DM2) revealed that 1QMZ has a larger ATP-binding pocket than 1DM2. Moreover, in 1DM2 a hydrogen bond exists between the carbonyl oxygen of the imidazolone moiety of the ligand and one of the two water molecules located behind K33 of the kinase. Also, D145 adopts a different conformation in 1DM2 and forms a hydrogen bond with the ligand. The result of these differences is that, when docked into 1QMZ, hymenialdisine cannot engage in all binding interactions it makes in 1DM2. It can wobble around and adopt several distinctly different, but energetically degenerated poses. This larger, suboptimal binding site explains the failure of some docking programs to dock hymenialdisine correctly. In order to test this hypothesis, hymenialdisine was docked into the ATP-binding pocket of 1DM2 using QXP. Two runs were carried out: one with the water molecule present and one without the water molecule. In both cases QXP was able to identify the correct binding mode of hymenialdisine and to score the correct pose highest. Of course a much more radical example of different conformations adopted by kinases is the transition between so-called DFG-in or DFG-out conformations, where a stretch of three residues (DFG) is either rotated into or away from the ATP-binding pocket [96].

In order to deal with the problem of flexible receptors in docking, several approaches have been proposed, which can be grouped roughly into the following categories: (1) letting the receptor or parts thereof move during docking; (2) docking the compounds into several different conformations of the same receptor and aggregating the results; and (3) docking into averaged receptor representations. The borders between these three approaches are sometimes fuzzy and some of the practical implementations known contain elements of more than one of these methods. Examples for each of the three categories are provided here.

Regarding the first category, one rather well-known program that allows **receptor flexibility during docking** is QXP [56]. Here the user can specify certain parts of the protein to move during the minimization step at the end of each Monte Carlo cycle during docking. In some cases this can alleviate clashes between the ligand and receptor that would otherwise occur and this can therefore lead to better pose prediction results. Also in the latest version of ICM-dock receptor flexibility is encoded [54, 55]. In this case the amino acid side chains are allowed to move during docking.

Another example for the incorporation of receptor flexibility has been provided by researchers at Schrodinger. In this case rigid receptor docking using Glide is iteratively combined with protein structure prediction using Prime [97]. The authors reported that using only Glide the average root-mean-square deviation (RMSD) to the crystal structure for 21 different complexes was 5.5Å, and application of their new induced-fit docking (IFD) procedure reduced the average RMSD to 1.4Å. The tradeoff here is that the IFD procedure requires a relatively large amount of time per ligand and is therefore not applicable to high-throughput docking. Nevertheless, this methodology could prove useful for building and refining homology models, detailed binding studies during lead optimization, and the generation of different conformational hypotheses prior to a larger docking exercise.

Regarding the second category, a number of approaches have been published where ligands are docked into **several different conformations of the same receptor**, in order to address the problem of receptor flexibility. Cavasotto and Abagyan have presented an algorithm to generate a discrete set of receptor conformations, and each of these structures is then used for rigid receptor docking [98]. Subsequently the results of the multiple docking runs are combined in order to improve enrichment. Combining the results is achieved by merging the hit lists for each of the docking runs and keeping the best rank for each compound. For several protein kinases this procedure led to a significant increase in hit rates compared to the individual results. In another study scoring functions that are more (soft) and less (hard) tolerant to bad geometries were compared in docking runs against one or more conformations of the same receptor [99]. The soft scoring function proved to be superior to the hard potential when a single receptor conformation was used. Conversely, when docking was performed into multiple receptor conformations the hard potential showed better performance. In this case multiple flexible regions of the binding site were treated independently, recombining them to generate different discrete conformations [100]. It was also noted that softer scoring functions can increase the likelihood of false positives. Using FlexX as docking program, a comparison of single versus multiple conformer docking was performed for protein tyrosine phosphatase-1B (PTP-1B) [101]. Different receptor structures had been created by considering different combinations of side-chain rotamers within the active site. The inhibitors were then docked against all active-site models and for each inhibitor the model with the best interaction energy was identified. This allowed for successful discrimination between correct and incorrect binding modes as well as for an improvement in the ranking of the inhibitors. The FlexE program [57] is based on a united protein description originating from different superimposed conformations of a protein. During the incremental construction of a ligand discrete protein conformations are sampled in a combinatorial fashion. The program was evaluated for 10 proteins represented by 105 crystal structures from the PDB and one modeled structure. For 83% of the ligands the correct pose was found. The results were of a quality comparable to the one obtained by sequentially docking into all conformations separately, but the run times for FlexE docking were much shorter than for the sequential docking.

Receptor averaging is another way of approaching the problem of receptor flexibility. Using AutoDock it was investigated how the interaction energy grids for different receptor conformations can be combined [52]. The study was carried out using complexes of 21 peptidomimetic inhibitors with human immunodeficiency virus-1 (HIV-1) protease. Four different schemes of combining the grids were tested. It turned out that the mean grids performed worst, whereas the energy-weighted methods gave much better results. That using simply an average structure representing several different conformations leads to inferior results has been noted by researchers at Eli Lilly [102]. Their results indicated that docking accuracy decreases significantly when an average structure is used.

McGovern and Shoichet also carried out a very interesting study [103]. The focus of this study was the information loss that occurs when the active-site conformation becomes less defined. To this end, 10 different enzyme-binding sites, represented by their holo, apo, and model structures, were investigated. The MDDR (MDL Drug Data Report) database of 95 000 small molecules containing at least 35 ligands for each of the 10 systems was docked against all 30 structures using the DOCK program [61,62]. The ability of each structure to enrich the known ligands for that enzyme over random selection was evaluated. In seven cases, there was clear superiority of the holo structures over the apo and model structures. However, the apo and model structures proved superior for two and one enzymes, respectively. For the latter cases it was postulated that the holo structure may in some cases be overspecialized by induced fit to a particular ligand, and therefore the apo or model structure may be a better choice for the docking experiment.

Summarizing, one may say that to date, protein flexibility remains one of the most challenging problems in docking and scoring. Progress has been made and interesting approaches have been proposed, but it is still an open question whether these techniques have advanced sufficiently to be of substantial help.

Challenge 2 - Water

Water molecules often play a key role in protein–ligand recognition. If one ignores water-mediated interactions during docking then the calculated interaction energy of a given ligand conformation may be too low. If, on the other hand, one retains crystallographically observed water molecules then the binding pose and affinity of a ligand that in reality replaces that water molecule will not be correct. It is notoriously difficult to treat water adequately, as first one needs to identify possible positions for water molecules where they could interact with the protein and ligand, and subsequently one must be able to predict whether a water molecule is indeed present at that position. Researchers at Astex and the Cambridge Crystallographic Data Centre recently implemented an elegant procedure in the latest version of GOLD to address both these issues [104]. The water positions they consider for a given target are taken from a set of complex structures of that target, but one could also use programs to predict potential water-binding sites [105, 106]. Each water molecule can then be present ('on') or absent ('off '). If a water molecule is on, it can make favorable interactions with the ligand and protein, but it pays an entropic penalty for loss of translational and rotational degrees of freedom [107]. The value of this penalty was optimized using a training set of 58 protein-ligand complexes. Considering both the training and test sets, on and off status are correctly predicted for 93% of the water molecules. This increases correct pose prediction rates of water-mediated complexes by 10–12 percentage points, but it decreases correct pose prediction rates for non-water-mediated complexes by 6–7 percentage points. This latter decrease is readily explained when one assumes that prediction of a water molecule where there should not be one leads to an incorrect binding mode. The expectation is that the correlation of calculated and measured affinities will improve with the inclusion of water molecules in the docking runs, which in turn should improve the enrichments obtained in VS experiments, but this remains to be investigated.

Another approach to dealing with water molecules involved in protein–ligand interactions has been incorporated in the FlexX docking program. This method, referred to as the particle concept, includes the calculation of favorable positions of water molecules inside the active site prior to docking. During the incremental construction phase these water molecules are allowed to occupy the precomputed positions if they can form additional hydrogen bonds with the ligand. The method was tested using a data set of 200 protein–ligand complexes and with pose prediction quality as an evaluation criterion. Similar to the observations made for GOLD, it was found that on average the improvement was minor. Nevertheless, in a number of cases the predicted waters corresponded to the crystallographically observed ones, which led to an improvement in the predictions [108].

Another program that needs to mentioned in this context is the program SLIDE [67]. Prior to docking a knowledge-based approach, CONSOLV [109], is applied in order to select those water molecules that are likely to remain in their positions upon ligand binding and to determine an energy penalty for their displacement. During docking, overlap between the docked ligand and these water molecules is resolved by iterative translations or annihilation of the water molecules, applying appropriate penalties in due course.

Challenge 3 – Tautomers and Protomers

Another challenge in docking is accounting for the various tautomeric and protomeric states the molecules can adopt. In many databases molecules such as acids or amines are stored in their neutral forms. Considering that they are ionized under physiological conditions it is necessary to ionize them prior to docking. However, while standard ionization is easy to achieve, the problem of tautomer generation is already much more challenging: which tautomer should one use? Or should one use more than one (or all possible) tautomers for a given molecule? Not only for tautomers, but also for different ionization states balanced equilibria between the different forms provide real challenges in docking. One (radical) approach to this would be to generate all possible forms, subsequently to dock all of them, and to choose the relevant form based on the scores. However, it remains to be seen whether such an approach would be beneficial or just generate a large number of false positives.

Improvement 1 - Multiple Active-Site Corrections

A possible way of improving docking results is the application of so-called multiple active-site corrections (MASC) [110]. Here the underlying idea is that scoring functions are biased towards certain ligand types or characteristics, such as large or hydrophobic ligands. This implies that some ligands are generally predicted to be good binders regardless of whether these ligands will bind to certain active sites or not. Therefore, a simple statistical correction has been introduced, which can be interpreted either as a statistical measure of ligand specificity or as a correction for ligand-related bias in the scoring function. In order to calculate the MASC scores, each ligand is docked into a number of unrelated binding sites of different binding site characteristics. The corrected score (or MASC score) S'_{ij} for molecule *i* in binding site *j* is calculated as follows:

$$S'_{ij} = \frac{\left(S_{ij} - \mu_i\right)}{\sigma_i}$$

where S_{ij} is the uncorrected score for this molecule. μ_i and σ_i represent mean and standard deviation of the scores for molecule *i* across the different binding sites. Thus, the MASC score S'_{ij} represents a measure of specificity of molecule *i* for binding site *j*, compared to the other binding sites. The MASC scores were tested using FlexX and GOLD and a data set of 15 protein–ligand complexes. Without corrections only in three (FlexX) or four (GOLD) cases the endogenous ligand was identified correctly. After application of the MASC the success rates rose to 11 for both docking programs. In another test, the MASC scores were applied to a database of 600 drug-like molecules mixed with 30 inhibitors of p38a MAP kinase and 30 inhibitors of protein tyrosine phosphatase-1B (PTP-1B). Interestingly, when docking the 660 compounds into the p38a active site, the uncorrected GOLD scores led to an enrichment of PTP-1B inhibitors over the p38a inhibitors. Application of the MASC scores led to a reversal of this trend and the enrichment of true p38 inhibitors was significantly improved. The authors concluded that MASC improves the detection of true positives and reduces the number of false positives in database enrichment studies.

However, in a recent study scientists at OpenEye found that the application of MASC can also lead to a deterioration of the results [111]. When the docking program FRED [59, 60] was used in enrichment studies comparing different scoring functions and different targets, it was found that the value of MASC scoring depends heavily on the type of target and the scoring function used. Regarding the change of enrichment for five scoring functions plus consensus scoring, with each scoring scheme averaged across five different targets, only for consensus scoring a significant improvement using MASC scores was observed. Considering each target separately, the picture was also ambiguous. Only for reverse transcriptase did MASC lead to an improvement for all scoring schemes investigated. Other targets exhibited a very mixed behavior, most notably HIV protease, where for three scoring schemes the results after MASC deteriorated significantly, while for two scoring schemes a significant improvement was observed. For another target – the estrogen receptor – MASC scoring either reduced the enrichment or had no effect.

Improvement 2 - Docking with Constraints

By introducing a bias during docking it is possible to influence the way poses are generated and which ones are preferentially kept. For example, in the DockIt program [112], one can apply distance constraints between ligand and protein atoms that are subsequently used during pose generation via a distance geometry approach. The genetic algorithm (GA) behind the GOLD program [53, 74, 75, 113] makes it easy to include different types of constraints in the fitness function, thus enabling the generation of biased poses. In the PhDock approach [114], as implemented in DOCK 4.0 [115], one can perform pharmacophore-based docking by overlaying precomputed conformers of molecules based on to their largest 3D pharmacophore. The pharmacophore is then matched to predefined site points representing putative receptor interactions. Subsequently, all conformers are docked corresponding to the pharmacophore match and the fit of each individual conformer is scored. The advantage of this approach is twofold: speed through a rapid preorientation of the molecules to be docked as well as introducing bias towards good solutions by defining pharmacophore points that represent favorable interactions with the target binding site. The use of another docking program where one can apply pharmacophoric constraints during docking (FlexX-Pharm [116]) is illustrated below in the section on application examples.

Improvement 3 – Postprocessing

One can in principle distinguish between two approaches for introducing bias after the docking: applying postdock filters and using tailor-made (re)scoring functions. The latter have already been described above. In many cases postdock filters are conceptually simple and may correspond to certain geometric criteria, such as the presence of certain interactions (e.g., a hydrogen bond with a selected residue or a polar interaction) or the filling of a specified pocket in the active site. Many researchers in the pharmaceutical industry have written their own filters, but nowadays they are an integral part of several commercially available docking programs, such as Glide [63-65] and FRED [59, 60]. Implementation and automation of these filters can also reduce the need for the frequently quoted visual inspection of poses after docking and scoring.

The idea of checking for certain receptor–ligand interactions can be extended to entire interaction patterns. These interaction patterns in turn can be compared to a target interaction pattern, as observed in the co-crystal structure of a highly active ligand and its receptor. An interesting method has recently been developed where a structural interaction fingerprint (SIFt) is calculated to provide a unique representation of a binding mode [117]. In this approach each binding site residue is encoded by a string of 7 bits that represent the different possible interactions with that residue. If such an interaction occurs the corresponding bit is set. All bits together constitute the SIFt of a binding mode. SIFts can then be used to evaluate similarities between different poses of the same or different molecules, to assess how close a pose is to a predefined (e.g., crystallographic) pose and, of course, to filter docking results to find compounds with desirable interactions. One of the appealing features of SIFt is that it is an automated procedure, which can also objectively compare dissimilar ligands.

Another way of postprocessing is to use the docking results as input to develop a Bayesian model with the aim of reducing the numbers of false positives and false negatives. Klon *et al.* [118] did this as follows: they rank-ordered the list of compounds at the end of a docking run and designated the top-scoring ones as ‘good’ and the remaining ones as ‘bad.’ In the next step fingerprints were calculated for all the compounds and a naive Bayesian classifier was trained. Subsequently all compounds were re-ranked according to the Bayesian model. Applying this procedure to the results of high-throughput docking into an HIV-1 protease model using Glide, FlexX, and GOLD improved the hit rates in all cases significantly. Despite the impressive results, two caveats are appropriate here. First, the data set investigated consisted of a large number of inactives (the Available Chemicals Directory data set) and a small set of potent HIV-1 inhibitors from in-house chemistry efforts, which is arguably the easiest scenario to achieve enrichment. Second, given that the docking programs have already enriched the ‘good’ compounds with HIV-1 inhibitor chemistry, it is conceivable that the Bayesian model just separates very similar HIV-1 inhibitor-like compounds from other, chemically very diverse ACD molecules. The authors indicate, however, that the Bayesian model requires a large number of features for this approach to work and that consequently just the similarity between the core structures of the inhibitors cannot explain its success. At any rate, one would suspect the results to be very data set-dependent and further validation of this approach (which is in principle very interesting) with a different data set composition (e.g., with compounds that span a wide range of activities but have the similar core structures) would be important.

Assessment of Docking Performance

A multitude of approaches and docking programs is available today. This poses the question which docking program one should use, and which docking approach might be most appropriate for a given problem. In line with the two main tasks carried out by docking programs, evaluation thereof normally revolves about pose prediction and affinity prediction. As also noted by others, comparison and assessment of docking programs are not easy tasks [119]. Quite a number of studies comparing docking programs and assessing their performance have been published [6, 23, 89, 120-128]. The general conclusions that can be drawn from these studies are that no docking program is consistently superior to all other programs, and that success and failure very much depend on the combination of ligand and receptor characteristics as well as docking algorithm and scoring scheme used. In some cases certain scoring function characteristics (e.g. 'hard' or 'soft') appeared to be more appropriate for certain target characteristics than others. It is beyond the scope of this review to elaborate on these studies in detail, and the remainder of this section will be devoted to discussing some general aspects of the assessment of docking programs in more detail.

Regarding **pose prediction**, it is desirable to evaluate docking programs with respect to their ability to reproduce experimentally known poses in a reliable manner. The traditional way to do this is to calculate the RMS deviation (RMSD) between a pose generated by a docking program and the experimentally observed binding mode. Despite the practical appeal of using RMSDs from a crystal structure to assess pose prediction accuracy, they do not do justice to the complex interactions ligands make with proteins. For that reason, a novel way to evaluate pose prediction accuracy was devised [129]. Here the correctness of a pose was determined by visually comparing the (hydrogen-bonded and other) ligand–protein interactions for that pose with the experimentally observed interactions, and the resulting evaluation scheme was termed interactions-based accuracy classification (IBAC). It was shown that RMSD to X-ray values do not always correlate with IBAC and that in some cases IBAC gives a better indication of the correctness of a pose. The method, however, requires optical inspection of each pose, and a more automated procedure would be desirable. The recent introduction of the SIFt method [117] would be one way to address this.

With respect to **affinity prediction** or scoring one important aspect that needs to be considered when performing an evaluation is the data set under consideration. It would seem natural to carry out tests with data sets that resemble the data sets that will be used in production mode, but it appears that this is not always the case. The desired and expected activity range for inhibitors depends on the stage of a drug discovery project. At the hit identification stage, molecules with even weak activity (IC_{50} >100nM or >1μM) represent a useful source to initiate a medicinal chemistry program, while ligands with nanomolar affinity are searched for during the lead optimization phase. This puts different demands on the computational tools. In practice one rarely finds nanomolar compounds when screening databases of commercially available compounds, so for VS programs to be applicable during hit identification, they should be able to identify micromolar hits among a large number of inactive compounds, c.f. the 'hit identification' scenario in Fig. (**2**). By contrast, if one considers using these programs for lead optimization (e.g., for the design of a combinatorial library built around a potent scaffold), one needs to be able to distinguish potent compounds (<100nM) from moderately/weakly active (100nM – 10μM) and inactive (>10μM) ones. Consistent with observations made by Charifson *et al.* [89], one can expect that the ability of docking tools to distinguish active from inactive compounds depends considerably on the activity profile considered. In a recent study this was explicitly considered by defining three activity intervals and examining the docking and scoring results for each activity interval separately [93]. A main conclusion of this study was that affinity prediction and enrichment strongly depend on the distribution of activities in the data set. In this study it was

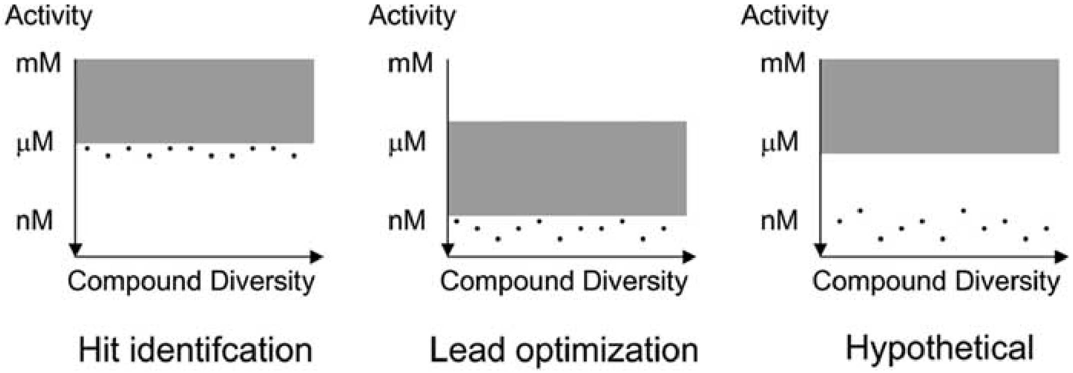


Fig. (2). Graphical representation of different activity distributions that can be used in the assessment of docking program performance. The few active compounds are represented by the black dots. The many inactives are represented by the gray areas. Under each distribution the corresponding real-life situation is indicated. The 'hypothetical' situation is rarely encountered in real life, but is often used to assess docking performance nonetheless, as in that situation docking programs/scoring functions exhibit their best performance in terms of separating actives from inactives.

also mentioned that quite often the performance of docking tools is assessed by spiking a large collection of supposedly inactive compounds with several fairly active compounds. Even if these tools are intended for hit identification only, such an evaluation procedure can be considered suboptimal, as commercial or proprietary compound collections rarely consist exclusively of nanomolar compounds against a given target on the one hand and fully inactive ones on the other. Similar conclusions were also drawn in another study [64].

Another point that needs to be considered for the assessment of docking program performance is the properties of the compounds, such as molecular weight (MW) or polarity. Some authors have argued that care must be taken that both active and inactive compounds represent an identical or very similar distribution of such properties. This is because charges can introduce an unwanted bias towards certain molecules [130] and because scoring functions in many cases correlate with MW [131]. Of course naturally there is a tendency for larger ligands to be more active, as the number of contacts with the receptor increases, but care must be taken nevertheless.

Often there is relatively low correlation between experimental and predicted activity. Docking has been successful in many cases nonetheless, because of its ability to select preferentially active compounds, thereby increasing the number of actives in a set of compounds that is selected for experimental testing. This increase is referred to as enrichment and has become a standard measure of quantifying the success of a docking campaign. In many cases enrichment is displayed in the form of enrichment curves, where the number (or percentage) of actives found is plotted against the number of rank-ordered molecules. Recently it was advocated to use receiver operating characteristic (ROC) curves instead [132]. The advantage of these curves is that they are independent of the proportion of actives in the test set. Also, they include information relating to false positives and false negatives in the same plot.

APPLICATION EXAMPLES

Even if the main application of structure-based VS probably lies in hit identification, there are many variations on this theme. In the following sections several examples are given of the successful application of docking and scoring to a variety of different problems. Each of these examples illustrates that docking and scoring is not a stand-alone technique, but that it is normally embedded in a workflow of different *in silico* as well as experimental techniques, and that careful evaluation before application is a prerequisite for success. For an extensive review of docking success stories, see reference [133].

Application Example 1 - Inhibiting Bcl-2 – Bak Interactions

Bcl-2 is an important factor in the apoptotic pathway and is overexpressed in many cancer types. The ability of Bcl-2 to form heterodimers with another regulatory protein called Bak confers it an anti-apoptotic role. Therefore, a virtual screen against Bcl-2 was performed with the idea to identify small molecules inhibiting the Bcl-2 Bak interaction [134]. To this end a homology model of Bcl-2 had been derived from the NMR three-dimensional structure of the complex of Bcl-XL with a Bak BH3 peptide. The National Cancer Institute (NCI) database of 206 876 organic molecules was docked against the BH3 binding pocket on Bcl-2 using the program DOCK. Ranking was performed with the energy scoring function implemented in the program and the non-peptidic molecules among the top 500 in the rank-ordered hit list were considered as potential Bcl-2 – Bak inhibitors. Out of 80 compounds 35 were available and finally tested. Seven of them had IC50 values between 1.6 and 14.0 μM. One of these hits also showed good anti-proliferative activity in a human myeloid leukemia cell line and induced apoptosis in cancer cells overexpressing Bcl-2.

Two important conclusions can be drawn from this study: Firstly, homology models can be suitable for virtual screening, if generated and evaluated with care. Another example for the successful use of a homology model is the virtual screen performed using a homology model of the alpha1A adrenergic receptor [135]. Secondly, the study demonstrated that virtual screening can be used also for identification of potential inhibitors of protein-protein interactions, which are thought to be notoriously difficult targets. Another example in this context has been provided recently by Trosset et al. in a virtual screen for inhibitors of the β-catenin–Tcf interaction [8].

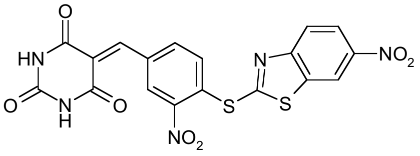
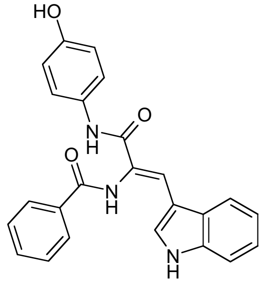
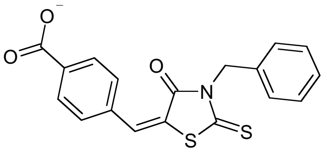
Application Example 2 - Virtual and Experimental Screening of Protein Tyrosine Phosphatase-1B (PTP-1B)

PTP-1B hydrolyzes phosphotyrosines on the insulin receptor, thereby deactivating it. Overproduction of this enzyme has been implicated in the onset of type-2 diabetes, and it is therefore a target for drug discovery [136, 137]. In a study by Doman et al. a comparison between HTS and docking against PTP-1B was performed [7]. For the HTS a library of approx. 400 000 compounds from a corporate collection was screened. Some 85 compounds were found with IC_{50} values between 100 and 1 μM, which corresponds to a hit rate of 0.021%.

In silico screening against PTP-1B was performed using 235 000 commercially available molecules from the ACD, BioSpecs and Maybridge databases. Docking was carried out with the Northwestern University version of DOCK [61, 62]. Target was the crystal structure of PTB-1B (PDB entry 1pty) and the site was defined by the locations of two bound phosphotyrosine molecules. After the docking the top-scoring 1000 molecules (500 for the ACD and 500 for the combined BioSpecs and Maybridge databases) were considered for further evaluation. A total of 889 molecules were actually available, and after visual inspection 365 compounds were chosen for testing. Of these, 127 molecules were found to be active with IC_{50} <100 μM, which corresponds to a hit rate of 34.8%.

Although the compound sets used in HTS and virtual screening were not identical and the experimental conditions for the HTS screen and the testing of the 365 compounds varied somewhat, the differences in hit rates were impressive and underpinned the usefulness of docking and scoring. However, it was noted that there was no clear correlation between the docking scores and the IC50 values. This is illustrated by the molecules listed in Table 2. Compounds **2**

Table 2. Selected Hits from the Structure-based Virtual Screen Against PTP-1B

Structure (Compound)	Docking rank	Docking score	IC_{50} (μM)
 (1)	406	-33.4	4.1
 (2)	440	-33.2	21.5
 (3)	11	-42.0	21.6

and **3** in Table **2** have very different scores (and corresponding ranks), but they possess equivalent activity. Compound **1**, in turn, is ranked relatively low, but has the highest activity.

In an absolute sense, a 35% hit rate resulting from a docking experiment is very high, but one sobering piece of information is that docking against the ‘open’ form of PTP-1B (PDB code 1BZH [139]) did not produce any experimentally confirmed hits, although admittedly only 15–20 of the docking hits were tested. One might think that the open form is not biologically relevant and that binding to it will not inhibit PTP-1B, but that is not the case, as 1BZH is actually a complex of the protein with an inhibitor.

Taken together, the high overall hit rate and the low correlation with experiment indicate that the docking program and scoring function are good at eliminating compounds that do not fit the active site well electrostatically or sterically, but that they are not able to differentiate reliably between two compounds that both fit (and presumably both exhibit measurable inhibition). When taking random samples of the corporate collection and of the docking database, some structural similarity exists. Surprisingly, no similarity existed between the docking hits and the HTS hits, while one would expect the similarity to increase as a result of the bias towards PTP1B inhibitors. This is also an indication that docking and HTS are complementary techniques and may be applied side-by-side.

Application Examples 3 and 4 - Layered Virtual Screening: Carbonic Anhydrase II and Checkpoint Kinase-1

In some cases structure-based VS can be part of a cascade of different computational techniques in the quest for binders of a given target. Two such examples are given here. The first example of the combination of different VS techniques is given by a study that led to the successful identification of subnanomolar inhibitors of **carbonic anhydrase II** (CAII) [9, 140]. A layered strategy including pharmacophore and ligand-based modeling was applied, where the last step consisted of docking and scoring of a number of compounds. The starting point of the study was the high-resolution crystal structure of CAII, and a set of 90 000 molecules (including 35 known CAII inhibitors) from the Maybridge and LeadQuest databases. As a first step, a binding-site analysis was performed, in order to identify key interactions between a putative compound and the receptor. These key interactions were transformed into a 3D pharmacophore model, which was used to search the database of 90 000 molecules. The database scan led to the identification of 3314 molecules. In the second step, these 3314 molecules were rank-ordered using the program FlexS that evaluates their potential binding affinities by comparison with a reference compound. In the third and final step the 100 best-ranking molecules were docked into the binding pocket of CAII using the docking program FlexX. FlexX score and DrugScore were used to predict the binding affinity of the docked molecules and to rank-order them accordingly. The 13 top-scoring compounds were chosen for testing, and three inhibitors with IC_{50} values of <1 nM were identified. Two of the subnanomolar hits were co-crystallized with CAII, and the binding mode generated by FlexX and ranked highest by DrugScore was found to be in good agreement with the experimental structures. The study yielded two important insights. First, it turned out that water molecules played an important role during the docking process. Four conserved water molecules had been identified by superposition of all complex and apo structures of CAII. Inclusion of these solvent molecules in the docking process added to the steric restriction of the binding pocket and led to better solutions. Secondly, despite the successful identification of high-affinity binders, it was observed that overall correlation between the IC50 values and the binding affinities predicted by FlexX score or DrugScore was rather poor.

The second example for the integration of structure-based VS into an *in silico* workflow is provided by a study on **checkpoint kinase-1** (Chk1) [141]. In this case, in the first stage the in-house compound collection was filtered by general physicochemical properties, such as MW and number of rotatable bonds, followed by removal of compounds with undesired chemical functionality. In the next step the remaining compounds were evaluated by their fit to a pharmacophore representing a minimal kinase binding motif, consisting of two hydrogen bonds (one acceptor, one donor) with the hinge region of the kinase, where the adenine moiety of ATP binds. Approximately 200 000 compounds passed this pharmacophore filter and were subsequently submitted to docking with FlexX-Pharm [116] with the same pharmacophore as constraint. Up to 100 poses were saved for each successfully docked compound. All saved poses were then rescored with a consensus scoring scheme that had been derived in a study for another kinase (CDK2), which included a combination of the FlexX and PMF scoring functions. Using this scheme and prior knowledge, 250 compounds were retained for visual inspection. After application of this final human filter, 103 compounds were assayed, which yielded 36 active compounds from four different chemical classes with activities ranging from 110nM to 68 mM. Several conclusions can be drawn from this study. Integration of structure-based VS in a general *in silico* workflow leads to a reduction of the number of molecules to be docked, thereby saving a significant amount of computing resources. A tailor-made scoring scheme that has been derived for a related protein can be successfully transferred to the target of interest. Despite all efforts to ensure that docking programs generate realistic poses, one notices an industry-wide, pervasive need to include visual inspection as the final filters in order to remove compounds whose poses are unrealistic.

CONCLUSIONS AND OUTLOOK

Virtual screening for the rapid identification of small-molecule ligands of macromolecular targets has become an established technology in drug discovery. High-throughput ligand docking or structure-based VS is a powerful technique to perform such a screen if a 3D structure of the target is available, and docking success stories are abundant. Remarkably, there is quite a number of successful applications using homology models as the target structure.

Many docking programs exist, but no single program has yet emerged that outperforms all others in all cases. Generally, programs do an adequate job searching conformational space and generating correct ligand poses (binding modes), but the scoring functions need improvement. This is evidenced by the fact that the correlation between calculated and observed binding affinities is often low and that separate scoring functions are frequently needed for pose evaluation and affinity prediction. Other problem areas are target flexibility, explicit water molecules and (de)solvation phenomena. Fortunately, active research is taking place to address all these issues and appreciable progress is being made. Tunable scoring functions, MM-PBSA and other rescoring approaches, explicit protein flexibility (as, e.g., in QXP and ICM) and explicit incorporation of water (as, e.g., in FlexX and GOLD) are just a few examples. Regarding protein flexibility one still needs to find ways to avoid false positives, by properly taking protein conformational energy into account. Another, yet not adequately addressed, challenge is the generation of different tautomer and protomer states for the ligand molecules and implementation of the corresponding scoring schemes. However, the ever increasing power of computer hardware will facilitate the implementation of more sophisticated methods and one can hope that docking tools will continue to improve significantly in the near future.

Despite the docking successes highlighted in this review, achieving success is not trivial. A docking campaign cannot be regarded as a black box that one feeds a general compound collection and that automatically produces a collection of high-affinity ligands. Preparing the protein and the ligands, selecting the docking programs and scoring functions, setting and tuning the parameters, and carrying out the postprocessing (including the often necessary visual inspection) require profound expertise. Docking is especially useful in reducing a collection of virtual compounds down to a manageable number to be synthesized and in selecting compounds from an external collection. Even when experimental HTS is envisioned, however, VS is important, as active compounds may be identified by one technique and not by the other. An added bonus is that VS is fast and inexpensive by any standard. It is recommended, if at all possible, to use docking in parallel with other techniques (experimental HTS, pharmacophore modeling, etc.), to dock to multiple conformations of the target, and to use the docking results to select as many compounds as possible for experimental confirmation.

In the light of the progress that has been made and considering the known successful applications and the ongoing developments, it is conceivable that the importance and impact of VS will continue to increase significantly.

ABBREVIATIONS

ACD	=	Available chemicals directory
AFMoC	=	Adaptation of fields for molecular comparison
ATP	=	Adenosine triphosphate
BS	=	Bayesian statistics

CAII	=	Carbonic anhydrase II
CDK2	=	cyclin-dependent kinase-2
Chk1	=	Checkpoint kinase-1
GA	=	Genetic algorithm
GB	=	Generalized Born
HTS	=	High-throughput screening
MASC	=	Multiple active site corrections
MD	=	Molecular dynamics
MM	=	Molecular mechanics
MW	=	Molecular weight
NMR	=	Nuclear magnetic resonance
PB	=	Poisson-Boltzmann
PDB	=	Protein data bank
PLS	=	Partial least squares
PMF	=	Potential of mean force
PTP-1B	=	Protein tyrosine phosphatase-1B
RMSD	=	Root-mean-square deviation
ROC	=	Receiver operating characteristics
SA	=	(Solvent accessible) surface area
SBVS	=	Structure-based virtual screening
SIFt	=	Structural interaction fingerprint
VS	=	Virtual screening

REFERENCES

[1] (**2000**) *Virtual Screening for Bioactive Molecules*, (Böhm, H.-J. and Schneider, G., Eds.), Wiley-VCH, Weinheim.

[2] Schneider, G. and Böhm, H.-J. (**2002**) *Drug Discov. Today*, *7*, 64–70.

[3] Waszkowycz, B. (**2002**) *Curr. Opin. Drug Discov.*, *5*, 407–413.

[4] Toledo-Sherman, L.M. and Chen, D. (**2002**) *Curr. Opin. Drug Discov. Dev.*, *5*, 414–421.

[5] Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K., Rejto, P.A., Arthurs, S., Colson, A.B., Freer, S.T., Larson, V., Luty, B.A., Marrone, T. and Rose, P.W. (**2002**) *J. Comput. Aided Mol. Des.*, *14*, 731–751.

[6] Stahl, M. and Rarey, M. (**2001***) J. Med. Chem.*, *44*, 1035–1042.

[7] Doman, T.N., McGovern, S.L., Witherbee, B.J., Kasten, T.P., Kurumbail, R., Stallings, W.C., Connolly, D.T., Shoichet, B.K. (**2002**) *J. Med. Chem.*, *45*, 2213–2221.

[8] Trosset, J.-Y., Dalvit, C., Knapp, S., Fasolini, M., Veronesi, M., Mantegani, S., Gianellini, L.M., Catana, C., Sundstrom, M., Stouten, P.F.W., Moll, J.K. (**2006**) *Proteins*, *64*, 60-67.

[9] Grüneberg, S., Stubbs, M.T. and Klebe, G. (**2002**) *J. Med. Chem.*, *45*, 3588–3602.

[10] Enyedy, I.J., Ling, Y., Nacro, K., Tomita, Y., Wu, X., Cao, Y., Guo, R., Li, B., Zhu, X., Huang, Y., Long, Y.Q., Roller, P.P., Yang, D. and Wang, S. (**2001**) *J. Med. Chem.*, *44*, 4313–4324.

[11] (**2001**) *Proteins*, *45* (S5), 1–199.

[12] Peitsch, M.C., Schwede, T., Diemand, A. and Guex, N. (**2002**) in *Current Topics in Computational Molecular Biology* (Jiang, T., Xu, Y. and Zhang, M.Q., Eds.). pp. 449–466. MIT Press, Cambridge, MA.

[13] Zimmer, R. and Lengauer, T. (**2002**) in *Bioinformatics – from Genomes to Drugs* (Lengauer, T., Ed.). pp. 237–313. Wiley-VCH, New York.

[14] Willis, R.C. (**2002**) *Mod. Drug Discov.*, *5*, 28–34.

[15] Sotriffer, C. and Klebe, G. (**2002**) *Farmaco*, *57*, 243–251.

[16] Bitetti-Putzer, R., Joseph-McCarthy, D., Hogle, J.M. and Karplus, M. (**2001**) *J. Comput. Aided Mol. Design*, *15*, 935–960.

[17] Abagyan, R. (**1992**) *J. Mol. Biol., 225*, 519–532.

[18] Smellie, A., Teig, S.L. and Towbin, P. (**1995**) *J. Comp. Chem.*, *16*, 171–187.

[19] (**1998**) *Tabu Search* (Glover, F.W., Ed). Kluwer Academic Publishers, Boston.

[20] Gohlke, H. and Klebe, G. (**2002**) *Angew. Chem. Int. Ed.*, *41*, 2644–2676.

[21] Ajay, Murcko, M.A. and Stouten, P.F.W. (**1997**) in *Practical Application of Computer-Aided Drug Design* (Charifson, P.S., Ed.). pp. 355-410. Marcel Dekker: New York.

[22] Muegge, I. and Martin, Y.C. (**1999**) *J. Med. Chem.*, *42*, 791–804.

[23] Gohlke, H., Hendlich, M. and Klebe, G. (**2000**) *J. Mol. Biol.*, *295*, 337–356.

[24] DeWitte, R.S. and Shakhnovich, E.I. (**1996**) *J. Am. Chem. Soc.*, *118*, 11733–11744.

[25] Mitchell, J.B.O., Laskowski, R.A., Alex, A., Forster, M.J., Thornton, J.M. (**1999**) *J. Comput. Chem.*, *20*, 1177–1185.

[26] Gohlke, H. and Klebe, G. (**2001**) *Curr. Opin. Struct. Biol.*, *11*, 231–235.

[27] Mark, A.E. and van Gunsteren, W.F. (**1994**) *J. Mol. Biol.*, *240*, 167–176.

[28] Williams, D.H., Maguire, A.J., Tsuzuki, W. and Westwell, M.S. (**1998**) *Science*, *280*, 711–714.

[29] Dill, K.A. (**1997**) *J. Biol. Chem*., *272*, 701-704.

[30] Reyes, C.M. and Kollman, P.A. (**2000**) *J. Mol. Biol.*, *297*, 1145–1158.

[31] Kuhn, B. and Kollman, P.A. (**2000**) *J. Am. Chem. Soc.*, *122*, 3909–3916.

[32] Wang, J.M., Morin, P., Wang, W. and Kollman, P.A. (**2001**) *J. Am. Chem. Soc.*, *123*, 5221–5230.

[33] Böhm, H.-J. and Stahl, M. (**1999**) *Med. Chem. Res.*, *9*, 445–462.

[34] Stahl, M. (**2000**) *Perspect. Drug Discov. Des.*, *20*, 83–98.

[35] Böhm, H.-J. (**1994**) *J. Comput.-Aided Mol. Des.*, *8*, 243–256.

[36] Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P. (**1997**) *J. Comput.-Aided Mol. Design*, *11*, 425–445.

[37] Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R. (**1996**) *J. Am. Chem. Soc.*, *118*, 3959–3969.

[38] Jones G., Willett, P. and Glen, R.C. (**1995**) *J. Mol. Biol.*, *245*, 43-53.

[39] Jones G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. (**1997**) *J. Mol. Biol.*, *267*, 727-748.

[40] Gehlhaar, D.K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.B., Fogel, L.J. and Freer, S.T. (**1995**) *Chem. Biol.*, *2*, 317-324.

[41] Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K., Reijto, P.A., Arthurs, S., Colson, A.B., Freer, S.T., Larson, V., Luty, B.A., Marrone, T. and Rose, P.W. (**2000**) *J. Comput.-Aided Mol. Des.*, *14*, 731-751.

[42] Rarey, M., Kramer, B., Lengauer, T. and Klebe G. (**1996**) *J. Mol. Biol.*, *261*, 470-89.

[43] Stahl, M. and Rarey, M. (**2001**) *J. Med. Chem.*, *44*, 1035-1042.

[44] Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J. (1998) *J. Comput. Chem.*, *19*, 1639-1662.

[45] Sotriffer, C.A., Ni, H. and McCammon, J.A. (**2000**) *J. Med. Chem.*, *43*, 4109-4117.

[46] Totrov, M. and Abagyan R. (**1999**) in *Proceedings of the Third Annual International Conference on Computational Molecular Biology* (Istrail, S., Pevzner. P. and Waterman, M., Eds.). pp. 312–320. ACM Press, New York.

[47] Welch, W., Ruppert, J. and Jain, A.N. (**1996**) *Chem. Biol.*, *3*, 449–462.

[48] Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E. (**1982**) *J. Mol. Biol.*, *161*, 269-288.

[49] Ewing, T.J.A. and Kuntz, I.D. (**1997**) *J. Comput. Chem.*, *18*, 1176–1189.

[50] Gschwend, D.A. and Kuntz, I.D. (**1996**) *J. Comput.-Aided Mol. Des.*, *10*, 123–132.

[51] Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. (**1996**) *J. Mol. Biol.*, *261*, 470–489.

[52] Âsterberg, F., Morris, G.M., Sanner, M.F. and Olson, A.J. (**2002**) *Proteins*, *46*, 34–40.

[53] Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R. (**1997**) *J. Mol. Biol.*, *267*, 727–748.

[54] Totrov, M. and Abagyan, R. (**1997**) *Proteins*, *S1*, 215–220.

[55] ICM-Dock: MolSoft, La Jolla, CA. http://www.molsoft.com/ docking.

[56] McMartin, C. and Bohacek, R.S. (**1997**) *J. Comput.-Aided Mol. Des.*, *11*, 333–344.

[57] Claußen, H., Buning, C., Rarey, M. and Lengauer, T. (**2001**) *J. Mol. Biol.*, *308*, 377–395.

[58] Luty, B.A., Wasserman, Z.R., Stouten, P.F.W., Hodge, C.N., Zacharias, M. and McCammon, J.A. (**1995**) *J. Comput. Chem.*, *16*, 454–464.

[59] McGann, M., Almond, H., Nicholls, A., Grant, J.A. and Brown, F. (**2003**) *Biopolymers*, *68*, 76-90.

[60] FRED: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com/products/applications/fred.html.

[61] Lorber, D.M. and Shoichet, B.K. (**1998**) *Protein Sci.*, *7*, 938–950.

[62] Shoichet, B.K., Leach, A.R. and Kuntz, I.D. (**1999**) *Proteins*, *34*, 4–16.

[63] Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., Shaw, D.E., Francis, P. and Shenkin, P.S. (**2004**) *J. Med. Chem.*, *47*, 1739–1749.

[64] Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., Frye, L.L., Pollard, W.T. and Banks, J.L. (**2004**) *J. Med. Chem.*, *47*, 1750–1759.

[65] Glide: Schroedinger, Portland, OR. http://www.schroedinger.com.

[66] Venkatachalam, C.M., Jiang, X., Oldfield, T. and Waldman, M. (**2003**) *J. Mol. Graph. Model.*, *21*, 289-307.

[67] Schnecke, V. and Kuhn, L.A. (**2000**) *Perspect. Drug Discov. Des.*, *20*, 171-190.

[68] Jain, A.N. (**2003**) *J. Med. Chem*., 46, 499-511.

[69] Böhm, H.-J. (**1992**) *J. Comput.-Aided Mol. Des.*, *6*, 61–78.

[70] Böhm, H.-J. (**1992**) *J. Comput.-Aided Mol. Des.*, *6*, 593–606.

[71] Klebe, G. (**1994**) *J. Mol. Biol.*, *237*, 221–235.

[72] Kramer, B., Rarey, M. and Lengauer, T. (**1999**) *Proteins*, *37*, 228–241.

[73] Böhm, H.-J. (**1994**) *J. Comput.-Aided Mol. Design*, *8*, 243–256.

[74] Jones, G., Willett, P. and Glen, R.C. (**1995**) *J. Mol. Biol.*, *245*, 43–53.

[75] Jones, G., Willett, P. and Glen, R.C. (**1995**) *J. Comput.-Aided Mol. Des.*, *9*, 532–549.

[76] Giordanetto, F., Cotesta, S., Catana, C., Trosset, J.-Y., Vulpetti, A., Stouten, P.F.W. and Kroemer, R.T. (**2004**) *J. Chem. Inf. Comput. Sci.*, *44*, 882–893.

[77] Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C. (**1995**) *J. Med. Chem.*, *38*, 2681–2691.

[78] Wang, T. and Wade, R.C. (**2001**) *J. Med. Chem.*, *44*, 961–971.

[79] Murcia, M. and Ortiz, A.R. (**2004**) *J. Med. Chem.*, *47*, 805–820.

[80] Gohlke, H. and Klebe, G. (**2002**) *J. Med. Chem.*, *45*, 4153–4170.

[81] Silber, K., Heidler, P., Kurz, T. and Klebe, G. (**2005**) *J. Med. Chem.*, *48*, 3547–3563.

[82] Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D.A. and Cheatham, T.E., III. (**2000**) *Acc. Chem. Res.*, *33*, 889–897.

[83] Wang, J., Kang, X., Kuntz, I.D. and Kollman, P.A. (**2005**) *J. Med. Chem.*, *48*, 2432–2444.

[84] Sitkoff, D., Sharp, K.A. and Honig, B. (**1998**) *J. Phys. Chem.*, *98*, 1978–1983.

[85] Rush, T.S., III, Manas, E.S., Tawa, G.J. and Alvarez, J.C. (**2005**) in *Virtual Screening in Drug Discovery* (Alvarez, J.C. and Shoichet, B.K., Eds.). pp. 249-277. Taylor & Francis, Boca Raton, FL.

[86] Blaney, J.M. (**2004**) Presentation at SMi workshop High-Throughput Molecular Docking, London, February 2004.

[87] Kuhn, B., Gerber, P., Schulz-Gasch, T. and Stahl, M. (**2005**) *J. Med. Chem.*, *48*, 4040–4048.

[88] Ferrara, P., Gohlke, H., Price, D.J., Klebe, G. and Brooks, C.L., III. (**2004**) *J. Med. Chem.*, *47*, 3032–3047.

[89] Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P. (**1999**) *J. Med. Chem.*, *42*, 5100–5109.

[90] Wang, R. and Wang, S. (**2001**) *J. Chem. Inf. Comput. Sci.*, *41*, 1422–1426.

[91] Terp, G.E., Johansen, B.N., Christensen, I.T. and Jørgensen, F.S. (**2001**) *J. Med. Chem.*, *44*, 2333–2343.

[92] Clark, R.D., Strizhev, A., Leonard, J.M., Blake, J.F. and Matthew, J.B. (**2002**) *J. Mol. Graph. Model.*, *20*, 281–295.

[93] Cotesta, S., Giordanetto, F., Trosset, J.-Y., Crivori, P., Kroemer, R.T., Stouten, P.F.W. and Vulpetti, A. (**2005**) *Proteins*, *60*, 629–643.

[94] Teague, S.J. (**2003**) *Nat. Rev. Drug Discov.*, *2*, 527–541.

[95] Murray, C.W., Baxter, C.A. and Frenkel, A.D. (**1999**) *J. Comput.-Aided Mol. Des.*, *13*, 547–562.

[96] Huse, M. and Kuriyan, J. (**2002**) *Cell*, *109*, 275–282.

[97] Sherman, W., Day, T., Jacobson, M.P., Friesner, R.A. and Farid, R. (**2006**) *J. Med. Chem.*, *49*, 534-553.

[98] Cavasotto, C.N. and Abagyan, R.A. (**2004**) *J. Mol. Biol.*, *337*, 209–225.

[99] Ferrari, A.M., Wei, B.Q., Costantino, L. and Shoichet, B.K. (**2004**) *J. Med. Chem.*, *47*, 5076–5084.

[100] Wei, B.Q., Weaver, L.H., Ferrari, A.M., Matthews, B.W. and Shoichet, B.K. (**2004**) *J. Mol. Biol.*, *337*, 1161–1182.

[101] Frimurer, T.M., Peters, G.H., Iversen, L.F., Andersen, H.S., Møller, N.P.H. and Olsen, O.H. (**2003**) *Biophys. J.*, *84*, 2273–2281.

[102] Erickson, J.A., Jalaie, M., Robertson, D.H., Lewis, R.A. and Vieth, M. (**2004**) *J. Med. Chem.*, *47*, 45–55.

[103] McGovern, S.L. and Shoichet, B.K. (**2003**) *J. Med. Chem.*, *46*, 2895–2907.

[104] Verdonk, M.L., Chessari, G., Cole, J.C., Hartshorn, M.J., Murray, C.W., Nissink, J.W.M., Taylor, R.D. and Taylor, R. (**2005**) *J. Med. Chem.*, *48*, 6504–6515.

[105] García-Sosa, A.T., Mancera, R.L. and Dean, P.M. (**2003**) *J. Mol. Model.*, *9*, 172–182.

[106] Goodford, P.J.A. (**1985**) *J. Med. Chem.*, *28*, 849–857.

[107] Clarke, C., Woods, R.J., Gluska, J., Cooper, A., Nutley, M.A. and Boons, G.-J. (**2001**) *J. Am. Chem. Soc.*, *123*, 12238–12247.

[108] Rarey, M., Kramer, B. and Lengauer, T. (**1999**) *Proteins*, *34*, 17–28.

[109] Raymer, M.L., Sanschagrin, P.C., Punch, W.F., Venkataraman, S., Goodman, E.D. and Kuhn, L.A. (**1997**) *J. Mol. Biol.*, *265*, 445-464.

[110] Vigers, G.P.A. and Rizzi, J.P. (**2004**) *J. Med. Chem.*, *47*, 80–89.

[111] McGann, M. (**2005**) Presentation at SMi meeting Drug Design, London, February 2005.

[112] DockIt: Metaphorics, Aliso Viejo, CA. http://www.metaphorics.com/products/dockit.

[113] Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W. and Taylor, R.D. (**2003**) *Proteins*, *52*, 609–623.

[114] Joseph-McCarthy, D., Thomas, B.E. IV., Belmarsh, M., Moustakas, D. and Alvarez, J.C. (**2003**) *Proteins*, *51*, 172–188.

[115] Ewing, T.J.A., Makino, S., Skillman, A.G. and Kuntz, I.D. (**2001**) *J. Comput.-Aided Mol. Des.*, *15*, 411–428.

[116] Hindle, S.A., Rarey, M., Buning, C. and Lengauer, T. (**2002**) *J. Comput.-Aided Mol. Des.*, *16*, 129–149.

[117] Deng, Z., Chuaqui, C., Singh, J. (**2004**) *J. Med. Chem.*, *47*, 337–344.

[118] Klon, A.E., Glick, M. and Davies, J.W. (**2004**) *J. Chem. Inf. Comput. Sci.*, *44*, 2216–2224.

[119] Cole, J.C., Murray, C.W., Nissink, J.W.M., Taylor, R.D. and Taylor, R. (**2005**) *Proteins*, *60*, 325–332.

[120] Westhead, D.R., Clark, D.E. and Murray, C.W. (**1997**) *J. Comput.-Aided Mol. Des.*, *11*, 209-228.

[121] Ha, S., Andreani, R., Robbins, A. and Muegge, I. (**2000**) *J. Comput.-Aided Mol. Des.*, *14*, 435-448

[122] Bissantz, C., Folkers, G. and Rognan, D. (**2000**) *J. Med. Chem.*, *43*, 4759-4767.

[123] Strynadka, N.C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B.K., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. and James, M.N. (**1996**) *Nat. Struct. Biol.*, *3*, 233-239.

[124] Schulz-Gasch, T. and Stahl, M. (**2003**) *J. Mol. Model.*, *9*, 47-57.

[125] Kontoyianni, M., Sokol, G.S. and McClellan, L.M. (**2005**) *J. Comput. Chem.*, *26*, 11-22.

[126] Kellenberger, E., Rodrigo, J., Muller, P. and Rognan, D. (**2004**) *Proteins*, *57*, 225–242.

[127] Perola, E., Walters, W.P. and Charifson, P.S. (**2004**) *Proteins*, *56*, 235-249.

[128] Kontoyianni, M., McClellan, L.M. and Sokol, G.S. (**2004**), *J. Med. Chem.*, *47*, 558-565.

[129] Kroemer, R.T., Vulpetti, A., McDonald, J.J., Rohrer, D.C., Trosset, J.-Y., Giordanetto, F., Cotesta, S., McMartin, C., Kihlen, M. and Stouten, P.F.W. (**2004**) *J. Chem. Inf. Comp. Sci.*, *44*, 871–881.

[130] Morley, S.D. and Afshar, M. (**2004**) *J. Comput.-Aided Mol. Des.*, *18*, 189–208.

[131] Verdonk, M.L., Berdini, V., Hartshorn, M.J., Mooij, W.T.M., Murray, C.W., Taylor, R.D. and Watson, P. (**2004**) *J. Chem. Inf. Comput. Sci.*, *44*, 793–806.

[132] Triballeau, N., Acher, F., Brabet, I., Pin, J.P. and Bertrand, H.-O. (**2005**) *J. Med. Chem.*, *48*, 2534–2547.

[133] Kubinyi, H. (**2006**) in *Computer Applications in Pharmaceutical Research and Development* (Ekins, S., Ed.). John Wiley, New York.

[134] Enyedy, I.J., Ling, Y., Nacro, K., Tomita, Y., Wu, X., Cao, Y., Guo, R., Li, B., Zhu, X., Huang, Y., Long, Y.-Q., Roller, P.P., Yang, D. and Wang, S. (**2001**) *J. Med. Chem.*, *44*, 4313-4324.

[135] Evers, A. and Klabunde T. (**2005**) *J. Med. Chem.*, *48*, 1088-1097.

[136] Elchebly, M., Payette, P., Michaliszyn, E., Cromlish, W., Collins, S., Loy, A.L., Normandin, D., Cheng, A., Himms-Hagen, J., Chan, C.C., Ramachandran, C., Gresser, M.J., Tremblay, M.L. and Kennedy, B.P. (**1999**) *Science*, *283*, 1544–1548.

[137] Møller, N.P., Iversen, L.F., Andersen, H.S. and McCormack, J.G. (**2000**) *Curr. Opin. Drug Discov. Dev.*, *3*, 527–540.

[138] Puius, Y.A., Zhao, Y., Sullivan, M., Lawrence, D.S., Almo, S.C. and Zhang, Z.Y. (**1997**) *Proc. Natl Acad. Sci. USA*, *94*, 13420–13425.

[139] Groves, M.R., Yao, Z.J., Roller, P.P., R Burke, T., Jr. and Barford, D. (**1998**) *Biochemistry*, *37*, 17773–17783.

[140] Grüneberg, S., Wendt, B. and Klebe, G. (**2001**) *Angew. Chem. Int. Ed.*, *40*, 389–393.

[141] Lyne, P.D., Kenny, P.W., Cosgrove, D.A., Deng, C., Zabludoff, S., Wendoloski, J.J. and Ashwell, S. (**2004**) *J. Med. Chem.*, *47*, 1962–1968.

Received: July 25, 2006 **Revised: December 03, 2006** **Accepted: December 04, 2006**