29 Protein–Ligand Docking as an Energy Optimization Problem

MAXIM TOTROV
Molsoft LLC, San Diego, CA, USA

RUBEN ABAGYAN
Genomics Institute of the Novartis Research Foundation and
The Scripps Research Institute, San Diego, CA, USA

Formation of non-covalent complexes is an essential part of almost any biological process. Remarkable complexity of the biochemical machinery of the living organisms would have been impossible without the ability of the participating molecules to recognize each other among thousands of other compounds simultaneously present in any cell. Specific binding between molecules is crucial in catalysis, signal transduction and molecular transport mechanisms, and determines the pharmacological effect of many drugs. Better knowledge of the nature of molecular recognition on the microscopic level is important for our understanding of the normal and pathological processes in the cell and may help in such practical applications as drug design. X-ray crystallography has revealed detailed atomic descriptions of many individual proteins, nucleic acids and small biological molecules, as well as a number of structures of complexes. The Protein Data Bank (PDB) (Bernstein et al. 1977), where solved protein three-dimensional (3D) structures are deposited, is growing by about 1000 new structures a year. Available structures of complexes can be analyzed to discover the basic interactions and principles of molecular recognition, whereas the individual structures can be used in the prediction of unknown or novel complexes. First attempts to predict molecular interactions and design novel ligands utilized hand-made physical models of receptor sites and ligands (Beddell et al. 1976). Because the manipulation of the systems containing hundreds or thousands of atoms is necessary to simulate the binding process, the progress in numerical computational approaches was essential for the advancement of the macromolecular association studies. Computer simulations of molecular recognition were first attempted more than 20 years ago (Kuntz et al. 1982). Considerable progress has been achieved in recent years but the reliability and precision of the existing complex prediction methods are still far from ideal.
MOLECULAR DOCKING SIMULATIONS

Prediction of the structure of a complex starting from the structures of individual molecules is commonly called a molecular docking problem. Structures of the protein–ligand and especially protein–protein complexes often show remarkable shape complementarity on the interface, suggesting the idea that the docking algorithms should search for such matching surfaces. Early approaches such as the original DOCK (Kuntz et al. 1982) exclusively used this geometric criterion. Both components of the complex were assumed to be rigid and the docking procedure searched for favourable mutual orientation using ‘sphere matching’ (DesJarlaists et al. 1986), least-squares fitting of the surface patterns (Bacon and Moult 1992; Leach and Kuntz 1992). Fourier transform (Katchalski-Katzir et al. 1992), distance–matrix matching (Helmer-Citterich and Tramontano 1994) or ‘geometric hashing’ (Fischer et al. 1995). Purely geometric approaches demonstrated certain success in recombining the structures of protein–protein complexes when the components were taken from the native complexed structure, which is a somewhat artificial starting point. In the more realistic cases where the individual structures of the constituents were used, these techniques often failed to distinguish the correct orientation (Bacon and Moult 1992). High complementarity of the interacting surfaces in the native complexes is due in part to the ‘induced fit’, e.g., the conformational change in the constituents of the complex upon binding, whereas individual structures often do not show the perfect matching expected in the complex. There are two general directions in which the simplistic geometric docking algorithms are being improved: the introduction of flexibility of ligand and/or receptor to reproduce or mimic the induced fit; and the inclusion of binding determinants other than pure surface complementarity. First attempts to introduce flexibility in protein–protein docking were limited to ‘softening’ of the geometric criteria, which would allow a certain degree of penetration between the two interacting surfaces (Jiang and Kim 1991; Walls and Sternberg 1992). Direct simulations with all-atom models may account for the flexibility more accurately and sometimes show promising results (Nilges and Brunger 1993; Abagyan et al. 1994; DiNoia et al. 1994), but are often extremely computationally expensive. Whichever way the flexibility is introduced, it results in much greater ambiguity of the results of geometric docking, because many approximate matches can be found. The multiplicity of solutions calls for additional criteria to select the correct answer. This leads to the inclusion in the docking protocol of the other binding determinants, such as estimates of solvation free energy change or molecular mechanics energy (Shoichet and Kuntz 1991), and ultimately the approximations of the free energy change upon binding (Bohm 1994a,b). Most methods, however, still use simplistic but faster measures during the generation of the bound conformations and then re-evaluate the putative solutions using more sophisticated potentials.

DOCKING AS AN ENERGY OPTIMIZATION PROBLEM

Complexes considered in the docking studies are, in general, thermodynamically stable systems. Thus, the native bound conformation should represent the global minimum of the free energy. Consequently, to find the docked conformation, the global minimum of the free energy function of the system has to be located. Because the precise evaluation of the free energy is difficult, one can try to use some approximation that would have a similar global minimum. From the energetic point of view, surface complementarity docking methods assume that the interaction energy is proportional to the contact area or other similar measure of the fit of two surfaces, possibly with some penalty for bad contacts (clashes). Although this assumption may account reasonably well for van der Waals interactions and, to some extent, for solvation, it obviously disregards the energy contributions from specific pairwise atomic interactions such as hydrogen bond formation and electrostatics. Many recent docking studies try to incorporate these terms, often as the additional criteria to select the answer from several solutions generated by geometric docking, either using force-field energy evaluation (Shoichet and Kuntz 1991) or elaborate scoring functions (Bohm 1994a; Jain 1996). In several works, physical energy terms were used throughout the algorithm (Abagyan et al. 1994; Totrov and Abagyan 1994).

Two major components are required for a successful prediction of the structure of the protein–ligand complex: an efficient global optimization procedure, which is capable of finding a global minimum for the strongly anisotropic function of dozens of variables; and a free energy approximation for the complex in solution, which is computationally inexpensive to be used in the search procedure yet sufficiently accurate to ensure the uniqueness of the native conformation. In the following we will review the energy calculations and global optimization methods.

ENERGY TERMS

Energy calculations are at the center of almost any molecular simulation technique. It is convenient and customary to divide the energy of the molecular system into a number of components, or energy terms. Below, five major terms of the molecular interaction energy will be considered in greater detail.

ELECTROSTATIC INTERACTIONS

Electromagnetism is the fundamental force of biochemistry (Davis and McCammon 1990). All processes on the molecular level can be described in terms of electromagnetic interaction combined with quantum mechanical and
thermodynamic effects. Although covalent and hydrogen bonding as well as van der Waals interaction all have electrostatic nature, these interactions are complicated by quantum mechanics and it is often convenient to separate them from the longer range electrostatic interactions. It is the latter type of interactions that are customarily referred to as electrostatics in biomolecular structure. All proteins and a large majority of ligands contain polar atoms interacting strongly with each other and the solvent in the wide range of distances. For a charged amino acid the strength of electrostatic forces may exceed by more than an order of magnitude the strength of van der Waals interaction (Warshel and Russell 1984).

The evaluation of electrostatic interactions in proteins was first attempted by Lingstrom-Lang in 1924 and a theory of electrostatics in macromolecules was proposed (Tanford and Kirkwood 1957). These macroscopic studies gave some qualitative insights, but only the availability of high-resolution protein structures and computer calculations allowed quantitative studies of protein electrostatics. The largest problem in electrostatic calculations is the presence of a highly polar solvent (water). In vacuum or in the uniform media the interaction between two charges can be described simply by Coulomb’s law.

\[ E = k \frac{q_1 q_2}{\varepsilon R_{12}} \]

where \( q_{1,2} \) are the charges, \( R_{12} \) is the distance between them, \( \varepsilon \) is the dielectric constant and \( k \) is 332.0 when the charges are electron units; distance is expressed in angstroms and energy in kcal mol\(^{-1}\). In an aqueous environment this relation has to be corrected to include the interaction of the charges under consideration with a large (virtually infinite) number of surrounding water molecules. Early attempts to simulate macromolecules without consideration of solvent screening ran into difficulties, for example a DNA double-helix would be torn apart by electrostatic forces unless the electric charges were drastically reduced (Harvey 1989).

The straightforward and rigorous approach is to include explicitly a sufficiently thick layer of water molecules into the calculations. Obviously, it makes the calculations heavier but the principal difficulty of the explicit methods is that liquid water is essentially a dynamic environment. Any static placement of water molecules around the system under consideration would result in large errors, because the physically observed interaction with water is the result of averaging over a large thermodynamic ensemble of the possible states of the solvent. Thus, to achieve accurate results one has to generate this ensemble by an extensive molecular dynamics simulation (Rastelli et al. 1995, Simmerling and Elber 1995). Although this might be the most rigorous approach to the solvation electrostatic calculations, in most cases it is impractical. Langevine dipoles were proposed (Rossky et al. 1978; Luzzhkov and Warshel 1992) to make implicit averaging over the water molecule's orientations, which eliminates the necessity of generation of a large ensemble of water configurations. The method was applied in protein–protein docking and gave promising results (Jackson et al. 1998).

The solvent effectively screens the interaction of the charges of the solute. Generally, the farther from each other and the more exposed to the solvent the charges are, the more their interaction is attenuated. This observation suggested simple corrections to the Coulomb law, such as a distance-dependent dielectric constant and charge scaling. Although it is somewhat \textit{ad hoc} and does not take into account the interaction of the individual charges with the solvent (self-energy), a distance-dependent dielectric constant \( \varepsilon = \varepsilon_0 R \) is widely used because of its simplicity (McCament et al. 1979; Pickersgill 1988). This expression actually accelerates calculations of the energy and forces because they become dependent only on \( R^2 \) instead of \( R \), eliminating costly square-root calculations. Charge scaling was shown to improve the simulation results for such systems as DNA. Although these crude approaches can hardly be used for quantitative evaluation of the properties of a macromolecule in solution, they keep the extra calculations to a minimum.

Alternatively, the solvent can be considered as a continuous medium of high dielectric constant. This treatment of solvent is more computationally tractable than the inclusion of explicit water molecules. The electric potential in the medium of the variable dielectric constant obeys the Poisson differential equation

\[ -\nabla [\varepsilon(r) \nabla \phi(r)] = \rho(r) \]

where \( \varepsilon(r) \) is the dielectric constant (permittivity) \( \phi \) is the electric potential and \( \rho \) is the charge density. If \( \varepsilon(r) \) \text{constant}, the Poisson equation is equivalent to the Coulomb law, but the solution becomes more complicated when the space is divided into the regions of various dielectric permittivity. Analytical results exist only for special cases such as a sphere. Certain methods utilize these analytical solutions to obtain relatively simple approximations of energy under an assumption that the protein has near-spherical shape (e.g. image method (Friedman 1975; Schaefer and Froemmel 1990; Abagyan and Totrov 1994). Similar assumptions are used in the generalized Born approximation (Still et al. 1990; Cramer and Truhlar 1992). The precision of these methods is rather limited. A much more rigorous approach is to solve the Poisson equation numerically. Several techniques based on this idea were developed and are widely used in the protein energy calculations (Zauhar and Morgan 1985; Juffer et al. 1991; Nicholls and Honig 1991; Zauhar and Varnek 1996). The main difficulty in their application to docking is high computational cost. A hybrid method was proposed recently and used in docking simulation, utilizing a single numerical solution of the Poisson equation for unbound receptor supplemented by generalized Born-type terms calculated for each specific bound ligand conformation (Majeux et al. 1999).
HYDROPHOBICITY

Transfer to the aqueous solution of a number of organic groups results in a free energy loss related to the ordering of water molecules around such groups, which is known as a hydrophobic effect. The concept of hydrophobic interaction was introduced by Kauzmann (1959). This effect is similar in nature to the macroscopic surface tension. Hydrophobic interaction is a major driving force in the formation of most ligand–receptor complexes. For some ligands, such as steroids, the interaction is almost exclusively hydrophobic and many other ligands are amphiphilic with hydrophobic groups binding into hydrophobic pockets of the receptor. By fitting the transfer free energies of hydrocarbons against the solvent-accessible surface, the hydrophobic contribution was shown (Chothia 1976) to be proportional to the solvent-accessible surface with fairly good precision. However, the coefficient of this proportionality is a subject of some controversy because it differs sharply from the microscopically observed value of the surface tension constant. The macroscopic surface tension value derived from the transfer energies of aliphatic compounds is close to $30 \text{cal A}^{-2}$ whereas the macroscopic hydrocarbon–water surface tension constant is $\sim 75 \text{cal A}^{-2}$. Some attempts were made to explain the discrepancy by taking into consideration the curvature dependence of the surface tension and the difference of the molar volume of solute and solvent (Sharp et al. 1991).

It remains to be seen if the division of the water–solute interaction into solvation electrostatics and hydrophobic components is the most adequate approach. Methods based on this partitioning were shown to reproduce successfully experimental data on transfer free energies for a large set of compounds (Sitkoff et al. 1994). However, alternative approaches to water–solute interaction evaluation were also developed, particularly a number of atomic solvation parameter (ASP)-based methods (Eisenberg and McLachlan 1986; Wesson and Eisenberg 1992). The ASP methods differentiate the atoms of the solute into a number of types, each with a particular value of solvation energy surface density, generalizing the surface tension. The underlying assumption is that the water–solute interaction can be partitioned into atomic contributions, which are proportional to the solvent-accessible surface areas of the atoms. The popularity of the ASP approach is due in part to the simplicity and computational efficiency, but the drawbacks are that neither the proportionality of the solvation energy to the accessible surface nor the partitioning of the solvation energy into atomic contributions can be rigorously justified and are largely ad hoc assumptions. Nevertheless, good agreement with experimental data can be achieved (Horton and Lewis 1992), which might be explained in part by the large number of adjustable parameters in the ASP models. It is questionable that these methods can perform well on a set of compounds that is much larger than the set used for the parameter adjustment.

VAN DER WAALS INTERACTIONS

The most generic type of interatomic force—van der Waals interactions—exhibits itself as a very strong repulsion at short distances and turns into relatively weak and quickly decreasing attraction as the distance between two atoms grows. It is commonly described by 6–12 potential

$$E_{\text{VW}}(R_{ij}) = - \frac{A_{ij}}{R_{ij}^6} + \frac{B_{ij}}{R_{ij}^{12}}$$

where $R_{ij}$ is the distance between the two atoms $i$ and $j$. Parameters $A_{ij}$ and $B_{ij}$ depend on the types of atoms and are usually calculated using combination rules from the parameters for the identical pairs of atoms, which in turn are evaluated from quantum mechanical or experimental data. Usually these parameters are derived along with the other components of the atomic interaction energy to form so-called molecular mechanics force fields, such as CHARMM (Brooks et al. 1983), AMBER (Weiner et al. 1984), MMFF (Halgren 1995) and ECEPP (Momany et al. 1975). Although the $1/R^6$ form of the attraction term has a strict quantum mechanical basis, rigorous description of the repulsion term is more complicated. Alternative forms of the repulsion term have been proposed (e.g. Halgren 1995). Fortunately, the interactions in biomolecular systems occur mostly in the range of interatomic distances where the attractive term is prevalent, and seem to avoid strong repulsion, thus alleviating the problem of the exact description of the repulsive term.

Still, the extreme sensitivity of van der Waals interactions to the small conformational changes makes its inclusion in the calculation of binding energy problematic. This led a number of authors simply to omit the van der Waals contribution in the binding energy, because it seems to introduce more noise than signal into the energy estimates (Krystek et al. 1993; Vajda et al. 1994). Such omission is partly justified by the cancellation of ligand–receptor interactions in the bound state and the ligand-solvent/receptor-solvent interactions in the unbound state. One can assume that the overall number of interatomic contacts in the system remains nearly constant upon binding, resulting in conservation of the total van der Waals interaction energy. Geometric docking can be used to achieve reasonably good packing. However, this approach leaves out entirely the dependence of the interaction energy on the quality of the interface. In the case of the docking of novel ligands, evaluation of the interface is essential for determination of the binding likelihood and the correct binding mode. Possible compromise is to modify the van der Waals potential so that it becomes less sensitive to the small deviations in atomic coordinates.
HYDROGEN BONDS

Hydrogen bond interaction is a specific attraction between polar hydrogens and a number of heavy atoms (primarily oxygen, nitrogen, sulfur) that have unshared electron pairs. The observations of large number of complexes with solved 3D structures show that many ligands form extensive networks of hydrogen bonds with their receptors, especially in cases of high-specificity and high-affinity binding. Hydrogen bonds also play an important role in protein folding, where their formation between the turns of the a-helices and between the b-strands stabilizes these essential secondary structure elements. Unfortunately, there seems to be no agreement so far about the adequate functional form for the hydrogen bonding interaction term, or even the energetic value of an average hydrogen bond. Because its origin lays in same electrostatic and quantum interactions as the origin of van der Waals and electrostatic terms, hydrogen bonding is often included in the force field as a modification to the van der Waals potential for specific atom pairs (Nemethy et al. 1992; Halgren 1995). The modification may only involve a change in the parameters (MMFF) or a different functional form (10–12 instead of the standard 6–12 van der Waals potential in ECEPP). Some force fields simply ignore hydrogen bonding in the hope that the electrostatic term will provide sufficient favorable contribution when positive hydrogen atoms and negative hydrogen bond acceptors are brought together. However, the charge distribution around the acceptor atoms is highly anisotropic because the unshared electron pairs occupy sp^3 orbitals, resulting in strong anisotropy of the hydrogen bond interaction. High directionality of the hydrogen bond interaction can also be observed in the solved structures of the proteins and protein complexes (Ippolito et al. 1990). This anisotropy is largely ignored by pairwise, atom-centric potentials used by the majority of the force fields. Such omission may not lead to large errors as long as only naturally occurring conformations are considered because they often already have optimal or suboptimal configuration of hydrogen bonds. However, in the course of a simulation, such as docking, it may result in erroneous formation of hydrogen bonds of physically impossible geometries. Several forms of the hydrogen bonding term with explicit angular dependence were proposed (Goodford 1985; Miller et al. 1994).

CONFORMATIONAL ENTROPY

Binding of the ligand to the receptor usually imposes strong constraints upon its conformational freedom. Also the surface side chains of the receptor that are in contact with the ligand may no longer access some of their rotameric states. There is a loss in translational and rotational degrees of freedom, which does not depend on the participating molecules and can be seen as constant as long as only 1:1 stoichiometry complexes are considered. Thus, binding may result in a substantial decrease in the entropy. As an illustration, one can consider the burial of one CH_3 group in an aliphatic chain. The loss of three rotameric states of the chain results in the entropy loss that adds R T ln 3 = 0.66 kcal mol^-1 to the free energy of the system, whereas the decrease in the hydrophobic term is around -0.88 kcal mol^-1 (Yang et al. 1992).

Exact determination of the entropy change would require extensive molecular dynamics simulations. Currently such simulations are too expensive computationally to use them routinely for a large number of putative complexed structures. Docking methods generally assume that upon binding the ligand is locked in a single conformation. Although in some cases this assumption might be far from true, it allows exclusion of the conformational entropy term from docking simulations as a constant.

CONFORMATIONAL SEARCH TECHNIQUES

An efficient global optimization procedure is a key component of the docking protocol. Many approaches treat both ligand and receptor as rigid bodies (Kuntz et al. 1982; Cherfils et al. 1991; Bacon and Moult 1992). Such treatment allows for rapid location of the optimal mutual orientation of the two molecules by special techniques (DOCK), but has limited applicability because the majority of small ligands are flexible and structural rearrangements occur in a number of receptors. To some extent, the limitations of the rigid-body docking can be circumvented if several low-energy conformations of the ligand are generated and then docked. The best solution then can be picked as an answer (Kearsley et al. 1994; Leach 1994). However, the number of conformations that have to be docked independently to achieve an accurate solution may become very large even for relatively small compounds. Therefore, many techniques try to treat the flexibility of the ligand more directly. Flexible ligand often can be partitioned into rigid fragments. For each fragment, rigid docking can produce a number of favorable orientations. Fragments are then reassembled into the original chemical structure (Gulukota et al. 1996; Hammerhead in Welch et al. 1996). Alternatively, one fragment is assumed to be essential for binding and placed in the active site first, then others are attached incrementally (Bohm 1992; Rarey et al. 1996).

Global optimization of the free energy function with respect to the orientation and the conformation of the ligand is, perhaps, the most strict approach. However, two features of the protein–ligand energy landscape complicate the problem of the energy optimization: high dimensionality and multiplicity of local minima. High dimensionality makes the exhaustive search of the conformational space very expensive computationally. Large numbers of local minima make rational determination of the global search direction virtually impossible and limit the usability of the derivatives to a
small vicinity of one local minimum. In order to deal with these difficulties, techniques such as Monte-Carlo minimization (Caflisch et al. 1992; Totrov and Abagyan 1997; Trosset and Scheraga 1998), Monte Carlo with simulated annealing (Goodsell and Olson 1990; Goodsell et al. 1996) and genetic algorithms (Jones et al. 1997) have been applied with various success. Some of these methods use internal coordinates to reduce the dimensionality of the search space.

**MONTE CARLO**

The term Monte Carlo has been introduced by Metropolis and Ulam (Metropolis et al. 1953), with an allusion to the essentially random nature of such simulations. Monte Carlo minimization consists of three repetitive steps:

(i) Random Jump: one or several variables in the system are changed randomly.

(ii) Local Minimization: the energy of randomized conformation is optimized using the conjugate gradient or quasi-Newton technique to achieve a new local minimum.

(iii) Evaluation: a new conformation is accepted or rejected according to the Metropolis criterion — if the energy of the new conformation \( E_{\text{new}} \) is lower than the energy of the old one \( E_{\text{old}} \), a new conformation is always accepted and used in the next iteration, otherwise it is accepted with the probability of

\[
P_{\text{acc}} = \exp\left[-(E_{\text{new}} - E_{\text{old}})/kT\right]
\]

where \( k \) is Boltzmann’s constant and \( T \) is the effective temperature of the simulation.

It has been established that a full local minimization after each random step greatly improves the efficiency of the procedure (Li and Scheraga, 1987; Abagyan and Argos 1992). However, some components of the energy, such as solvation electrostatic energy, might have no derivatives and/or might be too expensive computationally for local minimization. A double-energy Monte Carlo minimization scheme (Abagyan et al. 1994) circumvents this obstacle by using two sets of energy terms, one for the local gradient minimization stage and another one for the Metropolis criterion evaluation stage in the Monte Carlo step. Such division can be justified if the extra terms included for the Metropolis criterion are relatively ‘slow’, i.e. insensitive to small conformational changes.

**INTERNAL COORDINATES**

One of the principal difficulties in biomolecular simulations is the size of the system, which often contains thousands of atoms. As a consequence, the conformational space has a very high dimensionality, complicating the search for the global energy minimum. The use of internal coordinates substantially reduces the number of variables defining the conformation of the system. A cartesian description requires three variables \((x,y,z)\) per atom. An internal coordinates description uses bond lengths, planar angles and torsion angles instead. Because bond lengths and planar angles are essentially rigid at normal conditions, one can consider them as constants and only allow torsion angle changes (rotations around the bonds), reducing the dimensionality of conformational space by at least threefold. Practically, even greater reduction is achieved because at every branching point several atoms share the same torsion angle (Figure 29.1).

A formal geometrical description to allow efficient manipulations of the multi-molecular system in internal coordinates with arbitrary subsets of free and fixed variables was introduced (Abagyan et al. 1994). The technique represents the system as a directed tree-like graph imposed on all atoms as well as on some auxiliary virtual atoms (Figure 29.1). Each atom in this basic description has three geometric parameters determining its position with respect to the preceding part of the tree. The parameters are bond length \( b \), bond angle \( \omega \) and torsion \( \phi \) or phase \( \phi \) dihedral angles for the main branch and side branches, respectively. The sub-trees of different molecules join in the starting triple of virtual atoms, which are fixed at the origin of the coordinate system and allow for standard treatment of all real atoms including the root atoms of each molecular sub-tree. When several internal variables are fixed (considered constant) a group of atoms may form a so-called rigid body, where mutual positions of the atoms involved do not change with any changes of the remaining free variables. The concept of rigid bodies provides an important additional advantage for the energy calculations, because all pair-wise energy contributions from the atoms within a rigid body are constant. Such contributions often can be excluded from the calculations when only the relative energy change is important, improving the computational performance.

**OTHER APPROACHES**

Various global optimization techniques were applied to the docking problem. Among the more popular is the genetic algorithm (GA), which was widely applied in protein folding simulations (Clearwater et al. 1991; Unger and Moult 1993; Dandekar and Argos 1994). The idea of the GA is to mimic the evolution process by manipulating ‘chromosomes’, each containing a set of
fairly successfully to reconstitute a large number of known complexes (Jones et al. 1997), although no tests were undertaken to compare its performance with more conventional approaches such as Monte Carlo.

Notably, Fourier transform was also used to locate the optimal geometric fit (Katchalski-Katzir et al. 1992). The method is efficient and attractively simple conceptually. Unfortunately it seems to be applicable only to a rather simplistic fitness function and can only optimize efficiently the three translational degrees of freedom. Rotations still need to be sampled by other means, i.e. systematic or random search. The Fourier transform approach may be useful primarily in the cases where the interacting molecules are very big, making other methods too expensive computationally.

Molecular dynamics (MD) simulation can be used as an optimization method, and potentially it can provide a realistic picture of the binding process. However, MD is the most expensive approach computationally and so far it is impossible to simulate the whole progress of the system from unbound components to the complex. The use of MD in docking is now limited to the simulations of the already bound complexes, where it is used successfully to predict various thermodynamic properties (Miranker and Karplus 1991; DiNola et al. 1994; Luty et al. 1995; Rosenfeld et al. 1995). Somewhat better performance can be achieved using so-called Brownian dynamics (Rossky et al. 1978), which was applied to simulate long-range diffusion-like motions of the interacting macromolecules (Kozack and Subramaniam 1993).

AN EXAMPLE DOCKING STUDY ON A SET OF PROTEIN–LIGAND COMPLEXES WITH KNOWN 3D STRUCTURES

We tested the ability of an internal coordinate Monte-Carlo minimization docking procedure to predict the native conformations of protein–ligand complexes using a benchmark set of 51 high-resolution structures from the PDB. Ligands were diverse in size (from 12 to 84 atoms) with a broad range of chemical properties and included sugars, fatty acids, phosphates, bases, heterocyclic and other compounds, which ensured the applicability of the docking procedure to a large variety of receptor–ligand pairs.

METHODOLOGY

Energy

Our energy estimate used during the docking simulations consisted of the following terms:

$$E = E_{\text{FFint}} + E_{\text{VW}} + E_{\text{HB}} + E_{\text{HP}} + E_{\text{EL}}$$
where $\Delta E_{\text{FF}}$ is the force-field energy, which included internal van der Waals interactions and torsion energy for the ligand calculated with ECEPP/3 parameters (Nemethy et al. 1992). Because ECEPP/3 only has parameters for amino acid atom types, the atoms of ligands were assigned the closest chemically similar atom types. The rest of the terms refer to intermolecular interactions.

Because of its extreme rigidity, the van der Waals potential in its standard 6–12 form may introduce a large noise in the energy function. For intermolecular interactions, therefore, we used a modified smoother form of the potential with most of the repulsive part truncated. Truncation was achieved by the following transformation of the original value of van der Waals potential

$$E_{\text{VW}} = \begin{cases} E_{\text{VW}}^0 & \text{if } E_{\text{VW}} \leq 0 \\ \frac{E_{\text{VW}}^0 E_{\text{max}}}{E_{\text{VW}}^0 + E_{\text{max}}} & \text{if } E_{\text{VW}} > 0 \end{cases}$$

This expression ensures a smooth transition from the undistorted form of van der Waals potential in the negative range of values to the increasingly attenuated form in the positive range, asymptotically approaching the $E_{\text{max}}$ cut-off value. The $E_{\text{max}}$ was chosen on the basis of preliminary tests to be 1.5 kcal mol$^{-1}$. Lower values sometimes result in severely clashed docking solutions because the van der Waals repulsion is no longer able to compete with attractive terms, primarily electrostatics. This and other interaction potentials were precalculated on a grid to accelerate energy evaluation during the simulations. The grid cell size was set to 0.5 Å.

Parameter $\Delta E_{\text{HB}}$ is the hydrogen bonding term, which was calculated using a Gaussian-type potential positioned around the center of each lone electron pair of the hydrogen bond acceptors

$$E_{\text{HB}} = E_{\text{HB}}^0 e^{-r \cdot r_{\text{mp}}^{2} / 2 e_{\text{HB}}}$$

The peak interaction energy $E_{\text{HB}}^0$ was assumed to be 2.5 kcal mol$^{-1}$ as an average of various estimates, and the radius of the interaction sphere $d_{\text{HB}}$ was assumed to be 1.4 Å, allowing for about a 30–40° deviation from the ideal geometry in accordance with observations in X-ray structures. $r_{\text{mp}}$ is the radius vector of the interaction center, which was placed 1.7 Å from the atom. In the case of hydrogen atoms the center was placed along the axis of the covalent bond attaching the hydrogen to the rest of the molecule. In the case of heavy sp$^3$ atoms, one (for nitrogen) or two (for oxygen) centers were placed at an angle of 120° to the existing covalent bond. For sp$^3$ oxygen and sulfur, two centers were placed in tetrahedral geometry at 109° to the existing covalent bonds and to each other.

The electrostatic term $E_{\text{EL}}$ was calculated using a modified Coulomb law with distance-dependent dielectric constant $\varepsilon = 4r$. The hydrophobic term $E_{\text{HP}}$ was calculated as being roughly proportional to the buried hydrophobic surface with the free energy density of 30 kcal mol$^{-1}$ Å$^{-2}$. To accelerate calculations, a grid-based form of the hydrophobic potential was developed. The fragments of the solvent-accessible surface were generated using the modified Shrake and Rupley algorithm (Shrake and Rupley 1973, Abagyan et al. 1994). The algorithm produces dots that evenly cover the surface. The hydrophobic potential on the grid was then calculated as

$$E_{\text{HP}} = E_{\text{HP}}^0 e^{-\left(d_{\text{surf}}/d_{w}\right)^2}$$

where $d_{\text{surf}}$ is the distance to the closest point of the hydrophobic surface and $d_{w}$ is the effective radius of the hydrophobic interaction, which was set to the diameter of the water molecule 2.8 Å. The value of $E_{\text{HP}}^0 = 3\text{kcal mol}^{-1}$ was chosen to approximate the surface tension of 30 kcal mol$^{-1}$ Å$^{-2}$ for extended hydrophobic surfaces in test cases.

**Conformational Search Procedure**

All ligand–receptor pairs in the set were docked using a flexible Monte Carlo docking procedure with potential maps as implemented in ICM software (Abagyan and Totrov 1994; Abagyan et al. 1994; Totrov and Abagyan 1997). The ICM method describes both the relative positions of two molecules and their conformations by a uniform set of internal variables. Any subset of internal variables can be subjected to local or global energy minimization procedures. In this study, the global Monte Carlo minimization procedure, similar to that described previously (Abagyan et al. 1994; Totrov and Abagyan 1997), was used. It involved a random conformational change of two possible types: a positional Pseudo-Brownian random move or internal torsion modification, followed by local energy minimization (up to 100 steps of conjugate gradient minimization) and selection by the Metropolis criterion (temperature factor was set to 600 K). Pseudo-Brownian random moves changed the position of the ligand molecule as a whole with a certain amplitude (here we used 2 Å), as well as randomly rotated it around its center of gravity by an angle close to the translation amplitude over the radius of gyration. Internal torsion angles of the ligand were randomly changed one at a time, with an amplitude of 180°.

Geometrically different (as evaluated by the root-mean-square displacement of the ligand atoms) and low-energy conformations were accumulated in the conformational stack (Abagyan and Argos 1992). The adaptive length of the Monte-Carlo runs was used, with the limit on the total number of steps proportional to the size (the number of atoms) of the ligand: $N_{\text{Msteps}} = 50N_{\text{LigAtom}}$. Similarly, an adaptive length of local minimization during the Monte-Carlo run was used: $N_{\text{LocMinSteps}} = 25 + N_{\text{LigAtom}}$. The
factors in these relations were established empirically from the convergence and efficiency considerations.

Test Data Set

The set of 51 complexes (Table 29.1) was extracted from high-resolution PDB structures. The structures were selected according to a number of criteria: we discarded all structures at resolutions worse than 2.0 Å because large errors in the receptor coordinates could result in poor docking and recognition for reasons unrelated to our study. Some complexes had the ligand bound covalently to the receptor and were also discarded because the prediction of such chemical reactions is beyond the scope of our approach. We also omitted complexes where metal ions were directly involved in the protein–ligand interaction because the force field used in the simulations did not provide for adequate modeling of such atoms. For a number of receptors, structures of several complexes with different ligands were available. In such cases we used a single receptor structure in docking experiments with all ligands. Hydrogen atoms were added to all X-ray structures using the hydrogen placement algorithm of ICM software (Abagyan et al. 1994). Electric charges were assigned to the atoms of the ligands using the bond-charge increment algorithm from the MMFF94 force field (Halgren 1995).

RESULTS AND DISCUSSION

A total of 51 complexes with known structures were predicted. Only the best-energy conformation in each case was retained and compared with the experimental structure. A total of 35 predictions were within 3 Å from the native structure, producing correct overall positioning of the ligand, and 26 were within 2 Å, giving a fairly detailed picture of the receptor–ligand interaction (Table 29.1 and Figure 29.2, see plate section). As expected, good precision is achieved for tight, enclosed binding pockets, whereas for more loose, open binding sites such as in phospholipase, FK506 binding protein or fatty acid binding protein, the prediction quality is often marginal. Single simulation took 2–12 min of CPU time, which illustrates the advantage of precalculated grid potentials, because similar simulations with a full-atom receptor molecule take several hours (Totrov and Abagyan 1997).

The results show that docking techniques, such as flexible docking in internal coordinates using a grid potential representation of the receptor molecule, in the majority of cases can produce a model of protein–ligand interaction with a precision that allows its use in applications such as drug design. However, an important condition for the current docking methods is the relative rigidity of the binding site. Reliable ways to treat receptor flexibility are yet to be

<table>
<thead>
<tr>
<th>Source (PDB code)</th>
<th>Receptor</th>
<th>Ligand</th>
<th>Root-mean-square displacement (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1hna</td>
<td>Fatty acid binding protein</td>
<td>Elaidic acid</td>
<td>4.89</td>
</tr>
<tr>
<td>1hms</td>
<td>Fatty acid binding protein</td>
<td>Oleic acid</td>
<td>6.69</td>
</tr>
<tr>
<td>1hmt</td>
<td>Fatty acid binding protein</td>
<td>Stearic acid</td>
<td>3.49</td>
</tr>
<tr>
<td>1icn</td>
<td>Intestinal fatty acid binding protein</td>
<td>Myristate</td>
<td>1.79</td>
</tr>
<tr>
<td>1icn</td>
<td>Intestinal fatty acid binding protein</td>
<td>Oleic acid</td>
<td>1.46</td>
</tr>
<tr>
<td>4dfra</td>
<td>DHFR</td>
<td>Methotrexate</td>
<td>1.81</td>
</tr>
<tr>
<td>1dyj</td>
<td>DHFR</td>
<td>5,10-Dideazatetrahydrofolate</td>
<td>2.84</td>
</tr>
<tr>
<td>1dyl</td>
<td>DHFR</td>
<td>5-Deazaflavone</td>
<td>2.48</td>
</tr>
<tr>
<td>1dyi</td>
<td>DHFR</td>
<td>Folate</td>
<td>3.45</td>
</tr>
<tr>
<td>1joms</td>
<td>DHFR</td>
<td>Folic acid</td>
<td>5.00</td>
</tr>
<tr>
<td>2tbsa</td>
<td>Trypsin</td>
<td>Benazmidine</td>
<td>1.91</td>
</tr>
<tr>
<td>1tng</td>
<td>Trypsin</td>
<td>Aminomethylcyclohexane</td>
<td>1.27</td>
</tr>
<tr>
<td>1tnh</td>
<td>Trypsin</td>
<td>4-Fluorobenzylamine</td>
<td>1.94</td>
</tr>
<tr>
<td>1tni</td>
<td>Trypsin</td>
<td>4-Phenylbutylamine</td>
<td>3.05</td>
</tr>
<tr>
<td>1tnj</td>
<td>Trypsin</td>
<td>2-Phenylethylamine</td>
<td>2.79</td>
</tr>
<tr>
<td>1tnk</td>
<td>Trypsin</td>
<td>3-Phenylpropylamine</td>
<td>2.60</td>
</tr>
<tr>
<td>1tnl</td>
<td>Trypsin</td>
<td>Tranoylpyromine</td>
<td>2.11</td>
</tr>
<tr>
<td>1881a</td>
<td>Lysozyme mutant</td>
<td>O-Xylene</td>
<td>0.42</td>
</tr>
<tr>
<td>1851</td>
<td>Lysozyme mutant</td>
<td>Indole</td>
<td>1.33</td>
</tr>
<tr>
<td>1841</td>
<td>Lysozyme mutant</td>
<td>Isobutylbenzene</td>
<td>4.53</td>
</tr>
<tr>
<td>1871</td>
<td>Lysozyme mutant</td>
<td>p-Xylene</td>
<td>1.75</td>
</tr>
<tr>
<td>1861</td>
<td>Lysozyme mutant</td>
<td>n-Butylbenzene</td>
<td>1.62</td>
</tr>
<tr>
<td>1811</td>
<td>Lysozyme mutant</td>
<td>Benzene</td>
<td>0.80</td>
</tr>
<tr>
<td>1831</td>
<td>Lysozyme mutant</td>
<td>Indene</td>
<td>0.52</td>
</tr>
<tr>
<td>1821</td>
<td>Lysozyme mutant</td>
<td>Benzofuran</td>
<td>0.43</td>
</tr>
<tr>
<td>1erb</td>
<td>Retinol binding protein</td>
<td>n-Ethyl retinamide</td>
<td>0.99</td>
</tr>
<tr>
<td>1fel</td>
<td>Retinol binding protein</td>
<td>Retinoid</td>
<td>2.21</td>
</tr>
<tr>
<td>1fem</td>
<td>Retinol binding protein</td>
<td>Retinoic acid</td>
<td>2.35</td>
</tr>
<tr>
<td>1fen</td>
<td>Retinol binding protein</td>
<td>Axerophthane</td>
<td>0.93</td>
</tr>
<tr>
<td>1sre</td>
<td>Streptavidin</td>
<td>Haba</td>
<td>1.50</td>
</tr>
<tr>
<td>1srp</td>
<td>Streptavidin</td>
<td>3'-Methyl-haba</td>
<td>7.24</td>
</tr>
<tr>
<td>1sri</td>
<td>Streptavidin</td>
<td>3',5'-dimethyl-haba</td>
<td>7.53</td>
</tr>
<tr>
<td>1srj</td>
<td>Streptavidin</td>
<td>Naphthyl-haba</td>
<td>0.76</td>
</tr>
<tr>
<td>1gar</td>
<td>Glycinamidase ribonucleotide transformylase</td>
<td>Burroughs-Welch</td>
<td>2.19</td>
</tr>
<tr>
<td>1finder</td>
<td>Ferredoxin reductase</td>
<td>FAD</td>
<td>8.87b</td>
</tr>
<tr>
<td>1ake</td>
<td>Adenylate kinase</td>
<td>Inhibitor ap5a</td>
<td>0.98</td>
</tr>
<tr>
<td>1ref</td>
<td>Flavodoxin</td>
<td>Flavin mononucleotide</td>
<td>1.2</td>
</tr>
<tr>
<td>1mrg</td>
<td>α-Monomorcharin</td>
<td>Adenine</td>
<td>3.51</td>
</tr>
<tr>
<td>1mrg</td>
<td>α-Trichosanthin</td>
<td>Adenine</td>
<td>0.42</td>
</tr>
</tbody>
</table>

(continued)
developed. The growth in the available computer power and improvement in simulation techniques should ultimately allow detailed predictions of flexible receptor and ligand interaction.

REFERENCES


PROTEIN-LIGAND DOCKING: ENERGY OPTIMIZATION PROBLEM


Figure 29.2. (a) and (b) Comparison of predictions (solid lines) and experimentally determined structures (dotted lines) for 53 protein–ligand complexes.
Phospholipase c-d-1 complex with inositol triphosphate

Maltodextrin-binding protein complex with maltose

α-Monomorcharin complex with adenine

Neuraminidase complexes with sialic acid, 2,3-dehydro-2-deoxy-α-acetyl neuraminic acid and 4-(acylamino)-3-hydroxy-5-nitrobenzoic acid

α-Trichosanthin complex with adenine

Flavodoxin complex with flavin mononucleotide

Streptavidin complexes with 2-(4'-hydroxyphenyl)-azobenzoate (HABA), 3'-methyl-HABA, 3',5'-dimethyl-HABA and naphthyl-HABA

α-Ribose-binding protein complex with 3-c-ribose

Dihydrofolate reductase complexes with methotrexate, 5,10-dideazafolate, 5-dezaflolate, folate and folinic acid

Trypsin complexes with inhibitors benzamidine, aminomethylcyclohexane, 4-fluorobenzyamine 4-phenylbutylamine, 2-phenylethylamine, 3-phenylpropylamine and tranlycromine

Tyrosine kinase of FGF receptor complex with Sugen inhibitor

Figure 29.2. continued
Figure 33.10. Structure of the CD4 Phe 43 binding site in gp120 in the CD4 D1D2 – gp120 core – 17b Fab three-component complex determined by high-resolution crystallographic analysis. The structure is from Kwong et al. (1998). The arrow points to the critical Phe 43 residue as deduced originally by site-directed mutagenesis (Arthos et al. 1989; Wu et al. 1996).