14 Fragment-Based High Throughput Docking

Peter Kolb, Marco Cecchini, Danzhi Huang, and Amedeo Caflisch

14.1 INTRODUCTION

Structural genomics programs around the world are delivering an abundance of three-dimensional (3-D) structures of proteins, some of which are pharmacologically highly relevant. Hence, computer programs for automatic docking of libraries of compounds are being developed further and applied to design drugs against a plethora of diseases including AIDS, Alzheimer's disease, cancer, malaria, and sleeping sickness. In this chapter, we first review the most common approaches for structure-based flexible ligand docking. Some technical improvements for more efficient sampling and more appropriate scoring functions are then presented. Finally, a number of practical suggestions are given for high throughput docking (HTD) with special emphasis on our fragment-based approach.

14.2 OVERVIEW

The basic strategy of any docking approach is to generate a conformation of a putative ligand, which is then placed (or *docked*) in the binding site of a protein target (also referred to as *receptor*). The result of these two operations is usually called a *pose*. A *score* has to be assigned to each pose, thus producing a *ranking*, with the correct pose (i.e., the natural binding mode) at the first rank or at least as close as possible to it.

14.2.1 Defining the Binding Site

Prior to any attempt of docking, the approximate location of the binding site needs to be defined. It is easiest for the case in which the crystal structures of the receptor in complex with some ligands are already known. Usually, the binding site is then defined as the residues lying within a certain cutoff from the ligands.

A greater challenge is presented when only the 3-D structure of the protein is known. In that case, profound knowledge of the function of the protein is necessary. There are programs that analyze the protein surface and provide quantitative information on it, among them GRASP (Graphical Representation and Analysis of Structural Properties) [1] and HYDROMAP [2], which calculate

the electrostatic potential and hydrophobicity map, respectively. Alternatively, some programs use so-called "flood filling" algorithms that attempt to identify cavities on the protein surface. Basically, they fill the space that is not occupied by the protein with points and then roll a large "eraser" over the surface of the protein. All remaining points are said to be in protein pockets [3].

In general, the residues in the binding site are important because their interaction with the ligand is stronger and usually treated in more detail. The binding site residues are explicitly used during the computation of the score and they are sometimes also considered as entities providing anchor points for the positioning of a conformation. Therefore, they should be chosen according to the type and function of the receptor, as well as the program's strategy to determine ligand poses.

Recently, the program AutoDock [4,5] was tested on "blind" docking, that is without defining any selected portion of the protein as binding site [6]. Docking was successful for ligands with less than 10 rotatable bonds, but only at high computational cost (in the order of days). Hence, the definition of the binding site is necessary for virtual screening (VS) of large databases.

Another aspect is the selection of an appropriate protein (and thus binding site) conformation. McGovern and Shoichet have performed a comparative study [7], using the x-ray structures of the complexed and uncomplexed protein as well as conformations obtained by homology modeling of 10 different proteins. The highest enrichment of known ligands in a database was in most cases achieved with the complexed structure. Using a conformation from a complex introduces a bias toward known inhibitors, however, and should thus be complemented by other protein structures in a screening project.

14.2.2 Generating a Pose

Two main types of approaches to obtain a ligand pose have to be distinguished: the ones that use only the complete structure of the ligand and those that follow an incremental strategy. Section 14.2.2.1 and Section 14.2.2.2 refer to the first type; the incremental methods are described in the Section 14.2.2.3.

14.2.2.1 Generation of Ligand Conformations

Typically, docking programs modify only the torsional degrees of freedom of rotatable bonds to produce different ligand conformations. It is important to at least modify the torsional angles of groups carrying hydrogen bond donors (HDO) to allow optimization of this type of interaction. Torsional angles of bonds in rings, double or triple bonds, or single bonds to symmetrical groups (like methyl) are normally kept fixed. In one study with the focus on protein flexibility, the backbone of peptidic inhibitors was considered as being rigid and only "sidechain" flexibility was allowed [5]. A rigorous test of a docking program should consider full flexibility, however [8,9]. An important exception is the docking of small fragments (like benzene or benzamidine), for whom the rigid body approximation is an appropriate description of their limited flexibility [10,11]. Some programs do not

allow ligand flexibility, but the success rates in these cases are low if one does not use the conformation found in the crystal structure [12]. Clearly, such methods can hardly be used to predict the binding modes of "new" ligands. The program DOCK [13,14] also started as a rigid-body docking tool, but ligand flexibility was introduced in DOCK 4.0, using an exhaustive search and conformational refinement with the simplex method [15].

There are two common approaches for generating different ligand conformations:

1. In procedures that search the conformational space of the ligand outside of the binding site, a pool of relevant conformations with low internal energy is generated, and they are subsequently docked rigidly. The sampling of the ligand conformational space can be done exhaustively, modifying each torsional angle in discrete steps [16,17]. Alternatively, the procedure can employ rotamer libraries which assign the most probable values to torsions depending on the atom types [9,15,18].
2. The conformations can be subject to an optimization algorithm, where the torsional angles correspond to the variables of the optimizer. One can further distinguish between two optimizer types: Monte Carlo (MC) searches [3] (also used for *de novo* design by DeWitte et al. [19,20]) and genetic algorithms and other evolutionary approaches [4,8,21–23]. MC approaches use a single conformation that is randomly perturbed and improved. Genetic algorithms (GAs) employ a multitude of information-containing chromosomes (usually referred to as the *population*), which interact with each other and evolve to better solutions. These algorithms are more promising for docking [4], because the energy surfaces to be searched are rugged. MC methods tend to be rather slow, which is a disadvantage for large-scale library screening. Furthermore, if one uses MC–simulated annealing approaches, the additional problem of choosing an appropriate initial temperature and a cooling schedule arises.

14.2.2.2 Defining Ligand Positions

There are several strategies to position and orient the ligand in the binding site:

- The translational degrees of freedom can be encoded in an optimizer.
- The position can be determined by matching the shape of the ligand to the binding site.
- The conformation can be superimposed on a set of points that contain information about the binding site (for references see below).

As an example of approaches that follow the first strategy, the chromosomes in a GA can additionally carry genes for the translational degrees of freedom of the ligand and three (in the case of Euler angles) or four (when quarternions are used [4,24]) variables specifying the ligand orientation.

In approaches that follow the second strategy, the surface of the binding site is compared to the solvent accessible surface of the current ligand conformation. An optimal position is found based on some measure of similarity between those two. LigandFit [3] uses an algorithm developed by Oldfield [25,26], which treats both the binding site and the ligand as a collection of grid points. The shape of such a collection is characterized by a matrix. From the eigenvalues of these matrices, the shape discrepancy can be computed and used to assign a score to each conformation. FRED [17] employs a bump map, which is a Boolean grid representing the receptor, with true values where ligand atoms can potentially be placed. After this initial filtering step, several other scoring functions can be applied, among them Gaussian shape fitting. This function has favorable values when the ligand and the protein have high surface contact and little volume overlap.

DOCK [14,15] follows the third strategy by first filling the binding site with spheres of different sizes. The centers of these spheres are considered as anchors for atoms of the ligand. Variations of this approach at different levels of sophistication include the use of HDOs and HACs (hydrogen bond acceptors) as well as hydrophobic surface points as anchors [27,28]. An example of this is SEED, which was developed to dock small molecules with solvation [10,11]. It uses anchors on the surface of the receptor and performs an exhaustive search on a discrete space by matching donor and acceptor vectors (or vectors of hydrophobic interaction centers) and rotating the ligand around these axes. Other programs use information from the placement of predefined small molecular fragments to match their positions to similar entities in the ligand [16]. The Fragment-based Flexible Ligand Docking (FFLD) program utilizes the results from the docking of small and mainly rigid molecules that have been specifically chosen to match chemical moieties actually present in the ligand [8]. The underlying assumption for all these methods is that the interaction between a protein and a ligand is dominated by some key groups of the ligand. Hence, if the positions of these groups are determined correctly, the rest of the ligand will almost inevitably assume the correct pose.

14.2.2.3 Incremental Methods

Programs like FlexE [9] (an advanced version of FlexX [18]), SLIDE [28], or DOCK 4.0 [15] also try to optimize the interactions of the key groups, but do this individually for each group. The ligand is first split into several units (fragments), the first of which is placed as a seed. Usually, the determination of the pose of the first fragment is done with high accuracy. Sequentially, all the other fragments are connected in their due order, whereby each position is optimized, often exhaustively. At every step, the highest ranking solutions are retained and the next fragment is connected to each of them. It is important to carefully select only a small number of candidate solutions at every step (pruning) to control the exponential increase of possible solutions.

14.2.3 Ranking the Poses

At the beginning of this chapter, we distinguished between exhaustive searches and optimization techniques. The latter minimize an objective function that is usually computationally not too expensive, because it has to be called quite frequently, and a force-field-based binding energy is evaluated for the final ranking. Exhaustive searches use only one energy function.

14.2.3.1 Objective Function

The objective function approximates the interaction energy between ligand and receptor and the internal strains of the ligand and the protein, if the latter is also flexible. Typical components are the intermolecular van der Waals (vdW) and Coulombic energy, and sometimes a term for hydrogen bonds. The internal strain is usually estimated by the intraligand vdW energy and sometimes the dihedral energy. Most objective functions do not take into account terms for bond, angle, and torsional strains. It has been proposed to increase the chances of the optimizer by smoothing the energy landscape. Whitfield et al. [29] introduced a gravitational force that dominates all other forces in the initial steps of the search and then decreases over time. It is assumed that the position of the global optimum does not change due to the smoothing and that only the well depth is modified. Hansmann and Wille [30] developed energy landscape paving, which penalizes scores that are found repetitively. Searches can thus escape local minima and go into regions of different energy.

Most of the docking programs that use physics-based functions (like DOCK [13–15], AutoDock [4,5], and FFLD [8]) employ a grid-based approach for efficiency reasons. These grids contain the Coulombic potential and vdW potential of the protein and avoid the need for recalculating the full energy for every pose during a database screen. Trilinear interpolation [31] is often used to compute the interaction energies from the grid values of the potential.

Empirical-based functions (such as the one used in FlexX [18] and FlexE [9]) use additive approximations to estimate the binding free energy. They contain several terms corresponding to hydrogen bonding, hydrophobic interactions, entropic changes, and sometimes, interactions with metal ions. The coefficients of each term in the sum are obtained from a fit to known experimental binding energies for various protein–ligand complexes [32,33].

14.2.3.2 Binding Energy Function and Postprocessing

After a docking run, the best poses of the ligand can be reranked using a more accurate force field [34,35]. This often contains the same terms as the objective function, but takes longer ranging interactions and ligand and receptor desolvation into account. Sometimes, the ligand pose is also minimized within the receptor using a molecular mechanics force field [36,37]. In our group, ligand poses are normally minimized with CHARMm [36] using the CHARMm22 force field (Accelrys, Inc.), and often also with the TAFF-force-field (Tripos). Additionally, the score and rank of each pose can be redetermined using more accurate energy

functions that include electrostatic solvation like the one in SEED [10,11] or knowledge-based interaction fields like SuperStar [38], potential of mean force (PMF) [39], Small Molecule Growth (SMoG) [40], and DrugScore [41]. The energy rankings produced by the different scoring functions are usually compared, as a number of studies suggest that consensus scoring improves the chance of finding a true hit [42,43].

14.2.3.3 Solvation

The effects of solvation play a key role in molecular recognition events. To calculate the electrostatic contribution to solvation in the continuum dielectric approximation, one could solve the finite-difference Poisson–Boltzmann (PB) equation [44–47] for every new position of the ligand molecule. Considering the current computer power, this would be forbiddingly expensive, especially for HTS. Therefore, only a few docking programs take into account electrostatic solvation effects. The continuum dielectric approximation and the generalized Born (GB) approach [48,49] are used in SEED [10,11], Program to Engineer Peptides (PEP) [50,51], and DOCK [52]. Fairly recently, Arora and Bashford have presented a modified GB approach that estimates desolvation by an integral over the occluded volume [53].

Some docking programs treat solvation effects just with respect to the presence or absence of conserved water molecules that form interactions that are either essential for the protein conformation or necessary to mediate interactions between ligand and protein. Clearly, this approximation completely neglects the bulk properties of water (e.g., dielectric screening). Österberg et al. use grids that have been derived by averaging over several crystal structures, some of which can contain water molecules [5]. Although the method has mainly been developed to incorporate protein flexibility, heterogeneities in the presence of water molecules can be taken into account as well. Schnecke et al. consider water explicitly and have a term penalizing the replacement of water molecules by a hydrophobic group of the ligand [28]. Finally, Rarey et al. have described a method to precompute positions of water molecules and place them if they can form hydrogen bonds with the (partial) ligand during the incremental construction in FlexX [54].

14.2.4 Protein Flexibility

In principle, it would be ideal to allow full flexibility for the protein to model large displacements upon ligand binding. Such studies have already been undertaken [55], but because the computational time was in the order of days for a single ligand, this can clearly not be applied to the screening of large libraries of compounds. As a consequence, flexibility of the protein, if any, is mostly limited to the binding site and its vicinity. Three different approaches shall be highlighted here.

AutoDock [5] incorporates both protein mobility and structural water heterogeneity. It first generates the energy grids for a number of different protein

structures. The program then offers several ways to combine these grids into a single grid. It either computes simple point-by-point averages or weights the different grid points according to their energies and physico-chemical characteristics. This mean grid approach has the advantage that one can still dock to one rigid structure, which facilitates the analysis of the results compared to docking to several distinct conformations. On the other hand, it can only be used to approximate minor displacements. Moreover, the mean grid structure is the product of an averaging scheme and thus might not be observable in reality. Another drawback is the fact that no protein structure is present, but only its representation as a grid. One could thus not follow a multiple step approach (See Section 14.3.4) and do minimization with CHARMm [36], for example.

FlexE [9] is based on a so-called united protein description [56], which is derived from superimposing the backbones of an ensemble of different crystal structures. Variations of the structure in the binding site region are either maintained as distinct possibilities or are combined to one structure in case they are similar. During the incremental construction algorithm, the ligand is placed fragment by fragment into the active site of the united protein description. After each construction step, all possible interactions between the (partially) placed ligand and all instances of the united protein description are determined. The score is then assigned for the (partial) ligand in the best instance.

SLIDE [28] goes one step further and first docks a rigid scaffold into a rigid binding site. Gradually, the other parts of the ligand are attached to the scaffold. Clashes between the ligand and the protein are resolved by allowing rotations of bonds (both in the ligand and the protein) that have been defined as flexible beforehand. The bonds that should be rotated are determined with mean-field theory, which is capable of finding the minimum amount of rotations necessary to resolve all clashes [57–59]. Although flexibility is limited to the binding site residues, this approach comes close to an induced fit.

One of the most thorough approaches besides [55] has been undertaken by Lin et al. [60]. For their relaxed complex method, first long molecular dynamics (MD) simulations of 2 ns were conducted, with snapshots taken every 10 picoseconds (ps). Two candidate compounds were then docked to the ensemble of MD conformations. This technique recognizes the fact that ligands may bind tightly to conformations that appear only infrequently in the dynamics of a protein. However, every molecule has to be docked to a large number of different protein structure which strongly limits the size of the library.

14.3 TECHNICAL IMPROVEMENTS

14.3.1 CURRENT LIMITATIONS

As mentioned above, docking approaches can be described as a combination of two components—the search strategy and the scoring function. Because in most cases the objective function (See Section 14.2.3.1) is also used as the binding energy function (See Section 14.2.3.2), in the following, the term *scoring function*

will be employed. The critical element of the search procedure is the amount of time required to effectively sample the relevant conformational space. The scoring function has to be fast enough to allow its application to a large number of potential solutions and, in principle, be able to effectively distinguish the experimentally observed binding mode from all others explored in the search. Consequently, the scoring function should include and appropriately weight just the energetic contributions that are relevant in the binding process. Nevertheless, an accurate scoring function will generally be computationally expensive and so the function's complexity is often reduced at the expense of a loss in accuracy.

The proper combination of an effective search algorithm and an adequate scoring function, whose global minimum corresponds to the biologically relevant complex, will solve the docking problem in a reasonable amount of time. However, because the approaches published up to date can fail, especially in cross-docking, this ideal combination has obviously not been found yet. Therefore, improvements in the efficiency of the search strategy and the accuracy of the scoring function are required as they will increase the reliability of the docking predictions and reduce the computational requirements, which is important for screening large libraries.

Docking predictions are still prone to fail and often the proposed binding modes do not reproduce the crystal structure of the protein–ligand complex [6,9,35,61]. In case of failure, the predicted binding mode can have a worse or a better score than the x-ray structure of the ligand. In the first case, the search strategy adopted in the docking approach could have been not effective enough. The search algorithm was thus not able to generate a pose sufficiently close to the experimental binding mode. In the second case, the failure might arise from an inadequate scoring function that allows more favorable binding modes than the one in the crystal structure. In the first case, one should focus on the improvement of the search procedure; in the second case, one should concentrate on the optimization of the scoring function.

Unfortunately the situation is much more complicated because the components of a docking protocol are not separate entities and as such they should be improved together. In the first scenario, for example, the scoring function could have played an important role because the resulting energy landscape was not smooth enough to allow the search to proceed efficiently while avoiding premature convergence. Although the scoring function described the protein–ligand interactions well, it was not suitable for the applied search strategy. In the second scenario, it could have happened that the experimentally determined structure was not close to a minimum of the scoring function. In this case, any energy comparison is much less meaningful. Although a proper combination of an efficient search algorithm and an accurate scoring function are the keys for a successful docking protocol, it is certainly not clear what "proper," "efficient," and "accurate" mean. In Section 14.3.2. and Section 14.3.3, we describe some important requirements for both the search strategy and the scoring function and how they are embedded in our docking approach.

14.3.2 Search Strategy

Docking procedures belong to the category of global optimization techniques where the aim is finding the global minimum of the scoring function. A rigorous search algorithm would exhaustively investigate all possible binding modes between the ligand and the receptor. The degrees of translational and rotational freedom of the ligand would be explored along with the internal conformational degrees of freedom of both the ligand and the receptor. However, this is impractical because of the size of the search space, even when considering a rigid protein. Only a small amount of the total conformational space can be sampled and a balance must be reached between the computational expense and the amount of search space examined. A wide range of global optimization algorithms are currently available, but not all of them are suitable for docking. Most optimization algorithms for docking fall into one of three classes—gradient-based algorithms, combinatorial algorithms, and stochastic algorithms [62].

The strength of gradient-based methods is that they efficiently find a local minimum close to the initial conformation. Because gradient-based methods do not allow the system to escape from local minima they have to be combined with other search strategies, such as cycles of MC perturbations and gradient minimizations [63]. Moreover, most scoring functions do not have an analytical gradient.

Combinatorial algorithms have the potential advantage of being extremely fast and effective. The most successful combinatorial algorithms used for molecular docking [10,11,18,64,65] have set themselves apart in their ability to dock libraries of small molecules in a reasonable amount of time. Unfortunately, increasing the number of conformational degrees of freedom leads to an explosion of the dimension of the search space. To be able to sample such large spaces, the computational expense is usually controlled by a discretization of the space, which can restrict the effectiveness of the algorithm.

Stochastic algorithms have the advantage that, irrespective of the dimensionality of the problem and given enough time, they get arbitrarily close to the global minimum. On the other hand, they have the disadvantage that they require a large amount of central processing unit (CPU) time to achieve an acceptable degree of reliability [4,62]. Although computationally expensive, stochastic optimization algorithms seem to be the most suitable for flexible docking. In fact, the dimensionality of the search space and the ruggedness of the binding energy landscape make both gradient-based and combinatorial methods less effective. GAs are stochastic optimization methods that mimic the process of natural evolution by manipulating a population of data structures called chromosomes [66,67]. Although requiring rather large amounts of CPU time, GAs have been shown to effectively explore rough energy surfaces and to be suitable as search strategies for docking [4,8,21–23,68,69]. A GA was chosen as the search strategy for the original version of FFLD [8], the docking protocol developed in our group. During the FFLD evolution, a loop over generations is performed until the maximum number of steps is reached. Starting from an initial random population of chromosomes containing the dihedral angles of

the ligand as genes, the GA repeatedly applies two mutually exclusive evolutionary operators—one-point crossover and mutation. This yields new chromosomes (children) that replace appropriate members (parents) of the population. These nonlinear genetic operators help to overcome the barriers of the binding energy landscape and the search can proceed efficiently. Throughout the simulation, a constant evolutionary pressure is kept by selecting parent chromosomes with a bias toward the fittest. This pressure moves the population toward conformations related to the global minimum and increases the fitness of the individuals. The selection of the members of the population that should be replaced by new chromosomes is a crucial step. To avoid premature convergence, it is important to keep structural diversity. In the search strategy used in FFLD [8], both the energy difference and the conformational similarity are taken into account to determine if a given member of the population should be replaced by a new chromosome. At the end of each GA step, every new chromosome is compared with the old population by the following procedure: if a similar chromosome is found in the old population, it is replaced by the new chromosome only if the energy of the new one is more favorable; otherwise, the new chromosome is discarded. The similarity test significantly improves the efficiency of the search strategy and avoids premature convergence [50].

Following a comparative study of several search engines in AutoDock [4], a hybrid search procedure was introduced in the latest version of FFLD [35]. The hybrid search combines a global optimization procedure based on a GA with a local minimization algorithm to improve exploration of regions within energy basins. Local optimization has been shown to dramatically improve the success rate of the GA search without any loss in efficiency [4]. For the best 10% of the new individuals, a local optimization is performed to improve the ligand fitness before performing the similarity test. To evaluate the performance of the hybrid search procedure implemented in FFLD, it was compared with the GA of the original version [8]. The simulations showed that the hybrid search is more efficient than the canonical GA as it always reached a conformation with lower energy. The results of two docking experiments carried out with both search methods are presented in Figure 14.1. The first experiment, in which a ligand with 10 rotatable bonds was docked in human-immunodeficiency virus type 1 (HIV-1) protease (Figure 14.1, top), shows that the hybrid search procedure is more efficient than the genetic algorithm especially at the beginning of the simulation where the energy gap is large. At about 60% of the evolution the gap decreases and the performance of the two methods is comparable. Docking a ligand with 21 rotatable bonds in HIV-1 protease (Figure 14.1, bottom) shows that the hybrid search procedure performs better during the entire simulation and the energy gap increases until the end. Moreover, the standard deviation of the hybrid search evolutions (shown as error bars in Figure 14.1, bottom) is larger, indicating that it is less prone to converge prematurely. This comparison shows that the local search improves the quality of the docking predictions in case the conformational space of the ligand is large. This is mainly due to the fact that the random perturbations of binary strings performed by the GA during the evolution correspond to radical jumps in the energy landscape and may be too large. On the contrary, the local optimizer is able to refine the large perturbations due to crossover

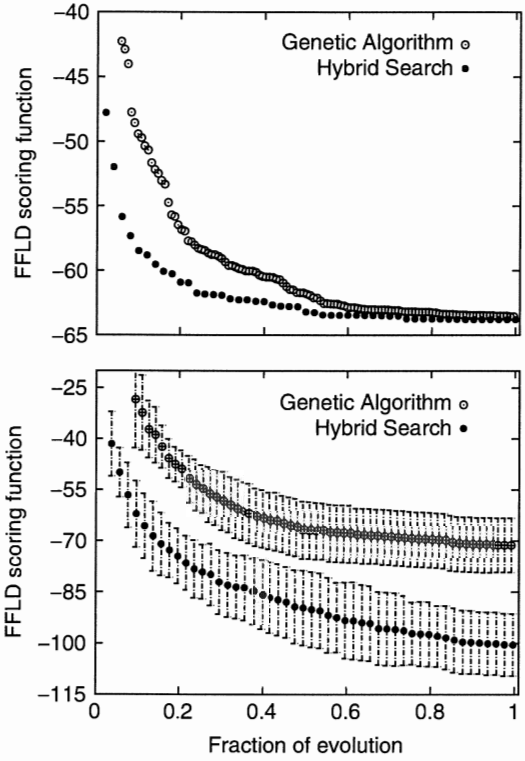

FIGURE 14.1 Evolution of the best individual of the population averaged over 10 docking runs for two different experiments. Empty and filled bullets indicate evolutions performed by GA and hybrid search procedure, respectively. Docking of HIV-1 protease ligands with 10 and 21 rotatable bonds are shown from top to bottom, respectively. In the bottom plot, the vertical bars show the standard deviation computed over 10 docking runs.

and mutations and leads to a better investigation of the energy landscape. The results of this docking study [35] suggests, in agreement with previous studies [4], that hybrid search methods should be preferred to canonical GAs.

The similarity test and the hybrid search procedure are just examples of possible means one can adopt in a protocol to increase the efficiency and accuracy of the search algorithm. However, the study clearly indicates that there is still room for improvement and that novel concepts can be effective. It is worth stressing again that the search algorithm is only half of the docking problem; the other factor to be incorporated into a successful protocol is the scoring function. In Section 14.3.3, the requirements for a scoring function that are suitable for docking are discussed.

14.3.3 Scoring Function

Underlying any docking approach is a model of ligand–protein interactions describing molecular recognition. In principle, a complete thermodynamic description of this process involves contributions from several balancing factors, including solvent reorganization, conformational entropy, and vdW and electrostatic interaction energies. For biomolecular systems, it is difficult to evaluate these terms with sufficient accuracy to permit quantitative predictions. Moreover, the complete energy function necessary for prediction of accurate binding affinities may not be suitable for docking simulations. The scoring function used in docking simulations should be a simple model of ligand–protein interactions rather than an estimation of the free energy of binding. It must be simple enough to permit a rapid evaluation and, more importantly, the resulting energy landscape must be smooth enough to allow the search to proceed efficiently without getting trapped in local minima. Nevertheless, a scoring function that is suitable for docking needs to be accurate, because it must be able to distinguish the experimental binding mode from all the other modes explored by the search algorithm.

With respect to this point, Verkhivker et al. [69] suggested that such an energy function should fulfill both a thermodynamic and a kinetic requirement. In other words, the energy related to the crystallographic structure of the ligand in the complex must be the global minimum of the binding energy landscape (*thermodynamic requirement*), but at the same time this conformation must be accessible during the search (*kinetic requirement*). The complexity of a complete and accurate force field that describes the binding process precisely, although it would fulfill the thermodynamic requirement, typically results in a rugged energy landscape and thus does not meet the kinetic criterion of the docking problem. The multitude of energetically similar but structurally different local minima inevitably leads to kinetic bottlenecks that dramatically reduce the frequency of successful structure predictions. This is the case for standard molecular mechanics force fields [36,37], because they have not been designed to reduce the ruggedness of the energy landscape. One of the critical factors that determines the success rate in predicting the structure of ligand–protein complexes is the roughness of the binding energy landscape [68,69]. Consequently, the applicability of standard force fields in docking is limited and simpler molecular recognition models that fulfill both the thermodynamic and kinetic requirements are to be designed and developed.

A fundamental component of models for molecular recognition is the steric energy function, which is based on surface complementarity. However, this term alone is not sufficient to distinguish effectively between alternative binding modes. Electrostatic interactions may provide additional specificity to discriminate between true and false solutions and they should be embedded in the scoring function. Finally, an intraligand energy term is also required; it largely reduces the conformational space to be investigated by preventing strained dihedrals and steric clashes among atoms of the ligand. Hence, the three key elements of a scoring function necessary for robust structural assessment during docking are:

1. Ligand–protein steric interactions
2. A simple description of ligand–protein electrostatics
3. An intraligand strain

In the FFLD docking approach developed in our group [8], the scoring function is

$$\Delta E_{total} = E_{dihedral}^{ligand} + E_{vdW}^{ligand} + E_{vdW}^{inter} + E_{polar}^{inter} \tag{14.1}$$

The dihedral energy of the ligand $\left(E_{dihedral}^{ligand}\right)$ has recently been implemented in FFLD (D. Huang, unpublished results) using the lowest order terms of a cosine expansion for each torsion. The second $\left(E_{vdW}^{ligand}\right)$ and the third $\left(E_{vdW}^{inter}\right)$ terms of Equation 14.1 are intraligand and ligand–receptor vdW energies, respectively. Both terms are described as the sum of an attractive dispersion and a steep repulsion term by the 6-12 Lennard-Jones potential. The last term in Equation 14.1 is the protein–ligand polar interaction energy $\left(E_{polar}^{inter}\right)$. The intermolecular polar term approximates electrostatic interactions and includes hydrogen bonds (HB) and unfavorable polar contacts (UP), namely two HAC (or HDO) atoms close to each other. Hence

$$E_{polar}^{inter} = \sum_{i=1}^{N_{HB}} E_i^{HB} + \sum_{i=1}^{N_{UP}} E_i^{UP} \tag{14.2}$$

where N_{HB} and N_{UP} are the number of hydrogen bonds and the number of unfavorable polar contacts, respectively. The energies E_{HB} and E_{UP} are approximated by constant values [35]. Distance- and angle-dependent criteria are considered for the definition of a hydrogen bond, but only a distance dependence is applied for unfavorable polar contacts. Originally, the distance dependence of both terms in Equation 14.2 and the directionality of the hydrogen bonds follow simple step functions (Figure 14.2, top left and top right, dashed lines) that are efficiently evaluated [8]. The steep repulsive part of the Lennard-Jones potential directly affects the height of the energy barriers and generates a rough energy surface. To reduce the steepness of this energy component, an intermolecular soft-core vdW term was implemented [8]. Following previous studies by Gehlhaar et al. [68], the repulsive part of the Lennard-Jones potential was linearized in FFLD, such that the functional form has a finite value when the interatomic distance approaches zero (Figure 14.2, bottom). The intermolecular soft-core vdW does not penalize binding modes with small atomic interpenetrations of the ligand with the protein and permits the formation of unphysical states that could open multiple pathways leading to the crystal structure. These states, otherwise forbidden by the presence of realistic energy barriers in standard force fields, may provide kinetically accessible routes to the global minimum.

In a recent study [35], a significant improvement with respect to the original version of our docking approach [8] has been observed by replacing the step

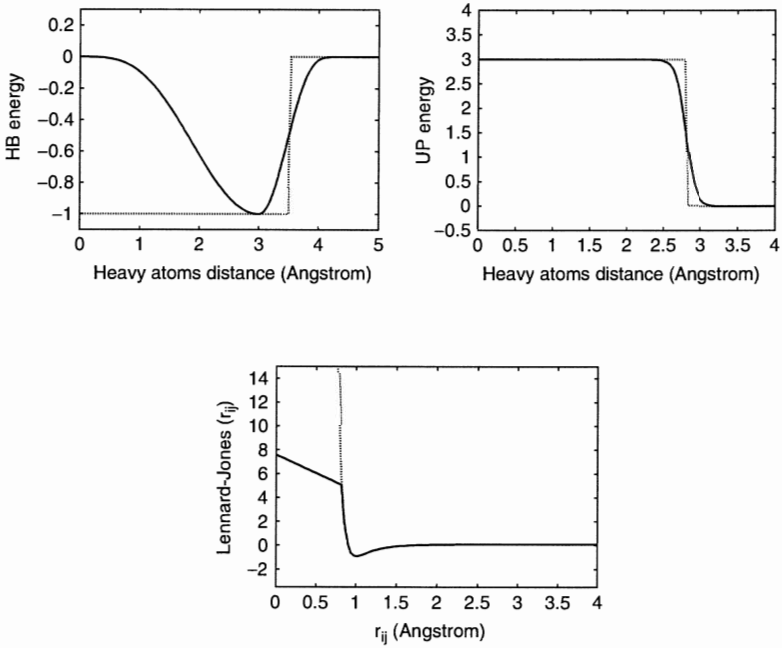

FIGURE 14.2 The distance dependence of hydrogen bonds (HB), unfavorable polar contacts (UP) and ligand-receptor vdW interactions is presented from left to right, respectively. The smooth functions (solid lines) [35] used for replacing the original stepwise functions (dashed lines) in the intermolecular polar interaction term are shown. On the bottom, the intermolecular soft-core vdW (solid line) [8] is compared with the 6-12 Lennard-Jones potential (dashed line). Values are in kcal/mol.

functions in the ligand–receptor polar interaction term $\left(E_{polar}^{inter}\right)$ with *smooth* functions. Smooth functions allow the optimization of the hydrogen bonding pattern avoiding discontinuities on the energy landscape. The continuous gradient can guide the search algorithm toward lower energy conformations at every point. In the latest version of the FFLD docking program, a sigmoidal function was used to describe the unfavorable polar contacts and bathtub-shaped functions were used for the distance dependence and the directionality of the hydrogen bonds (Figure 14.2, top left and top right, solid lines). Furthermore, it was observed that the distance- and angle-dependence in the polar term significantly reduced the noise arising from the energy degeneration of structurally different ligand conformations and improved the convergence of the docking runs [35].

Previous works by Gehlhaar and Verkhivker [68,69] suggested that a dynamical modification of the scoring function is helpful. In their docking experiments, an adaptive scoring function based on a piecewise linear potential was used. During docking, the height of the energy barriers had been continuously augmented by

increasing the repulsive term of the potential. Thus, in the later stages of the simulation, this adaptive procedure narrowed the search to only a few energetically favorable binding modes, funneling the algorithm to the global minimum. According to the authors, the adaptive softness of the energy function facilitated the conformational search both by promoting escape from local minima and by destabilizing alternative solutions. Increasing the repulsive term of the potential yields a rougher energy landscape, but the energy function becomes more and more accurate and leads the search to the global minimum. Similar dynamical modifications of the energy function have been adopted to mimic the docking funnel [29,30,55]. Although the essential idea is rather simple, no general rules for adapting the potential are available and the optimal way for scaling the barriers may be strictly dependent on the system explored. Moreover, if the scaling is not accomplished in a proper way, the adaptive scoring function might not fulfill the kinetic requirement. Because of these limitations, we and others [35,70–73] have chosen an alternative approach. This is described in Section 14.3.4.

14.3.4 Multiple-Step Docking

Combining different scoring schemes into a single docking approach is a useful method to increase the effectiveness of a docking protocol. A two-step strategy makes use of a simple molecular recognition model based on the minimal frustration principle [68,69], followed by a more accurate energy evaluation to rank the docking solutions. When using multiple-step procedures, there is a clear distinction between the objective function, which is fast but approximative, and the binding energy function (See Section 14.2.3.1 and Section 14.2.3.2).

The basic assumption behind multiple-step approaches is that there is at least one low-lying minimum of the objective function inside the global minimum basin of the binding energy. The fast objective function is then thought of as a coarse-grained description of the more accurate binding energy function. The first step intends to overcome the kinetic bottlenecks of the accurate energy function by using a simpler and much less frustrated energy model. After the first step of the procedure, the final set of ligand conformations can undergo a gradient-based minimization with a standard force field. The minimized conformations are then ranked according to their energy. Multiple-step docking approaches are widely used and have been published [70–73]. A multiple-step procedure was also applied in the most recent version of our docking approach [35]. The results of FFLD [8] were postprocessed by CHARMm minimization [36] of the flexible ligand in the rigid receptor. The docking study showed the effectiveness of a multiple-step strategy. It was possible to correctly reproduce the binding mode of highly flexible inhibitors (up to 22 rotatable bonds) of HIV-1 protease, if the strain in their covalent geometry upon binding was not too large. Moreover, it was observed that the postprocessing step led to more reliable predictions and improved the success rate of the docking experiments [35].

14.4 PROTOCOLS

In this section, we will explain the use of our docking approach. However, many of the guidelines and recommendations introduced here will also hold true when using other docking programs.

14.4.1 Our Docking Approach

The SEED/FFLD approach uses a GA to optimize ligand conformations and previously docked fragments to place the ligand in the binding site. It relies on the assumption that the most significant interactions with the protein are formed by three or more fragments of the ligand. Hence, it should be possible to first investigate the binding modes of the fragments and then use this information to place the whole molecule. This docking approach consists of four separate steps, the principles of which shall be described below. A more detailed protocol can be found in the following subsections and the original articles [8,10,11,35].

At first, those parts of the ligand that are supposed to account for most of the interactions (the fragments, Figure 14.3) have to be defined. This choice is rather important, for example, fragments that are too small will yield anchor positions that cannot discriminate the physicochemical characteristics of the binding site. A computer program has been developed to automatically choose at least three fragments (P. Kolb et al., unpublished), because the matching algorithm employed in the last step uses triangles. In the second step, the selected fragments are minimized with a force field to obtain low energy conformations. Subsequently, they are docked as rigid molecules with SEED [10,11] (Figure 14.4). As described before, SEED uses polar and hydrophobic vectors as anchors. The polar vectors are distributed around HDOs and HACs, whereas apolar vectors are used to mark hydrophobic regions. The latter are obtained by placing a low dielectric sphere (methane) at equal intervals on the solvent accessible surface of the protein. Points that have a favorable interaction energy are retained and the vectors are defined by joining each point with the corresponding atom center. During docking, every vector is matched to the complementary vectors on the fragments and the fragments are rotated exhaustively around these vector-defined axes. For each fragment position on each SEED point, a binding energy, which

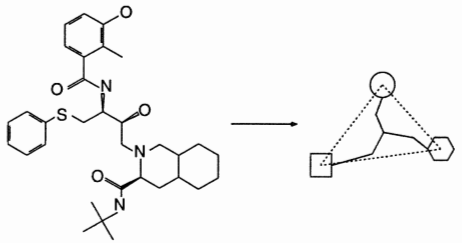

FIGURE 14.3 Schematic depiction of the fragment selection process. The molecule is Viracept (Agouron/Pfizer).

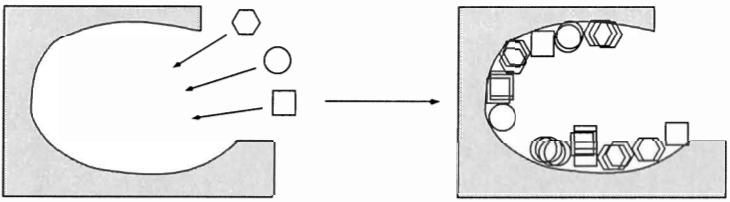

FIGURE 14.4 Schematic depiction of the docking process of the small fragments.

includes electrostatic solvation, is evaluated. Thus, if the fragments chosen are rigid (which is the case for small molecules and aromatic systems), the ranking is determined with high reliability. The information obtained from SEED consists of the 3-D coordinates of the geometrical centers of the fragment poses as well as their binding energies. Each fragment pose is one possible corner point of the placement triangle used in the last step. On average, a SEED run yields up to 100 poses per fragment type.

In the third step, this number is reduced to obtain a manageable number of possible triangle combinations. In practice, we reduce it to 20, using a clustering method which is based both on geometric proximity and the value of the binding energy for each pose [35]. For each fragment, the 20 points define a map that contains the important information from SEED and is still diverse enough to offer useful anchor points (Figure 14.5). Diversity is especially important because using only the top-ranked poses of the fragments does not always lead to the solution. This is due to the fact that the binding mode of the entire ligand is a compromise that tries to satisfy most of the fragments.

The fourth and last step is the docking of the complete putative ligand. This is done with the program FFLD [8], which uses a scoring function consisting of ligand dihedral and vdW energy, and protein–ligand polar and vdW contributions (See Section 14.3.3). Ligand conformations are generated and optimized by a GA, which encodes the torsional angle values of the rotatable bonds. For each conformation, the geometrical centers of the three fragments define a triangle. Based on the side lengths of the ligand triangle, FFLD finds those SEED points that form triangles of approximately the same shape. It then

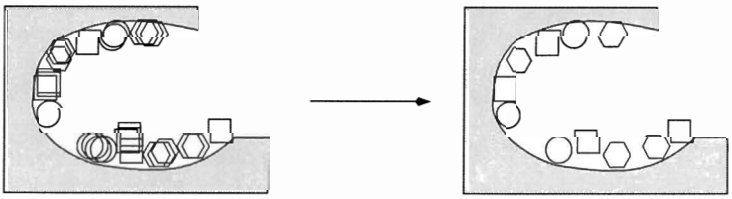

FIGURE 14.5 Schematic depiction of the clustering procedure. The different fragment types are shown for clarity.

tries to match the ligand triangle with each of the possible SEED triangles using a least-squares-fitting method (a variant of the Kabsch algorithm [74]) and assigns the score of the best placement to this conformation (Figure 14.6). The output of FFLD consists of the final poses for all conformations, usually 100 to 200 in total. It is worth noting that, because every conformation yields multiple poses at each step, FFLD will not only find the best binding mode, but also a number of alternative binding modes of comparable score. The alternative binding modes, in fact, can be used as a starting point for further postprocessing with more accurate energy functions.

14.4.2 Preparation of the Library of Compounds

The first and most basic requirement is that the ligand is a chemically complete molecule (i.e., all valences must be satisfied). Special care must be taken to specify the correct bond types, because this will be the basis for the definition of the bonds that are rotatable. Another main concern is the correct assignment of the partial charges. These are needed for the calculation of the interaction energy in SEED and the postprocessing step. We use the modified partial equalization of orbital electronegativity (MPEOE) method developed by No et al. [75,76] as implemented in WITNOTP (A. Widmer, Novartis Pharma AG, Basel, unpublished), which yields partial charges consistent with those of the protein atoms in the CHARMm22 force field. Other implementations should also give reliable partial charges, but we have not tested them.

As a prerequisite to docking, one has to consider the state of ionizable groups in the protein (see below) and the ligand. Because the physiological conditions for protein–ligand complexes are in most cases close to pH 7, acidic groups are usually in a deprotonated and basic groups in a protonated state. A pK_a calculation could be done with a finite-difference Poisson solver in case of uncertainties. For a heterogenous library of compounds, it is much more difficult to assign formal charges. We usually check for groups where the assignment is evident (e.g., primary, secondary or tertiary amines, which are positively charged). Afterward, atom types for the CHARMm22 force field have to be assigned. Any ligand should furthermore be minimized with an accurate force field to obtain a low-energy conformation.

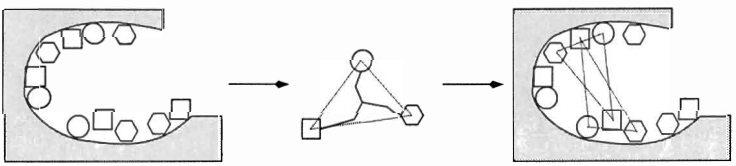

FIGURE 14.6 The docking process: FFLD tries to place the triangle defined by a conformation of the ligand (generated with the GA) on the anchor points computed by SEED.

14.4.3 Fragment Choice

The decomposition of a ligand into fragments and the choice of the anchor fragments have been automatized recently (P. Kolb et al., unpublished). We will list the major rules here as they can be of general interest. The decomposition is guided by the fact that SEED treats all molecules as rigid. Hence, preference is given to aromatic rings and other small rings and molecules that contain several amidic, double, or triple bonds. The fact that nonaromatic ring systems might have several distinct conformations can be accounted for by the ability of SEED to dock multiple (predefined) conformations at the same time. If one of these conformations can be docked with a lower binding energy than the others, it will automatically be chosen in the subsequent steps, because it will receive higher ranks.

The selection follows a few simple rules:

1. All atoms in a fragment must be connected by rigid or terminal bonds (for the definition of rigid bonds see above).
2. Large fragments are preferred because there are more steric constraints for large entities, as a consequence these should be positioned first.
3. Cyclic fragments are preferred because they usually are more rigid than acyclic moieties.
4. Because the fragments should be involved in the most significant interactions, those that contain HDOs and HACs are selected. Charged groups usually do not make such good anchors, because they tend to be positioned at the borders of the binding site, which are more exposed to the solvent. (However, there are exceptions as in the case of thrombin, where a favorable electrostatic interaction is provided by a charged aspartic acid in the specificity pocket [8].
5. Fragments that are close to the center of the molecule are omitted, especially if they have a high number of substituent groups. Such central or scaffold fragments will hardly ever form specific interactions.
6. Finally, fragments should not overlap (i.e., one atom should not be part of two fragments), because this would mean that there are no rotatable bonds in between, so their relative position cannot be changed.

These rules can be exemplified with the molecule XK263 (DuPont Merck, Figure 14.7). In principle, there are three fragment types that could be chosen—naphthalene, benzene, and the cyclic urea in the center. The largest fragment would be the cyclic urea. According to Rule 5, this is not a good choice as it is the core fragment and has four substituents. Furthermore, it is the most flexible of the three types, which is another point against its choice according to Rule 2. The remaining two types are aromatic and thus a recommended choice (Rule 1). Finally, it is better to select the two naphthalenes, because they are larger than the benzenes (Rule 2).

A more difficult choice is presented by acetyl-pepstatin (Figure 14.8), because it has no rings and almost no rigid bonds. All the fragments obtained by the application of Rule 1 are therefore small. All the larger fragments with a rigid

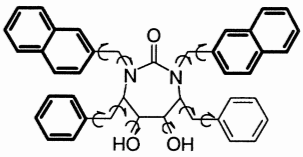

FIGURE 14.7 XK263 (Dupont Merck) is a nanomolar inhibitor of HIV-1 aspartic protease (PDB accession code of the complex: 1HVR). Selected fragments are bold. Curly arrows denote rotatable bonds.

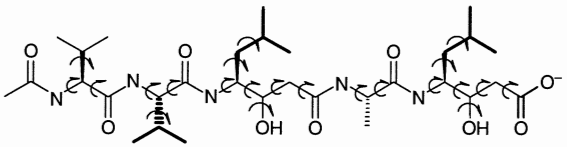

FIGURE 14.8 Acetyl-pepstatin is a micromolar inhibitor of HIV-1 aspartic protease (PDB accession code of the complex: 5HVP). Selected fragments are bold. Curly arrows denote rotatable bonds.

bond (the amide groups) are located in the backbone and will not make good anchors (Rule 5). One of the few choices remaining is to select three i-butanes (the sidechains), which are preferable with respect to the terminal carboxylic group, because this group is charged (Rule 4).

14.4.4 Protein Preparation

It has to be emphasized that the preparation of the protein is a crucial step in the protocol and should be done carefully. It is not advisable to use automatic methods, as they cannot take into account all eventualities and special cases.

14.4.4.1 First Checks

The attention of the experimenter should be turned to all specific and unusual details, like nonstandard amino acids (e.g., cysteine-sulfonic acid, selenomethionine, etc.). Furthermore, the protein can contain prosthetic groups, cofactors, or other small molecules. Prosthetic groups should be kept for the docking run in all cases, because they will most probably be present in the protein in its native environment. Whether or not cofactors should be considered, depends on the system. Most probably, they can be removed, since they will not compete with an inhibitor, unless they have a strong affinity to the protein by themselves and will be present in the binding site most of the time. In general, small molecules (such as polyethylene glycol) are due to the crystallization conditions and should be removed. The final decision, however, has to be taken *ad hoc* for every system.

In any case, one should check in the pdb-file that no atoms are missing in the aforementioned residues and molecules, because most structure manipulation programs do not check on nonstandard residues automatically. Quite frequently,

crystal structures will lack even whole parts of the protein due to poor electron density in disordered regions. This fact is usually commented on in the pdb-file or in the paper. It is then up to the researcher to decide if this is negligible or not. Judging from our experience, in the majority of cases, these incomplete regions are far away from the binding site. Thus, they will not have a great influence on the binding energy evaluation. Unless there are only one or two amino acids missing, it is not advisable to rebuild the protein in those regions. The error introduced by guessing the conformation without proper equilibration will probably be larger than the error due to the absence of the residues.

Another special case are ions. Those that are required for the stability of the protein should be kept, especially if they are close to the binding site. An ion in the binding site should always make a favorable interaction with an oppositely charged group in the ligand. It is advisable to determine the charged warhead for the candidate ligands *a priori* and discuss the simpleness of synthesis of the resulting compounds with a medicinal chemist.

Lastly, the presence of disulfide bonds has to be investigated. Information whether or not there are any should be listed in the pdb-file in a line commencing with "SSBOND." However, it is safer to visualize all cysteine residues. If the sulfur–sulfur distance between two cysteine residues is around 2 Å and the relative geometry is right, they will most likely form a disulfide bond.

14.4.4.2 Charged Residues

Special care should be exercised when treating residues with ionizable groups. The most sophisticated approach is to solve the finite-difference Poisson equation to calculate the pK_a of all titratable groups. If the *in vitro* tests are done at physiological pH, we normally assume both basic and acidic sidechains as well as the terminal carboxyl and amino group as ionized.

The situation for histidine residues is more complicated. First, one has to select a protonation state and then, in the case of monoprotonation, also which nitrogen (δ or ε) should be protonated. To properly assign the protonation state of the histidines, it is important to consider the local environment of these residues in the folded structure of the protein. At low pH ($pH \leq 6$), a diprotonated state should be assigned to histidines partially or fully exposed to the solvent. For calculations at physiological pH, a monoprotonated state is commonly preferred and we assign a monoprotonated state to the histidines irrespective of their position. If the environment does not indicate a clear preference for one of the two variants because of potential HACs or steric hindrance, we arbitrarily choose the δ-protonated variant.

Related to the issue of the charged residues is the choice of the interior dielectric constant of the protein, which is necessary for SEED. The value of this constant influences the strength of the coulombic interactions and can lead to significantly different results, as model calculations have shown (Majeux et al., unpublished results). Previously, values ranging from 1 to 4 have been used [10,11]. It is useful to perform preliminary docking runs with interior dielectric values of 1, 2, and 4 and compare the results with available crystal structures.

14.4.4.3 Adding Hydrogens

It is necessary to add hydrogens, because files in the Protein Data Bank (PDB) usually do not contain any. This should be done with a program like CHARMm [36] using the HBUILD module, which first places those hydrogens whose positions can be determined unambiguously, such as hydrogens connected to a peptidic nitrogen, and afterwards performs exhaustive searches to place hydroxyl hydrogens on serine, threonine, and tyrosine. To assign atom types, we use the atom type definition of the CHARMm22 force field. It has proven useful to recheck on all nonstandard residues to verify the correct assignment. Finally, the hydrogens should be minimized with an appropriate force field while keeping the protein backbone rigid.

14.4.4.4 Binding Site Definition

As mentioned above, this step is of high importance. To begin with, one should have a look at the publication describing the crystal structure and the interactions. The basis for the selection of the residues belonging to the binding site will most often be the pose of a known ligand. If such information is not available, one has to select the binding site by hand. In that case, in-depth knowledge of the function of the protein or crystal structures of closely related proteins of the same family are necessary.

We select the binding site by first determining all protein atoms that are within a cutoff radius of 5 Å from any ligand atom. It is important that there is a clear inside and outside of the binding site to avoid the positioning of anchors in solvent-exposed regions of the protein. Hence, selecting residues whose sidechains point away from the binding site have to be avoided. To achieve this, only residues which have at least 50% of their atoms within the cutoff distance are marked as members of the binding site. The cutoff should not be too small, as the bias toward the binding mode of the known ligand would be too big and no alternative ones could be detected. On the other hand, because the binding site residues are providing the anchor points for SEED, the number of anchors correlates with the number of residues. Thus, docking would take increasingly long as the binding site becomes larger and would additionally yield too many solutions, which are then difficult to rank. If a large binding site is really needed, it is probably better to split it into several (overlapping) sectors. Sometimes, it is advisable to manually alter the definition until one is satisfied with the distribution and the number of the anchor points. In this case, one has to remember that the binding mode (and consequently the ranking of a library of compounds) might be affected by the human intervention, which is usually based on previous knowledge. This bias might preclude interesting surprises like alternative binding modes [77].

As was mentioned before, SEED puts anchor vectors on atoms of the binding site residues. Clearly, only vectors pointing inside the binding site should be used. For that reason, the latest version of SEED employs a cutoff based on the angle between the vector and predefined points in the binding site (usually the

heavy atoms of a native ligand) for choosing the most suitable ones [35]. Using the atoms of a ligand from a known complex to define the binding site does not introduce a bias, though, and corresponds to the situation in an advanced drug design program, where one or more crystal structures of protein/ligand complexes have already been solved.

Another critical issue is the ionization state of groups in the binding site. This is probably best illustrated by the case of the aspartic proteases, which contain an aspartyl dyad in the cleavage site. Piana et al. [78] have shown that, besides the pH, the ligand has an influence, as it can stabilize either the neutral, negatively or dinegatively charged form of the dyad state. Consequently, the charge state of the dyad can influence the types of ligands that will receive a high ranking.

14.4.4.5 Conserved Water Molecules

In many proteins, water molecules located at distinct positions can play a crucial role because they provide important interactions with the ligand. Wrongly positioned water molecules, on the other hand, can impede docking and make the detection of the correct binding mode impossible. Deciding which water molecules to keep is not trivial. Evidence can come from multiple x-ray structures with different ligands. If a water molecule is repeatedly found at the same position and also forms hydrogen bonds with the ligand, it is likely to be conserved because of structural relevance. Additional help is offered by prediction programs such as ConSolv [79], which compares the ligand-free form of the protein with the complex.

Our example, HIV-1 protease, for which numerous x-ray structures are available, normally contains a water molecule bridging the two flaps and the inhibitor. This water is necessary if one wants to reproduce the binding mode of acetyl-pepstatin in its native protein structure, 5HVP. The structure of 1HVR, however, does not contain a water molecule at that specific position. During binding, the carbonyl group of the cyclic urea displaces this water and directly stabilizes the two flaps of the protease. Therefore, docking the ligand XK263 in 1HVR requires the water site to be empty. It is possible to reproduce its binding mode only after removal of the water. However, it is not possible to know this *a priori* for every molecule in a large database for screening. Hence, in the absence of further information, we suggest removing all water molecules from the binding site.

14.4.4.6 Reference Structure

For every new project, the setup of the approach chosen for docking should be validated. The most common way to do this is by redocking a ligand to the corresponding protein structure from the complex. However to judge the performance of the method, it is crucial not to use the exact pose of the ligand from the crystal structure. This pose is the time-average over the ligand poses during the collection of the diffraction data (as is the case for the conformation of the protein).

Thus, it is likely that, according to the parameters of the applied scoring function, some atom positions have clashes with the protein. This problem can be solved by minimizing the ligand within the binding site with a gradient-based method applying the same scoring function as will be used for docking, while keeping the protein rigid. The minimized ligand then offers an appropriate reference structure for redocking calculations.

The ligand conformation which is used as input structure for the docking experiments should have been minimized with a force field outside of the binding site to remove any geometrical bias. However, one has to bear in mind that the force field will not only modify the torsion angles, but also bond lengths and bond angles. If the strain in the ligand conformation is large upon binding, the minimization outside of the receptor might yield a covalent geometry that is not compatible with the binding site. Therefore, because in the docking search only torsional degrees of freedom are considered, the docking approach might not be able to reproduce the experimental binding mode [35].

14.4.5 Running SEED

SEED provides the anchors for the final docking procedure. Thus, it is worth analyzing the SEED results in detail. One should have a close look at the binding site with a molecular viewer to see the distribution of the polar and apolar vectors used by SEED to dock the fragments. If a project is in an advanced stage and a considerable amount of structural information is available, the user should eventually change the number of the polar and apolar vectors as well as the definition of the binding site or the interior dielectric constant.

14.4.6 Running FFLD

The only parameters that should be modified in FFLD are the input values for the hybrid search algorithm. It has to be emphasized that optimal input values depend on the shape of the energy hypersurface and can thus hardly be predicted. As the limiting factor rather is the computer power, the user might want to select fewer chromosomes or fewer steps (which results in fewer energy evaluations) or a smaller frequency for the local search.

It is important, however, to perform multiple runs with different seeds for the random generation of the initial population. As with any stochastic search method, the hybrid search can be trapped in local minima. This is only detectable by comparing the results of many runs, therefore we typically perform 10 runs with different random seed numbers per ligand. Moreover, to judge the quality of the predictions, it is important to have a look at the convergence rate (i.e., which percentage of the different runs reach a similar conformation) [35]. This finding was obtained in a cross-docking study (which corresponds to the situation in a screening project) on 5 complexes of HIV-1 protease. Each of the 5 ligands was docked into all protein structures except its native one, which resulted in a total number of 20 docking experiments. For each docking experiment, convergence toward the lowest energy conformation (which is not

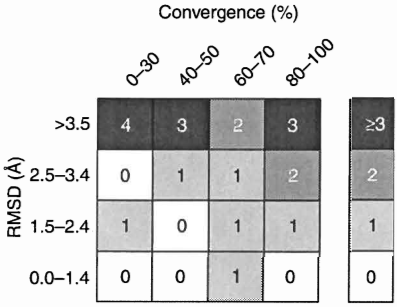

FIGURE 14.9 The density plot with the frequency of a certain rmsd from the experimentally determined structure for a given amount of convergence in 10 GA runs with different seeds. As an example, the "3" in the top right corner means that in 3 of the 20 docking experiments between 8 and 10 runs converged to the same conformation and this conformation has a rmsd larger than 3.5 Å from the experimental structure.

necessarily identical to the experimental structure) in 10 FFLD runs with different seeds was determined. The convergence values were then used to build a density plot that reports the frequency of finding a binding mode with a certain root-mean-square deviation (rmsd) from the experimental structure for a given amount of convergence (Figure 14.9). This density plot is almost upper triangular, which implies that experiments with less than 60% of convergence have probably failed to locate the global minimum. Consequently, these runs should not be relied on. On the other hand, a high convergence is no guarantee for successful docking, as is shown by the high number of runs that fully converged on a wrong structure (Figure 14.9, top right corner). The reason for this is probably to be searched for in the oversimplified nature of the energy function, which precludes an accurate detection of the solution. Taken together, these results suggest that a high convergence rate in multiple GA runs may be a necessary, although not sufficient, criterion for a good prediction.

ACKNOWLEDGMENTS

We thank Dr. Nicolas Majeux for interesting discussions. We also thank Fabian Dey for comments on this chapter. The development of our docking programs has been financially supported by Novartis, Aventis, and the Swiss National Center of Competence in Research (NCCR) in Structural Biology.

REFERENCES

[1] A. Nicholls, K. Sharp, B. Honig, Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons, *Proteins: Structure, Function and Genetics* 11:281–296, 1991.

[2] M. Scarsi, N. Majeux, and A. Caflisch, Hydrophobicity at the surface of proteins, *Proteins: Structure, Function and Genetics* 37:565–575, 1999.

[3] C.M. Venkatachalam, X. Jiang, T. Oldfield, M. Waldman, LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, *J. Mol. Graphics Modelling* 21:289–307, 2003.

[4] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, Automated docking using a Lamarckian Genetic Algorithm and an empirical binding free energy function, *J. Comput. Chem.* 19:1639–1662, 1998.

[5] F. Österberg, G.M. Morris, M.F. Sanner, A.J. Olson, and D.S. Goodsell, Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock, *Proteins: Structure, Function, and Genetics* 46:34–40, 2002.

[6] C. Hetényi, and D. van der Spoel, Efficient docking of peptides without prior knowledge of the binding site, *Protein Sci.* 11:1729–1737, 2002.

[7] S.L. McGovern and B.K. Shoichet, Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes, *J. Med. Chem.* 46:2895–2907, 2003.

[8] N. Budin, N. Majeux, and A. Caflisch, Fragment-based flexible ligand docking by evolutionary optimization, *Biol. & Chem.* 382:1365–1372, 2001.

[9] H. Claussen, C. Buning, M. Rarey, and T. Lengauer, FlexE: Efficient molecular docking considering protein structure variations, *Algorithmica* 308:377–395, 2001.

[10] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, and A. Caflisch, Exhaustive docking of molecular fragments with electrostatic solvation, *Proteins: Structure, Function, and Genetics* 37:88–105, 1999.

[11] N. Majeux, M. Scarsi, and A. Caflisch, Efficient electrostatic solvation model for protein-docking, *Proteins: Structure, Function, and Genetics* 42:256–268, 2001.

[12] A. Fahmy and G. Wagner, TreeDock: a tool for protein docking based on minimizing van der Waals energies, *J. Am. Chem. Soc.* 124:1241–1250, 2002.

[13] E.C. Meng, B.K. Shoichet, and I.D. Kuntz, Automated docking with grid-based energy evaluation, *J. Comput. Chem.* 13:505–524, 1992.

[14] I.D. Kuntz, E.C. Meng, S.J. Oatley, R. Langridge, and T.E. Ferrin, A geometric approach to macromolecule–ligand interactions, *J. Mol. Biol.* 161:269–288, 1982.

[15] T.J.A. Ewing, S. Makino, A.G. Skillman, and I.D. Kuntz, DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, *J. Computer-Aided Mol. Design* 15:411–428, 2001.

[16] A.N. Jain, Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine, *J. Med. Chem.* 46:499–511, 2003.

[17] OpenEye Software, FRED, 2002. http://www.eyesopen.com/products/applications/fred.html.

[18] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, A fast flexible docking method using an incremental construction algorithm, *J. Mol. Biol.* 261:470–489, 1996.

[19] R. DeWitte, and E. Shakhnovich, SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence, *J. Am. Chem. Soc.* 118:11733–11744, 1996.

[20] R. DeWitte, A. Ishchenko, and E. Shakhnovich, SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates: 2. Case studies in molecular design, *J. Am. Chem. Soc.* 119:4608–4617, 1997.

[21] M. Thormann and M. Pons, Massive docking of flexible ligands using environmental niches in parallelized genetic algorithms, *J. Comput. Chem.* 22:1971–1982, 2001.

[22] G. Jones, P. Willett, and R.C. Glen, Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation, *J. Mol. Biol.* 245:43–53, 1995.

[23] C.M. Oshiro, I.D. Kuntz, and J.S. Dixon, Flexible ligand docking using a genetic algorithm, *J. Computer-Aided Mol. Design* 9:113–130, 1995.

[24] P.G. Maillot, *Graphics Gems*, London: Academic Press, p. 498, 1996.

[25] T.J. Oldfield, A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. *Acta Cryst.* D57:82–94, 2001.

[26] T.J. Oldfield, X-ligand: an application for the automated addition of flexible ligands into electron density, *Acta Cryst.* D57:696–705, 2001.

[27] G. Jones, P. Willett, R.C. Glen, A.R. Leach, and R. Taylor, Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.* 267:727–748, 1997.

[28] V. Schnecke and L. Kuhn, Virtual screening with solvation and ligand-induced complementarity, *Persp. Drug Discov. Des.* 20:171–190, 2000.

[29] T.W. Whitfield, and J.E. Straub, Gravitational smoothing as a global optimization strategy, *J. Comput. Chem.* 23:1100–1103, 2002.

[30] U.H.E. Hansmann and L.T. Wille, Global optimization by energy landscape paving, *Phys. Rev. Lett.* 88:068105, 2002.

[31] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in Fortran*, Cambridge, UK: Cambridge University Press, 1992.

[32] H.J. Bohm, The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure, *J. Computer-Aided Mol. Design* 8:243–256, 1994.

[33] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee, Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, *J. Computer-Aided Mol. Design* 11:425–445, 1997.

[34] G.M. Verkhivker, D. Bouzida, D.K. Gehlhaar, P.A. Rejto, S. Arthurs, A.B. Colson, S.T. Freer, V. Larson, B.A. Luty, T. Marrone, and P.W. Rose, Binding energy landscapes of ligand–protein complexes and molecular docking: principles, methods, and validation experiments. In A.K. Ghose, and V.N. Viswanadhan, Eds., *Combinatorial Library Design and Evaluation: Principles, Software, Tools, and Applications in Drug Discovery*, New York: Marcel Dekker, pp. 157–195, 2001.

[35] M. Cecchini, P. Kolb, N. Majeux, and A. Caflisch, Automated docking of highly flexible ligands by genetic algorithms: a critical assessment, *J. Comput. Chem.* 25:415–422, 2004.

[36] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, CHARMm: a program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.* 4:187–217, 1983.

[37] W. Cornell, P. Cieplak, C. Bayly, I. Gould, K. Merz, Jr., D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.* 117:5179–5197, 1995.

[38] M.L. Verdonk, J.C. Cole, P. Watson, V.J. Gillet, and P. Willett, SuperStar: improved knowledge-based interaction fields for protein binding sites, *J. Mol. Biol.* 307:841–859, 2001.

[39] I. Muegge, A knowledge-based scoring function for protein–ligand interactions: probing the reference state, *Persp. Drug Discov. Des.* 20:99–114, 2000.

[40] A.V. Ishchenko and E.I. Shakhnovich, Small Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein–ligand interactions, *J. Med. Chem.* 45:2770–2780, 2002.

[41] H. Gohlke, M. Hendlich, and G. Klebe, Knowledge-based scoring function to predict protein–ligand interactions, *J. Mol. Biol.* 295:337–356, 2000.

[42] P.S. Charifson, J.J. Corkery, M.A. Murcko, and W.P. Walters, Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins, *J. Med. Chem.* 42:5100–5109, 1999.

[43] R.D. Clark, A. Strizhev, J.M. Leonard, J.F. Blake, and J.B. Matthew, Consensus scoring for ligand/protein interactions, *J. Mol. Graphics Modelling* 20:281–295, 2002.

[44] J. Warwicker and H.C. Watson, Calculation of the electric potential in the active site cleft due to α-helix dipoles, *J. Mol. Biol.* 157:671–679, 1982.

[45] M.K. Gilson and B.H. Honig, Energetics of charge-charge interactions in proteins, *Proteins: Structure, Function, and Genetics* 3:32–52, 1988.

[46] D. Bashford, and M. Karplus, pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model, *Biochem.* 29:10219–10225, 1990.

[47] M.E. Davis, J.D. Madura, B.A. Luty, and J.A. McCammon. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian dynamics program, *Comput. Phys. Comm.* 62:187–197, 1991.

[48] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, *J. Am. Chem. Soc.* 112:6127–6129, 1990.

[49] M. Scarsi, J. Apostolakis, and A. Caflisch, Continuum electrostatic energies of macromolecules in aqueous solutions, *J. Phys. Chem.* A101:8098–8106, 1997.

[50] N. Budin, S. Ahmed, N. Majeux, and A. Caflisch. An evolutionary approach for structure-based design of natural and non-natural peptidic ligands, *Comb. Chem. High Throughput Screen.* 4:695–707, 2001.

[51] N. Budin, N. Majeux, C. Tenette-Souaille, and A. Caflisch, Structure-based ligand design by a build-up approach and genetic algorithm search in conformational space, *J. Comput. Chem.* 22:1956–1970, 2001.

[52] X. Zou, Y. Sun, and I.D. Kuntz, Inclusion of solvation in ligand binding free energy calculations using the generalized-Born model, *J. Am. Chem. Soc.* 121:8033–8043, 1999.

[53] N. Arora, and D. Bashford, Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions, *Proteins: Structure, Function, and Genetics* 43:12–27, 2001.

[54] M. Rarey, B. Kramer, and T. Lengauer. The particle concept: placing discrete water molecules during protein–ligand docking predictions, *Proteins: Structure, Function, and Genetics* 34:17–28, 1999.

[55] J. Apostolakis, A. Plückthun, and A. Caflisch, Docking small ligands in flexible binding sites, *J. Comput. Chem.* 19:21–37, 1998.

[56] R.M.A. Knegtel, I.D. Kuntz, and C.M. Oshiro, Molecular docking to ensembles of protein structures, *J. Mol. Biol.* 266:424–440, 1997.

[57] R.M. Jackson, H.A. Gabb, and M.J.E. Sternberg, Rapid refinement of protein interfaces incorporating solvation: application to the docking problem, *J. Mol. Biol.* 276:265–285, 1998.

[58] P. Koehl and M. Delarue, Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy, *J. Mol. Biol.* 239:249-275, 1994.

[59] P. Koehl and M. Delarue, Mean-field minimization methods for biological macromolecules, *Curr. Opin. Struct. Biol.* 6:222–226, 1996.

[60] J.H. Lin, A.L. Perryman, J.R. Schames, and J.A. McCammon, Computational drug design accommodating receptor flexibility: the relaxed complex scheme, *J. Am. Chem. Soc.* 124:5632–5633, 2002.

[61] E. Yuriev and P.A. Ramsland, Mcg light chain dimer as a model system for ligand design: a docking study, *J. Mol. Recognit.* 15:331–340, 2002.

[62] J.D. Diller and C.L.M.J. Verlinde, A critical evaluation of several global optimization algorithms for the purpose of molecular docking, *J. Comput. Chem.* 20:1740–1751, 1999.

[63] A. Caflisch, P. Niederer, and M. Anliker, Monte Carlo docking of oligopeptides to proteins, *Proteins: Structure, Function, and Genetics* 13:223–230, 1992.

[64] M. Miller, S.K. Kearsley, D.J. Underwood, and M.D. Sheridan, FLOG — a system to select quasi-flexible ligands complementary to a receptor of known 3-dimensional structure, *J. Computer-Aided Mol. Design* 8:153–174, 1994.

[65] S. Makino and I.D. Kuntz, Automated flexible ligand docking method and its application for database search, *J. Comput. Chem.* 18:1812–1825, 1997.

[66] D.E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Reading, MA: Addison-Wesley, 1989.

[67] L. Davis, Ed., *Handbook of Genetic Algorithms*, New York: Van Nostrand Reinhold, 1991.

[68] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, and S.T. Freer, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming, *Chem. Biol.* 2:317–324, 1995.

[69] G.M. Verkhivker, P.A. Rejto, D.K. Gehlhaar, and S.T. Freer, Exploring the energy landscape of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKBP-12 complexes, *Proteins: Structure, Function, and Genetics* 25:342–353, 1996.

[70] L. Schaffer and G.M. Verkhivker, Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization, *Proteins: Structure, Function, and Genetics* 33:295–310, 1998.

[71] D. Hoffman, B. Kramer, T. Washio, T. Steinmetzer, M. Rarey, and T. Lengauer, Two-stage method for protein–ligand docking, *J. Med. Chem.* 42:4422–4433, 1999.

[72] J. Wang, P.A. Kollman, and I.D. Kuntz, Flexible ligand docking: a multistep strategy approach, *Proteins: Structure, Function, and Genetics* 36:1–19, 1999.

[73] M.L.P. Price, and W.L.J. Jorgensen, Analysis of binding affinities for celecoxib analogues with COX-1 and COX-2 from combined docking and Monte Carlo simulations and insight into the COX-2/COX-1 selectivity, *J. Am. Chem. Soc.* 122:9455–9466, 2000.

[74] W. Kabsch, A solution for the best rotation to relate two sets of vectors, *Acta Cryst.* A32:922–923, 1976.

[75] K. No, J. Grant, and H. Scheraga, Determination of net atomic charges using a modified partial equalization of orbital electronegativity method: 1. Application to neutral molecules as models for polypeptides, *J. Phys. Chem.* 94:4732–4739, 1990.

[76] K. No, J. Grant, M. Jhon, and H. Scheraga. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method: 2. Application to ionic and aromatic molecules as models for polypeptides, *J. Phys. Chem.* 94:4740–4746, 1990.

[77] K. Hilpert, J. Ackermann, D.W. Banner, A. Gast, K. Gubernator, P. Hadvary, L. Labler, K. Müller, G. Schmid, T. Tschopp, and H. van de Waterbeemd, Design and synthesis of potent and highly selective thrombin inhibitors, *J. Med. Chem.* 37:3889–3901, 1994.

[78] S. Piana, D. Sebastiani, P. Carloni, and M. Parrinello, *Ab initio* molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site, *J. Am. Chem. Soc.* 123:8730–8737, 2001.

[79] M.L. Raymer, P.C. Sanschagrin, W.F. Punch III, S. Venkataraman, E.D. Goodman, and L.A. Kuhn, Predicting conserved water and water-mediated ligand interactions in proteins using a k-nearest-neighbor genetic algorithm, *J. Mol. Biol.* 265:445–464, 1997.