FABAD J. Pharm. Sci., 29, 81-92, 2004

SCIENTIFIC REVIEW

INVITED REVIEW

Making the Best Account of Molecular Docking in Drug Design

Pavel POSPISIL*°, Gerd FOLKERS*

Dr. Pavel Pospisil

Pavel Pospisil received his Bachelor of Science in Biochemistry from the Université Joseph Fourier in Grenoble, France in 1996 and his Masters in Biochemistry Engineering from the Institute of Chemical Technology in Prague, Czech Republic in 1999. He was awarded his Ph.D. in natural sciences from the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland in 2002. Dr. Pospisil completed his Postgraduate Fellowship at Arpida AG (Basel) and Pharmaceutical Chemistry ETH (Zurich), where his focused of research included applied docking of clinical trials phase I derivatives, tautomerism in drug design, and water molecule positioning during the docking process.

Dr. Pospisil was awarded the BAYER Award for Excellence in Computational Chemistry (Germany, 2003); the HLAVKA Award (Czech Republic, 1998), for the best thesis in applied biochemistry; and the SVOC Prize (3rd place) in a local competition of young researchers (ICT Prague, 1997).

Between 1999-2003, Dr. Pospisil has published numerous articles in the fields of biochemistry and pharmaceutical sciences, including his valuable research in molecular modeling and docking.

Making the Best Account of Molecular Docking in Drug Design

Summary

Computer-aided design of drugs has become an essential scientific field of pharmaceutical sciences. Automated industrial procedures of drug discovery have been accelerated by use of computers. Virtual screening of electronic chemical databases through a structure of an important target protein is an example of such an application. This article presents in a simple way the actual techniques and challenges of receptor-based molecular docking. Namely, the contemporary issues of scoring functions, flexibility of molecules and other problems owing to the specific active site environments are discussed.

Key Words : *Docking, energy of binding, database filtering, flexibility, crystal water molecules, crystal metal ions, covalent binding.*

Received : 21.06.2004

INTRODUCTION

In the last decade, molecular modeling has become a central technique in drug discovery. Faster computers, progress in programming, new modeling software, the World Wide Web and greater molecular visualization has brought modeling to practically every desktop computer. Parallel to this, in chemical laboratories and the pharmaceutical industry, new techniques have appeared. Among them are high-throughput screening (HTS) of chemical libraries, combinatorial chemistry and automatized crystallization of proteins and other drug targets. For example, HTS is a well-established method in lead finding among the compounds from the chemical library[1]. While HTS is a powerful technique allowing the testing of millions of compounds per day, its costs are very high, whereas the hit rate is relatively low[2].

* Institute of Pharmaceutical Sciences, Swiss Federal Institute of Technology, ETH Zurich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland.

° Corresponding author e-mail: pavel.pospisil@pharma.ethz.ch

Computer-aided drug design can reduce these expenses. Such a cost-cut is based on the hypothesis that it is not necessary to test the full chemical library but rather only a small set of compounds with suitable characteristics[3-5]. Under this premise, virtual screening is used to reduce the size and focus of the library to be tested. Virtual screening is a computational method that selects the most promising compounds – hits – from an electronic database[5]. The best of them becomes a lead structure if its binding to a given target is verified experimentally and can be further developed to a drug. Virtually, the chemical database of small compounds can be screened based on similarity search or fitting the known pharmacophore[6] – a so-called ligand-based screening. Screening of the compounds through the three-dimensional (3D) structure of the macromolecular target[7] is the receptor-based application of the virtual screening.

Since more and more 3D structures have become available due to X-ray crystallography, NMR spectroscopy and homology modeling, software tools using this information in drug design projects become more important[2-4]. Virtual screening has been mainly developed for docking of small molecules (generally called ligands) to proteins[5,7]. Docking can be understood as the placing of a small molecule into the active site of the protein and predicting the energetically favorable complex. There are several aspects of the docking issue that make the prediction of the ligand-protein complex more difficult. Among these aspects belong:

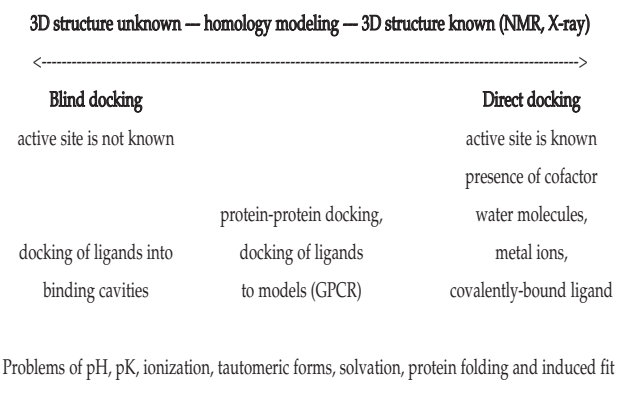
a) Accuracy of the so-called scoring function, i.e. the function (equation) used to calculate the binding affinity of the protein-ligand complex. Scientists calculate with several types of scoring functions, and each of them has certain strengths and weaknesses.

b) Docking and flexibility of molecules. Although the flexibility of small molecules is nowadays a standard procedure in docking, large macromolecular targets, as protein, membrane and DNA are still considered rigid in docking. Recently, new docking softwares simulating a certain flexibility of the proteins have been developed. Furthermore, approaches using molecular dynamic simulations are used to account for the flexibility and partial movements of protein domains.

c) Contemporary issues of molecular docking. Docking *per se* represents a system with the large number of degrees of freedom, among which can be counted the orientation, conformation and move of a small ligand, the relative orientation of the ligand and the protein, and the presence of water molecule, cofactor, metal ions and other molecules in the active site. Microenvironment of the active site on ligand binding is also one of the important aspects. For example, dielectric effect within the binding site, can change the protonation or the tautomeric state of molecules[8]. In addition, several specific aspects complicate molecular docking, such as: docking into the targets that are not well described or understood. This can be docking of the intercalators into DNA (dynamic and linear structure of DNA), docking of a covalently binding inhibitor (docking programs cannot simulate the creation of a covalent bond), docking on the membrane, docking into the transmembranal receptors or docking into the modeled proteins, in which the 3D structure is not known, etc.

Problematic aspects and contemporary issues of molecular docking can be seen in Figure 1.

Figure 1. Contemporary issues of molecular docking



Scoring function

Scoring function answers the question: what is the ligand's free energy of binding (ΔG) to the target? The scoring function is a fundamental element of docking software and estimates the affinity of the compound for the given protein. Computer-assisted docking methods are far from accurate in calculating ΔG unless the advanced and time-consuming techniques are used, among them for example the free energy perturbation or dynamic simulation of the ligand-protein complex. However, virtual screening requires fast docking and thus, the free energy of binding is calculated with simple functions[9]. The calculation of ΔG is simplified, and the enthalpy-entropy compensation, water solvation before and after the binding and conformational changes in the protein are omitted or very roughly approximated[7]. There are three types of scoring functions: force field scoring, empirical scoring and knowledge-based scoring, also called PMF (potentials of mean force).

Force field scoring function is based on non-bonded terms of a classical molecular mechanics. A Lennard-Jones potential describes Van der Waals interactions and Coulomb energy the electrostatic interactions. Making the sum of these potentials over all atom-atom interactions first requires a pre-calculated grid of points within the active site, for every point of which the energy is calculated. The grid maps the active site and contains all energetic information on the functional groups that may be involved in binding. The program DOCK[10,11] is an example of a program in which the energy scoring component is force field based. While force field scoring performs well on selecting docking modes, the comparison of different molecules is difficult because the entropic and solvation effects are not considered.

Empirical scoring is based on known measured free binding energies correlated to geometric parameters of the protein-ligand complex. From the set of protein-ligand complexes, for which both the binding affinity and 3D structure are known, the multiple linear regression or partial least squares is plotted. This is used to optimize the coefficients, to weigh the computed terms and to construct the function, also called master equation[12]. Terms of the master equation represent different physico-chemical effects contributing to the binding energy. It includes polar interactions as hydrogen bonds and ionic interactions, and non-polar interactions as lipophilic contact and aromatic interactions. In addition, entropic term is simulated by the flexibility of the ligand (number of rotatable bonds). The first function of this type was developed by Böhm for the de novo design program LUDI[13-15]. The empirical scoring function is implemented in the program FlexX.

The knowledge-based scoring function was originally developed for protein structure prediction and has been successfully applied[16-18]. The principle of this function is that situations frequently seen in 3D structures are energetically favorable. Potential of mean force encodes structural information from X-ray coordinates of protein-ligand complex into Helmholtz free interaction energies of protein-ligand atom pairs. For each atom pair, the number of occurrences is counted depending on the distance. The advantage of the knowledge-based function is that there is no need for experimentally measured free energies, as is the case in empirical scoring, and that solvation and entropic terms are treated implicitly.

The single scoring function has both strengths and weaknesses in predicting ΔG. Some of the weaknesses are the over/underestimation of hydrophobic interactions in relation to the electrostatic ones. Some of the scoring functions are more suited for the scoring of the databases containing more or less diverse compounds; others perform better for the ones with large number of interaction sites. Could it help to use several scoring functions and to average the results? "Consensual" scoring, therefore, refers to the use of several scoring functions. For a given docking orientation of a compound (whatever algorithm is used), several master equations are calculated[19]. One of the popular consensual scorings is CScore™[19]. CScore includes five scoring functions, F, G, D, Chem and PMF, originating from docking

tools FlexX[20], Gold[21], Dock[10,11], Chemscore[22] and Potential Mean Force[17], respectively. F score, G score and Chemscore are empirical scoring functions, D is a force-field based function, and PMF is knowledge-based potential of mean force. CScore ranks from 0 to 5, with 5 representing a compound ranked in the top 5% of all scorings normalized.

Docking and flexibility of molecules

Docking programs differ in several aspects, such as the algorithm used to place the ligand in the active site, the type of the scoring function and the flexibility of the ligand and the protein. Certainly, there are other differences, such as the possibility to choose the parameters, the user-friendliness of the program, and the critical parameter to the industry: the run-time per molecule. One of the recently discussed aspects with a high impact on docking results is the flexibility of molecules[7]. From the point of view of algorithms, docking programs can be separated as rigid-body docking algorithms and flexible ligand docking algorithms.

Flexibility of ligands. Rigid-body docking algorithms have been historically the first approaches for screening. The protein as well as ligand is fixed in the conformational space and the docking consists in matching of orientations. Main rigid-body approaches are: clique-search based approach, geometric hashing, pose clustering and superposition of point sets[9]. Clique-search based method was the first and until today most widely used software tool implemented in DOCK. The idea is based on searching distance-compatible matches. A set of spheres is created as the reverse image of the active site of the protein and the ligand spheres or directly atoms are matched to this image. The ligand position is further optimized and scored. DOCK also evolutes in several mixed docking algorithm and scoring approaches[21, 23-25], for example by chemical score, which is based on labeling of spheres by chemical properties[26], or the matching of several ligand conformations to one or several protein conformations in order to imitate the flexibility of the system. All mentioned rigid-body docking algorithms based on matching, pattern recognition and superposition are reviewed in *Bioinformatics - From Genomes to Drugs* [9].

Flexible ligand docking algorithms, in contrast to rigid-body, consider several low energy conformations of the ligand. Since drugs are smaller than proteins, docking of flexible ligands is computed faster. Flexible algorithms can be summarized as follows: use of conformational ensembles, docking based on fragmentation of the ligand, place and join algorithms, incremental construction algorithms, genetic algorithms, distance geometry, taboo (or tabu) search and random searches[9,27]. In addition to these basically combinatorial optimization algorithms, there are approaches tackling the problem by simulation techniques, meaning that a calculation starts with a certain ligand conformation, which is then locally moved towards lowest energy binding modes. Along with these, we include docking using molecular dynamics, simulated annealing and Monte-Carlo simulations[7].

Flexibility of proteins. Fully flexible docking programs that treat both ligand and protein as flexible molecules have appeared recently. For a long time, it was computationally impossible to take into account the protein flexibility. Development of these docking algorithms was aided by the exponential growth of computer speed, RAM and disk capacity.

Depending upon the extent of the movement, flexibility can be divided into three groups. First, only motion involving side-chains and some backbone atoms, which is common in the active site of the protein. Second, the hinge-bending movements of joint domains, which are believed to allow the "induced fit" of the ligand. Third is the renaturation upon ligand binding. However, as the ligand stabilizes the complex, only conformational changes of side-chains of the active site and the hinge-bending movements of the protein can be considered.

Protein flexibility is handled via several techniques, some of which are mentioned here: using the rotamer library of side-chain[26,28,29], free movement of domains[30] (but still kept rigid)[28,29], and optimizati-

on after docking between the side-chains and the placed ligand[31,32]. Other programs include the flexibility as an ensemble of several overlapping protein conformations. Knegtel et al.[33] used a composite grid from multiple crystal and NMR structures of protein-ligand complexes. FlexE uses unified protein description from superimposed structures, where only dissimilar areas are separately treated and rotamery library of moving side-chains is incorporated[26].

Finally, with the advances of proteomics and structural genomics, flexibility of proteins in docking is becoming imperative[4]. Schafferhans et al.[34] developed the approach named DragHome to dock ligands into approximate models, combining homology modeling and data from 3D-QSAR of ligands. Wojciechowski et al.[35] treated the flexibility of both protein and ligand using their approximate models and matched them by means of their sterical and chemical complementarity.

Program algorithms. Performance of programs depends on the algorithm used for docking. Programs such as DOCK and FlexX use incremental construction algorithm. In program DOCK, reverse image of the site is filled with a minimum set of overlapping spheres. The ligand is partitioned and a single anchor fragment of the ligand is matched to spheres of the site. The anchor fragment is scored in terms of the interactions to protein, and the best anchor is used for growing of the ligand. Several examples of DOCK testing of the efficacy and selectivity have shown that Dock version 4.0 is fast and effectively able to prioritize molecules from large database screening[36], which can be further used for mid-sized database screening. FlexX is a fully automated approach developed for screening of chemical databases. Similar to DOCK, FlexX divides the ligand along its rotational bonds into fragments, docks first a largest fragment and then reattaches the remaining fragments[37]. However, in contrast to DOCK, FlexX places the base fragment rather than on matching the shape-based points on interaction points. This allows the handling of much smaller fragments, down to the size of single functional groups. FlexX is suitable for screening larger databases containing thousands of compounds as well as the single ligand docking.

AutoDock is a program based on the Lamarckian genetic algorithm that mimics the evolution process while optimizing the docking[25]. As in evolution where individuals change due to the genetic information, crossing of chromosomes, mutations of genes and influence from the environment, genetic algorithms use ligand conformations as individuals, and decide which survive and produce offspring. Each chromosome encodes a possible ligand-protein complex conformation. A chromosome is assigned a fitness function, which is closely related to scoring functions for molecular docking with one extension, the genotype-to-phenotype conversion. AutoDock contains also local search, which is based on simulation annealing and is used for the optimization of the ligand conformation in the complex with the protein. Genetic algorithms are now common for searching conformational space and find their use even in virtual screening of larger databases. The docking program GOLD[21] is an example.

Databases and their filtering. Preparation of the chemical database consists in retrieving suitable compounds from the large chemical database and formatting them appropriately for a screening program. Databases typically contain only the 2D structural formulae of the individual molecules. A commonly used database is ACD (Available Chemicals Directory), which is the most current comprehensive structure searchable database, with over 400,000 research-grade and bulk chemicals[38]. Besides the structural information stored in 2D format, it contains additional information, such as address of the supplier and price. Another commonly used database is NCI Database from the National Cancer Institute of NIH[39] (US) with over 250,000 compounds, which can be downloaded free of charge. ChemNavigator is an example of a commercialized database containing more than four million chemical structure records using the iResearch System[40]. Furthermore, there are in-house databases of pharmaceutical companies that are usually kept confidential and are

the result of years of ad-hoc synthesis, enlarged by compounds derived from combinatorial chemistry. Compound libraries used in lead finding programs can be generally reduced in size by filtering the unsuitable compounds which, based on experience, would not pass clinical trials due to undesired properties. We can use different filters to exclude these "non-drug-like" compounds and keep only those that resemble drugs. The Lipinski rule of five is the well known method of evaluating the drug-likeness profile of the molecules[41]. Lipinski et al. suggest that poor absorption or permeation is more likely when the molecular weight is over 500, the calculated octanol/water partition coefficient (clogP) is higher than 5, and when there are more than 10 H-bond acceptors and more than 5 H-bond donors. Often, this rule is misinterpreted and it is incorrectly assumed that the molecule cannot be drug-like if one of these rules is broken. Lipinski claims that if a compound has more than two of these mentioned properties it shows poor permeability and thus should be removed from the database[41].

An additional filter consists in the removal of highly reactive and toxic compounds according to their reactive moieties, such as acyl-halides, sulfonyl-halides, Michael acceptors, etc[42]. Nevertheless, filtering also depends on the purpose of the screening. In our opinion, it is useful to keep compounds with some of these reactive groups in virtual screening, because some hits of the screening may give a valid indication towards the lead structure finding. Finally, other filters for specific absorption, distribution, metabolism, and excretion properties (ADME), such as filters for prediction of aqueous solubility[43], membrane permeation[44] and metabolic clearance[45], are being developed. Furthermore, a universal filter that automatically distinguishes between drugs and chemicals has been designed from known databases. This filtering assigns to compounds a drug-likeness score. Approaches have been published that are based either on neural networks[46], genetic algorithms[47], decision trees[48] or pharmacophoric description[49].

Once the database is retrieved in 2D, the 3D structure of the molecules has to be generated. For this step, powerful tools like Concord[50] and Corina[51] exist. Both programs are based on known geometry of atoms and generate three-dimensional atomic coordinates. They process a large variety of structural data file formats, e.g. CORINA uses MDL SDFile, SMILES linear code, SYBYL MOLFILE and MOL2, and PDB file formats[52]. Programs consider the stereochemistry of the compounds. In addition to automatically generated databases, it is important to add known reference ligands. They can be drawn manually or, if their binding to the protein is structurally known, they can be retrieved from the Protein Data Bank (PDB)[52], the standard databank of available 3D structures of protein-ligand complexes. Ligands from PDB serve as a reference for docking accuracy. Ligands known from the literature and with unknown binding complex with the macromolecule are manually drawn. Similarly, when preparing the 3D compounds to be included in virtual screening, it is appropriate to check their 3D structure conformation in the Cambridge Structural Database (CSD)[53]. We would like to stress that laboratory experimental results determining the characteristics and activity of the proteins should be presented side by side with the structural data.

Contemporary issues of molecular docking

Docking is targeted not only to proteins. Drugs are small molecules interacting with macromolecules as proteins, nucleic acids, polysaccharides and lipophilic membranes. Some pioneering studies using molecular docking are performed on these targets, e.g. those of Chen et al. and Louch et al.[54,55]. Another growing field is the docking to transmembranal proteins as G-protein coupled receptors (GPCR) or ion channels[56,57]. Determination of 3D-structures of receptors anchored in the membrane is difficult; rhodopsin is the first low-resolution structure of a GPCR, solved in 2000[58]. Models of GPCR built by specific homology modeling of receptors can serve as docking targets, and these modeled targets have become a new field of genomics and proteomics.

However, the main docking studies are performed

on cytoplasmic enzymes. These proteins represent a majority of molecules stored in the PDB. From the point of view of knowledge of the active site, the issues of docking can generally be separated according to the following sections.

Blind docking. When the active site of the protein is not known, the search for both the binding site and the binding mode of the ligand is called a "blind docking"[59]. Development in blind docking is also valuable for search of protein-protein interactions, when both macromolecules are supposed to complement together by their shape and intermolecular interactions. As an example of blind docking, we can mention the work of Hetenyi et al., who tested the docking of peptides to proteins (peptidases and proteases) without prior knowledge of the binding site or the binding mode of the ligand[59]. In blind docking, it is assumed that lowest-energy conformation with the evident shape complementarity is the correct structure[59]. They used the program AutoDock, which proved to be efficient in finding the binding pockets for protein-peptide complexes[59].

The blind docking application can also be used in cases when the structure of the targeted protein is not resolved and the homology model of the protein has to be built. In actual time, the number of protein sequences that can be modeled is increasing steadily because of the growth in the number of known protein structures (June 2004 – 25,960)[52], and the accuracy of the predictions is improving because of the improvements in the modeling software[60]. In the model, we assume that the active site is surrounding the ligand, similar to the template molecule. Nevertheless, it does not mean that the site should be at the same location and of the same form as in the template. It is suitable to submit such a model to molecular dynamic simulation and, if possible, to use the docking which treats the protein as flexible.

If the 3D-structure of the protein is not known, a model may be built by means of homology modeling. It consists usually in the structural alignment of the protein of interest to the resolved 3D-structure of the isoenzyme isolated from another species or organ. Higher similarity leads to a more reliable model. 30-50% of sequence identity is considered enough for appropriate determination of the binding site (medium accuracy). Models built on more than 50% of sequence identity (high accuracy) approach low resolution X-ray structures (3Å resolution) or medium resolution NMR structures (10 distance restraints per residue)[60,61]. High-accuracy models can be used directly for docking of small ligands unless the medium-accuracy model is built and refined by molecular dynamic (MD) simulation[62]. Final conformation is obtained after MD simulation by energy optimization of the averaged structure retrieved over the stable dynamic plateau. The final model reveals the fold of the protein backbone, the position of the side-chains of key amino acid residues, and the position of water molecules in the active site, while the analysis of the trajectories shows the hinge-bending movements and the range of flexibility of the side-chains.

In the post-genomic era, models are needed and docking procedures will have to rely on them. Recently, scientists have begun building a variety of models with, for example, enzymes with or without known substrates (orphan target), proteins involved in signaling pathways, messengers, and transporters, and models of ions channel or G-protein coupled receptors[57,63]. Also, automatic creation of models is on run, and the database MODBASE containing annotated comparative protein structure models is steadily increasing[64], counting approximately 500,000[64] models.

Direct docking. When the active site is known, the situation becomes easier. If the crystallized enzyme or receptor is complexed with the ligand, the site of docking is the pocket surrounding the ligand. The challenge is to determine the binding mode of a new ligand. It is essential to understand the catalytic process carried out by the enzyme or the function of the receptor. Crystal structures often contain water molecules in the active site. Concrete discrete waters may mediate the binding of the ligand to the enzyme, or may directly participate as a co-substrate molecule on the catalysis. When running virtual scre-

ening, it is a dilemma as to whether or not the water molecules should be removed. Many docking manuals recommend the removal of water molecules from the site. The screening is mainly performed to an empty, water-devoid active site. However, attempts to screen the site with the water molecules included in the site[59,65,66] have resulted in a positive contribution to the binding prediction.

The limitation of docking programs is the neglect of solvation effect or use of solvent models in a snapshot method where docking poses are first generated in vacuo and then ranked with a scoring function that includes a solvent model. The search function of docking favors therefore the *in vacuo* conformations. Furthermore, the bound solvent molecules are not considered, yet in the HIV-1 protease[67], HSV1 thymidine kinase[65] or heat-labile and cholera toxin[66], for example, explicit waters play an important role in ligand binding. It can be noticed that waters can be included in the docking also upon the grid-based calculations, using the probe water molecule, as it can be done by program GRID[68]. GRID finds for a given probe atom or group of atoms its favorable positions.

Waters. Waters included in the docking procedure can be positioned at the concrete place in the binding site and as a part of the protein. This is possible in many docking programs (FlexX, Gold, Dock, AutoDock). In AutoDock, for example, parameters for waters (position of hydrogens and partial charges of the atoms) can be defined by the user as, for example, token from AMBER[69]. In FlexX[70], presence of water molecules can be defined in two ways: the user can select from the FlexX menu to "customize" as either concrete waters in the site or as the "particle concept"[70]. In the particle concept, waters are positioned during the screening (during the complex construction phase) in the site between the ligand and the protein. This concept has not led to significant improvement of the docking predictions[70], but it may change in the future. FlexX is also an example of a docking program with its own parameters, which can be modified according to the needs of the user. Another example of a program considering water during docking is SLIDE[32,71]. Waters are token from the crystal structure of the ligand-free protein, and the program Consolv[72] is used to predict if the waters are likely to be conserved upon ligand docking and mediate interactions between the two molecules[73]. Only those waters that are predicted as being conserved are kept, and there is a penalty coupled to the confidence of the prediction, assessed for displacement of waters that are predicted as conserved. The fact that we can keep discrete water molecules in the site as a part of the docking target can lead to new unique hits in a large database screening.

Presence of metal ions in the active site. Human PDE4 is an enzyme regulating the level of cyclic nucleotides in many physiological processes. PDE4 is a representative example for the combination of docking problems. First, the site is known but devoid of the ligand, and second, the water molecules and metals are present.

Many enzymes need for their catalysis the presence and participation of metal ions (Zn2+, Mg2+...) (kinases, cytochrome P450-based enzyme, metalloproteases, etc.). Metal ions are not parameterized for different metal coordination in the site. The bonding around metals is more varied than two or four bonds in an organic molecule, and it is also more than one geometrical arrangement, which is possible around one metal with a given number of ligands. Another problem encountered in the metal system is the lack of well-defined bonds.

With respect to the treatment of metal coordination, docking programs are clearly deficient. Nevertheless, using some approximation, they are sufficient for molecular docking. For example, recognizing metals like zinc or magnesium and attributing them the charge 2+ is a basic approximation. Metals corresponding to the artifact of the crystallization have to be removed from the docking and metals known as essential for catalysis may be kept, depending on their role in binding the ligand. It was found that metal deletion is essential for binding of imidazole derivatives (e.g. miconazole) to lanosterol demetyla-

se (cytochrome P450-based enzyme) and its inhibition[74]. The metal is not essential for binding of tyrosine kinase inhibitors such as Gleevec™[75]. If no experimental information is available, calculated affinities compared to the experimental can be a good hint to decide upon the method of screening, e.g. with or without waters and metals.

Docking of the covalently bound inhibitor. The issue of covalently binding ligands is that docking programs are not built for this purpose. Docking programs are based on molecular mechanics for non-covalent, intermolecular interactions. There is a missing drive for highly energetic creation of the covalent bond.

On the γ-chymotrypsin protein, Hetenyi et al. showed the docking of the covalently bound ligand[59]. With a certain success, they were able to reproduce the identification of the binding pocket for the tripeptide Gly-Ala-Trp. Due to only one fragment of the tripeptide – the aromatic side-chain of tryptophan – program AutoDock positioned the tryptophan into the hydrophobic pocket, the most important part of the protein. The binding location was found even when only a part of the ligand or of the site was defined properly. The rest of the molecule could be localized as the tail-part of the ligand and as its flexibility allowed. However, the position of the bond and atoms, which have to be cleaved and create later the covalent bond with the protein, is at best, at its non-bonded intermolecular distance.

Another example concerns searching for the inhibitors of methionyl-aminopeptidases (MetAPs). MetAPs catalyze the hydrolytic cleavage of the starter methionine of newly synthesized polypeptides and proteins. Inhibitors of the bacterial MetAP and several analogs, which have all been derived from the natural product fumagillin, contain an epoxide moiety, which forms a covalent bond to the enzyme[76]. The case of the docking to MetAP is a hard task in two respects: need for proper active site elucidation, and the covalent, irreversible binding of the inhibitors (Klein et al., personal communication). MetAPs are metal-dependent enzymes, with metal cations present in the active site. The understanding of the catalytic mechanism of the enzyme, particularly the function of the metal ion, is a prerequisite to any docking. Because the irreversible inactivation is a very desirable mode of action of enzyme inhibitors - as long as high selectivity is given - the reactivity of the epoxide group in these molecules must be finely tuned to "fit" the MetAP active site and eventually make a covalent bond to one of the active site histidines.

CONCLUDING REMARKS

Modern docking programs are efficient in searching for the active site and in docking flexible ligands. Docking into models of homologous proteins, which has become a vital field in the post-genomic era, can prove the accuracy of the model and lead quickly and efficiently to putative drug candidates. If water, metals and cofactors are present in the binding site of the macromolecular target, they have to be considered in docking based upon their physiological role. Position of these elements in the docking site can be read from the 3D-structure or can be calculated before the docking process. Although the parameterization of the metals in the active site and the position of favorable water molecules for binding are unknown, modern docking programs allow several simplified methods of considering these elements in enzymatic reaction.

It is worth noting that visual inspection of the results of virtual screening remains imperative. As all docking and screening projects show, one cannot avoid looking at the predicted 3D complexes to ascertain that the ligand has been truly docked in the binding site and interacts with key residues of the active site. Experimental measurements of binding and mutagenetic studies of the target protein should be in accordance with the results of structure-based predictions of binding. Interestingly, one forgets that manual docking and moving of the ligand in the active site, as the chemical and modeler intuition tell us, is still a valuable tool of ligand-protein complex prediction.

REFERENCES

1. Mander T. Beyond uHTS: ridiculously HTS?, *Drug Discovery Today*, 5, 223-225, 2000.
2. Lahana R. How many leads from HTS?, *Drug Discovery Today*, 4, 447-448, 1999.
3. Schneider G, Bohm HJ. Virtual screening and fast automated docking methods, *Drug Discovery Today*, 7, 64-70, 2002.
4. Waszkowycz B, Perkins TDJ, Sykes RA, Li J. Large-scale virtual screening for discovering leads in the postgenomic era, *Ibm Systems Journal*, 40, 360-376, 2001.
5. Walters WP, Stahl MT, Murcko MA. Virtual screening - an overview, *Drug Discovery Today*, 3, 160-178, 1998.
6. Pickett SD, McLay IM, Clark DE. Enhancing the hit-to-lead properties of lead optimization libraries, *J. Chem. Inf. Comput. Sci.*, 40, 263-272, 2000.
7. Halperin I, Ma BY, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions, *Proteins-Structure Function and Genetics*, 47, 409-443, 2002.
8. Pospisil P, Ballmer P, Scapozza L, Folkers G. Tautomerism in computer-aided drug design, *J. Recept. Signal Transduct. Res.*, 23, 361-371, 2003.
9. Lengauer T. *Bioinformatics - From Genomes to Drugs*, Wiley CH ed., Wiley-VCH, Weinheim, 442, 2002.
10. http://www.cmpharm.ucsf.edu/kuntz/ DOCK, 2002.
11. Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, *J. Comput. Aided Mol. Des.*, 15, 411-428, 2001.
12. Rognan D, personal communication, 2000.
13. Bohm HJ. On the use of Ludi to search the fine chemicals directory for ligands of proteins of known 3-dimensional structure, *J. Comput. Aided Mol. Des.*, 8, 623-632, 1994.
14. Bohm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure, *J. Comput. Aided Mol. Des.*, 8, 243-256, 1994.
15. Bohm HJ. The computer-program Ludi - a new method for the de novo design of enzyme-inhibitors, *J. Comput. Aided Mol. Des.*, 6, 61-78, 1992.
16. Mitchell JBO, Laskowski RA, Alex A, Thornton JM. BLEEP - potential of mean force describing protein-ligand interactions: I. generating potential, *J Comput Chem*, 20, 1165-1176, 1999.
17. Muegge I, Martin YC. A general and fast scoring function for protein-ligand interactions: a simplified potential approach, *J. Med. Chem.*, 42, 791-804, 1999.
18. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions, *J. Mol. Biol.*, 295, 337-356, 2000.
19. http://www.tripos.com/sciTech/inSilicoDisc/virtual Screening/cscore.html CScore, 2002.
20. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.*, 261, 470-489, 1996.
21. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.*, 267, 727-748, 1997.
22. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions. 1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, *J. Comput. Aided Mol. Des.*, 11, 425-445, 1997.
23. http://www.cmpharm.ucsf.edu/kuntz/ FRED docking program.
24. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles, *Protein Sci*, 7, 938-950, 1998.
25. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, *J Comput Chem*, 19, 1639-1662, 1998.
26. Claussen H, Buning C, Rarey M, Lengauer T. FlexE: efficient molecular docking considering protein structure variations, *J. Mol. Biol.*, 308, 377-395, 2001.
27. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods, *J. Comput. Aided Mol. Des.*, 16, 151-166, 2002.
28. Leach AR. Ligand docking to proteins with discrete side-chain flexibility, *J. Mol. Biol.*, 235, 345-356, 1994.
29. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm, *Proteins-Structure Function and Genetics*, 33, 227-239, 1998.
30. Sandak B, Wolfson HJ, Nussinov R. Flexible docking allowing induced fit in proteins: insights from an open to closed conformational isomers, *Proteins*, 32, 159-174, 1998.
31. Schnecke V, Swanson CA, Getzoff ED, Tainer JA, Kuhn LA. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility, *Pro-*

teins-Structure Function and Genetics, 33, 74-87, 1998.
32. Schnecke V, Kuhn LA. Virtual screening with solvation and ligand-induced complementarity, *Perspect. Drug Discov. Design*, 20, 171-190, 2000.
33. Knegtel RMA, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures, *J. Mol. Biol.*, 266, 424-440, 1997.
34. Schafferhans A, Klebe G. Docking ligands onto binding site representations derived from proteins built by homology modelling, *J. Mol. Biol.*, 307, 407-427, 2001.
35. Wojciechowski M, Skolnick J. Docking of small ligands to low-resolution and theoretically predicted receptor structures, *J. Comput Chem*, 23, 189-197, 2002.
36. Knegtel RM, Wagener M. Efficacy and selectivity in flexible database docking, *Proteins*, 37, 334-345, 1999.
37. Rarey M, Kramer B, Lengauer T. Time-efficient docking of flexible ligands into active sites of proteins, *Ismb*, 3, 300-308, 1995.
38. http://www.library.wisc.edu/help/quickguide/acd.htm ACD.
39. http://www.nci.nih.gov/ NCI NIH, 2002.
40. http://www.chemnavigator.com/cnc/chemInfo/iResearchSys.asp iResearch System, 2002.
41. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.*, 23, 3-25, 1997.
42. Oprea TI. Property distribution of drug-related chemical databases, *J. Comput. Aided Mol. Des.*, 14, 251-264, 2000.
43. Huuskonen J, Rantanen J, Livingstone D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices, *Eur J Med Chem*, 35, 1081-1088, 2000.
44. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, *J. Med. Chem.*, 43, 3714-3717, 2000.
45. Zuegge J, Schneider G, Coassolo P, Lave T. Prediction of hepatic metabolic clearance - comparison and assessment of prediction models, *Clin. Pharmacokinet.*, 40, 553-563, 2001.
46. Sadowski J, Kubinyi H. A scoring scheme for discriminating between drugs and nondrugs, *J. Med. Chem.*, 41, 3325-3329, 1998.
47. Gillet VJ, Willett P, Bradshaw J. Identification of biological activity profiles using substructural analysis and genetic algorithms, *J. Chem. Inf. Comput. Sci.*, 38, 165-179, 1998.
48. Wagener M, van Geerestein VJ. Potential drugs and nondrugs: prediction and identification of important structural features, *J. Chem. Inf. Comput. Sci.*, 40, 280-292, 2000.
49. Muegge I, Heald SL, Brittelli D. Simple selection criteria for drug-like chemical matter, *J. Med. Chem.*, 44, 1841-1846, 2001.
50. http://www.tripos.com/sciTech/inSilicoDisc/chemInfo/concord.html Concord, 2002.
51. Sadowski J, Gasteiger J. From atoms and bonds to 3-dimensional atomic coordinates - automatic model builders, *Chem. Rev.*, 93, 2567-2581, 1993.
52. http://www.rcsb.org Protein Data Bank, 2004.
53. http://www.ccdc.cam.ac.uk/prods/csd/csd.html Cambridge Structural Database, 2002.
54. Chen IJ, Neamati N, MacKerell AD Jr. Structure-based inhibitor design targeting HIV-1 integrase, *Curr Drug Targets Infect Disord*, 2, 217-234, 2002.
55. Louch HA, Buczko ES, Woody MA, Venable RM, Vann WF. Identification of a binding site for ganglioside on the receptor binding domain of tetanus toxin, *Biochemistry*, 41, 13644-13652, 2002.
56. Schapira M, Raaka BM, Samuels HH, Abagyan R. In silico discovery of novel retinoic acid receptor agonist structures, *BMC Struct Biol*, 1, 1, 2001.
57. Schapira M, Raaka BM, Samuels HH, Abagyan R. Rational discovery of novel nuclear hormone receptor antagonists, *Proc Natl Acad Sci U S A*, 97, 1008-1013, 2000.
58. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. Crystal structure of rhodopsin: a G protein-coupled receptor, *Science*, 289, 739-745, 2000.
59. Hetenyi C, Van Der Spoel D. Efficient docking of peptides to proteins without prior knowledge of the binding site, *Protein Sci*, 11, 1729-1737, 2002.
60. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes, *Annu Rev Biophys Biomol Struct*, 29, 291-325, 2000.
61. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3, *Proteins*, Suppl 1, 50-58, 1997.
62. Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE. Structure-based inhibitor de-

sign by using protein models for the development of antiparasitic agents, *Proc Natl Acad Sci U S A*, 90, 3583-3587, 1993.

63. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A. Protein structure modeling for structural genomics, *Nat Struct Biol*, 7 Suppl, 986-990, 2000.
64. Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A. MODBASE, a database of annotated comparative protein structure models, *Nucleic Acids Res*, 30, 255-259, 2002.
65. Pospisil P, Scapozza L, Folkers G. The role of water in drug design: thymidine kinase as case study, *Rational approaches to drug design: 13th European Symposium on Quantitative Structure-Activity Relationship*, Prous Science, Barcelona-Philadelphia, 92-96, 2001.
66. Minke WE, Diller DJ, Hol WG, Verlinde CL. The role of waters in docking strategies with incremental flexibility for carbohydrate derivatives: heat-labile enterotoxin, a multivalent test case, *J. Med Chem*, 42, 1778-1788, 1999.
67. Lam PY, Jadhav PK, Eyermann CJ, Hodge CN, Ru Y, Bacheler LT, Meek JL, Otto MJ, Rayner MM, Wong YN, et al. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors, *Science*, 263, 380-384, 1994.
68. http://www.moldiscovery.com/index.html *GRID20*; 20 ed.
69. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz K, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J Am Chem Soc*, 117, 5179-5197, 1995.
70. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking, *Proteins-Structure Function and Genetics*, 37, 228-241, 1999.
71. http://www.bch.msu.edu/labs/kuhn/web/projects/slide/home.html SLIDE, 2002.
72. http://www.bch.msu.edu/labs/kuhn/web/software/consolv/doc.html Consolv, 2002.
73. Raymer ML, Sanschagrin PC, Punch WF, Venkataraman S, Goodman ED, Kuhn LA. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm, *J Mol Biol*, 265, 445-464, 1997.
74. Wright GD, Honek JF. Effects of iron binding agents on Saccharomyces cerevisiae growth and cytochrome P450 content, Can. *J. Microbiol.*, 35, 945-950, 1989.
75. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase, *Science*, 289, 1938-1942, 2000.
76. Liu S, Widom J, Kemp CW, Crews CM, Clardy J. Structure of human methionine aminopeptidase-2 complexed with fumagillin, *Science*, 282, 1324-1327, 1998.