STRUCTURE-BASED DRUG DESIGN: Computational Advances

Tami J. Marrone, James M. Briggs, and J. Andrew McCammon
Department of Chemistry and Biochemistry and Department of Pharmacology,
University of California San Diego, La Jolla, California 92093-0365

KEY WORDS: structure-based drug design, free energy perturbation, docking, computational chemistry

ABSTRACT
Structure-based computational methods continue to enhance progress in the discovery and refinement of therapeutic agents. Several such methods and their applications are described. These include molecular visualization and molecular modeling, docking, fragment methods, 3-D database techniques, and free-energy perturbation. Related issues that are discussed include the use of simplified potential energy functions and the determination of the positions of tightly bound waters. Strengths and weaknesses of the various methods are described.

INTRODUCTION
Computing is used in various ways in drug discovery. Important examples include QSAR, artificial intelligence, and structure-based methods. Here, we focus on the structure-based methods. These show increasing utility for the discovery of lead compounds, and especially for the refinement of lead compounds and for the re-engineering of drugs to overcome certain types of resistance. The structure-based methods are becoming increasingly important due in part to the rapid growth in structural data (1004 structures released in the Protein Data Bank in 1995 alone) (1), and the particularly high speed with which structures can be determined as part of a focused drug-discovery effort with a well-characterized target.

This review is divided into two major sections: the discovery of leads, and the refinement and re-engineering of leads. The methods and topics described here include molecular visualization and modeling, docking from structural databases, assembly of leads from fragments, 3-D database methods, simplified potential energy functions, detailed calculations of equilibrium constants, and
methods to allow for water-binding sites. Since many of the methods used to discover new classes of lead compounds can also be used to refine them, Figure 1 shows schematically where each of these methods can be used in the structure-based drug design process.

**DISCOVERY OF LEAD COMPOUNDS**

*Basic Molecular Visualization and Modeling*

Advances in the ability to visualize molecular structure and properties have led to a revolution in computer modeling. Many commercial (InsightII (2), Quanta (2), Cerius2 (2), Sybyl (3), CAChe (4), etc) and academic (MacroModel (5), Grasp (6), etc) programs are available for this task, and many provide interfaces to computational codes. Molecular structure and property visualization are important in all phases of the molecular design process, although they are used more heavily in some phases than in others. Some determinations of the “goodness” of a fit of a model compound to a target binding site are made via visual inspections of the docked structures, with interactive feedback of interatomic distances and energy components. Since most modeling efforts utilize visualization, there are countless applications in the literature. However, only a few of these applications are described here as examples of the importance of molecular visualization and modeling in structure-based drug design.

In a recent study directed at the rational design of nonpeptide-based inhibitors of the HIV protease, molecular visualization and modeling played a significant
role (7). The work began with the knowledge of X-ray structures of HIV protease/inhibitor complexes. It was clear from those structures that a tetracoordinated water molecule was present in the active site of the enzyme where it was hydrogen bonded to the backbone amide hydrogens of two residues (Ile50 and Ile50') and donated two hydrogen bonds to the inhibitor. A goal of the design process was to incorporate the function of this water molecule into a new inhibitor. The design cycle started with the development of a pharmacophore model based on the available X-ray data. The initial models included two symmetric hydrophobic groups to fit into the S1 and S1' hydrophobic pockets in the enzyme and one hydrogen bonding site that would bind to the catalytic aspartates. The pharmacophore model was used to search a database of 3-D molecular structures and resulted in a molecule with a phenyl group that included an oxygen to take the place of the structural water molecule. Visual analysis of the structure of the HIV protease with the phenyl-containing molecule suggested that a benzene ring might not be able to place all of the groups where they needed to be simultaneously. A cyclohexanone-containing molecule was next suggested and then subsequently modified to a synthetically more accessible seven-membered-ring cyclic urea, with the added benefit that the nitrogens could be easily functionalized. Two hydroxyl groups were used to hydrogen bond to the two aspartates in the active site of the protease. Further modeling suggested appropriate stereochemistries for all of the chiral sites and predicted that the cyclic urea should bind quite well. Later crystallographic work confirmed the conclusions of the modeling work. Subsequent use of medicinal chemistry to enhance the oral bioavailability resulted in a subnanomolar clinical trial candidate with good pharmacokinetics.

Another picomolar inhibitor of the HIV-1 protease was discovered by using energy minimization of new molecular structures within the active site of the enzyme (8). Only atoms in the inhibitors were allowed to move. Primarily through a visual inspection of the docked complexes, it was noticed that there was unoccupied volume near Asp 29/30 and Asp 129/130 and that the backbone amide hydrogens could hydrogen-bond with appropriate acceptors on groups in those pockets. Hydroxyl groups were placed on the phenyl rings in those pockets, which resulted in an increase in binding by three orders of magnitude in the best case.

Visualization was used recently to help develop a simplified potential function to quickly estimate the relative free energies of binding for inhibitors to the HIV-1 protease (9). The function consists of a molecular-mechanics-based enthalpy term and a hydrophobic term. New potential inhibitors were docked into the active site by overlaying them with positions of known inhibitors from X-ray structures and then relaxing the whole system using the AMBER (10) program. An atomic hydrophobic interaction energy based on the molecular
surface was used; the hydrophobicity parameters were derived from a large database of 1-octanol-water partition coefficients for a variety of molecules. Molecular hydrophobicity maps were visualized in efforts to understand the critical interactions needed between inhibitors and the active-site hydrophobic pockets; the maps helped in the rationalization of binding differences. Trends, but not magnitudes, in the calculated relative free energies of binding match those from experiment and from free-energy-perturbation calculations.

**Docking from Structural Databases**

Docking compounds from databases to targets of known structure can be utilized to discover or refine new leads. Structures from a database of small-molecule compounds are fit into the target structure using a docking program. The energies of the resulting complexes are evaluated and those that show the most promise can be experimentally tested as possible lead compounds. Docking has been reviewed extensively (11–15), so this section only briefly describes a few aspects and applications of docking as a method for lead generation in structure-based drug design.

The scoring method is critical in ranking the docked structures. Energy evaluations can be computationally expensive, so simplified energy functions are used to rank the structures. Some of these energy functions are described in more detail below. Besides using a simplified energy function, methods are employed to make the energy evaluation more efficient. An example is the use of grid-based methods (16–19) for docking of flexible or rigid compounds to a rigid target. In these, the interaction energy of the target with appropriate probes is precomputed and mapped out onto a three-dimensional grid.

Initial docking algorithms (20, 21) involved fitting a rigid compound with a rigid target. However, conformational flexibility has been incorporated into several docking algorithms (16, 17, 22–28) at additional computational cost. Including conformational flexibility increases the chance of finding the lowest-energy structure of the complex, because compounds and targets are capable of altering their structures upon binding. However, even where using conformational flexibility is not cost efficient for screening compounds in a database, it could be useful in further refinement of leads.

Several new lead compounds have been generated using docking from databases (29–31). One good example, which demonstrates how docking from structural databases can be used to generate and refine leads, is the screening of the Fine Chemicals Directory for possible new inhibitors of thymidylate synthase (31). Using a crystal structure of thymidylate synthase as the target, an average of $10^4$ orientations of each compound were docked to the target using a steric fit. The electrostatic energy was used to score the complexes, and a solvation correction was applied to the top-scoring compounds to further screen
the candidates. Based on these scores and lack of similarity to the thymidylate synthase substrate, several possible inhibitors were proposed and screened for inhibition. The crystal structure of the complex of one of the leads, sulisobzone, with the enzyme showed another possible site that could be exploited. Subsequent docking simulations with molecules sterically similar to sulisobzone identified phenolphthalein analogs that were found to have inhibition constants in the 1–30 \( \mu \)M range. The crystal structure of phenolphthalein with thymidylate synthase showed that phenolphthalein bound in the alternate binding site. This application nicely demonstrates one approach to structure-based drug design.

Faster computers and more efficient algorithms will permit the use of conformational flexibility during the screening process to identify leads that may be missed within the context of rigid docking. More detailed and accurate potential functions could also be used in the scoring phase of the docking process. Databases of commercially available compounds are sometimes preferred for docking applications to avoid possible difficulty in the syntheses of the lead compounds generated by the docking simulations. The number of chemical libraries has exploded with the advent of combinatorial synthesis, and perhaps databases of these libraries will offer a larger selection of commercially available compounds to investigate as leads in docking simulations.

Assembly of Leads from Fragments

When the structure of the target is known, novel leads can be generated using fragment methods. The interactions of several chemical groups with the target are calculated, and possible binding sites are identified. Databases can be searched to match small molecules or chemical groups with the possible binding site. These candidates can then be assembled into one compound using linking groups.

One of the earliest attempts to model the interaction of chemical groups with a protein was made by Goodford using GRID (19). This method was developed to quickly determine possible binding sites of a protein, and it is the basis for several of the de novo design methods. The interaction energies of various probes representing chemical groups are mapped onto a grid. The interaction energy was originally described using a coulombic term, a Lennard-Jones term, and a hydrogen-bonding term. The hydrogen-bonding term was subsequently refined by fitting to experimental data (32–34). Possible binding sites for a particular chemical group are identified as the sites with the most favorable interaction energies for a chemical group.

MCSS/HOOK The MCSS (Multiple Copy Simultaneous Search) (35) program is used to find energetically favorable binding sites and orientations for peptide
fragments, and the HOOK (36) program connects them together. This work is related to the GROW (37) and LUDI (38) methods. The procedure involves choosing a sphere large enough to encompass the entire region of interest. In this sphere, thousands of copies of the “protein backbone” fragment N-methylacetamide (NMA) are placed along with thousands of copies of organic molecules representing the terminal portions of amino acid side chains (e.g. methanol, acetate, methyl, etc). The NMA backbone pieces are connected together based on an energy function that has one term for bonding and one for steric contacts.

LUDI The LUDI algorithm represents a molecular-fragment–based approach to identifying, de novo, potential leads (38). Fragments are identified such that hydrogen bonds are made and hydrophobic pockets are filled with hydrophobic fragments. The fragments are then connected with standard linkers to make single molecules. Four kinds of directional interaction sites are identified with the overall binding site: lipophilic-aliphatic, lipophilic-aromatic, hydrogen donor, and hydrogen acceptor. The fragment library is searched for fragments that have the proper number and 3-D arrangement of atom types that match a set of descriptors in the binding site. The matching fragment is then docked such that the matching atoms and descriptor sites are maximally aligned. Conformational flexibility can be included by using multiple entries for various fragments representing different conformations. The linking of fragments is similar to that used in the CAVEAT program (39). The nearest hydrogen atoms are found and replaced with one of the linker fragments depending on the distance between the remaining heavy atoms. The method was tested with dihydrofolate reductase (DHFR) and trypsin and was found to generate fragments that aligned favorably with known inhibitors such as methotrexate and benzamidine, for DHFR and trypsin, respectively.

GROW The GROW method (37) starts with the coordinates of a binding site and one seed fragment (an acetyl group) and grows a peptide-based molecule of specified length. The peptide can be grown in the N- or C-terminal direction, or both. The successful design of a peptide is greatly dependent on a good initial placement of the seed fragment. Ideally this would be done by overlaying the seed with a suitable portion of an available X-ray or NMR structure of a compound-target complex, although semi-automated methods are available. The scoring function consists of intermolecular van der Waals and electrostatic energies as well as conformational and solvation energies. The AMBER (10) force field is used for the van der Waals, electrostatic, and conformational energies along with a solvation term based on changes in surface area with terms to favor the hydration of polar groups and disfavor that of hydrophobic ones. The
fragment library is constructed with the use of this scoring function, after which the fragments are arranged according to their score (including the solvation energy). Fragments from the previously generated library of peptide fragments are then fused to the appropriate end of the seed and the energy of the complex is evaluated (scored). Peptides with favorable scores are carried through the process for a second round of fusions, and so on. Once a collection of peptides of desired length is obtained, a subset is subjected to energy minimization in the binding site (the peptides and a selected portion of the binding site are permitted to move). The peptides are also energy minimized outside of the binding site to estimate whether the conformational change that is required for binding will cause the binding energy to become unfavorable. This procedure was carried out for two aspartyl proteases, rhizopuspepsin, and a model of the renin active site. In these cases, after relaxation of the most promising leads in the active sites of the X-ray complexes, structures were obtained that closely matched the crystal structures. As with most computational methods, graphical analyses are used heavily in the first and final stages.

MCDNLG  The Monte Carlo De Novo Ligand Generator (MCDNLG) (40) uses a binding cavity from an X-ray or NMR structure that is then filled with a densely packed array of atoms of random type. Each atom has three properties: element type, hybridization, and the number of implicit hydrogens. The pseudo-atoms can have a maximum of 12 bonds simultaneously. The energy function is composed of intra- and intermolecular components that control ligand geometry and ultimate shape and chemical complementarity to the binding site. Some intramolecular terms include bond energy, angle and dihedral angle strain, and valence strain (this allows for bond making and breaking). The intermolecular components include dispersion, repulsion, hydrogen bonding, and a desolvation penalty for polar atoms. A Monte Carlo procedure is employed to randomly select atoms to appear or disappear or change type or change bonding arrangements or rotate or translate. A novel ligand can be created in \( \sim 300,000 \) Monte Carlo steps. Quite diverse ligands result from different initial conditions (random number seeds). As a test, the method was applied to the identification of inhibitors of dihydrofolate reductase, thymidylate synthase, and HIV-1 protease. Sets of compounds were generated, which fell into families that compared well with molecules known to bind. It should be noted that this method is most useful for coming up with novel leads, not for finding either the final product or molecules that are synthetically accessible.

SPROUT  The SPROUT program (41) requires information about the binding site from either X-ray or NMR data, or from a pharmacophore model from the overlay of known inhibitors in their bioactive conformations. Target sites in the
binding pocket are identified and labeled by type (primarily hydrogen-bonding, electrostatic, and hydrophobic). The binding site dimensions are conferred by a fixed boundary. A fragment library is used that has been presorted according to atomic and molecular properties (shape, hydrogen-bond, etc). Fragments of appropriate shape are selected and overlaid on a target site. The fragment can be rotated around a pivot point to find an ideal “docking” arrangement (i.e. one that satisfies the steric requirements but does not violate the boundary). Once fragments have been docked into all of the target sites, fragments are joined in a manner dependent on the identity of the fragments (acyclic, cyclic, fusion, bridging, and spiro). In the second phase of the sprouting, atom types are interchanged with others of the same hybridization in order to find a combination with optimal interactions with the binding site. The scoring function is composed of terms accounting for the loss of translational, rotational, and torsional degrees of freedom on binding, hydrophobic, van der Waal, and solvation free energies, and an enthalpy term to account for the energy required to obtain the bioactive conformation. The methods were applied to trypsin and HIV-1 protease and generated ligands that resembled known inhibitors.

3-D Database Methods

CLIX The CLIX program (42) uses a combination of the GRID (19) program and a search and docking from the Cambridge Crystallographic Database (CCDB) in order to identify small molecules that have both the steric and chemical likelihood of binding. The GRID program is used to find discrete binding sites for a large variety of small representative groups. Once identified, a database search is undertaken of the CCDB for ligands that contain pairs of interaction sites identified by the GRID search. An iterative search is performed to ensure that selected molecules from the CCDB contain groups representative of the GRID interaction sites. Each molecule is then placed into the binding site such that two of the matching groups in the molecule are overlaid on their respective GRID interaction sites. The molecule is rigidly rotated about the line that connects the two interaction sites to see whether an overlap can be found that overlays other functionalities in the molecule with other GRID interaction sites but does not cause steric interactions with the binding site. The GRID energy-scoring method is used. Further analyses with the GRID program are used to identify places in the binding site that are not satisfied with a particular small molecule. Corresponding portions in the small molecule are changed to more appropriate groups (i.e. those that correspond to the GRID probes that have high scores in those regions). Users of CLIX try to ensure that chemically reasonable choices are made by considering adjacent functionalities. A test of the procedure was made with hemagglutinin for the identification of the binding mode of sialic acid. Of the seven resulting structures with the lowest
energy, three were very close to the crystal structure (0.42–0.52 Å), while the remaining four were far off (4–6 Å).

CAVEAT 3-D DATABASE SEARCHING  The CAVEAT program (39) represents a novel approach to 3-D database searching (43). Most 3-D database searching methods involve the overlaying of molecules in a database with a known inhibitor or a pharmacophore model. The CAVEAT program operates under the assumption that it is enough, at least initially, to focus on the orientation of bonds and not on the placement of atoms. In this way, a 3-D database can be prescreened for pairs of bonds and their relative orientations. User-specified bond types are used to construct a sorted CAVEAT database of bond unit vector pairs. A major goal of the CAVEAT work was to generate a method that could give results in an interactive fashion (i.e. fast). This is accomplished by doing most of the processing work during the construction of the CAVEAT bond vector database.

Queries are constructed via the consideration of pairs of bond vectors. One seeks to address the smallest set of relevant pairs. Tolerances are assigned to each query pair so that close matches can be retained. The final phase of the query is to group the results into families. Graph-theoretical methods are used, among others, for the clustering of similar/related structures. A CAVEAT database consisting of ~50,000,000 bond pairs can be searched to generate hundreds to thousands of matches in 0.5–2.0 minutes.

Simplified Potential Energy Functions

Simplified empirical potential energy functions are used to rank docked structures or predict binding affinities. This section briefly discusses molecular mechanics potential functions. Some methods used to include solvation implicitly in molecular mechanics calculations are discussed in detail along with their strengths and weaknesses. Applications of these methods are also described.

Traditional energy refinement, molecular dynamics, and Monte Carlo simulations of molecules often use molecular-mechanics-based potential functions. These functions consist of bonded interactions (bonds, bond angles, dihedrals) and nonbonded interactions (coulombic, Lennard-Jones). The parameters and functional forms used to describe the energetics of the system are often called a force field. Several force fields have been developed for proteins (44–49) and nucleic acids (44–46, 48, 50). The number of nonbonded interactions among the atoms usually increases on the order of $n^2$, where $n$ is the number of atoms in the system. Energy evaluations become more computationally expensive as the size of the compound-target complex increases or if explicit waters are added to the system to describe solvation. Quantitative calculation of the relative free energies of binding of two compounds to
the same target is quite difficult using molecular-mechanics potential functions and including explicit solvation. However some success has been obtained with this approach, as discussed in detail in the section below on computational alchemy. To obtain qualitative relative free energies of binding or to rank several structures of a compound-target complex, simpler energy functions have been developed to predict trends without the cost of a detailed simulation. These functions describe the free energies of binding or the energy of a conformation in simplified, empirical terms for rapid screening. Several of these energy functions have been described in recent reviews (11) on docking and usually include an electrostatic term and a hydrophobic term that can be based on the surface area of the system.

Some energy functions use molecular-mechanics-based force fields and simple models of solvation. One common empirical method makes use of the solvent-accessible surface area of each amino acid or other group (51–54). Atomic solvation parameters for these groups are derived from empirical vapor-to-water or water-to-octanol transfer free energies for appropriate analogues. A hydrophilic group has a favorable solvation energy, while a hydrophobic group has an unfavorable solvation energy. Derivatives of the accessible surface area with respect to the atomic positions can be used in molecular-mechanics and molecular-dynamics simulations. A similar solvation term was developed (55) where the interaction between a protein and solvent is described as a function of atom occupancies. Here the occupancy of an atom is the volume around the atom that is available for water to occupy. The overlap of the solute atoms and the occupancy is described as a Gaussian function that is easily differentiated. This allows the method to be used in dynamics (55, 56) and mechanics. Others (57) have also parameterized atomic solvation parameters to be compatible with molecular mechanics calculations. Such solvation models have been compared in the literature and shown to be very sensitive to the number and choice of atomic solvation parameters (58, 59). These methods have a major drawback in that they essentially include only the first hydration shells of water and do not take into account any longer-range forces such as electrostatics (59). Another way to describe solvation is through the use of continuum models, where the solute is treated as a set of point charges in a dielectric continuum. Solvation energies are calculated by solving the Poisson-Boltzmann (PB) equation, and the derivatives of the associated electrostatic forces can be used in stochastic dynamics simulations (60). The PB approach only calculates the electrostatic interactions of the solutes and the solvent. Lennard-Jones interactions as well as the energy to form the solute cavity in the continuum must be added-in where necessary. However, in comparing a series of similar compounds, the Lennard-Jones and cavity terms would be similar and could be neglected. This approach was utilized (61) to show that PB electrostatics calculations better
estimated the experimental relative free energies of solvation of organophosphorous molecules than explicit free energy calculations. In the case that electrostatics dominate the binding of a compound to a target, the relative free energies of binding can be calculated for a series of compounds using the PB approach (62–64). The PB approach has been successful in predicting rate constants of diffusion-controlled enzymatic reactions from Brownian dynamics simulations (65). These calculations have also shown success in estimating pKₐs and ionization states of ionizable groups in proteins (66).

To account for the missing cavitation and Lennard-Jones terms in continuum models, hydrophobic solvation can be described as an apolar surface area–dependent term. It has been shown that the vapor-to-water and water-to-octanol transfer free energies of alkanes are approximately linear with respect to surface area, although there is a debate on how to extract the magnitude of the hydrophobic interaction with water. A new parameter (67) set has been derived that includes a surface area–dependent term for use in continuum models to predict the free energies of solvation of organic molecules. The stabilities of loops (68), helices (69), and beta sheets (70) have been examined with the continuum-plus-apolar model. Structures have been successfully ranked from docking procedures (71), and the effects of mutation in the λ repressor-operator complex have been studied (72) using this model.

The PB-plus-apolar model has several advantages. The solvent is a continuum, so computational time is saved by avoiding sampling of solvent configurations (73). Also, long-range electrostatics and ionic strength are included in the calculation via the PB equation. Polarization can be included by varying the dielectric coefficients of the solute. However, this method may not be well suited for electrically neutral solutes (i.e. no partial charges on the solute model) because the hydrophobic effect is taken into account in a very approximate way through the surface term.

REFINEMENT AND RE-ENGINEERING

*Computational Alchemy*

This section briefly reviews the thermodynamic cycle perturbation method, which in principle allows exact calculations of the relative binding affinities of compounds for a given target. A general description of the method is given, followed by examples of its use in drug discovery efforts. The concluding discussion describes some of the limitations, recent advances, and future prospects of this method.

The binding constant of a compound to a target is related of course to the free energy of binding. Many of the computational tools described above work by providing rapid estimates of these free-energy changes. The exact relation
between binding constants and free-energy changes is given by the familiar expression

\[ K = \exp(-\Delta G^\circ / RT), \]

where \( K \) is the thermodynamic equilibrium constant, \( \Delta G^\circ \) is the standard change in free energy, \( R \) is the gas constant, and \( T \) is the absolute temperature. In principle, it would be possible to determine which of two compounds, \( C_1 \) and \( C_2 \), would bind more strongly to a target \( M \) by simulating the two binding processes in the appropriate solvent and calculating the corresponding changes in free energy, \( \Delta G_1 \) and \( \Delta G_2 \). Such calculations are in fact possible for very simple solutes, using either molecular dynamics or Monte Carlo simulations (74).

For solutes of the complexity of pharmacologic compounds and targets, direct calculations of the free energies of binding are seldom possible. This is due to the significant structural changes involved, which may include desolvation of charged groups on the compound and of recessed binding sites in the target, and changes in the conformation and titration states of either molecule. The potential energy functions used in current simulation studies generally do not allow for automatic adjustment of titration states. But a more general problem encountered in efforts to simulate binding is that the solvation and conformational changes can not be simply accounted for because they occur on time scales that are long compared to typical simulations. With current supercomputers, simulations of an enzyme-inhibitor complex in water can be extended to perhaps 10 ns. Actual binding processes may require many orders of magnitude more time, and attempts to simulate them on shorter time scales can lead to artifactual results.

To overcome this problem, an indirect approach to calculating relative free energies of binding was introduced in 1984 (75). This makes use of the thermodynamic cycle shown in Figure 2. Because free energy is a state function, the relative free energy of binding, \( \Delta \Delta G = \Delta G_2^\circ - \Delta G_1^\circ = \Delta G_4^\circ - \Delta G_3^\circ \). The determination of \( \Delta G_3^\circ \) and \( \Delta G_4^\circ \) involves simulations in which the molecule \( C_1 \) is gradually transformed into the molecule \( C_2 \). In the first biological application of the method, for example, benzamidine was transformed into parafluorobenzamidine, both in water, to provide \( \Delta G_3^\circ \), and in the substrate recognition pocket of trypsin, to provide \( \Delta G_4^\circ \) (76). Because these nonphysical transformations often involve “transmutation” of one type of atom into another, such calculations have been colorfully described as “computational alchemy” (74). Comparison of the magnitudes of \( \Delta G_3^\circ \) and \( \Delta G_4^\circ \) provides insight into the relative contributions of solvation and compound-target interactions in the analysis of the origins of the binding selectivity.

To carry out an alchemical simulation, one must gradually change the potential energy function from that for \( C_1 \) and its surroundings to that for \( C_2 \) and its surroundings, during the course of a molecular-dynamics or Monte Carlo
Simulation. Two general approaches have been widely used for calculating the free-energy changes. In the “perturbation” method, a simulation is run for $C_1$ in its surroundings, and then the changes in potential energy upon making a finite but small change in $C_1$ (e.g. changing the $C$-$H$ bondlength 25% of the way toward the $C$-$F$ bondlength, in the benzamidine example) are calculated for each of a sequence of snapshots from the simulation. Properly averaged, these energy changes provide an estimate of the free-energy change for the small change in $C_1$. Then a simulation is run for the partly changed molecule in its surroundings and analyzed as above to estimate the next increment in the free-energy change. This procedure is continued until one has a set of finite differences that can be added to provide an estimate of the total free-energy change, $\Delta G_3^1$ or $\Delta G_4^4$. The other widely used approach for calculating alchemical free-energy changes is the “thermodynamic integration” method. This method provides estimates of derivatives of the free energy with respect to a parameter $\lambda$ that defines the extent of the change in $C_1$, again using a sequence of simulations corresponding to different values of $\lambda$ in a standard implementation. A simple integral of these derivatives from $\lambda = 0$ to $\lambda = 1$ then provides the estimate of $\Delta G_3^1$ or $\Delta G_4^4$. An excellent account of the technical details of alchemical simulations has been published very recently (77).

Computational alchemy methods have been applied to a number of systems of pharmaceutical interest in recent years, primarily to explore ways in which compounds might be modified to increase their binding to soluble molecular targets. The early availability of crystallographic structures of the HIV protease complexed with a variety of inhibitors has stimulated many such studies, several of which have been described in the literature. A good example, which
includes references to earlier studies, is a very recent analysis of modifications of a parent $N,N$-disubstituted benzene sulfonamide (78). The parent compound, which was arrived at by structure-based drug design, was modified in ways intended to increase its solubility (and bioavailability) while maintaining its charge, molecular weight, and inhibitory potency. The modifications involved the replacement of an aromatic $H$ by $OH$ and $NH_2$. The calculated changes in the free energies of hydration and of binding were found to be in good agreement with experimental results, although the experimental hydration data were estimated from the solubilities of the corresponding solids. In another good example, significantly larger changes were made to evaluate the prospective affinities of a series of inhibitors based on a hydroxyethylene backbone (79). This work was actually used to guide the synthesis of compounds for testing, and the predicted affinities were found to be in general agreement with the subsequent experiments. This study also emphasized the value of the computational alchemy approach in accounting for differences in the desolvation contributions to the binding constants.

Other systems of pharmacologic interest that have been studied by computational alchemy methods include the following enzymes and sets of related inhibitors: thymidylate synthase (80), acetylcholinesterase (81), adenosine deaminase (82), and elastase (83).

Experience obtained from the simulations described above, and from applications of computational alchemy to other systems (74, 84), point to a number of strengths and limitations of the method. The primary strength, of course, is the rigorous foundation of the method. As the speed of computers and the quality of the potential functions increase, the calculations will become increasingly accurate. Indeed, for small- or medium-sized solutes (e.g. synthetic host-guest systems), discrepancies between theoretical and experimental measures of molecular recognition have been resolved by corrections of the latter in some cases (85). Such accuracy is not yet routinely possible in problems of pharmacological interest, as discussed below.

Perhaps the greatest challenge in attempting to calculate the relative free energy of binding of two or more compounds to a target is adequately sampling the relevant configurations of the atoms in the system (86). Even where a high-resolution experimental structure of an initial compound-target complex is available, this does not describe the large number of slightly different conformations that are reflected in the thermodynamics of the system (87–89). This distribution of configurations will in general change somewhat when one compound is replaced by another in the target complex. Owing to the large size and inherent flexibility of biopolymers, it is generally not possible to sample a representative set of the configurations involved (90). In fact, the sampling problem is seen to be even more challenging when one recognizes that target
proteins, for example, contain many titratable side chains, and the protonation states of these may change when different compounds are bound.

Another challenge facing the practical application of computational alchemy to drug discovery is insuring the adequacy of the underlying potential energy function. Although existing potential functions that assume pairwise additivity of nonbonded interatomic interactions appear to be sufficient in many cases (91), the description of certain interactions may require explicit treatment of electronic polarizability (92) or special techniques to avoid the usual truncation of long-ranged electrostatic interactions (93). Also, existing potential functions are largely limited to amino acid residues, nucleic acid bases, and other commonly occurring groups. The development of potential functions that are of the same quality for prosthetic groups or model compounds is a continuing challenge (94).

The development of new methods to meet challenges such as those listed above is an increasingly active area of research. Some notable lines of progress include the development of parallel computing tools that will increase the speed of the calculations (95); new approaches for choosing reference states in alchemical simulations that will speed convergence of the free energy calculations (96, 97); improved methods for assigning protonation states (98, 99); and improvements in the potential energy functions, both for specific chemical groups (92, 94) and for the treatment of long-range electrostatic interactions (93, 100).

Computational alchemy is already contributing to drug discovery efforts, particularly in the late stages of structure-based design projects and where the sampling and potential-function difficulties are not too large (78, 79). The utility of these methods will increase as the basic research described above yields faster, more generally reliable tools.

**Structural Water**

Refining and re-engineering compounds can include designing the compound to either mimic tightly bound waters or to utilize waters as bridging groups. These design methods are currently being explored by computational methods. Several computational methods, such as grand canonical ensemble simulations and neural networks, are used to determine water binding sites in targets. Grand canonical methods allow water to be inserted/deleted during a Monte Carlo or molecular-dynamics simulation. The probability of a water molecule’s occupying a position in the target can be calculated and used to predict the positions of tightly bound waters. In particular, grand canonical Monte Carlo has been used successfully to predict water binding sites in hyaluronic acid (101) and DNA (102) crystal structures. In these simulations, the target was held fixed in order to compare with the experimentally determined crystal waters. However,
conformational flexibility is needed to better represent the distribution of tightly bound waters in solution, because tightly bound crystal waters may be a product of the crystallization method and may not be present in solution. Methods such as grand canonical molecular dynamics are being developed to address the conformational flexibility issue (103). Neural networks are also being used to identify possible binding sites of water (104). By examining the protein data bank it is possible to identify sites and train the neural network to recognize the primary or secondary structures in a protein where a water could be found. Knowledge of the positions of tightly bound waters would permit a deeper consideration of the enthalpic (increased H-bonding) and entropic (desolvation penalties, release of water into bulk, and so on) contributions to binding that should prove useful in the refinement of first-generation inhibitors.

CONCLUSIONS

Structure-based drug design has shown success in the design and refinement of compounds of pharmacologic interest. Many of the challenging aspects of computational methods (e.g. the quality of potential functions and the efficiency of the algorithms) are active areas of research. As computers become faster, problems encountered due to limited computer time will be resolved. As these challenges are met, computational chemistry will become a more powerful tool and will contribute even more routinely to the design of therapeutic agents.

ACKNOWLEDGMENTS

This work was supported in part by the NSF, NIH, and the Metacenter Program of the NSF Supercomputer Centers. TJM is an NIH postdoctoral fellow.

Literature Cited

1. Statistics obtained from the Web site of the Protein Data Bank: http://www.pdb.bnl.gov
2. Quanta/InsightII/Cerius2. Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752
3. Sybyl. Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2913
4. CAChe. CAChe Scientific, Inc., P.O. Box 4003, Beaverton, OR 97076
5. MacroModel. Department of Chemistry, Columbia University, New York, NY 10032
6. Grasp. Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032


