
1

MODELING AND INFORMATICS IN DRUG DESIGN

PRASAD V. BHARATAM,* SMRITI KHANNA, AND SANDREA M. FRANCIS
National Institute of Pharmaceutical Education and Research (NIPER), S.A.S. Nagar, India

Contents

- 1.1 Introduction
- 1.2 Computational Chemistry
 - 1.2.1 *Ab Initio* Quantum Chemical Methods
 - 1.2.2 Semiempirical Methods
 - 1.2.3 Molecular Mechanical Methods
 - 1.2.4 Energy Minimization and Geometry Optimization
 - 1.2.5 Conformational Analysis
- 1.3 Computational Biology
 - 1.3.1 *Ab Initio* Structure Prediction
 - 1.3.2 Homology Modeling
 - 1.3.3 Threading or Remote Homology Modeling
- 1.4 Computational Medicinal Chemistry
 - 1.4.1 Quantitative Structure–Activity Relationship (QSAR)
 - 1.4.2 Pharmacophore Mapping
 - 1.4.3 Molecular Docking
 - 1.4.4 *De Novo* Design
- 1.5 Pharmacoinformatics
 - 1.5.1 Chemoinformatics
 - 1.5.2 Bioinformatics
 - 1.5.3 Virtual Screening
 - 1.5.4 Neuroinformatics
 - 1.5.5 Immunoinformatics
 - 1.5.6 Drug Metabolism Informatics
 - 1.5.7 Toxicoinformatics
 - 1.5.8 Cancer Informatics
- 1.6 Future Scope
- References

**Corresponding author.*

Preclinical Development Handbook: ADME and Biopharmaceutical Properties,
edited by Shayne Cox Gad
Copyright © 2008 John Wiley & Sons, Inc.

1.1 INTRODUCTION

Modeling and informatics have become indispensable components of rational drug design (Fig. 1.1). For the last few years, chemical analysis through molecular modeling has been very prominent in computer-aided drug design (CADD). But currently modeling and informatics are contributing in tandem toward CADD. Modeling in drug design has two facets: modeling on the basis of knowledge of the drugs/leads/ligands often referred to as ligand-based design and modeling based on the structure of macromolecules often referred to as receptor-based modeling (or structure-based modeling). Computer-aided drug design is a topic of medicinal chemistry, and before venturing into this exercise one must employ computational chemistry methods to understand the properties of chemical species, on the one hand, and employ computational biology techniques to understand the properties of biomolecules on the other. Information technology is playing a major role in decision making in pharmaceutical sciences. Storage, retrieval, and analysis of data of chemicals/biochemicals of therapeutic interest are major components of pharmacoinformatics. Quite

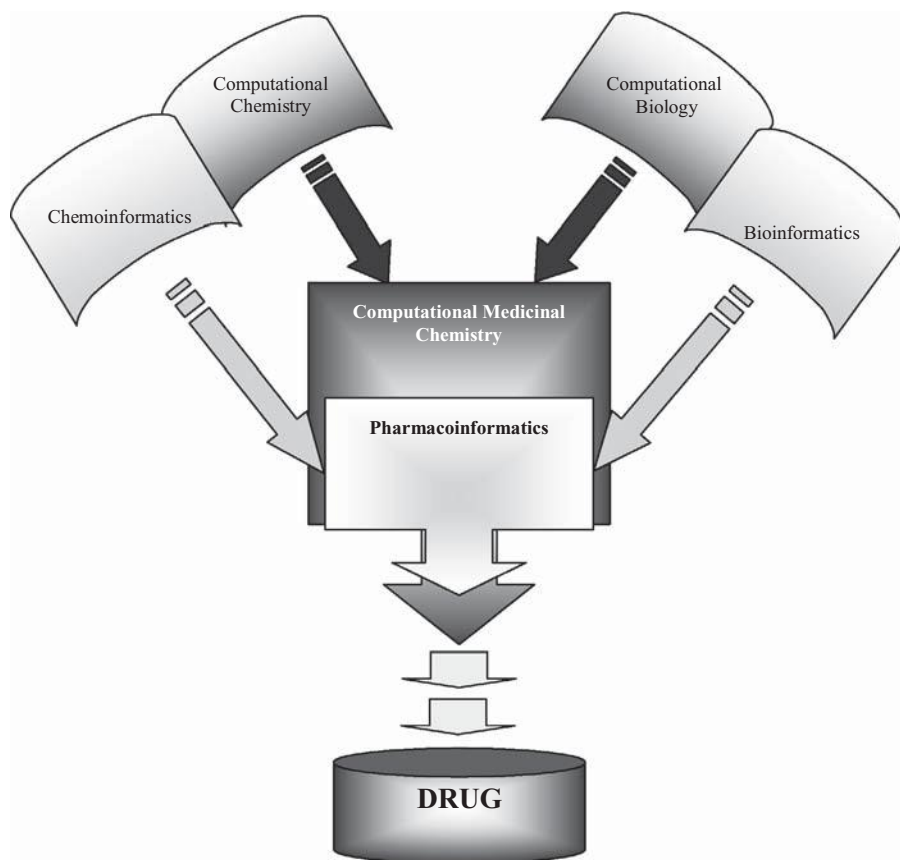


FIGURE 1.1 A schematic diagram showing a flowchart of activities in computer aided drug development. The figure shows that the contributions from modeling methods and informatics methods toward the drug development are parallel and in fact not really distinguishable.

often, the efforts based on modeling and informatics get thoroughly integrated with each other, as in the case of virtual screening exercises. In this chapter, the molecular modeling methods that are in vogue in the fields of (1) computational chemistry, (2) computational biology, (3) computational medicinal chemistry, and (4) pharmacoinformatics are presented and the resources available in these fields are discussed.

1.2 COMPUTATIONAL CHEMISTRY

Two-dimensional (2D) structure drawing and three-dimensional (3D) structure building are the important primary steps in computational chemistry for which several molecular visualization packages are available. The most popular of these are ChemDraw Ultra and Chem3D Pro, which are a part of the ChemOffice suite of software packages [1]. ACD/ChemSketch [2], MolSuite [3], and many more of this kind are other programs for the same purpose. Refinement has to be carried out on all the drawings and 3D structures so as to improve the chemical accuracy of the structure on the computer screen. Structure refinement based on heuristic rules/cleanup procedures is a part of all these software packages. However, chemical accuracy of the 3D structures still remains poor even after cleanup. Further refinement can be carried out by performing energy minimization using either molecular mechanical or quantum chemical procedures. By using these methods, the energy of a molecule can be estimated in any given state. Following this, with the help of first and second derivatives of energy, it can be ascertained whether the given computational state of the molecules belongs to a chemically acceptable state or not. During this process, the molecular geometry gets modified to a more appropriate, chemically meaningful state – the entire procedure is known as geometry optimization. The geometry optimized 3D structure is suitable for property estimation, descriptor calculation, conformational analysis, and finally for drug design exercise [4–6].

1.2.1 Ab Initio Quantum Chemical Methods

Every molecule possesses internal energy (U), for the estimation of which quantum chemical calculations are suitable. Quantum chemical calculations involve rigorous mathematical derivations and attempt to solve the Schrödinger equation, which in its simplest form may be written as

$$H\Psi = E\Psi \quad (1.1)$$

$$\hat{H}_{el} = \sum_i \left(-\frac{1}{2} \nabla_i^2 \right) - \sum_i \sum_a \frac{Z_a}{|r_i - d_a|} + \frac{1}{2} \sum_i \sum_{j \neq i} \frac{1}{|r_i - r_j|} + \frac{1}{2} \sum_a \sum_{b \neq a} \frac{Z_a Z_b}{|d_a - d_b|} \quad (1.2)$$

where ψ represents the wavefunction, E represents energy, ∇ represents the kinetic energy operator for electrons, r_i defines the vector position of electron i with vector components in Bohr radii, Z_a is the charge of fixed nucleus a in units of the elementary charge, and d_a is the vector position of nucleus a with vector components in Bohr radii.

Exact solutions to Schrödinger equation cannot be provided for systems with more than one electron. Several *ab initio* molecular orbital (MO) and *ab initio* density functional theory (DFT) methods were developed to provide expectation value for the energy. This energy can be minimized and thus the geometry of any molecule can be obtained, with high confidence level, using quantum chemical methods. During this energy estimation, the wavefunctions of every molecule can be defined, which possess all the information related to the molecule. Thus, properties like relative energies, dipole moments, electron density distribution, charge distribution, electron delocalization, molecular orbital energies, molecular orbital shapes, ionization potential, infrared (IR) frequencies, and chemical shifts can be estimated using *ab initio* computational chemistry methods. For this purpose, several quantum chemical methods like Hartree–Fock (HF), second order Moller–Plesset perturbation (MP2), coupled cluster (CCSD), configuration interaction (QCISD), many-body perturbation (MBPT), multiconfiguration self-consistent field (MCSCF), complete active space self-consistent field (CASSCF), B3LYP, and VWN were developed. At the same time to define the wavefunction, a set of mathematical functions known as *basis set* is required. Typical basis sets are 3-21G, 6-31G*, and 6-31+G*. Combination of the *ab initio* methods and basis sets leads to several thousand options for estimating energy. For reliable geometry optimization of drug molecules, the HF/6-31+G*, MP2/6-31+G*, and B3LYP/6-31+G* methods are quite suitable. When very accurate energy estimation is required, G2MP2 and CBS-Q methods can be employed. Gaussian03, Spartan, and Jaguar are software packages that can be used to estimate reliable geometry optimization and very accurate energy estimation of any chemical species. In practice, quantum chemical methods are being used to estimate the relative stabilities of molecules, to calculate properties of reaction intermediates, to investigate the mechanisms of chemical reactions, to predict the aromaticity of compounds, and to analyze spectral properties. Medicinal chemists are beginning to take benefit from these by studying drug–receptor interactions, enzyme–substrate binding, and solvation of biological molecules. Molecular electrostatic potentials, which can be derived from *ab initio* quantum chemical methods, provide the surface properties of drugs and receptors and thus they offer useful information regarding complementarities between the two [7–10].

1.2.2 Semiempirical Methods

The above defined *ab initio* methods are quite time consuming and become prohibitively expensive when the drugs possess large number of atoms and/or a series of calculations need to be performed to understand the chemical phenomena. Semiempirical quantum chemical methods were introduced precisely to address this problem. In these methods empirical parameters are employed to estimate many integrals but only a few key integrals are solved explicitly. Although these calculations do not provide energy of molecules, they are quite reliable in estimating the heats of formation. Semiempirical quantum chemical methods (e.g., AM1, PM3, SAM1) are very fast in qualitatively estimating the chemical properties that are of interest to a drug discovery scientist. MOPAC and AMPAC are the software packages of choice; however, many other software packages also incorporate these methods. Qualitative estimates of HOMO and LUMO energies, shapes of molecular orbitals, and reaction mechanisms of drug synthesis are some of the applications of

semiempirical analysis [5, 10]. When the molecules become much larger, especially in the case of macromolecules like proteins, enzymes, and nucleic acids, employing these semiempirical methods becomes impractical. In such cases, molecular mechanical methods can be used to estimate the heats of formation and to perform geometry optimization.

1.2.3 Molecular Mechanical Methods

Molecular mechanical methods estimate the energy of any drug by adding up the strain in all the bonds, angles, and torsions due to the energy of the van der Waals and Coulombic interactions across all atoms in the molecule. It reflects the internal energy of the molecule; although the estimated value is nowhere close to the actual internal energy, the relative energy obtained from these methods is indicative enough for chemical/biochemical analysis. It is made up of a number of components as given by

$$E_{\text{mm}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{vdw}} + E_{\text{torsion}} + E_{\text{charge}} + E_{\text{misc}} \quad (1.3)$$

Molecular mechanical methods are also known as force field methods because in these methods, the electronic effects are estimated implicitly in terms of force fields associated with the atoms. In Eq. 1.3, the energy (E) due to bonds, angles, and torsional angles can be estimated using the simple Hooke's law and its variations, whereas the van der Waals (vdw) interactions are estimated using the Lennard-Jones potential and the electrostatic interactions are estimated using Coulombic forces. The energy estimation, energy minimization, and geometry optimization using these methods are quite fast and hence suitable for studying the geometries and conformations of biomolecules and drug-receptor interactions. Since these methods are empirical in nature, parameterization of the force fields with the help of available spectral data or quantum chemical methods is required. AMBER, CHARMM, UFF, and Tripos are some of the force fields in wide use in computer-aided drug development [5, 10].

1.2.4 Energy Minimization and Geometry Optimization

Drug molecules prefer to adopt equilibrium geometry in nature, that is, a geometry that possesses a stable 3D arrangement of atoms in the molecule. The 3D structure of a molecule built using a 3D builder does not represent a natural state; slight modifications are required to be made on the built 3D structure so that it represents the natural state. For this purpose, the following questions need to be addressed: (1) Which minimal changes need to be made? (2) How much change needs to be made? (3) How does one know the representation at hand is the true representation of the natural state? To provide answers to these questions we can depend on energy, because molecules prefer to exist in thermodynamically stable states. This implies that if the energy of any molecule can be minimized, the molecule is not in a stable state and thus the current representation of the molecule may not be the true representation of the natural state. This also implies that we can minimize the energy and the molecular structure in that energy minimum state probably represents a true natural state. Several methods of energy minimization have been developed by

computational chemists, some of which are nonderivative methods (simplex method) but many of which are dependent on derivative methods (steepest descent, Newton–Raphson, conjugate gradient, variable metrics, etc.) and involve the estimation of the gradient of the potential energy curve [4–6]. The entire procedure of geometry modification to reach an energy minimum state with almost null gradient is known as geometry optimization in terms of the structure of the molecule and energy minimization in terms of the energy of the molecule. All computational chemistry software packages are equipped with energy minimization methods—of which a few incorporate energy minimization based on *ab initio* methods while most include the semiempirical and molecular mechanics based energy minimization methods.

1.2.5 Conformational Analysis

Molecules containing freely rotatable bonds can adopt many different conformations. Energy minimization procedures lead the molecular structure to only one of the chemically favorable conformations, called the local minimum. Out of the several local minima on the potential energy (PE) surface of a molecule, the lowest energy conformation is known as the global minimum. It is important to note all the possible conformations of any molecule and identify the global minimum before taking up a drug design exercise (Fig. 1.2). This is important because only one of the possible conformations of a drug, known as its bioactive conformation, is responsible for its therapeutic effect. This conformation may be a global minimum, a local minimum, or a transition state between local minima. As it is very difficult to identify

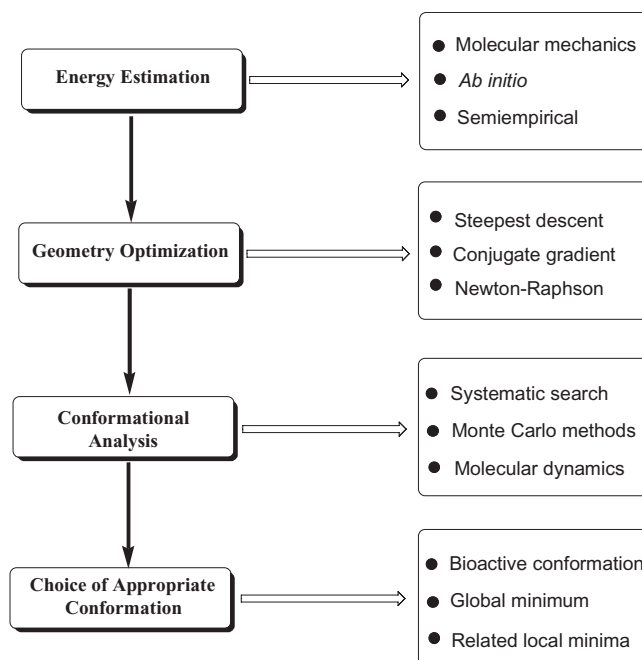


FIGURE 1.2 Flowchart showing the sequence of steps during molecular modeling of drug molecules.

the bioactive conformation of many drug molecules, it is common practice to assume the global minima to be bioactive. The transformation of drug molecules from one conformer to another can be achieved by changing the torsional angles. The computational process of identifying all local minima of a drug molecule, identifying the global minimum conformation, and, if possible, identifying the bioactive conformation is known as conformational analysis. This is one of the major activities in computational chemistry.

Manual conformational search is one method where the chemical intuition of the chemist plays a major role in performing the conformational analysis. Here, a chemist/modeler carefully chooses all possible conformations of a given drug molecule and estimates the energy of each conformation after performing energy minimization. This procedure is very effective and is being widely used. This approach allows the application of rigorous quantum chemical methods for the conformational analysis. The only limitation of this method arises from the ability and patience of the chemist. There is a possibility that a couple of important conformations are ignored in this approach. To avoid such problems, automated conformational analysis methods were introduced.

Various automated methods of conformational analysis include systematic search, random search, Monte Carlo simulations, molecular dynamics, genetic algorithms, and expert systems (Table 1.1) [4, 5]. The systematic conformational search can be performed by varying systematically each of the torsion angles of the rotatable bonds of a molecule to generate all possible conformations. The step size for torsion angle change is normally 30–60°. The number of conformations across a C—C single bond would vary between 6 and 12. With an increase in the number of rotatable bonds, the total number of conformations generated becomes quite large. The “bump check” method reduces the number of possibilities; still, the total number of conformations generated can be in the tens of thousands for drug molecules. Obviously, most of the conformations are chemically nonsignificant.

The random conformational search method employs random change in torsional angle across rotatable bonds and performs energy minimization each time; thus, a handful of chemically meaningful conformations can be generated [11–18].

Molecular dynamics is another method of carrying out conformational search of flexible molecules. The aim of this approach is to reproduce time-dependent motional behavior of a molecule, which can identify bound states out of several possible

TABLE 1.1 Different Methods of Conformational Analysis

Methods for Conformational Analysis	Remarks
Systematic search	Systematic change of torsions
Random search	Conformations picked up randomly
Monte Carlo method	Supervised random search
Molecular dynamics	Newtonian forces on atoms and time dependency incorporated in conformational search
Genetic algorithm	Parent–child relationship along with survival of the fittest techniques employed
Expert system	Heuristic methods based on rules and facts employed

states. The user needs to define step size, time of run, and the temperature supplied to the system at the beginning of the computational analysis. A simulated annealing method allows “cooling down” of the system at regular time intervals by decreasing the simulation temperature. As the temperature approaches 0 K, the molecule is trapped in the nearest local minimum. It is used as the starting point for further simulation and the cycle is repeated several times [19].

1.3 COMPUTATIONAL BIOLOGY

Computational biology is a fast growing topic and it is really not practical to distinguish this topic from bioinformatics. However, we may broadly distinguish between the two topics as far as this chapter is concerned. Molecular modeling aspects of computational biology, which lead to structure prediction, may be discussed under this heading, whereas the sequence analysis part, which leads to target identification, may be discussed under the section of pharmacoinformatics. Structure prediction of biomolecules (often referred to as “structural bioinformatics”) adopts many aspects of computational chemistry. For example, energy minimization of protein receptor structure is one important step in computational biology. Molecular mechanics, molecular simulations, and molecular dynamics are employed in performing conformational analysis of macromolecules.

A rational drug design approach is very much dependent on the knowledge of receptor protein structures and is severely limited by the availability of target protein structure with experimentally determined 3D coordinates. Proteins exhibit four tiered organization: (1) primary structure defining the amino acid sequence, (2) secondary structure with α -helical and β -sheet folds, (3) tertiary structure defining the folding of secondary structure held by hydrogen bonds, and (4) quaternary structure involving noncovalent association between two or more independent proteins. Methods for identifying the primary amino acid sequence in proteins are now well developed; however, this knowledge is not sufficient enough to understand the function of the proteins, the drug–receptor mutual recognition, and designing drugs. Various experimental techniques like X-ray crystallography, nuclear magnetic resonance, and electron diffraction are available for determining the 3D coordinates of the protein structure; however, there are many limitations. It is not easy to crystallize proteins and even when we succeed, the crystal structure represents only a rigid state of the protein rather than a dynamic state. Thus, the reliability of the experimental data is not very high in biomolecules. Computational methods provide the alternative approach—although with equal uncertainty but at a greater speed. Homology modeling and *ab initio* methods are being employed to elucidate the tertiary structure of various biomolecules. The 3D structures of proteins are useful in performing molecular docking, *de novo* design, and receptor-based pharmacophore mappings. The computational methods of biomolecular structure prediction are discussed next (Fig. 1.3) [20, 21].

1.3.1 Ab Initio Structure Prediction

This approach seeks to predict the native conformation of a protein from the amino acid sequence alone. The predictions made are based on fundamental understanding of the protein structure and the predictions must satisfy the requirements of free-

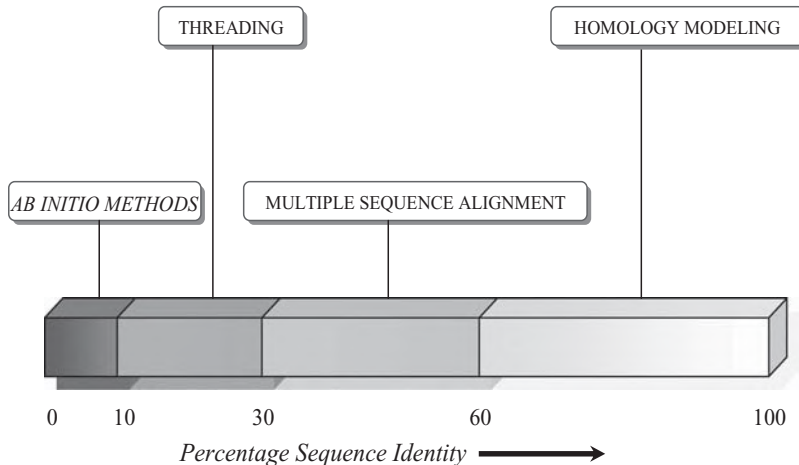


FIGURE 1.3 A list of computer-aided structure prediction methods with respect to their suitability to the available sequence similarity.

energy function associated with lowest free-energy minima. The detailed representation of macromolecules should include the coordinates of all atoms of the protein and the surrounding solvent molecules. However, representing this large number of atoms and the interactions between them is computationally expensive. Thus, several simplifications have been suggested in the representations during the *ab initio* structure prediction process. These include (1) representation of side chains using a limited set of conformations that are found to be prevalent in structures from the Protein Data Bank (PDB) without any great loss in predictive ability [22] and (2) restriction of the conformations available to the polypeptides in terms of phi-psi (ϕ - ψ) angle pairs [23]. Building the protein 3D structure is initiated by predicting the structures of protein fragments. Local structures of the protein fragments are generated first after considering several alternatives through energy minimization. A list of possible conformations is also extracted from experimental structures for all residues. Protein tertiary structures are assembled by searching through the combinations of these short fragments. During the assembling process, bump checking and low energy features (hydrophobic, van der Waals forces) should be incorporated. The final suggested structure is subjected to energy minimization and conformational analysis using molecular dynamics simulations. *Ab initio* structure prediction can be used to guide target selection by considering the fold of biological significance. The *ab initio* macromolecular structure prediction methods, if successful, are superior to the widely used homology modeling technique because no *a priori* bias is incorporated into the structure prediction [24].

1.3.2 Homology Modeling

Homology or comparative modeling uses experimentally determined 3D structure of a protein to predict the 3D structure of another protein that has a similar amino acid sequence. It is based on two major observations: (1) structure of a protein is uniquely determined by its amino acid sequence and (2) during evolution, the structure is more conserved than the sequence such that similar sequences adopt

practically identical structure and distantly related sequences show similarity in folds. Homology modeling is a multistep method involving the following steps: (1) obtaining the sequence of the protein with unknown 3D structure, (2) template identification for comparative analysis, (3) fold assignment based on the known chemistry and biology of the protein, (4) primary structure alignment, (5) backbone generation, (6) loop modeling, (7) side chain modeling, (8) model optimization, and (9) model validation.

The methodology adopted in homology modeling of proteins can be described as follows. The target sequence is first compared to all sequences reported in the PDB using sequence analysis. Once a template sequence is found in the data bank, an alignment is made to identify optimum correlation between template and target. If identical residues exist in both the sequences, the coordinates are copied as such. If the residues differ, then only the coordinates of the backbone elements (N, C α , C, and O) are copied. Loop modeling involves shifting all insertions and deletions to the loops and further modifying them to build a considerably well resembling model. Modeling the side chains involves copying the conserved residues, which also includes substitution of certain rotamers that are strongly favored by the backbone. Model optimization is required because of the expected differences in the 3D structures of the target and the template. The energy minimizations can be performed using molecular mechanics force fields (either well defined and/or self-parameterizing force fields). Molecular dynamic simulations offer fast, more reliable 3D structure of the protein. Model validation is a very important step in homology modeling, because several solutions may be obtained and the scientific user should interfere and make a choice of the best generated model. Often, the user may have to repeat the process with increased caution [20, 24].

1.3.3 Threading or Remote Homology Modeling

Threading (more formally known as “fold recognition”) is a method that may be used to suggest a general structure for a new protein. It is mainly adopted when pairwise sequence identity is less than 25% between the known and unknown structure. Threading technique is generally associated with the following steps: (1) identify the remote homology between the unknown and known structure; (2) align the target and template; and (3) tailor the homology model [24].

1.4 COMPUTATIONAL MEDICINAL CHEMISTRY

Representation of drug molecular structures can be handled using computational chemistry methods, whereas that of macromolecules can be handled using computational biology methods. However, finding the therapeutic potential of the chemical species and understanding the drug–receptor interactions *in silico* requires the following well developed techniques of computational medicinal chemistry.

1.4.1 Quantitative Structure–Activity Relationship (QSAR)

QSAR is a statistical approach that attempts to relate physical and chemical properties of molecules to their biological activities. This can be achieved by using easily

TABLE 1.2 Different Dimensions in QSAR

1D QSAR: Affinity correlates with pK_a , $\log P$, etc.
2D QSAR: Affinity correlates with a structural pattern.
3D QSAR: Affinity correlates with the three-dimensional structure.
4D QSAR: Affinity correlates with multiple representations of ligand.
5D QSAR: Affinity correlates with multiple representations of induced-fit scenarios.
6D QSAR: Affinity correlates with multiple representations of solvation models.

calculatable descriptors like molecular weight, number of rotatable bonds, and $\log P$. Developments in physical organic chemistry over the years and contributions of Hammett and Taft in correlating the chemical activity to structure laid the basis for the development of the QSAR paradigm by Hansch and Fujita. Table 1.2 gives an overview of various QSAR approaches in practice. The 2D and 3D QSAR approaches are commonly used methods, but novel ideas are being implemented in terms of 4D–6D QSAR. The increased dimensionality does not add any additional accuracy to the QSAR approach; for example, no claim is valid which states that the correlation developed using 3D descriptors is better than that based on 2D descriptors.

2D QSAR Initially, 2D QSAR or the Hansch approach was in vogue, in which different kinds of descriptors from the 2D structural representations of molecules were correlated to biological activity. The basic concept behind 2D QSAR is that structural changes that affect biological properties are electronic, steric, and hydrophobic in nature. These properties can be described in terms of Hammett substituent and reaction constants, Verloop sterimol parameters, and hydrophobic constants. These types of descriptors are simple to calculate and allow for a relatively fast analysis.

Most 2D QSAR methods are based on graph theoretical indices. The graph theoretical descriptors, also called the molecular topological descriptors, are derived from the topology of a molecule, that is, the 2D molecular structure represented as graphs. These topological connectivity indices representing the branching of a molecule were introduced by Randić [25] and further developed by Kier and Hall [26, 27]. The graph theoretical descriptors include mainly the Kier–Hall molecular connectivity indices (χ) and the Wiener [28, 29], Hosoya [30], Zagreb [31], Balaban [32], kappa shape [33], and information content indices [32]. The electrotopological state index (E-state) [34] combines the information related to both the topological environment and the electronic character of each skeletal atom in a molecule. The constitutional descriptors are dependent on the constitution of a molecule and are numerical descriptors, which include the number of hydrogen bond donors and acceptors, rotatable bonds, chiral centers, and molecular weight (1D) [35]. Apart from that, several indicator descriptors, which define whether or not a particular indicator is associated with a given molecule, are also found to be important in QSAR. The quantum chemical descriptors include the molecular orbital energies (HOMO, LUMO), charges, superdelocalizabilities, atom–atom and molecular polarizabilities, dipole moments, total and binding energies, and heat of formation. These are 3D descriptors derived from the 3D structure of the molecule and are electronic in nature [36]. These parameters are also often clubbed with the 2D QSAR analysis.

Statistical data analysis methods for QSAR development are used to identify the correlation between molecular descriptors and biological activity. This correlation may be linear or nonlinear and accordingly the methods may be divided into linear and nonlinear approaches. The linear approaches include simple linear regression, multiple linear regression (MLR), partial least squares (PLS), and genetic algorithm–partial least squares (GA-PLS). Simple linear regression develops a single descriptor linear equation to define the biological activity of the molecule. MLR is a step ahead as it defines a multiple term linear equation. More than one term is correlated to the biological activity in a single equation. PLS, on the other hand, is a multivariate linear regression method that uses principal components instead of descriptors. Principal components are the variables found by principal component analysis (PCA), which summarize the information in the original descriptors. The aim of PLS is to find the direct correlation not between the descriptors and the biological activity but between the principal component and the activity. GA-PLS integrates genetic algorithms with the PLS approach. Genetic algorithms are an automatic descriptor selection method that incorporates the concepts of biological evolution within itself. An initial random selection of descriptors is made and correlated to the activity. This forms the first generation, which is then mutated to include new descriptors, and crossovers are performed between the equations to give the next generation. Equations with better predictability are retained and the others are discarded. This procedure is continually iterated until the desired predictability is obtained or the specified number of generations have been developed. The nonlinear approaches include an Artificial Neural Network (ANN) and machine learning techniques. Unlike the linear approaches, nonlinear approaches work on a black box principle; that is, they develop a relation between the descriptors and the activity to predict the activity, but do not give the information on how the correlation was made or which descriptors are more contributing. The ANN algorithm uses the concept of the functioning of the brain and consists of three layers. The first layer is the input layer where the structural descriptors are given as an input; second is the hidden layer, which may be comprised of more than one layer. The input is processed in this part to give the predicted values to the third output layer, which gives the result to the user. The user can control the input given and the number of neurons and hidden layers but cannot control the correlating method [37–40].

The QSAR model developed by any statistical method has to be validated to confirm that it represents the true structure–activity relationship and is not a chance correlation. This may be done by various methods such as the leave-one-out and leave-multiple-out cross-validations and the bootstrap method. The randomization test is another validation approach used to confirm the adequacy of the training set. Attaching chemical connotation to the developed statistical model is an important aspect. A successful QSAR model not only effectively predicts the activity of new species belonging to the same series but also should provide chemical clues for future improvement. This requirement, as well as the recognition that the 3D representation of the chemicals gives more detailed information, led to the development of 3D QSAR.

3D QSAR 3D QSAR methods are an extension of the traditional 2D QSAR approach, wherein the physicochemical descriptors are estimated from the 3D struc-

tures of the chemicals. Typically, properties like molecular volume, molecular shape, HOMO and LUMO energies, and ionization potential are the properties that can be calculated from the knowledge of the 3D coordinates of each and every atom of the molecules. When these descriptors of series of molecules can be correlated to the observed biological activity, 3D QSAR models can be developed. This approach is different from the traditional QSAR only in terms of the descriptor definition and, in a sense, is not really 3D in nature.

Molecular fields (electrostatic and steric), which can be estimated using probe-based sampling of 3D structure of molecules within a molecular lattice, can be correlated with the reported numeric values of biological activity. Such methods proved to be much more informative as they provide differences in the fields as contour maps. The widely used CoMFA (comparative molecular field analysis) method is based on molecular field analysis and represents real 3D QSAR methods [41]. A similar approach was adopted in developing modules like CoMSIA (comparative molecular similarity index analysis) [42], SOMFA (self-organizing molecular field analysis) [43], and COMMA (comparative molecular moment analysis) [44]. Utilization and predictivity of CoMFA itself has improved sufficiently in accordance with the objectives to be achieved by it [45]. Despite the formal differences between the various methodologies, any QSAR method must include some identifiers of chemical structures, reliably measured biological activities, and molecular descriptors. In 3D QSAR, alignment (3D superimposition) of the molecules is necessary to construct good models. The main problems encountered in 3D QSAR are related to improper alignment of molecules, greater flexibility of the molecules, uncertainties about the bioactive conformation, and more than one binding mode of ligands. While considering the template, knowledge of the bioactive conformation of any lead compound would greatly help the 3D QSAR analysis. As discussed in Section 1.2.5, this may be obtained from the X-ray diffractions or conformation at the binding site, or from the global minimum structure. Alignment of 3D structures of molecules is carried out using RMS atoms alignment, moments alignment, or field alignment. The relationship between the biological activity and the structural parameters can be obtained by multiple linear regression or partial least squares analysis. Given next are some details of the widely used 3D QSAR approach CoMFA.

CoMFA (Comparative Molecular Field Analysis) DYLOMMS (dynamic lattice-oriented molecular modeling system) was one of the initial developments by Cramer and Milne to compare molecules by aligning in space and by mapping their molecular fields to a 3D grid. This approach when used with partial least squares based statistical analysis gave birth to the CoMFA approach [46]. The CoMFA methodology is a 3D QSAR technique that allows one to design and predict activities of molecules. The database of molecules with known properties is suitably aligned in 3D space according to various methodologies. After consistently aligning the molecules within a molecular lattice, a probe atom (typically carbon) samples the steric and electrostatic interactions of the molecule. Charges are then calculated for each molecule using any of the several methods proposed for partial charge estimation. These values are stored in a large spreadsheet within the module (SYBYL software) and are then accessed during the partial least squares (PLS) routine, which attempts to correlate these field energy terms with a property of interest by the use of PLS with cross-validation, giving a measure of the predictive power of the model.

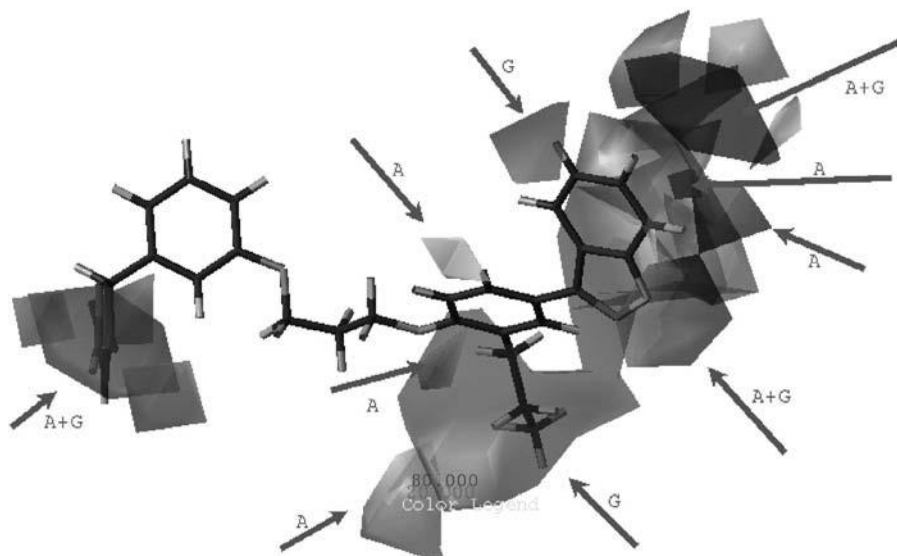


FIGURE 1.4 Steric and electrostatic contour map for the dual model showing the contributions from each model. “A” depicts the contributions made by the α -model and “G” depicts the contributions made by the γ - and model. (Reproduced with permission from The American Chemical Society; S. Khanna, M. E. Sobhia, P. V. Bharatam *J Med Chem* 2005;48:3015.)

Electrostatic maps are generated, indicating red contours around regions where high electron density (negative charge) is expected to increase activity, and blue contours where low electron density (partial positive charge) is expected to increase activity. Steric maps indicate areas where steric bulk is predicted to increase (green) or decrease (yellow) activity [41, 45]. Figure 1.4 shows a typical contour map from CoMFA analysis. CoMSIA [42], CoMMA [44], GRID [47], molecular shape analysis (MSA) [48], comparative receptor surface analysis (CoRSA) [49], and Apex-3D [50] are other 3D QSAR methods that are being employed successfully.

4D QSAR 4D QSAR analysis developed by Vedani and colleagues incorporates the conformational alignment and pharmacophore degrees of freedom in the development of 3D QSAR models. It is used to create and screen against 3D-pharmacophore QSAR models and can be used in receptor-independent or receptor-dependent modes. 4D QSAR can be used as a CoMFA preprocessor to provide conformations and alignments; or in combination with CoMFA to combine the field descriptors of CoMFA with the grid cell occupancy descriptors (GCODs) of 4D QSAR to build a “best” model; or in addition to CoMFA because it treats multiple alignments, conformations, and embedded pharmacophores, which are limitations of CoMFA [51].

5D QSAR The 4D QSAR concept has been extended by an additional degree of freedom—the fifth dimension—allowing for multiple representations of the topology of the quasi-atomistic receptor surrogate. While this entity may be generated using up to six different induced-fit protocols, it has been demonstrated that the

simulated evolution converges to a single model and that 5D QSAR, due to the fact that model selection may vary throughout the entire simulation, yields less biased results than 4D QSAR, where only a single induced-fit model can be evaluated at a time (software Quasar) [52, 53].

6D QSAR A recent extension of the Quasar concept to sixth dimension (6D QSAR) allows for the simultaneous consideration of different solvation models [54]. This can be achieved explicitly by mapping parts of the surface area with solvent properties (position and size are optimized by the genetic algorithms) or implicitly. In Quasar, the binding energy is calculated as

$$E_{\text{binding}} = E_{\text{ligand-receptor}} - E_{\text{desolvation,ligand}} - T \Delta S - E_{\text{internal strain}} - E_{\text{induced fit}} \quad (1.4)$$

1.4.2 Pharmacophore Mapping

A pharmacophore may be defined as the spatial arrangement of a set of key features present in a chemical species that interact favorably with the receptor leading to ligand-receptor binding and which is responsible for the observed therapeutic effect. It is the spatial arrangement of key chemical features that are recognized by a receptor and are thus responsible for biological response. Pharmacophore models are typically used when some active compounds have been identified but the 3D structure of the target protein or receptor is unknown. It is possible to derive pharmacophores in several ways, by analogy to a natural substrate, by inference from a series of dissimilar biologically active molecules (active analogue approach) or by direct analysis of the structure of known ligand and target protein.

A pharmacophoric map is a 3D description of a pharmacophore developed by specifying the nature of the key pharmacophoric features and the 3D distance map among all the key features. Figure 1.5 shows a pharmacophore map generated from the DISCO software module of SYBYL. A pharmacophore map may be generated from the superimposition of the active compounds to determine their common features. Given a set of active molecules, the mapping of a pharmacophore involves two steps: (1) analyzing the molecules to identify pharmacophoric features, and (2) aligning the active conformations of the molecules to find the best overlay of the corresponding features. Various pharmacophore mapping algorithms differ in the way they handle the conformational search, feature definition, tolerance definition, and feature alignment [55]. During pharmacophore mapping, generation and optimization of the molecules and the location of ligand points and site points (projections from ligand atoms to atoms in the macromolecule) are carried out. Typical ligand and site points are hydrogen bond donors, hydrogen bond acceptors, and hydrophobic regions such as centers of aromatic rings. A pharmacophore map identifies both the bioactive conformation of each active molecule and how to superimpose and compare, in three dimensions, the various active compounds. The mapping technique identifies what type of points match in what conformations of the compounds.

Besides ligand-based automated approaches, pharmacophore maps can also be generated manually. In such cases, common structural features are identified from a set of experimentally known active compounds. Conformational analysis is carried out to generate different conformations of the molecules and interfeature distances

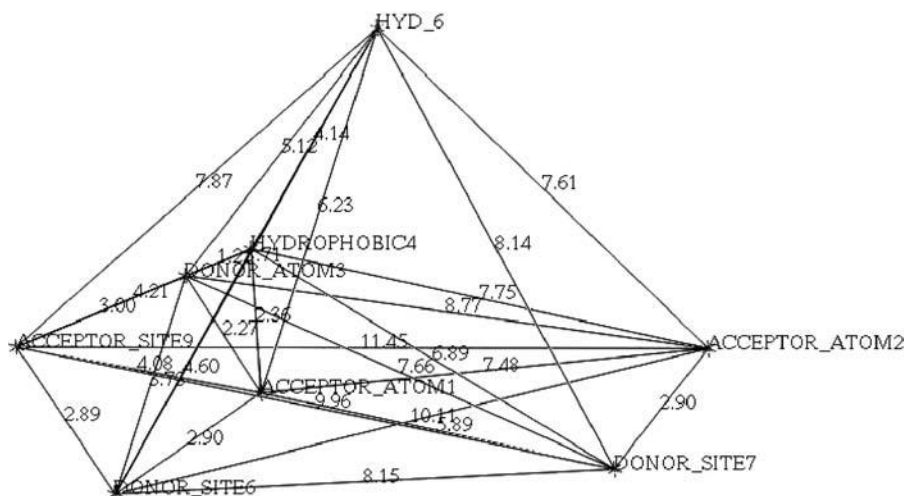


FIGURE 1.5 A pharmacophore map developed from a set of GSK3 inhibitors. The pharmacophore features include hydrogen bond acceptor atoms, hydrogen bond donor atoms, hydrogen bond donor site, hydrogen bond acceptor site, and hydrophobic centers. This 3D picture also shows the distance relationship between various pharmacophoric features present in the map.

are inferred to develop the final models. The receptor mapping technique is also currently in practice to develop pharmacophore models. The important residues required for binding the pharmacophores are identified, which are employed for generating the receptor-based pharmacophores. The structure of protein can be used to generate interaction sites or grids to characterize favorable positions for ligands.

After a pharmacophore map has been derived, there are two ways to identify molecules that share its features and thus elicit the desired response. First is the *de novo* drug design, which seeks to link the disjoint parts of the pharmacophore together with fragments in order to generate hypothetical structures that are chemically novel. Second is the 3D database searching, where large databases comprising 3D structures are searched for those that match to a pharmacophoric pattern. One advantage of the second method is that it allows the ready identification of existing molecules, which are either easily available or have a known synthetic route [56, 57]. Pharmacophore mapping methods are described next.

Distance Comparison Method (DISCO) The various steps involved in DISCO-based generation of a pharmacophore map are conformational analysis, calculation of the location of the ligand and site points, finding potential pharmacophore maps, and graphics analysis of the results. In the process of conformational search, 3D structures can be generated using any building program like CONCORD, from crystal structures, or from conformational searching and energy minimization with any molecular or quantum mechanical technique. Comparisons of all the duplicate conformations are excluded while comparing all the conformations. If each corresponding interatomic distance between these atoms in the two conformations is less

than a threshold (0.4 \AA), then the higher energy conformation is rejected. DISCO calculates the location of site points, which can be the location of ligand atoms, or other atom-based points, like centers of rings or a halogen atom, which are points of potential hydrophobic groups. The other point is the location of the hydrogen bond acceptors or donors. The default locations of site hydrogen bond donor and acceptor points are based on literature compilations of observed intermolecular crystallographic contacts in proteins and between the small molecules. Hydrogen bond donors and acceptors such as OH and NH_2 groups can rotate to change the locations of the hydrogen atom.

During the process of performing pharmacophore mapping in DISCO, the user may input the tolerance for each type of interpoint distance. The user may direct the DISCO algorithm to consider all the potential points and to stop when a pharmacophore map with a certain total number of points is found. Alternatively, the user may specify the types of points, and the maximum and minimum number of each, that every superposition must include. It can also be directed to ignore specific compounds if they do not match a pharmacophore map found by DISCO. The user may also specify that only the input chirality is used for certain molecules and that only certain conformations below a certain relative energy should be considered.

The DISCO algorithm involves finding the reference molecule, which is the one with the fewest conformations. The search begins by associating the conformations of each molecule with each other. DISCO then calculates the distances between points in each 3D structure. Then it prepares the corresponding tables that relate interpoint distances in the current reference conformation and distances in every other 3D structure. Distances correspond if the point types are the same. These distances differ by no more than the tolerance limits. The clique-detection algorithm then identifies the largest clique of distances common between the reference XYZ set and every other 3D structure. It then forms union sets for the cliques of each molecule. Finally, the sets with cliques that meet the group conditions are searched [58, 59].

CATALYST According to the pharmacophore mapping software CATALYST, a conformational model is an abstract representation of the accessible conformational space of a ligand. It is assumed that the biologically active conformation of a ligand (or a close approximation thereof) should be contained within this model. A pharmacophore model (in CATALYST called a hypothesis) consists of a collection of features necessary for the biological activity of the ligands arranged in 3D space, the common ones being hydrogen bond acceptor, hydrogen bond donor, and hydrophobic features. Hydrogen bond donors are defined as vectors from the donor atom of the ligand to the corresponding acceptor atom in the receptor. Hydrogen bond acceptors are analogously defined. Hydrophobic features are located at the centroids of hydrophobic atoms. CATALYST features are associated with position constraints that consist of the ideal location of a particular feature in 3D space surrounded by a spherical tolerance. In order to map the pharmacophore, it is not necessary for a ligand to possess all the appropriate functional groups capable of simultaneously residing within the respective tolerance spheres of the pharmacophoric features. However, the fewer features an inhibitor maps to, the poorer is its fit to them and the lower is its predicted affinity [60–63].

1.4.3 Molecular Docking

There are several possible conformations in which a ligand may bind to an active site, called the binding modes. Molecular docking involves a computational process of searching for a conformation of the ligand that is able to fit both geometrically and energetically into the binding site of a protein. Docking calculations are required to predict the binding mode of new hypothetical compounds. The docking procedure consists of three interrelated components—identification of the binding site, a search algorithm to effectively sample the search space (the set of possible ligand positions and conformations on the protein surface), and a scoring function. In most docking algorithms, the binding site must be predefined, so that the search space is limited to a comparatively small region of the protein. The search algorithm effectively samples the search space of the ligand–protein complex. The scoring function used by the docking algorithm gives a ranking to the set of final solutions generated by the search. The stable structures of a small molecule correspond to minima on the multidimensional energy surface, and different energy calculations are needed to identify the best candidate. Different forces that are involved in binding are electrostatic, electrodynamic, and steric forces and solvent related forces. The free energy of a particular conformation is equal to the solvated free energy at the minimum with a small entropy correction. All energy calculations are based on the assumption that the small molecule adopts a binding mode of lowest free energy within the binding site. The free energy of binding is the change in free energy that occurs upon binding and is given as

$$\Delta G_{\text{binding}} = G_{\text{complex}} - (G_{\text{protein}} + G_{\text{ligand}}) \quad (1.5)$$

where G_{complex} is the energy of the complexed protein and ligand, G_{protein} is the free energy of noninteracting separated protein, and G_{ligand} is the free energy of noninteracting separated ligand.

The common search algorithms used for the conformational search, which provide a balance between the computational expense and the conformational search, include molecular dynamics, Monte Carlo methods, genetic algorithms, fragment-based methods, point complementary methods, distance geometry methods, tabu searches, and systematic searches [64].

Scoring functions are used to estimate the binding affinity of a molecule or an individual molecular fragment in a given position inside the receptor pocket. Three main classes of scoring functions are known, which include force field-based methods, empirical scoring functions, and knowledge-based scoring functions. The force field scoring functions use molecular mechanics force fields for estimating binding affinity. The AMBER and CHARMM nonbonded terms are used as scoring functions in several docking programs. In empirical scoring functions, the binding free energy of the noncovalent receptor–ligand complex is estimated using chemical interactions. These scoring functions usually contain individual terms for hydrogen bonds, ionic interactions, hydrophobic interactions, and binding entropy, as in the case of SCORE employed in DOCK4 and Böhm scoring functions (explained in detail in Section 1.4.4) used in FlexX. In empirical scoring functions, less frequent interactions are usually neglected. Knowledge-based scoring functions try to capture the knowledge about protein–ligand binding that is implicitly stored in the Protein Data Bank by means of statistical analysis of structural data, for example, PMF and

DrugScore functions, Wallqvist scoring function, and the Verkhivker scoring function [5, 37, 65–67]. Various molecular docking software packages are available, such as FlexX [68], Flexidock [58], DOCK [69], and AUTODOCK [70].

FlexX FlexX is a fragment-based method for docking which handles the flexibility of the ligand by decomposing the ligand into fragments and performs the incremental construction procedure directly inside the protein active site. It allows conformational flexibility of the ligand while keeping the protein rigid. The base fragment or the ligand core is selected such that it has the most potential interaction groups and the fewest alternative conformations. It is placed into the active site and joined to the side chains in different conformations. Placements of the ligand are scored on the basis of protein–ligand interactions and ranked after the estimation of binding energy. The scoring function of FlexX is a modification of Böhm’s function developed for the *de novo* design program LUDI. Figure 1.6 shows details of the interaction between a ligand and a receptor, obtained from FlexX molecular docking.

DOCK DOCK is a simple minimization program that generates many possible orientations of a ligand within a user selective region of the receptor. DOCK is a program for locating feasible binding orientations, given the structures of a “ligand” molecule and a “receptor” molecule [69]. DOCK generates many orientations of one ligand and saves the best scoring orientation. The docking process is handled

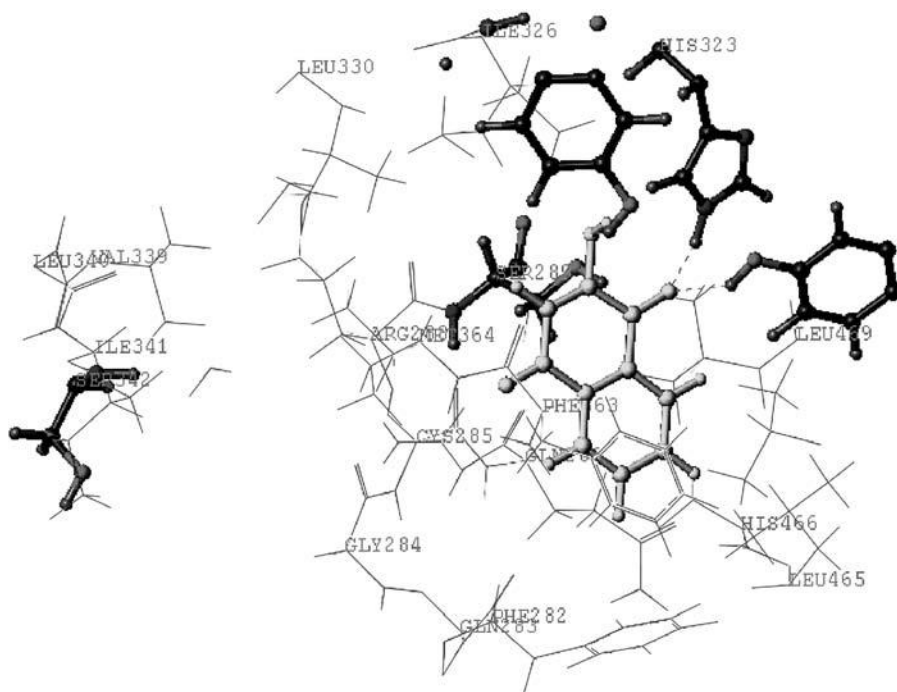


FIGURE 1.6 The result of docking a ligand in the active site of PPAR γ . The ligand has a hydrogen bonding interaction with histidine and tyrosine.

in four stages—ligand preparation, site characterization, scoring grid calculation, and finally docking. Site characterization is carried out by constructing site points, to map out the negative image of the active site, which are then used to construct orientations of the ligand. Scoring grid calculations are necessary to identify ligand orientations. The best scoring poses may be viewed using a molecular graphics program and the underlying chemistry may be analyzed.

There are many other widely used molecular docking software packages, like Flexidock (based on genetic algorithm), Autodock (based on Monte Carlo simulations and annealing), MCDOCK (Monte Carlo simulations), FlexE (ensemble of protein structures to account for protein flexibility), and DREAM++ (to dock combinatorial libraries).

1.4.4 De Novo Design

De novo design is a complementary approach to molecular docking: whereas in molecular docking already known ligands are employed, in *de novo* design, ligands are built inside the ligand binding domain. This is an iterative process in which the 3D structure of the receptor is used to build the putative ligand, fragment by fragment, within the receptor groove. Two basic types of algorithms are being widely used in *de novo* design. The first one is the “outside-in-method,” in which the binding site is first analyzed to determine which specific functional groups might bind tightly. These separated fragments are then connected together with standard linker units to produce the ligands. The second approach is the “inside-out-method,” where molecules are grown within the binding site so as to efficiently fit inside. *De novo* design is the only method of choice when the receptor structure is known but the lead molecules are not available. This method can also be used when lead molecules are known but new scaffolds are being sought. There are several programs developed by various researchers for constructing ligands *de novo*. GROW [71], GRID [72], CAVEAT [73], LUDI [74–77], LEAPFROG [58], GROUPBUILD [78], and SPROUT [79] are some of the *de novo* design programs that have found wide application.

GRID The GRID program developed by Goodford [72] is an active site analysis method where the properties of the active site are analyzed by superimposition with a regular grid. Probe groups like water, methyl group, amine nitrogen, carboxyl oxygen, and hydroxyl are placed at the vertices of the grid and its interaction energy with the protein is calculated at each point using an empirical energy function that determines which kind of atoms and functional groups are best able to interact with the active site. The array of energy values is represented as a contour, which enables identification of regions of attractions between the probe and the protein. It is not a direct ligand generation method but positions simple fragments.

LUDI LUDI, developed by Böhm [74–77], is one of the most widely used automated programs available for *de novo* design. It uses a knowledge-based approach based on rules about the energetically favorable interaction geometries of nonbonded contacts like hydrogen bonds and hydrophobic contacts between the functional groups of the protein and ligand. In LUDI the rules derived from statistical analysis of crystal packings of organic molecules are employed. LUDI is

fragment based and works in three steps. It starts by identifying the possible hydrogen bonding donors and acceptors and hydrophobic interactions, both aliphatic and aromatic, in the binding site represented as interaction site points. The site points are positions in the active site where the ligand could form a nonbonded contact. A set of interaction sites encompasses the range of preferred geometries for a ligand atom or functional group involved in the putative interaction, as observed in the crystal structure analyses. LUDI models the H-donor and H-acceptor interaction sites and the aliphatic or aromatic interaction sites. The interaction sites are defined by the distance R , angle α , and dihedral angle ω . The fragments from a 3D database of small molecules are then searched for positioning into suitable interaction sites such that hydrogen bonds can be formed and hydrophobic pockets filled with hydrophobic groups. The suitably oriented fragments are then connected together by spacer fragments to the respective link sites to form the entire molecule. Figure 1.7 shows LUDI generated fragment interaction sites inside the iNOS substrate binding domain.

An empirical but efficient scoring function is used for prioritizing the hit fragments given by LUDI. It estimates the free energy of binding (ΔG) based on the

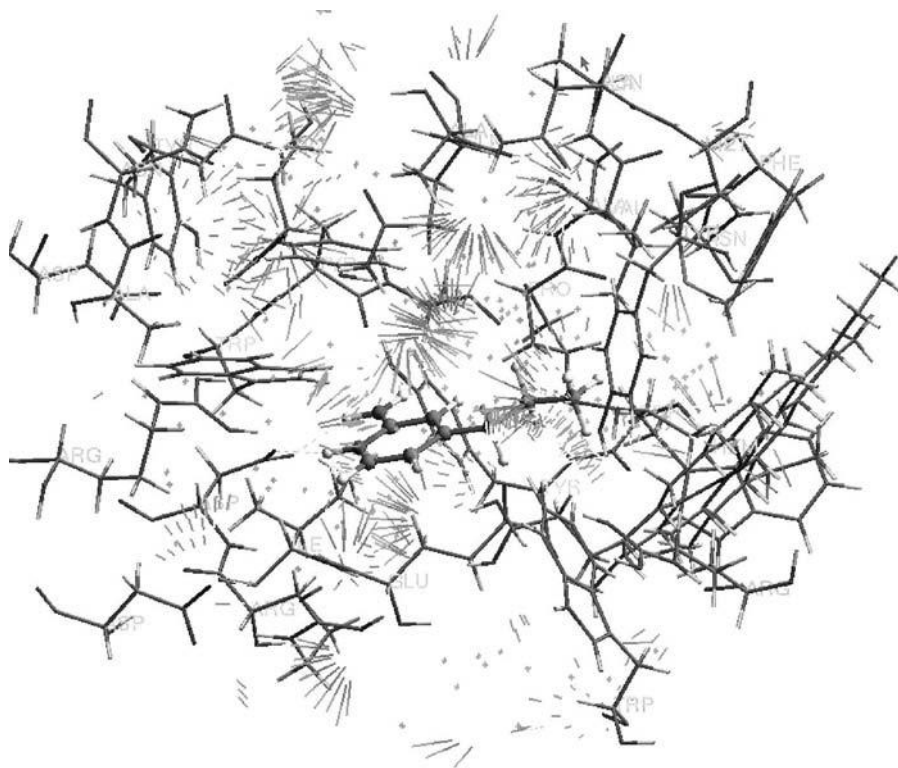


FIGURE 1.7 An example of *de novo* design exercise. In the substrate binding domain of inducible nitric oxide synthase, the stick representation shows the protein structure; ball-and-stick representation belongs to the designed ligand, and the gray sticks point out the interaction sites.

hydrogen bonding, ionic interactions, hydrophobic contact areas, and number of rotatable bonds in the ligand. The LUDI scoring function is given as

$$\begin{aligned} \Delta G = & \Delta G_o + \Delta G_{\text{hb}} \sum f(\Delta R)f(\Delta\alpha) + \Delta G_{\text{ion}} \sum f(\Delta R)f(\Delta\alpha) \\ & + \Delta G_{\text{lipo}} A_{\text{lipo}} + \Delta G_{\text{rot}} NR + \Delta G_{\text{aro/aro}} \Delta N_{\text{aro/aro}} \end{aligned} \quad (1.6)$$

ΔG_o represents the constant contribution to the binding energy due to loss of translational and rotational entropy of the fragment. ΔG_{hb} and ΔG_{ion} represent the contributions from an ideal neutral hydrogen bond and an ideal ionic interaction, respectively. The ΔG_{lipo} term represents the contribution from lipophilic contact and the ΔG_{rot} term represents the contribution due to the freezing of internal degrees of freedom in the fragment. NR is the number of acyclic $\text{sp}^3\text{-sp}^3$ and $\text{sp}^3\text{-sp}^2$ bonds.

1.5 PHARMACOINFORMATICS

Information technology provides several databases, data analysis tools, and knowledge extraction techniques in almost every facet of life. In pharmaceutical sciences, several successful attempts are being made under the umbrella of pharmacoinformatics (synonymously referred to as pharmainformatics) (Fig. 1.8). The scope and limitations of this field are not yet understood. However, it may be broadly defined as the application of information technology in drug discovery and development. It encompasses all possible information technologies that eventually contribute to drug discovery. Chemoinformatics and bioinformatics contribute directly to drug discovery through virtual screening. Topics like neuroinformatics, immunoinformatics, vaccine informatics, and biosystem informatics contribute indirectly by providing necessary inputs for pharmaceutical design in this area. Topics like metabolomics, toxicoinformatics, and ADME informatics are contributing to this field by providing information regarding the fate of a NCE/lead *in vitro* and *in vivo* conditions. In this chapter some important aspects of these topics are presented. It is not easy to offer a comprehensive definition of this field at this stage owing to the fact that several bold attempts are being made in this field and initial signals related to a common platform are only emerging. *Drug Discovery Today* made initial efforts in this area by bringing out a supplement on this topic in which it was mainly treated as a scientific discipline with the integration of both bioinformatics and chemoinformatics [80, 81]. Recent trends in this area include several service-oriented themes including healthcare informatics [82], medicine informatics [83], and nursing informatics [84]. Here we present an overview of the current status.

1.5.1 Chemoinformatics

Chemoinformatics deals with information storage and retrieval of chemical data. This has been pioneered principally by the American Chemical Society and Cambridge Crystallographic Databank. However, the term chemoinformatics came into being only recently when methods of deriving science from the chemical databases was recognized. The integration of back-end technologies (for storing and representing chemical structure and chemical libraries) and front-end technologies (for assessing and analyzing the structures and data from the desktop) provides

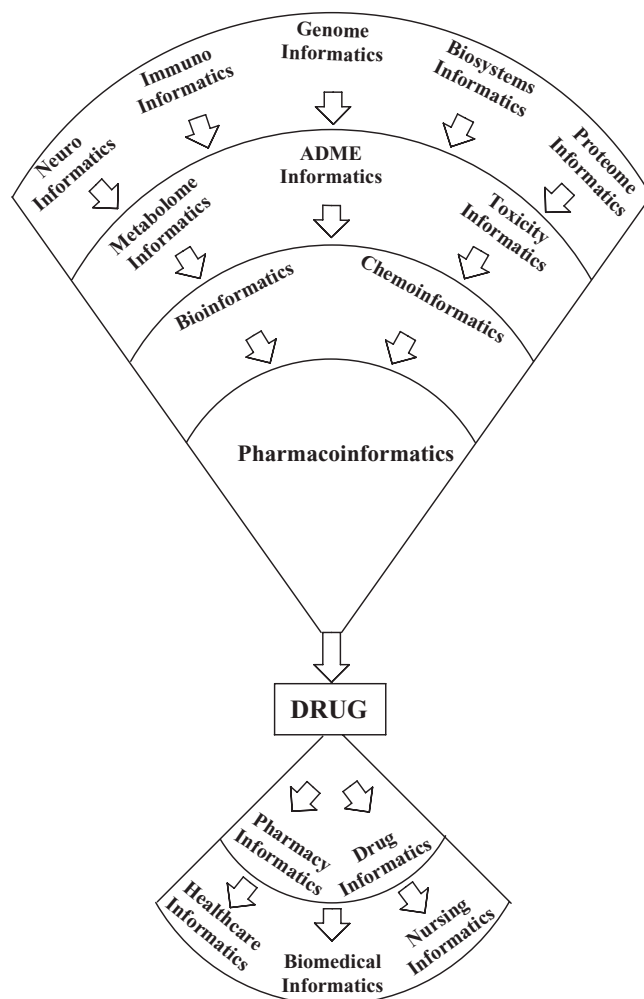


FIGURE 1.8 Flowchart of informatics-based activities in pharmaceutical sciences.

opportunities in chemoinformatics. Virtual screening and high-throughput (HTS) data mining are some of the important chemoinformatics methods.

Chemical data is mostly considered as heuristic data, because it deals with names, properties, reactions, and so on. However, chemoinformatics experts devised many ways of representing chemical data (Fig. 1.9). (1) One-dimensional information of the chemicals can be represented in terms of IUPAC names, molecular formulas, Wiswesser line notation (WLN), SMILES notation, SYBYL line notation (SLN), and so on in addition to numerical representation of physicochemical parameters like molecular weight, molar refractivity, surface area, $\log P$, and pK_a . (2) Two-dimensional chemical information consists of the chemical structural drawings, corresponding hashing, hash codes, connectivity tables, bond matrix, incidence matrix, adjacency matrix, bond-electron matrix, and so on. Graph theoretical procedures are extensively employed in this approach. To get a unique representation of

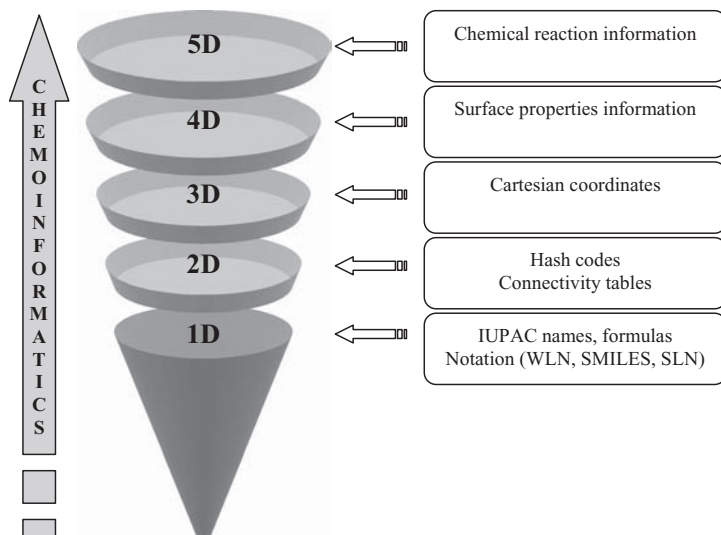


FIGURE 1.9 Information flowchart in chemistry. In this context the dimensionality is not a geometrical dimensionality but an information complexity dimensionality.

chemicals, several algorithms (e.g., Morgan algorithm) were devised, which are found to be extremely important while performing chemical data mining. (3) Three-dimensional representation of structures involves the definition of the Cartesian coordinates of each and every element in the molecule. Along with connection tables, distance matrix, bond angle matrix, and torsional angle matrix, the representations of chemical 3D structures are being made in the form of formatted flat files like .mol and .sdf. (4) Fourth-dimensional information of chemicals is also available in the form of surface properties (e.g., Connolly surface, solvent accessible surface, electron density surfaces, molecular electrostatic potential (MESP), internal structure representation (molecular orbitals)). (5) Fifth-dimensional information about the chemical reactions is of course the most important information about chemical species. This involves defining chemical reaction parameters like inductive effect, resonance effect, polarizability effect, steric effect, and stereochemical effect, on the one hand, and chemical bond formation representations like Hendrickson scheme, Ugi scheme, and InfoChem scheme on the other. Some of the useful chemoinformatics databases include CAS, CSD, STN, MARPAT, UNITY-3D, MACCS-3D, CONCORD, MAYBRIDGE, NCI, and CASReact, which include chemical data in all the possible dimensions as discussed earlier [6, 10, 85–88].

Searching chemical data also requires distinctive methodologies. Search based on hash codes, graph theoretical indices (e.g., Weiner index), charge-related topological indices, Tanimoto coefficients, Carbo coefficient, Hamming distances, Euclidean distance, clique detection, and pharmacophore map searching are some of the important techniques uniquely required in chemoinformatics. Virtual generation of synthesizable chemicals is an important component of this technique. The principles of similarity and diversity need to be employed simultaneously while performing this exercise. Virtual screening is often considered under chemoinformatics, which is discussed in Section 1.5.3 as the most important topic of pharmacoinformatics.

Chemical reaction informatics involves the exploration of synthetic pathways and the designing of new experiments. About 15–20 million reactions are currently available in chemical reaction databases (CASReact, ChemReact, CrossFire Plus, etc.). The chemical reaction informatics would essentially assist the chemist in giving access to reaction information, deriving knowledge, predicting the course and outcome of chemical reactions, and designing syntheses. The main tasks include (1) storing information on chemical reactions, (2) retrieving the information, (3) comparing and analyzing sets of reactions, (4) defining the scope and limitations of a reaction type, (5) developing models of chemical reactivity, (6) predicting the course of the reactions, (7) analyzing reaction networks, and (8) developing methods for the design of syntheses. Chemical reaction databases consist of the following information: (1) reactants and products; (2) atom mapping, which allows you to determine which atom becomes which product atom through the reaction; (3) information regarding reacting center(s); (4) the catalyst used; (5) the atmosphere, including temperature, pressure, and composition; (6) the solvent used; (7) product yield; (8) optical purity, and (9) references to literature.

1.5.2 Bioinformatics

The area of bioinformatics encompasses various fields of molecular biology requiring data handling like genomics, proteomics, sequence analysis, and regulatory networks. Results of the Human Genome Project have triggered several activities in bioinformatics as a result of the complete sequencing of the human genome consisting of approximately 30,000 genes. The data generated in the biology laboratories is being stored in data banks like GenBank (United States), EMBL (Europe), DDBJ (Japan), and Swiss-Prot, as primary data sources. Secondary data banks like PROSITE, Profiles, and Pfam contain the fruits of analyses of the sequences in the primary databases (Table 1.3). These databases are available on the Web as well as on some specialized networks like EMBnet and NCBIInet. Apart from the sequence data, data related to gene expression, gene products, and protein interactions are also available, whose management, analysis, and storage are the objectives of bioinformatics. Sequence analysis is the most important aspect of bioinformatics. There are many sequence analysis packages available like the Genetics Computing Group (GCG) package from Accelrys and the EMBOSS suite (European Molecular Biology Open Software Suite) from European Molecular Biology Laboratory (EMBL). Pairwise and multiple alignment algorithms were developed for determining the similarity between the sequences. Dotplot, which gives a dot matrix plot of any two sequences with respect to the amino acid comparison, is the simplest pairwise sequence analysis tool. Sequence identity, if present, becomes evident along diagonal areas in a dotplot. Dynamic programming methods like the Needleman–Wunsch algorithm (global alignment) and Smith–Waterman algorithm (local alignment) provide additional information after inserting necessary gaps. The sequence alignment program GAP from the GCG software and NEEDLE from EMBOSS are for global alignment, whereas BESTFIT from GCG and WATER from EMBOSS are for local alignment. Dynamic programming techniques provide optimal alignment between the amino acid sequences defined by the highest score. The number of matched pairs, mismatched pairs, and gaps can be used in estimating the score of a given pairwise comparison of two sequences. Percent Accepted Mutation matrix

TABLE 1.3 Some Important Bioinformatics Resources

Category	Name	Description	Source
Sequence databases	GenBank	Genetic sequence database	National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov
	EMBL	Nucleic acid and protein databases	European Molecular Biology Laboratory http://www.ebi.ac.uk/embl/index.html
	DDBJ UniProt/Swiss-Prot PIR	Nucleic acid database Protein database Protein Information Resource	http://www.ddbj.nig.ac.jp http://www.uniprot.org http://pir.georgetown.edu/pirwww/
Genome databases	dbEST	Expressed Sequence Tags database	http://www.ncbi.nlm.nih.gov/dbEST/index.html
	GDB	Human Genome Database	http://www.gdb.org/
	Ensembl	Genome database	http://www.ensembl.org/index.html
Secondary protein databases	Pfam	Protein family database with multiple sequence alignments and hidden Markov models	http://www.sanger.ac.uk/Software/Pfam/
	Prodom	Protein family and domain database	http://protein.toulouse.inra.fr/prodom/current/html/home.php
	PROSITE	Protein family and domain database	http://us.expasy.org/prosite/details.html
Protein interaction databases	BIND	Biomolecular Interaction Network Database	http://www.bind.ca
	DIP	Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/
	HPRD	Human Protein Reference Database	http://www.hprd.org/

250 (PAM250) or the BLOsum SUBstitution Matrix 62 (BLOSUM62) for protein sequences are the well accepted scoring matrices. The word k -tuple method is also a widely used sequence search tool, with heuristic algorithms like BLAST (Basic Local Alignment Search Tool) and FASTA. Multiple Sequence Alignment (MSA) methods like CLUSTALW and PILEUP have also been developed for alignment of three or more sequences. Highly conserved regions can be identified from such types of alignments, which is important for identifying members of the same protein family or for studying evolutionary relationships. Most of these tools are available

from the National Center for Biotechnology Information (NCBI) website (<http://www.ncbi.nlm.nih.gov>). An efficient search engine, ENTREZ, is also provided by NCBI for searching sequences or the related literature. The sequences are stored in various formats like GenBank, EMBL, Swiss-Prot, FASTA, and PIR. Some commercial software packages like the GCG Wisconsin package incorporate the databases and analysis tools, which can be customized for the purpose of end users [89].

Proteomics involves a detailed study of all the proteins present in a cell, their expression, post-translational modifications, interactions with drugs and proteins, and so on. Major technologies in the field of proteomics are mass spectrometry and gel electrophoresis. Proteome informatics deals with the application of informatic tools for proteome analysis. Several databases are available that contain information about the interactions between the proteins—for example, Database of Interacting Proteins (DIP) [90], the General Repository for Interaction Datasets (GRID) [91], the Biomolecular Interaction Network Database (BIND) [92], and the Human Protein Reference Database (HPRD) [93]. The general methods for extracting interaction information from the literature are natural language processing (NLP), naïve Bayes [94], decision trees [95], neural networks, nearest neighbor, and support vector machine (SVM). The prediction of protein functions can also be carried out by a comparative genome analysis. Domain fusion studies, chromosomal proximity studies, and phylogenetic profiling are some methods attempted in the postgenomic scenario to take the sequence information beyond, to the annotation of the proteome. In genome informatics, tools for genome comparison like PipMaker and Artemis Comparison Tool (ACT) and tools from NCBI like HomoloGene and LocusLink are included. Databases that integrate different computational methods to predict functional associations have also been developed, like STRING [96, 97] and POINT [98].

Target identification and target validation are the two major aspects of modern bioinformatics which are relevant to drug discovery. A relatively small number of targets (200–500) have been considered until now for the development of drugs. The traditional target identification process follows the path: assay observation → identification of the key protein → cloning of its gene. This process was bogged down by the dearth of information on several fronts, for example, structural data of proteins. With the recent advances in protein crystallography combined with the 2D- and 3D-NMR experiments, the available structural data improved drastically. This is further supplemented by the *ab initio* fold prediction from the primary sequence data. Thousands of structures can now be determined on an industrial scale. The trends in target identification made a paradigm shift in the recent past from a “deductive” approach involving assay-to-genes path to an “inductive” approach involving genes-to-assay path. The greatest challenge in bioinformatics is to facilitate this paradigm shift by providing automated tools for the identification of new genes as potential targets. Target validation involves the identification of the function of targets in disease states, such that “drugable” targets can be recognized. Experimental efforts in this context can be effectively complemented by computational methods by understanding drug–target interactions. For example, the identification of PPAR γ as a target for insulin resistance could be carried out with the effective integration of the results from experimental, computational biology and bioinformatics methods. Initially, a model of *trans*-retinoid acid receptor (RXR) was

built on the basis of available crystal structures. Multiple sequence alignment was then carried out on two human retinoic acid receptors and three human PPAR subtypes. Secondary structure prediction and homology modeling were carried out to understand the structure and binding characteristics of PPAR γ . Several computational techniques like pattern recognition, artificial neural networks, genetic algorithms, and alignment techniques are being employed in analyzing the data stored in databases. The advances in the area of bioinformatics should lead to an explosion in the number of available target molecules, and contribute tremendously to pharmacoinformatics.

1.5.3 Virtual Screening

Virtual screening is one of the most important technologies of pharmacoinformatics. It employs the databases and analysis tools discussed in the previous sections and pharmacophore mapping and molecular docking discussed under computational medicinal chemistry. This topic is considered under pharmacoinformatics because this technology is being heavily used for hit/lead identification through data search rather than studying molecular interactions and chemical principles. Virtual screening (also called *in silico* screening) provides a fast and cost effective tool for computationally screening compound databases in the search for novel drug leads. Various changes have occurred in the methods of drug discovery, the major ones taking place in the field of high throughput synthesis and screening techniques. The basic goal of virtual screening is the reduction of the huge chemical space of synthesized/virtual molecules and to screen them against a specific target protein virtually. Thus, the field of virtual screening has become an important component of drug discovery programs. Substantial efforts in this area have been made by large pharmaceutical companies. However, there are no well defined standards in virtual screening as yet [37].

Virtual Libraries Two types of virtual libraries can be generated. One is produced by computational design, with the basic idea to design synthesizable compounds computationally. The other is a virtual library of compounds that are already synthesized. It is possible to extract millions of compounds from these sources and create databases. Virtual libraries are useful in the absence of knowledge about specific drug targets for virtual screening. Focused virtual libraries are important sometimes to save resources as the hit rate is higher in such cases [62].

Various Approaches of Virtual Screening Virtual screening methods can be roughly divided into target structure-based and small molecule-based approaches, as shown in Table 1.4. When no structural information about the target protein is given, pharmacophore models can be used as filters for screening. These are mostly used when a set of active compounds is known that bind in a similar way to the protein. A pharmacophore capturing these features should be able to identify novel compounds with a similar pattern from the database. These pharmacophores may be generated manually or by automated software packages like CATALYST (HipHop, HypoGen) or DISCO, as described earlier. If the 3D structure of the receptor is known, a pharmacophore model can be derived based on the receptor active site. Usually all the pharmacophore searches are done in two steps: first the

TABLE 1.4 Computational Methods and Tools for Virtual Screening

Target structure-based approaches
Protein–ligand docking
Active site-directed pharmacophores
Molecule-based queries
2D substructures
3D pharmacophores
Complex molecular descriptors (e.g., electrotopological)
Volume- and surface-matching algorithms
Molecular fingerprints
Keyed 2D fingerprints (each bit position is associated with a specific chemical feature)
Hashed 2D fingerprints (properties are mapped to overlapping bit segments)
Multiple-point 3D pharmacophore fingerprints
Compound classification techniques
Cluster analysis
Cell-based partitioning (of compounds into subsections of n -dimensional descriptor space)
3D/4D QSAR models
Statistical methods
Binary QSAR/QSPR
Recursive partitioning

software confirms whether or not the screened compound has the required atom types or functional groups, and then it checks whether the spatial arrangement of these elements matches the query. In actual practice, receptor-based and small molecule-based approaches can be used in combination, taking into account as much information as possible.

Structure based virtual screening is carried out by docking and scoring techniques. Database of thousands of compounds can be screened against the specific target protein. FlexX [68], Flexidock [58], DOCK [52], and AUTODOCK [70] are the various docking programs employed. Molecular fingerprinting is another technique for performing virtual screening. These search tools consist of varying numbers of bits and encode different types of molecular descriptors and their values. Calculation of descriptors for a given molecule produces a characteristic bit pattern that can be quantitatively compared to others, applying a similarity metric [37].

Various filters can be used during the screening process which improves the chances of obtaining reliable hits. One such filter is the substructure filter. The database may contain certain functional groups in molecules that interact at unwanted sites and are “false positives.” Many of these features are called as the substructures that are used to filter datasets. Other filters applied are the molecular weight, the number of rotatable bonds, and the calculated $\log P$. All these considerations have led to the formulation of *Lipinski's rule of five* [99]. These rules suggest whether or not a molecule is going to be absorbed. The criteria for choosing better absorbed molecules are: (1) molecular weight <500, (2) $\log P < 5$, (3) hydrogen bonds donors <5, and (4) hydrogen bond acceptors <10. The rule of five has been derived following a statistical analysis of known drugs. More sophisticated computational models of “drug-likeness” have been developed by employing the techniques of artificial neural networks, decision trees, and genetic algorithms [6].

It is important to realize that virtual screening is expected to produce hits, not leads. The identification of nanomolar inhibitors by database mining is an extremely rare probability. Lead optimization should be taken up to further modulate the structural features. A virtual screening exercise cannot distinguish the agonists; also, in several cases, the inactives are picked up. Hence it is better not to be very narrow toward the end of any virtual screening procedure [62, 63].

Virtual screening can be applied to target-based subset selection from the databases. Statistical approaches like binary QSAR or recursive partitioning can be applied to process HTS results and develop predictive models of biological activity. The developed models can then be employed to select candidate molecules from databases. Similarly, hits from HTS are used in fingerprint searches or compound classification analysis to identify sets of similar molecules. Based on these results, a few compounds are selected for additional testing. Many assumptions are made in virtual screening, and a positive outcome cannot be guaranteed every time. However, the overall process is extremely cost effective and fast. Virtual screening has the ability to produce leads that otherwise may not have been identified. Hence, virtual screening is emerging as a major component of pharmacoinformatics.

1.5.4 Neuroinformatics

Neuroinformatics may be defined as the organization and analysis of neuroscientific data using the tools of information technology. The information sources in neuroinformatics include behavioral sciences (psychological description) and medicinal (including drugs and diagnostic images) and biological (membranes, neurons, synapses, genes, etc.) aspects. The aim of neuroinformatics is to unravel the complex structure–function relationship of the brain in an integrative effort. Neuroscientists work at multiple levels and are producing enormous amounts of data. Distributed databases are being prepared and novel analytical tools are being generated with the help of information technology. Producing digital capabilities for web-based information management systems is one of the major objectives of neuroinformatics. Apart from data sharing, computational modeling of ion channels, neurons and neural networks, second messenger pathways, morphological features, and biochemical reaction are also often included in neuroinformatics. The initial ideas on neuroinformatics can be traced to the work of Hodgkin and Huxley, who initiated computational neuronal modeling. Current efforts in the direction include studies related to modeling the neuropsychological tests, neuroimaging, computational neuroscience, brain mapping, molecular neuroimaging, and magnetic resonance imaging.

Important Databases in Neuroinformatics A probabilistic atlas and reference system for human brain is being developed as a neuroinformatics and neuroscience tool. Such a system is required because of the vast variance observed in the structure, function, and organization of human brain. It is a data source of digital images of human brain along with information on racial and ethnic conditions, education, handedness, personal traits, habits, and so on. It allows one to examine the relationship and distribution of the macro- and microscopic structure and function of human brain [100, 101]. Surface management system (SuMS) is another database that mainly deals with studies of the structure and function of cerebral cortex. All the data generated during reconstruction of cerebral cortex and subsequent flattening procedures is included in this database. SuMS (1) provides a systematic framework

for classification and storage, (2) serves as a version control system for the surface and volume datasets, (3) is an efficient data retrieval module, and (4) acts as a service request broker [102, 103]. Similarly, there are many other database systems with data analysis tools available in this field, a few of which are listed in Table 1.5.

TABLE 1.5 Selected List of Important Neuroinformatics Tools and Databases

Databases	Brief Description	URL
Brain Architecture Management System (BAMS)	Repository of brain structure information; contains to date around 40,000 connections	http://brancusi.usc.edu/bkms/
BrainMap	For meta-analysis of human functional brain-mapping literature	http://brainmap.org/
BrainInfo	Information about the brain and its functions	http://braininfo.rprc.washington.edu/
Brede Database	Neuroimaging data	http://hendrix.imm.dtu.dk/services/jerne/brede/
Surface Management System (SuMS)	A surface-based database to aid cortical surface reconstruction, visualization and analysis	http://sumsdb.wustl.edu/sums/index.jsp
fMRIDC	Functional neuroimaging (fMRI) data (fMRI Data Center)	http://www.fmridc.org
LGICdb	Ligand Gated Ion Channel database	http://www.ebi.ac.uk/compneursrv/LGICdb/
ModelDB	Neuronal and Network Models	http://senselab.med.yale.edu/senselab/ModelDB
CoCoMac	Collation of Connectivity data on the Macaque brain	http://cocomac.org/home.htm
L-Neuron	Computational Neuroanatomy Database	http://www.krasnow.gmu.edu/LNeuron
NeuroScholar	MySQL Database frontend with management of bibliography, histological and tracing data	http://www.neuroscholar.org
Catacomb	Components and Tools for Accessible Computer Modeling in Biology (Modeling Software for Neuroscience)	http://www.compneuro.org/catacomb/index.shtml
GENESIS	Neural Simulator	http://www.genessim.org/GENESIS/
Channel Lab	Single channel modeling program	http://www.synaptosoft.com/Channelab/index.html
NEURON	Simulation of individual neurons and networks of neurons	http://www.neuron.yale.edu
HHsim	Graphical Hodgkin–Huxley Simulator	http://www.cs.cmu.edu/~dst/HHsim/
NEOSIM	Neural Open Simulation—for modeling of networks	http://www.neosim.org/
NANS	Neuron and Network Simulator	http://vlsi.eecs.harvard.edu/research/nans.html
SNNAP	Simulator for Neural Networks and Action Potentials	http://snnap.uth.tmc.edu/

Several software tools have been made available over the past few years in the field of neuroinformatics. Neuroscholar [104] allows scientists to interact with information in the literature in a modular fashion. This tool permits the user to isolate fragments of data from a source and then bring them together to prepare a fact base for analysis and interpretation in a knowledge base environment. GENESIS (General Neural Simulation System) is a tool that helps in the simulation of neurosystems including subcellular components, biochemical reactions, single neurons, large neural networks, and system level models.

The Human Brain Project is one of the major initiatives under neuroinformatics. This is a broad based effort by neuroscientists and information scientists whose objective is to produce interoperable databases and analysis tools. This project included several tools for modeling simulation, information retrieval from multidisciplinary data, graphical interfaces, and integration of data analysis tools through electronic collaboration [105].

1.5.5 Immunoinformatics

Immunoinformatics is another major area in biomedical research where computational and informational technologies are playing a major role in the development of drugs and vaccines. This field is still in its infancy and it covers both modeling and informatics of the immune system and is the application of informatics technology to the study of immunological macromolecules, addressing important questions in immunobiology and vaccinology. Data sources for immunoinformatics include experimental approaches and theoretical models, both demanding validation at every stage. Major immunological developments include immunological databases, sequence analysis, structure modeling, modeling of the immune system, simulation of laboratory experiments, statistical support for immunological experimentation, and immunogenomics [106, 107]

There are many databases of relevance to immunologists, some of which are given in Table 1.6. IMGT, the international ImMunoGeneTics information system created in 1989, is one of the important ventures in immunoinformatics and is a knowledge resource comprising databases, tools, resources on immunoglobulins, T cell receptors, major histocompatibility complex (MHC), and related proteins of the immune system. IMGT includes sequence and genome databases with different interfaces, 3D structure database, web resources, and interactive tools for sequence and genome analysis. IMGT ONTOLOGY, which is a semantic specification of the terms to be used in immunogenetics and immunoinformatics, is available for IMGT users in the IMGT Scientific chart formalized in IMGT-ML (XML) schema [108–114]. The HIV Molecular Immunology Database is a database containing sequence and epitope maps of HIV-1 cytotoxic and helper T-cell epitopes.

Computer-aided vaccine design (CAVD) or computational vaccinology is another application of immunoinformatics involving prediction of immunogenicity. Immune interactions can be modeled using artificial neural networks (ANNs), hidden Markov models, molecular modeling, binding motifs, and quantitative matrices, of which ANN models have proved to be superior for the prediction of MHC-binding peptides [115–117]. Table 1.6 also gives a list of some of the tools for predicting whether or not a peptide would bind to a major histocompatibility complex. In addition to immunoinformatics, theoretical immunology is another related discipline and is the

TABLE 1.6 Some Selected Immunoinformatics Databases and Tools

Databases and Tools	Brief Description	URL
IMGT, the international ImMunoGeneTics information system	A sequence, genome, and structure database for immunogenetics data	http://imgt.cines.fr
HIV Molecular Immunology Database	A database of HIV-specific B-cell and T-cell responses	http://www.hiv.lanl.gov/content/immunology/index.html
MHCBN	Comprehensive database of MHC-binding, nonbinding peptides and T-cell epitopes	http://bioinformatics.uams.edu/mirror/mhcbn/
MHCPEP	Database of MHC-binding peptides	http://wehih.wehi.edu.au/mhcpep/
ADABase (Mutation Registry for Adenosine Deaminase Deficiency)	Contains information on diseases and mutations associated with adenosine deaminase	http://bioinf.uta.fi/ADABase/
BCIPep	Database of B-cell epitopes	http://bioinformatics.uams.edu/mirror/bcipep/
FIMM	Database of Functional Immunology	http://research.i2r.a-star.edu.sg/fimm/
IPD (Immuno Polymorphism Database)	Database for the study of polymorphism in genes of the immune system	http://www.ebi.ac.uk/ipd/
DHR (The Database of Hypersensitive Response)	Definition, source, and sequence of hypersensitive response (HR) proteins, etc.	http://sdbi.sdut.edu.cn/hrp/
VBASE2	Database of variable genes from the immunoglobulin loci	http://www.vbase2.org/
BIMAS	Bioinformatics and Molecular Analysis Section (MHC peptide-binding prediction)	http://bimas.dcrt.nih.gov/molbio/hla_bind/
SYFPEITHI	Database and prediction server of MHC ligands	http://www.syfpeithi.de/
ProPred	MHC Class I and II Binding Peptide Server	http://www.imtech.res.in/raghava/propred/ http://www.imtech.res.in/raghava/propred1/
nHLAPred	A neural network-based MHC Class I Binding Peptide Prediction Server	http://bioinformatics.uams.edu/mirror/nhlapred/
NetMHC	Peptide-binding prediction using artificial neural networks (ANNs) and weight matrices	http://www.cbs.dtu.dk/services/NetMHC/
MHCPred	Predict binding affinity for MHC I and II molecules	http://www.jenner.ac.uk/MHCPred/

application of mathematical modeling to diverse aspects of immunology ranging from T-cell selection in the thymus to the epidemiology of vaccination. The results from immunoinformatics are being heavily employed in defining another emerging science called Artificial Immune Systems (immunological computation or immunocomputing). The AIS computation is also of interest in modeling the immune system and solving immunological problems [118].

1.5.6 Drug Metabolism Informatics

An understanding of the pharmacokinetics of a drug can play a major role in reducing the probability of bringing a new chemical entity (NCE) with inappropriate ADME/Toxicity profile to the market. Drug metabolism and toxicity in the human body are primarily assessed during clinical trials, and preclinical assessment of the same involves study on *in vivo* and *in vitro* systems. *In silico* models for predicting pharmacokinetic properties based on the experimental results can greatly reduce the cost and time required for the experiments. These methods range from modeling approaches such as QSARs, to similarity searches as well as informatics methods like ligand–protein docking and pharmacophore modeling. Several ADME properties can be explained by simple molecular descriptors derived from the 2D chemical structure and can be used for the development of QSAR models. Such *in silico* prediction methods help chemists in judging whether or not a potential candidate may continue in the drug discovery pipeline. Metabolic biotransformation of any NCE may profoundly affect the bioavailability, activity, distribution, toxicity, and elimination of a compound; the effects of probable metabolism are now considered in the early stages of drug discovery with the help of computer-aided methods.

In silico prediction of metabolic biotransformation occurring at the liver cytochrome enzymes (CYP450 enzymes) are being studied [119]. Many databases and software systems are available in this field for the early prediction of substrates of CYP450 enzymes. Some of the databases and predictive systems for metabolic information of drugs are given in Table 1.7. The Human Drug Metabolism Database (hDMdb) project is a nonprofit, internet database of xenobiotic metabolic transformations that are observed in humans [120]. Other databases like MDL Metabolite contain xenobiotic transformation information that can also be linked to a toxicity database like MDL Toxicity database. Thus, it can give toxicity, if present, in the metabolite shown by the database. The predictive systems available for metabolism are mainly expert systems based on experimental data representing the metabolic effects (database) and/or rules derived from such data (rule-base). The rules may either be induced rules, which are quantitative, derived from a statistical analysis of the metabolic data, or knowledge-based rules derived from expert judgment. Some of the expert systems are MetabolExpert, METEOR, and META, as given in Table 1.7. These can also be linked to the corresponding toxicity prediction modules. METEOR covers both phase I and phase II biotransformation reactions and can analyze mass spectrometry data from metabolism studies. It can be linked to the DEREK software for toxicity prediction of the metabolites. Likewise, MetabolExpert can be linked to HazardExpert and META can be interfaced with MULTICASE, both of which are toxicity prediction modules. MetabolExpert is an open knowledge base, where the user can add his/her own rules. The META program operates from dictionaries of transformation operators, created by experts to rep-

TABLE 1.7 Databases and Tools for Metabolism Informatics

Databases and Tools	Brief Description	URL
Human Drug Metabolism Database (hDMdb)	IUPAC project for a web-based model database for human drug metabolism information	http://www.iupac.org/projects/2000/2000-010-1-700.html
MDL Metabolite	Comprised of a database, registration system, and browsing interface	http://www.mdl.com/products/predictive/metabolite/index.jsp
Accelrys' Metabolism Database	Biotransformations of organic molecules in a variety of species	http://www.accelrys.com/products/chem_databases/databases/metabolism.html
Biofrontier/P450	Human cytochrome P450 information and predictive system	http://www.fqs.pl/
Metabolism and Transport Drug Interaction database	Database developed by University of Washington	http://www.druginteractioninfo.org/
MetabolExpert (CompuDrug, Inc.)	Predictive system for metabolic fate of a drug	http://www.compudrug.com/
MEXAlert (CompuDrug, Inc.)	Rule-based prediction for first-pass metabolism	http://www.compudrug.com/
META (Multicase, Inc.)	Uses dictionaries to create metabolic paths of query molecules	http://www.multicase.com/products/prod05.htm
METEOR (LHASA Ltd., Leeds, UK)	Predictions presented as metabolic trees	http://www.lhasalimited.org/

resent known metabolic paths, and is capable of predicting the sites of potential enzymatic attack and the metabolites formed [121].

1.5.7 Toxicoinformatics

Early prediction of toxicological parameters of new chemical entities (NCEs) is an important requirement in the drug discovery strategy today. This is being emphasized in the wake of many drug withdrawals in the recent past. Computational methods for predicting toxicophoric features is a cost effective approach toward saving experimental efforts and saving animal life. Current efforts in toxicoinformatics are mainly based on QSTR (quantitative structure–toxicity relationships) and rule-based mechanistic methods [122–125]. QSTR is a statistical approach, in which a correlation is developed between structural descriptors of a series of compounds and their toxicological data. In this approach, a model can be trained with the help of a set of known data, validated using many approaches, and then used for the prediction of toxicological parameters. The only limitation of this approach is that the predictive power of these models gets reduced when chemicals belonging to a class outside the series of molecules is used for the construction of the model. Toxicity prediction tools using this approach include TOPKAT and CASE/M-CASE.

TOPKAT mainly employs electrotopological descriptors based on graph theory for the development of QSTR models. TOPKAT uses linear free-energy relationships in statistical regression analysis of a series of compounds. In this software, the continuous/dichotomous toxicity end points are correlated to the structural features like electronic topological descriptors, shape descriptors, and substructure descriptors. CASE (Computer Automated Structure Evaluation) and M-CASE are toxicoinformatics software packages that have the capability to automatically generate predictive models. A hybrid QSTR artificial expert system-based methodology is adopted in CASE-based systems. A linear scale called "CASE units" is defined, which segregate the given set of molecules as active/inactive/marginally toxic species. Molecular fragments are classified in terms of biophores (fragments associated with activity) and biophobes (fragments associated with inactivity). One advantage of these packages is that they also include experimental data that is not released by the FDA, which increases the applicability of this set of packages.

Toxicity prediction tools based on "mechanistic approaches" are knowledge based-systems, where a fact base and rule base can be effectively analyzed to give qualitative information regarding the toxicity of chemical species. The expert rules included in the knowledge base are generally derived from the molecular mechanism of the drug action and hence they are known as "mechanistic approaches." Well known software packages in this category are DEREK (Deductive Estimation of Risk from Existing Knowledge), HazardExpert, and Oncologic. DEREK is a widely used toxicity prediction system. This program not only predicts the potential toxicity of a query chemical but also provides details of the logical process that leads to the predicted results. Table 1.8 gives a list of known resources in toxicoinformatics.

1.5.8 Cancer informatics

Application of information technology has been extended to specific subtopics of pharmaceutical sciences like cancer, diabetes, and AIDS. One important example, wherein information technology is being extensively used, is cancer informatics. The necessity of such focused subtopics of pharmacoinformatics was required because of the overflow of information and lack of integrated data formats. Major cancer informatics initiatives are being undertaken by the National Cancer Institute Center for Bioinformatics (NCICB), National Institutes of Health (United States), National Cancer Research Institute (NCRI) (United Kingdom), and the National Cancer Center (NCC) (Japan).

A web-based environment called CaCORE (Cancer Common Ontologic Reference Environment) was established, which helps in the management, redistribution, integration, and analysis of data arising from studies involving cell and molecular biology, genomics, histopathology, drug development, and clinical trials. The structural and functional components of CaCORE include (1) Enterprise Vocabulary Services (EVS), which provide terminology development, dictionary, and thesaurus services like the description-logics based NCI Thesaurus and the NCI Metathesaurus, which is a collection of biomedical vocabularies based on National Library of Medicine (NLM) and the Unified Medical Language System (UMLS); (2) The Cancer Data Standards Repository (CaDSR), which is a metadata registry for

TABLE 1.8 Toxicoinformatics Tools

Tools	Description	URL	Predicted Endpoints
TOPKAT	QSAR	Accelrys http://www.accelrys.com/products/topkat	Mutagenicity, carcinogenicity, mammalian acute and chronic toxicities, developmental toxicities
M-CASE, CASE, ToxAlert, Casetox	Hybrid QSAR and expert system	Multicase Inc. http://www.multicase.com	Carcinogenicity, mutagenicity, teratogenicity, mammalian acute and chronic toxicities
DEREK	Knowledge based, structural rules	Lhasa Limited http://www.lhasalimited.org/	Mutagenicity, carcinogenicity, teratogenicity/developmental toxicity, skin sensitization, acute toxicity, etc.
OncoLogic	Knowledge based	www.oncologic.net	Carcinogenicity
HazardExpert	Knowledge based	CompuDrug http://www.compudrug.com	Carcinogenicity, mutagenicity, teratogenicity, membrane irritation, neurotoxicity
COMPACT	QSAR; mechanistic supported by molecular modeling studies	Surrey	Cytochrome P450 metabolism

common data elements that have been identified to simplify and standardize the data collection requirements and eligibility/exclusion criteria; and (3) the Cancer Bioinformatics Infrastructure Objects (CaBIO) module, which provides the data interface architecture [126–128].

Apart from the web-based facilities, several machine learning techniques like artificial neural networks and decision trees are also being utilized for cancer detection and diagnosis and more recently for prediction and prognosis. Several bioinformatics resources that are helpful in gene function prediction, in protein structure and function predictions, and in studies of protein–protein interactions are being employed in cancer informatics as well [129]. Several specific databases like PDQ, CGED, and FaCD (Table 1.9) are being heavily used by cancer informatics specialists. The NCI chemical databank is a source of all the chemicals tested for anticancer effects.

TABLE 1.9 Cancer Informatics Resources

Databases	Brief Description	URL
PDQ (Physician Data Query)	NCI's Comprehensive Cancer Database	http://www.cancer.gov/cancertopics/pdq/cancerdatabase
CGED (Cancer Gene Expression Database)	Database of gene expression profile and accompanying clinical information	http://cged.hgc.jp/
Mouse Retroviral Tagged Cancer Gene Database	Retroviral and transposon insertional mutagenesis in mouse tumors	http://rtcgd.ncifcrf.gov
The Familial Cancer Database (FaCD)	Assists in the differential diagnosis in familial cancer	http://facd.uicc.org/
International Agency for Research on Cancer (IARC) p53 Mutation database	Information on TP53 gene mutations	http://www.iarc.fr/p53/
The Tumor Gene Database	Information on genes that are targets for cancer-causing mutations	http://condor.bcm.tmc.edu/oncogene.html
SNP500Cancer Database	Central Resource for Sequence verification of SNPs	http://snp500cancer.nci.nih.gov

1.6 FUTURE SCOPE

The field of computer-aided drug development has undergone a paradigm shift in the past five years. Earlier, this subject was dominated by chemistry and physics of drug discovery, mainly through molecular modeling, QSAR. A dramatic increase has been noted in these two topics—several practical solutions are being offered and novel concepts are being introduced. At the same time an additional component is emerging in CADD through informatics. The current status of CADD includes both modeling and informatics with equal and synergistic contributions. Although a lot has been done in this area over the past twenty years, there is still a lot of scope left for future growth. First and foremost is winning the confidence of the experimental colleague, who is still skeptical about the value of these efforts. Second is the proper integration of modeling efforts and informatics efforts. Although virtual screening has proved to be highly successful, the methodologies being adopted lack common elements. Often the methods adopted in CADD are considered as technological components, although several fundamentals exist. In fact, the fundamentals of the field of drug discovery are only emerging. CADD methods should strongly contribute in establishing these fundamentals; efforts should be concentrated in this direction.

REFERENCES

1. <http://www.cambridgesoft.com>.
2. <http://www.acdlabs.com/download/chemsk.html>.

3. <http://www.chemsw.com>.
4. Goodman JM. *Chemical Applications of Molecular Modeling*. Cambridge, UK: Royal Society of Chemistry; 1998.
5. Holtje HD, Sippl W, Rognan D. *Molecular Modeling Basic Principles and Applications*, 2nd ed. Weinheim: Wiley-VCH Verlag GmbH; 2003.
6. Leach AR, Gillet VJ. *An Introduction to Chemoinformatics*. Dordrecht: Kluwer Academics; 2003.
7. Levine IN. *Quantum Chemistry*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall; 1991.
8. Szabo A, Ostlund S. *Modern Quantum Chemistry*. New York: MacMillan; 1982.
9. Foresman JB, Frisch AE. *Exploring Chemistry with Electronic Structure Methods*, 2nd ed. Pittsburgh: Gaussian Inc.; 1995.
10. Gasteiger J, Engel T. *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH Verlag GmbH; 2003.
11. Howard AE, Kollman PA. An analysis of current methodologies for conformational searching of complex molecules. *J Med Chem* 1998;31:1669–1675.
12. Lipton M, Still WC. The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. *J Comput Chem* 1988;9:343–355.
13. Dammkoehler RA, Karasek SF, Shands EF, Marshall GR. Constrained search of conformational hyperspace. *J Comput Aided Mol Des* 1989;3:3–21.
14. Saunders M. Stochastic exploration of molecular mechanics energy surfaces. Hunting for the global minimum. *J Am Chem Soc* 1987;109:3150–3152.
15. Saunders M. Stochastic search for the conformations of bicyclic hydrocarbons. *J Comput Chem* 1989;10:203–208.
16. Ferguson DM, Raber DJ. A new approach to probing conformational space with molecular mechanics: random incremental pulse search. *J Am Chem Soc* 1989;111:4371–4378.
17. Chang G, Guida WC, Still WC. An internal-coordinate Monte Carlo method for searching conformational space. *J Am Chem Soc* 1989;111:4379–4386.
18. Saunders M, Houk KN, Wu YD, Still WC, Lipton M, Chang G, Guida, WC. Conformations of cycloheptadecane. A comparison of methods for conformational searching. *J Am Chem Soc* 1990;112:1419–1427.
19. Lybrand TP. Computer simulation of biomolecular systems using molecular dynamics and free energy perturbation methods, in *Reviews in Computational Chemistry*. Hoboken, NJ: Wiley-VCH; 1990, pp 295–320.
20. Bourne PE, Weissig H. *Structural Bioinformatics (Methods of Biochemical Analysis)*. Hoboken, NJ: Wiley-Liss; 2003.
21. Clote P, Backofen R. *Computational Molecular Biology: An Introduction*. Hoboken, NJ: Wiley; 2000.
22. Dunbrack RL Jr, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol* 1994;1:334–340.
23. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol* 1995;249:493–507.
24. Sternberg MJE. *Protein Structure Prediction: A Practical Approach (The Practical Approach Series)*. New York: Oxford University Press; 1997.
25. Randić M. Characterization of molecular branching. *J Am Chem Soc* 1975;97:6609–6615.
26. Kier LB, Hall LH. *Molecular Connectivity in Chemistry and Drug Research*. New York: Academic Press; 1976.

27. Kier LB, Hall LH. *Molecular Connectivity in Structure–Activity Analysis*. Hoboken, NJ: Wiley; 1986.
28. Müller WR, Szymanski K, Knop JV, Trinajstić N. An algorithm for construction of the molecular distance matrix. *J Comput Chem* 1987;8:170–173.
29. Wiener H. Structural determination of paraffin boiling points. *J Am Chem Soc* 1947;69:17–20.
30. Hosoya H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull Chem Soc Jpn* 1971;44:2332–2339.
31. Bonchev D. *Information Theoretic Indices for Characterization of Chemical Structures*. Hoboken, NJ: Wiley; 1983.
32. Balaban AT. Highly discriminating distance-based topological index. *Chem Phys Lett* 1982;89:399–404.
33. Kier LB. A shape index from molecular graphs. *Quant Struct-Activity Relat* 1985;4:109–116.
34. Kier LB, Hall LH, Frazer JW. An index of electrotopological state for atoms in molecules. *J Math Chem* 1991;7:229–241.
35. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH Verlag GmbH; 2000.
36. Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev* 1996;96:1027–1044.
37. Abraham DJ. *Burger's Medicinal Chemistry & Drug Discovery*, 6th ed. Hoboken, NJ: Wiley-Interscience; 2003.
38. Hansch C, Leo A, Heller SR. *Exploring QSAR: Volume 1: Fundamentals and Applications in Chemistry and Biology*. Washington DC: American Chemical Society; 1995.
39. Kubinyi H. QSAR: Hansch analysis and related approaches. In *Methods and Principles in Medicinal Chemistry*. Weinheim: VCH; 1993.
40. Kubinyi H. *QSAR In Drug Design. Theory, Methods and Applications*. Leiden: ESCOM; 1993.
41. Cramer RDI, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 1988;110:5959–5967.
42. Klebe G, Abraham U, Meitzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* 1994;37:4130–4146.
43. Robinson DD, Winn PJ, Lyne PD, Richards WG. Self-organizing molecular field analysis: a tool for structure–activity studies. *J Med Chem* 1999;42:573–583.
44. Silverman BD, Platt DE. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J Med Chem* 1996;39:2129–2140.
45. Kubinyi H. Comparative molecular field analysis (CoMFA), in *The Encyclopedia of Computational Chemistry*. Chichester: Wiley; 1998, pp 448–460.
46. Wise M, Cramer RD, Smith D, Exman I. Progress in three-dimensional drug design: the use of real-time colour graphics and computer postulation of bioactive molecules in DYLOMMS. In *Quantitative Approaches to Drug Design*. Amsterdam: Elsevier; 1983, pp 145–146.
47. Pastor M, Cruciani G, Watson KAA. Strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure–activity relationship analysis. *J Med Chem* 1997;40:4089–4102.

48. Polanski J, Walczak B. The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput Chem* 2000;24:615–625.
49. Ivanciuc O, Ivanciuc T, Cabrol-Bass D. Comparative receptor surface analysis (CoRSA) model for calcium channel antagonists. *SAR QSAR Environ Res* 2002;12:93–111.
50. Hariprasad V, Kulkarni VM. A proposed common spatial pharmacophore and the corresponding active conformations of some peptide leukotriene receptor antagonists. *J Comput Aided Mol Des* 1996;10:284–292.
51. Vedani A, Briem H, Dobler M, Dollinger H, McMasters DR. Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J Med Chem* 2000;43:4416–4427.
52. Vedani A, Dobler M, Dollinger H, Hasselbach KM, Birke F, Lill MA. Novel ligands for the chemokine receptor-3 (CCR3): a receptor-modeling study based on 5D-QSAR. *J Med Chem* 2005;48:1515–1527.
53. Vedani A, Dobler M. 5D-QSAR: the key for simulating induced fit? *J Med Chem* 2002;45:2139–2149.
54. Vedani A, Dobler M, Lill MA. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J Med Chem* 2005;48:3700–3703.
55. Martin YC, Bures M, Dahaner, E. A fast approach to pharmacophore mapping and its applications to dopaminergic and benzodiazepine agonists. *J Comput Aided Mol Des* 1993;7:83–102.
56. Marriott DP, Dougall IB, Meghani P, Liu Y-J, Flower DR. Lead generation using pharmacophore mapping and three-dimensional database searching: application to muscarinic M(3) receptor antagonists. *J Med Chem* 1999;42:3210–3216.
57. Sutter J, Güner OF, Hoffmann R, Li H, Waldman M. *Pharmacophore, Perception Development and Use In Drug Design*. La Jolla, CA: International University Line, 2000, pp 504–506.
58. SYBYL7.0, Tripos Inc, 1699 South Hanley Rd, St Louis, MO 631444, USA. <http://www.tripos.com>.
59. Gardiner EJ, Artymiuk PJ, Willett P. Clique-detection algorithms for matching three-dimensional molecular structures. *J Mol Graph Model* 1997;15:245–253.
60. Sprague PW. Automated chemical hypothesis generation and database searching with CATALYST. In *Perspectives in Drug Discovery and Design*. Leiden: ESCOM Science; 1995, pp 1–20.
61. CATALYST4.10, Biosyn-MSI, San Diego, CA, USA, 2005. <http://www.accelrys.com>.
62. Bajorath J. Virtual screening in drug discovery: methods, expectations and reality. *Curr Drug Discov* 2002, pp 24–28.
63. Klebe G. *Virtual Screening: An Alternative or Complement to High Throughput Screening?* Berlin: Springer; 2000.
64. Leach AR. *Molecular Modelling: Principles and Applications*, 2nd ed. Boston: Addison Wesley Longman, Harlow; 1996.
65. Cohen N. *Guidebook on Molecular Modeling in Drug Design*. San Drego, CA: Academic Press; 1996.
66. Charifson PS. *Practical Application of Computer-Aided Drug Design*. New York: Marcel Dekker; 1997.
67. Schneider G, Bohm H-J. Virtual screening and fast automated docking methods. *Drug Discov Today* 2002;7:64–70.
68. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking methods using an incremental construction algorithm. *J Mol Biol* 1996;261:470–489.

69. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule–ligand interactions. *J Mol Biol* 1982;161:269–288.
70. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock2.4. *J Comput Aided Mol Des* 1996;10:293–304.
71. Moon JB, Howe WJ. Computer design of bioactive molecules: a method for receptor-based *de novo* ligand design. *Proteins* 1991;11:314–328.
72. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985;28:849–857.
73. Lauri G, Bartlett PA. CAVEAT: a program to facilitate the design of organic molecules. *J Comput Aided Mol Des* 1994;8:51–66.
74. Böhm HJ. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* 1992;6:593–606.
75. Böhm HJ. The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J Comput Aided Mol Des* 1992;6:61–78.
76. Böhm HJ. A novel computational tool for automated structure-based drug design. *J Mol Recognit* 1993;6:131–137.
77. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* 1994;8:243–256.
78. Rotstein SH, Murcko MA. GroupBuild: a fragment-based method for *de novo* drug design. *J Med Chem* 1993;36:1700–1710.
79. Gillet V, Johnson AP, Mata P, Sike S, Williams P. SPROUT: a program for structure generation. *J Comput Aided Mol Des* 1993;7:127–153.
80. Gund P, Sigal N. Applying information systems to high-throughput screening and analysis. In *Pharmainformatics: A Trend Guide. Drug Discor Today 1999; (Suppl)*, 25–29.
81. Bharatam PV. Pharmacoinformatics: IT solutions for drug discovery and development, *CRIPS (Current Research and Information on Pharmaceutical Sciences)*, 2003;4:2–5.
82. Hanson CW. *Healthcare Informatics*. New York: McGraw-Hill Professional; 2005.
83. Goodman KW. *Ethics, Computing, and Medicine: Informatics and the Transformation of Health Care*, 1st ed. Cambridge, UK: Cambridge University Press; 1998.
84. Saba VK, McCormick KA. *Essentials of Nursing Informatics*, 4th ed. New York: McGraw-Hill Medical; 2005.
85. Gadre SR, Shirsat RN. *Electrostatics of Atoms and Molecules*. Hyderabad: Universities Press (India) Limited; 2000.
86. Oprea TI. Chemoinformatics in drug discovery. In *Methods and Principles in Medicinal Chemistry*, Vol 23. Weinheim: Wiley-VCH Verlag GmbH; 2005.
87. Gasteiger J. Handbook on chemoinformatics: from data to knowledge, In *Representation of Molecular Structures*. Hoboken, NJ: Wiley; 2003.
88. Bajorath J. Chemoinformatics: concepts, methods, and tools for drug discovery. In *Methods in Molecular Biology*. Totowa, NJ: Humana Press; 2004.
89. Mount DW. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2001.
90. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002;30:303–305.
91. Breitkreutz BJ, Stark C, Tyers M. The GRID: the General Repository for Interaction Datasets. *Genome Biol* 2003;4:R23.

92. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003;31:248–250.
93. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, Rashmi BP, Shanker K, Padma N, Niranjana V, Harsha HC, Talreja N, Vrushabendra BM, Ramya MA, Yatish AJ, Joy M, Shivashankar HN, Kavitha MP, Menezes M, Choudhury DR, Ghosh N, Saravana R, Chandran S, Mohan S, Jonnalagadda CK, Prasad CK, Kumar-Sinha C, Deshpande KS, Pandey A. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* 2004;32:D497–501.
94. Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein–protein interactions. *Bioinformatics* 2001;17:359–363.
95. Wilcox A, Hripcsak G. Classification algorithms applied to narrative reports. *Proc AMIA Symp* 1999;455–459.
96. *STRING*, <http://string.embl.de/>.
97. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;31:258–261.
98. Tien AC, Lin MH, Su LJ, Hong YR, Cheng TS, Lee YC, Lin WJ, Still IH, Huang CY. Identification of the substrates and interaction proteins of aurora kinases from a protein–protein interaction model. *Mol Cell Proteomics* 2004;3:93–104.
99. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 1997;23:3–25.
100. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Le Goualher G, Boomsma D, Cannon T, Kawashima R, Mazoyer B. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos Trans R Soc Lond B Biol Sci* 2001;356:1293–1322.
101. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B, Holmes C, Collins L, Thompson P, MacDonald D, Iacoboni M, Schormann T, Amunts K, Palomero-Gallagher N, Geyer S, Parsons L, Narr K, Kabani N, Le Goualher G, Feidler J, Smith K, Boomsma D, Hulshoff Pol H, Cannon T, Kawashima R, Mazoyer B. A four-dimensional probabilistic atlas of the human brain. *J Am Med Inform Assoc* 2001;8:401–430.
102. Dickson J, Drury H, Van Essen DC. The surface management system (SuMS) database: a surface-based database to aid cortical surface reconstruction, visualization and analysis. *Philos Trans R Soc Lond B Biol Sci* 2001;356:1277–1292.
103. Van Essen DC. Windows on the brain: the emerging role of atlases and databases in neuroscience. *Curr Opin Neurobiol* 2002;12:574–579.
104. Burns GA. Knowledge management of the neuroscientific literature: the data model and underlying strategy of the NeuroScholar system. *Philos Trans R Soc Lond B Biol Sci* 2001;356:1187–1208.
105. Shepherd GM, Mirsky JS, Healy MD, Singer MS, Skoufos E, Hines MS, Nadkarni PM, Miller PL. The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci* 1998;21:460–468.
106. Rammensee HG. Immunoinformatics: bioinformatic strategies for better understanding of immune function. Introduction. *Novartis Found Symp* 2003;254:1–2.
107. Brusci V, Zeleznikov J, Petrovsky N. Molecular immunology databases and data repositories. *J Immunol Methods* 2000;238:17–28.

108. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SG. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res* 2003;31:311–314.
109. Lefranc MP. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* 2003;31:307–310.
110. Giudicelli V, Lefranc MP. Ontology for immunogenetics: the IMGT-ONTOLOGY. *Bioinformatics* 1999;15:1047–1054.
111. Petrovsky N, Schonbach C, Brusica V. Bioinformatic strategies for better understanding of immune function. *In Silico Biol* 2003;3:411–416.
112. Lefranc MP. IMGT, The international ImMunoGeneTics Information System. *Methods Mol Biol* 2004;248:27–49.
113. Lefranc MP, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, Combres K, Girod D, Jeanjean S, Protat C, Yousfi-Monod M, Duprat E, Kaas Q, Pommie C, Chaume D, Lefranc G. IMGT-ONTOLOGY for immunogenetics and immunoinformatics. *In Silico Biol* 2004;4:17–29.
114. Lefranc MP. IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics. *Mol Immunol* 2004;40:647–660.
115. Brusica V, Zeleznikow J. Artificial neural network applications in immunology, in *Proceedings of the International Joint Conference on Neural Networks*, 1999;5:3685–3689.
116. De Groot AS, Bosma A, Chinai N, Frost J, Jesdale BM, Gonzalez MA, Martin W, Saint-Aubin C. From genome to vaccine: *in silico* predictions, *ex vivo* verification. *Vaccine* 2001;19:4385–4395.
117. De Groot AS, Sbai H, Aubin CS, McMurry J, Martin W. Immuno-informatics: mining genomes for vaccine components. *Immunol Cell Biol* 2002;80:255–269.
118. de Castro LN, Timmis JI. Artificial immune systems: a novel paradigm to pattern recognition. In *Artificial Neural Networks in Pattern Recognition*. Paisley, UK: University of Paisley; 2002, pp 67–84.
119. Korolev D, Balakin KV, Nikolsky Y, Kirillov E, Ivanenkov YA, Savchuk NP, Ivashchenko AA, Nikolskaya T. Modeling of human cytochrome P450-mediated drug metabolism using unsupervised machine learning approach. *J Med Chem* 2003;46:3631–3643.
120. Erhardt PW. A human drug metabolism database: potential roles in the quantitative predictions of drug metabolism and metabolism-related drug–drug interactions. *Curr Drug Metab* 2003;4:411–422.
121. Langowski J, Long A. Computer systems for the prediction of xenobiotic metabolism. *Adv Drug Deliv Rev* 2002;54:407–415.
122. Barratt MD, Rodford RA. The computational prediction of toxicity. *Curr Opin Chem Biol* 2001;5:383–388.
123. Pearl GM, Livingston-Carr S, Durham SK. Integration of computational analysis as a sentinel tool in toxicological assessments. *Curr Top Med Chem* 2001;1:247–255.
124. Dearden JC, Barratt MD, Benigni R, Bristol DW, Combes RD, Cronin MTD, Judson PN, Payne MP, Richard AM, Tichy M, Worth AP, Yourick JJ. The development and validation of expert systems for predicting toxicity. *ATLA* 1997;25:223–252.
125. Schultz TW, Cronin MTD, Walker JD, Aptula AO. Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *J Mol Structure (Theochem)* 2003;622:1–22.
126. Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, Gustafson S, Buetow KH. caCORE: a common infrastructure for cancer informatics. *Bioinformatics* 2003;19:2404–2412.

127. Silva JS, Ball MJ, Douglas JV. The Cancer Informatics Infrastructure (CII): an architecture for translating clinical research into patient care. *Medinfo* 2001;10:114–117.
128. Hubbard SM, Setser A. The Cancer Informatics Infrastructure: a new initiative of the National Cancer Institute. *Semin Oncol Nurs* 2001;17:55–61.
129. Kihara D, Yang YD, Hawkin T. Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. *Cancer Informatics* 2006; pp 25–35.

2

COMPUTER TECHNIQUES: IDENTIFYING SIMILARITIES BETWEEN SMALL MOLECULES

PETER MEEK, GUILLERMO MOYNA, AND RANDY ZAUHAR

University of the Sciences in Philadelphia, Philadelphia, Pennsylvania

Contents

- 2.1 Introduction
- 2.2 Computer-Aided Drug Design (CADD)
- 2.3 Harvesting Data from Small Molecules
- 2.4 Representing Molecules for Interpretation by Computers
 - 2.4.1 Importance of Continuity Within Molecular Representations
- 2.5 Defining Similarity
 - 2.5.1 Why Do We Wish to Compare Molecules?
 - 2.5.2 Utilizing External Sources of Information
- 2.6 Detecting Similarity with *in silico* Techniques
 - 2.6.1 HQSAR
 - 2.6.2 QSAR
 - 2.6.3 Superimposition
 - 2.6.4 Program Suites
 - 2.6.5 Comparative Molecular Field Analysis (CoMFA) and Related Approaches
 - 2.6.6 Shape Signatures
- 2.7 Conclusion
 - Acknowledgments
 - References

2.1 INTRODUCTION

In this chapter, current computational methodologies available for the comparison of structurally similar compounds are presented and analyzed in detail. Problems encountered throughout these types of endeavors are summarized first, giving particular emphasis to the suitability of *in silico*, or computer-based, methods for evaluating molecular similarity among members of large compound libraries. The chapter then concentrates on the description of different techniques, starting from the simplest ones, then more complex, and finishing with emerging approaches. Salient examples of a technique application found in the literature are discussed, and brief guidelines for readers interested in employing these computational tools are outlined.

2.2 COMPUTER-AIDED DRUG DESIGN (CADD)

Implementing CADD techniques depends heavily on what molecular information is provided or accessible. However, of equal or even greater importance is how the information that is searched and utilized is organized and constructed. These, of course, are the databases containing the small molecules. The format of how a molecular entry is recorded and stored is pivotal and dictates what transformations of the entries are required before any kind of search or CADD application can be performed. In fact, this chapter might be described as a series of illustrations of the different ways in which molecules can be represented in a computer, each with its potential benefits and drawbacks.

2.3 HARVESTING DATA FROM SMALL MOLECULES

Over the past few decades there has been an unprecedented explosion of information in the chemical and life sciences. Development of new biological and chemical tools has been documented and new pharmaceutical agents are discovered with ever-increasing frequency year by year [1]. At the same time, it has become more and more difficult to gain regulatory approval of new chemical entities [2] due to new and stricter legislation. There are millions of molecules that are now known to possess bioactivity against one or more targets, or to be useful as reagents in academic and industrial research. Before any computer-based technique can be implemented, these vast numbers of isolated and synthesized molecules need to be organized. As might be expected, the need to catalog this abundant information was recognized early on, and many molecules are organized into indexes (e.g., CAS system) and/or are held in commercial [3–6], industrial [7, 8], or public [9–11] databases. Given the vast volume of “chemical space” that has already been explored, it is to the advantage of the investigator to consult these knowledge bases before planning a new synthesis, and at the outset of any efforts to identify new therapeutic or bioactive compounds. In particular, synthesizing a new compound often represents an enormous effort in the laboratory, work carried out in vain if the target compound (or one similar to it) is already available from a commercial source.

Given the sheer volume of information, there is considerable need to employ computers to carry out these investigations as few human minds past, present, or future can compare one molecule to a previously memorized database of 10,000,000 molecules. Harnessing computer power is not without drawbacks; we discuss here the importance and careful considerations required to represent chemical structures. Chemical database storage and searching relies on electronic definitions of chemical structures, and the details of the methodology used are vital to defining and measuring similarity between molecules.

2.4 REPRESENTING MOLECULES FOR INTERPRETATION BY COMPUTERS

Molecules are represented based on their constituent parts and the way in which these constituent parts are connected. The components of a small molecule are atoms, which are minimally identified by element type, but which may be characterized by a number of additional descriptors, including but not limited to van der Waals radius, valence, and/or charge state. The next step beyond simply enumerating the atoms in a molecule is to provide detailed connectivity information, typically in terms of a list of bonds of recognized chemical type; in this way the molecule is defined mathematically as a graph, with the atoms as vertices and the bonds as edges of specified type. The atom and bond definitions act as the building blocks of molecules that can be represented in one of three major forms: one-dimensional (1D), two-dimensional (2D), or three-dimensional (3D). Over the years several standard computer file formats have been developed for each dimension class (Table 2.1), and these have steadily evolved to allow incorporation of additional data types. For instance, many applications require a detailed distribution of partial atomic charges (as opposed to simply specifying ionization state); such charges are required by some force fields, such as MMFF94 [19], and are routinely computed using rapid semiempirical methods (Gasteiger [20] and Gasteiger and Huckel [21]). This information is readily included in many molecular file formats, including the structure data file sdf [22] and mol2 [14] specifications. In addition, some current formats (such as sdf) admit the possibility of “tagging” a molecule with arbitrary data, a boon in the sense of providing open-ended flexibility, but a peril in that descriptors may not be added uniformly across a database, and/or the additions may not be clearly annotated.

2.4.1 Importance of Continuity Within Molecular Representations

One of the most critical issues when applying various molecular databases is to recognize the varying levels of faithfulness between the representation of the molecule in the computer and the actual material available to the investigator. This is a function of both the database and the source of the compound. For example, suppose that a sample of a potentially bioactive compound is available in the form of a purified enantiomer. Many molecular databases store connectivity (bonding) information but do not distinguish between different stereoisomers. If the database entry corresponding to this compound is used to generate a molecular model with

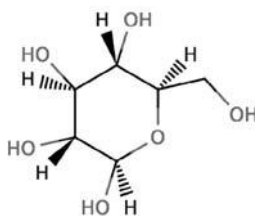
TABLE 2.1 Representation of the Small Molecule β -D-Glucopyranose

1D^a

SMILES string representation of β -D-glucopyranose (@ = S, @@ = R):
[H][C@@]1(O)O[C@]([H])(CO)[C@@]([H])(O)[C@]([H])(O)[C@@]1([H])O
 SLN string representation of β -D-glucopyranose:
O[1]CH(CH(CH(CH@1OH)OH)OH)OH)CH2OH

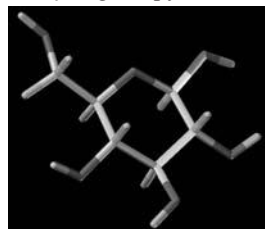
2D^b

MarvinSketch representation of
 β -D-glucopyranose



3D^c

SYBYL representation of
 β -D-glucopyranose



^aTwo examples: a SMILE string notation [12] and the SLN (SYBYL Line Notation [13], by Tripos [14]).

^bRepresentation using the classic bond-wedge diagram (MarvinSketch [15]).

^cThe image of β -D-glucopyranose created from a mol2 file (SYBYL by Tripos). The 2D and 3D structures are collapsible to linear notation and converted into a required file format by CORINA [16]; stereoisomers are generated with STERGEN [17] and tautomers with TAUTOMER [18].

detailed atomic positions, the resulting model may not correspond to the material sample. Conversely, the investigator may have a racemic mixture, whereas the database entry assumes a specific enantiomeric form. In either case, incorrect predictions may result, since biological activity is often remarkably sensitive to the details of 3D molecular structure.

Generally, it is often the case that small changes in chemical structure or physicochemical properties can have profound implications for the specificity and magnitude of the biological effects associated with a compound. Speaking first to the effects of structural variations, we can identify classic examples of chemical isomerism, including chiral, cis/trans (*E/Z*), functional group, and positional isomerism. Small variations such as substitution of atoms and/or isomeric rearrangements can have an important impact on both molecular shape and physical properties (e.g., dipole moment, hydrophobicity) that are a function of 3D structure. Hence, it is imperative for the investigator to be aware of the details in the information stored in a database, particularly when data is absent (e.g., no representation of stereoisomerism) as well as when data is present but may not correspond to the state of the physical sample (e.g., *R* chiral form where *S* is found in the material sample, or a neutral database entry corresponding to a molecule that may be ionized at the pH of interest). It is therefore imperative to use, derive, and develop uniform molecular representation systems meticulously, without which predictions or application of *in silico* techniques would be impossible.

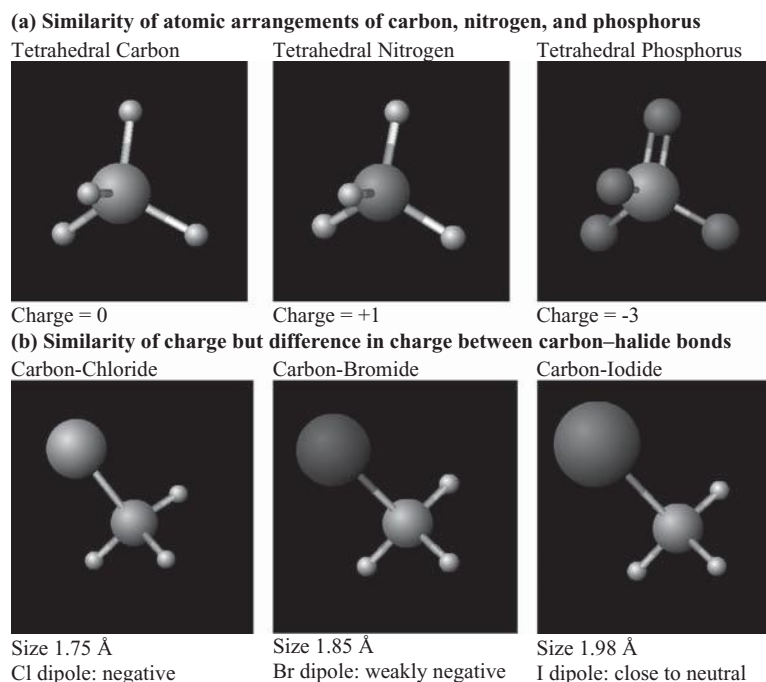


FIGURE 2.1 (a) Depiction of three tetrahedral atom arrangements: (**left**) the central **carbon** in methane, (**center**) the central protonated **nitrogen** in ammonium, and (**right**) the **phosphorus** in phosphate all look similar but carry different charges. (b) Atomic arrangements of a tetrahedral carbon bonded with a single halide of increasing size: (**left**) chloromethane, (**center**) bromomethane, and (**right**) iodomethane. All molecule constructions were created with MarvinSketch [15].

To highlight this concept further, we present some examples of small molecular frameworks that are at first glance similar but that in fact differ significantly (Fig. 2.1). Certain atoms have similar sizes and the same bond orientation (e.g., tetrahedral carbon, phosphorus, and protonated nitrogen, Fig. 2.1a) yet the molecular charge distribution is considerably different. In contrast, substituting one halide for another at a specified location in a framework has a small effect on total dipole moment, and yet these atoms differ significantly in size, with the largest, iodine, occupying about the same volume as a methyl group (Fig. 2.1b). This raises the question: When similarity is measured, what governs how similar an atom is to another atom? In other words, how similar is O to S, or N to P or C, or an atom to a small group (e.g., I to CH₃)? The answer depends heavily on the context of the atom or group in question. Moreover, it is important to recognize that the physical influences of atom or small group in determining molecular properties depend on its context; such factors include hybridization state and the local chemical environment, an insight first proposed by Hammett [23] and Hansch et al. [24]. These researchers developed mathematical representations for predicting molecular properties on the basis of structure.

Yet more complexity arises if we introduce the possibility of molecular flexibility. Most molecules of biological interest admit some degree of flexibility, including bonds about which free rotation is possible and saturated ring systems that can

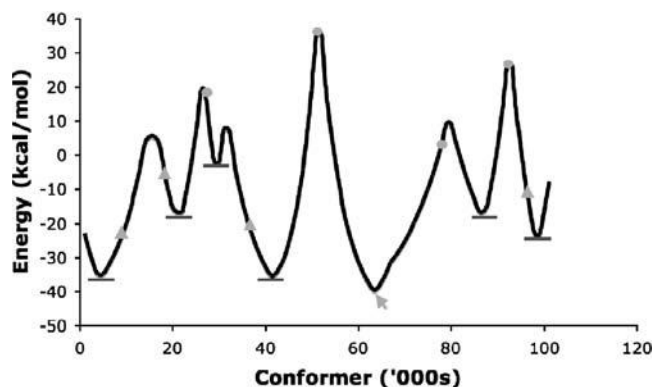


FIGURE 2.2 Graph of molecular conformation versus energy. An energy profile plot is shown for each possible conformation of a hypothetical molecule in 3D space. The troughs denote local energy minima (short gray lines) and the global energy minimum (gray arrow). When exploring molecular space, finding the global energy minimum is not guaranteed. For example, if the four starting points were the gray triangles, the output would not detect the global energy minimum, but maybe five local energy minima at best. Using the start points with the gray circles would almost certainly guarantee the global energy minima and possibly all the local minima too. The problem is inherent to all experimental investigations: if the sample size is not large enough, the minima determined will not be representative of the whole. Likewise, in energy calculations if the energy barrier is too high to climb, the pockets or valleys over the hills cannot be entered into and located. This is why small molecules bound to their targets are highly sought after; a more active molecule will have a bound structure not too dissimilar to its gas phase global minima.

adopt alternate conformations (e.g., chair vs. boat forms for six-membered rings, various classes of puckering observed in pyranoses). This leads to the possibility of multiple conformers for a molecule, each representing a local minimum on the energy surface of the compound. While a molecule will often have a single global energy minimum representing the conformation most favored in terms of energy, there may be other important minima of similar energy (Fig. 2.2), and there is no guarantee that the global minimum corresponds to the form of the molecule that is active against a particular biological target [25]. A number of techniques are available to identify low energy conformers [14, 26]. Although many databases include a representative low energy conformer for each compound, detailed considerations of molecular flexibility are usually relegated to the final stages of analysis, after a relatively small set of interesting molecules have been identified. Nonetheless, consideration of flexibility is often critical for rationalizing differences in activity among a panel of bioactive molecules.

2.5 DEFINING SIMILARITY

Defining similarity between molecules is an inherently ambiguous undertaking. What makes two molecules similar? Should we focus on the volumes occupied by molecules, their patterns of chemical bonding, the presence of particular functional

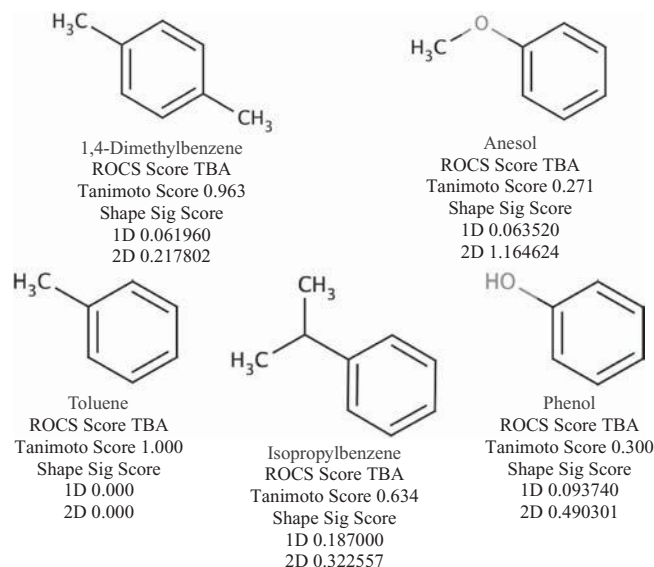


FIGURE 2.3 A simple similarity example using different comparison methods. A collection of five benzene derivatives is displayed and an assessment is made of their similarity to a toluene reference using Tanimoto index [27], 1D Shape Signatures, (shape comparison), and 2D Shape Signatures (shape and mean electrostatic potential comparison) [28–30]. All three types of comparison required 3D coordinate input files (mol2). A Tanimoto score of 1.000 implies a perfect match, with 0.000 implying a complete mismatch; a Shape Signatures score of 0.000 is a perfect match and 2.000 is a complete mismatch.

groups, ionization state, hydrophobicity, dipole moment, or any or all of a host of other topological and physicochemical descriptors? In most cases the desire is to find molecules that “look like” a particular query molecule, or that are compatible with a chemical/spatial pattern (corresponding perhaps to a pharmacophore model). Note that we are not typically interested in stringent criteria; rather, the goal is to cast a fairly wide net and identify molecules that have the potential to interact with a particular target of interest. Examples of comparisons (Fig. 2.3) demonstrate that different tools, each employing its own valid approach to defining molecular similarity, can produce very different results. This is not in itself surprising but does highlight the need for care and caution when applying *in silico* methods to scan chemical databases for lead compounds. It is important to understand what features a particular method is “looking for,” and it is just as critical to recognize what the method ignores! Similarity is itself an arbitrarily defined concept (with appreciable overlaps between techniques) and will depend entirely on the tool(s) used.

Essentially, *in silico* search techniques prove most useful when the similarity measurement (scoring index) reliably returns molecules with the same desired characteristics as the query. Such a technique will take a large dataset of unrelated molecules and reduce it to a smaller set, but with proportionally more molecules similar to the query. This process is called *enrichment*. Methods are available to quantitate enrichment [29, 31], and a measure of enrichment may provide a useful figure of merit when comparing and assessing computational search tools.

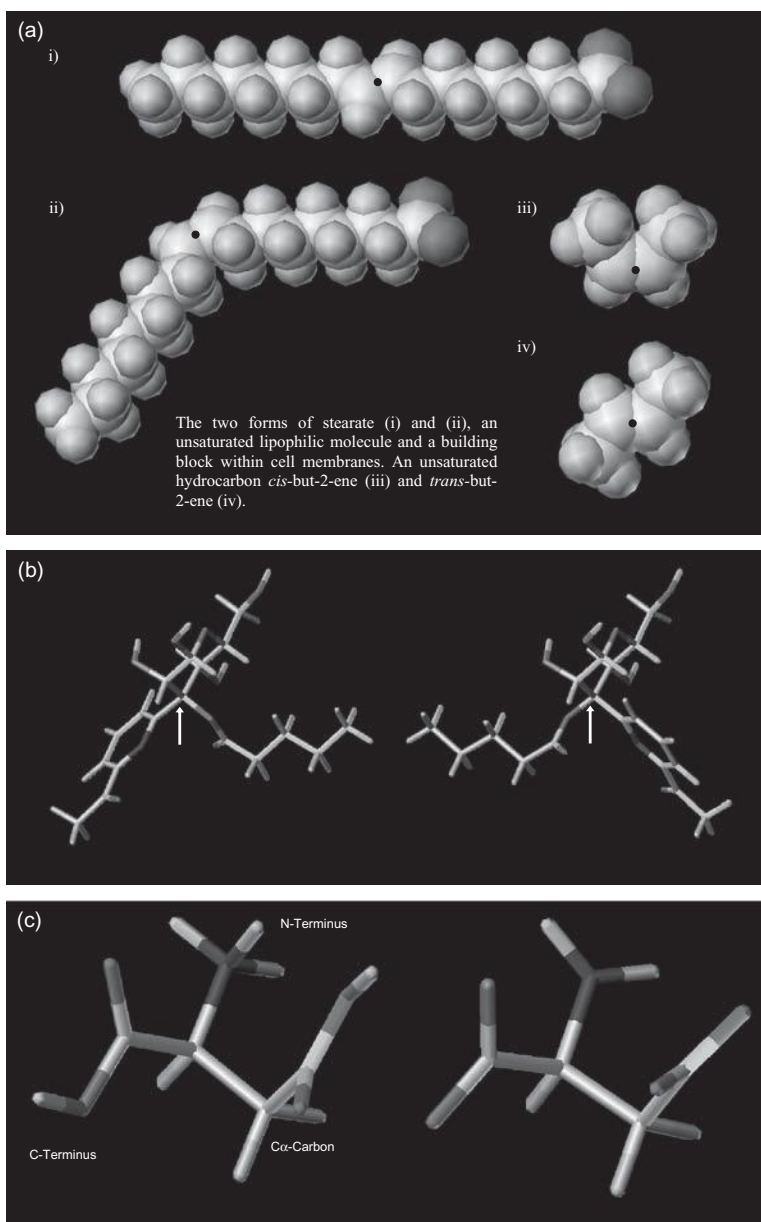
2.5.1 Why Do We Wish to Compare Molecules?

In the context of drug development (or identification of bioactivity in general), there are two motivations for comparing molecules. One is to locate molecules in a database that are similar to a known active, which might also be active against the target, and furthermore could conceivably provide better compounds as determined by any number of measures (e.g., activity, bioavailability, ease of synthesis). The second is to rationalize the efficacy of a set of known active compounds, to identify the features that are directly related to activity, and thus to determine how existing compounds can be improved. Of course, the two motivations can be combined: one might first use a library search to locate interesting lead compounds and then improve the leads by making modifications in line with established models that relate structure to activity. Clearly, at the very least, the database search component of this strategy needs to be automated, since humans do a poor job of scanning through hundreds of structures, let alone the hundreds of thousands that are available in contemporary libraries. But the central point is that drug design methods that employ measures of molecular similarity have the potential to rapidly locate lead compounds, to support the finetuning of structures to improve activity and other critical features, and to dramatically reduce costs by identifying compounds that may be readily available or easily modified.

Similarity searching is not a straightforward task: small alterations of a molecule (e.g., *cis/trans* stereoisomerism (Fig. 2.4a) and chirality (Fig. 2.4b)) can have a profound and marked effect on activity. Indeed, the protonation state of the molecule

FIGURE 2.4 (a) Shape and *cis/trans* isomerism. This part demonstrates how a subtle change in the arrangement of a carbon–carbon double bond ($sp^2=sp^2$) can dramatically affect the appearance of a molecule. The bond (black dot) is either in the *cis* conformation (same side), (i) and (iii), or the *trans* conformation (opposite sides), (ii) and (iv). One should take particular note that as small molecules are usually designed to biological targets, it is inevitable that *cis/trans* isomerism can dramatically affect the activity of a potential drug, and consequently is a major headache in chemical synthesis. (b) Shape and chiral isomerism. Controlling chirality poses a huge challenge for synthetic chemists. There are several chiral centers in this molecule; the only differing one is marked with an arrow, being *S* on the isomer to the left and *R* on the one to the right. Usually one optical isomer, or enantiomer, will have activity against a biological target (known as the eutomer), while the other lacks the desired activity. Due to the expense associated with the separation or enantiomers from a mixture, chirality is of paramount importance when searching molecular databases. (c) Shape and distribution of charge. Shown are two representations of aspartic acid—an amino acid: (left) the fully protonated amino acid and (right) the fully deprotonated amino acid. Usually the nitrogen is tetrahedral at physiological pH and the C terminus and side chain (connected to the C α -carbon) are deprotonated. The state of the proton donors and acceptors modifies the local charge and overall shape of the molecule; hence, the small molecule being investigated needs to mimic the donor and acceptor sites to be truly similar. As a general rule, the termini are not involved in donor and acceptor sites on a protein, as the vast majority are lost in the peptide bond formation. The side chains dictate the number of donor and acceptor sites.

can affect the structure and surface charge distribution (Fig. 2.4c), with flexibility and rotation further complicating the matter. Hence, one could define this problem as akin to a box of left-threaded and right-threaded nuts and bolts. The different classes of bolts may have many properties in common, such that even a trained eye might sort them together, yet they have distinct and incompatible functions. This paradigm is clearly evident in the biological context and highlighted with respect to



enantiomer specificity (HIV serine proteases made of exclusively either D or L amino acids) [32] and the toxic effects observed with thalidomide [33].

Consequently, while attempts to measure similarity between molecules have an objective basis, the evaluations depend heavily on the algorithms employed, sometimes leading to the appearance of subjectivity (the results obtained depend on the specific tools selected). That said, it has been evident for a long while that molecular similarity provides the foundation for rational drug design. As far as the choice of tools and algorithms is concerned, these are of course ultimately evaluated at the laboratory bench, and it is empirical studies that in the end provide the validation of existing approaches, and point out new directions for improvement and development on the computational side.

2.5.2 Utilizing External Sources of Information

As discussed earlier, several scenarios admit the use of similarity searching. In some cases, a relatively small set of molecules are compared, but in others the goal is to scan a potentially enormous library of compounds. To a large extent, the form of the investigation will be determined by the molecular descriptors maintained in the target database. These vary considerably from one database to another; for example, many commonly used compound databases (National Cancer Institute [10]) include only chemical formulas in the form of SMILES strings [12], with no 3D structures. On the other hand, these same libraries may be offered by other vendors with atomic coordinates appended (Tripos version of NCI), and furthermore a number of fast tools are available to generate coordinates from SMILES strings (CORINA [16], CONCORD [34]), so this data can be added by the end user. Some databases include a range of molecular descriptors for each entry, such as $\log P$, molar refractivity, molecular weight, or custom topological descriptors (e.g., chemical fingerprints [35]). Again, these descriptors, when missing, can sometimes be added by third-party tools (MOE [36], DayLight [37]). At the same time, whether these newly added atomic coordinates and/or descriptors can readily be used with available search software is a question that needs to be addressed at the outset; if a tool requires a proprietary database format, it is an open question as to whether any modifications can be introduced by the end user. An extreme case is found when the database is accessible only via a website maintained by a company or academic laboratory, and the raw data is inaccessible except through the supplied interface. The most attractive approach, rarely realized, is to maintain a flexible database format and something akin to a relational database that can be used to construct arbitrary queries of the information held in the library.

Databases for use in similarity searching are difficult to procure or expensive to obtain [3–8]. Recently, several have become available that are utilizable and relevant to small molecule comparison [9–11]. The research conducted at the University of the Sciences in Philadelphia aims to provide a readily searchable database to the global academic community (see Section 2.6.6).

2.6 DETECTING SIMILARITY WITH *IN SILICO* TECHNIQUES

There is a vast expanse of *in silico* techniques that “detect” similarity between small molecules, far too many to cover all in detail here due to space constraints. However,

the aim is to give a flavor of techniques that are in common use and applied in contemporary research. We begin with established techniques that are the least complex and require the least amount of input data and move through to more complicated data intensive methods and finally emerging technologies.

2.6.1 HQSAR

HQSAR (hologram quantitative structure–activity relationships) is a powerful approach that can claim important advantages over many of the traditional QSAR (quantitative structure–activity relationship) techniques we will describe later. This is especially true when little is known about the receptor target of interest. A key advantage of HQSAR is that it requires only 2D chemical structures combined with experimental binding data; there is no need to generate atomic coordinates or alignments of structures, as is required by some 3D methods like CoMFA (comparative molecular field analysis, discussed in Section 2.6.5). At the same time, predictions from HQSAR can be combined with other methods to sometimes improve the overall accuracy of predicted activities. Models of high statistical quality and predictive value can be rapidly and easily generated by HQSAR and can often be used in conjunction with chemical databases to select molecules with superior biological activity, providing the model is robust. HQSAR can be used to rapidly scan chemical libraries but is also useful with small datasets.

HQSAR requires a set of training molecules in order to construct predictive models. As with any machine learning algorithm, the reliability of the model generated is largely a function of the quality of the training set; it is imperative that the training molecules be well characterized, with a common binding site and accurate IC_{50} measurements.

HQSAR Application HQSAR is a rapid and inexpensive means to test compounds using a computer algorithm. The computer “learns” from the input provided from real data derived from actual experiments and generates a model (Fig. 2.5a). This model in turn serves as the basis for predicting the activity of compounds that have been proposed as new leads, or for scanning a chemical library for compounds likely to be active (Fig. 2.5b). While we will describe the HQSAR methodology in more detail below, the basic idea is straightforward. It is usually true that a biologically active compound has three or four groups that are crucial to activity, and these are held at specific distances by a skeleton/backbone structure. HQSAR aims to identify these groups and to correlate the arrangement of these key components with activity.

HQSAR: A Computational Perspective To understand HQSAR, an appreciation of the computational algorithm is required, and for this we need to consider the underlying representation of the data. Linear notation (more specifically SLN) is utilized to represent molecules (Table 2.1), [12, 13, 38], but the molecule is finally represented as a fingerprint [35]. The key importance of HQSAR fingerprint representations (Fig. 2.6) is that fragments are generated based on the chemical structure of each compound and need not correspond to a predefined list of functional groups and structural components. This flexibility leads to more robust database

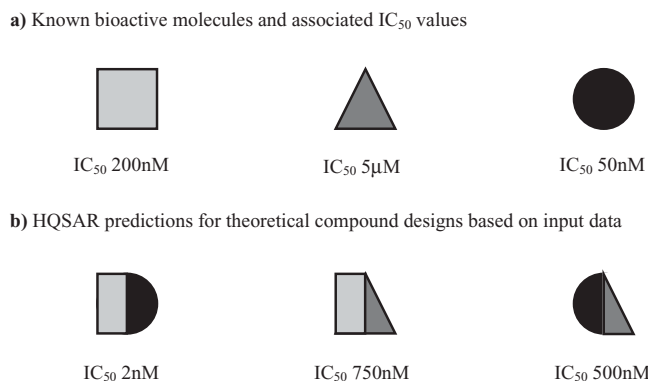


FIGURE 2.5 Simplified representation of HQSAR with a small molecule training and test set. The methodology behind HQSAR consists of (a) the training set of known molecules and (b) the predictions for theoretically designed molecules. The points at the very extremities of the shapes can be considered as active groups, and the overall shape as a molecular skeleton. When an HQSAR run is conducted, the molecules in the training set (a) define whether particular groups contribute to or negate activity in the HQSAR model (in this case experimentally derived IC_{50} values are used). Theoretical compounds incorporating variations in the component parts (b) are then tested with this HQSAR model. In many incidences results make sense (**lower center** and **lower right**) being between the IC_{50} values of the two component parts. There are cases where the IC_{50} can appear to be greater than the sum of the component parts (**lower left**), and in this case it is in a more favorable direction: 2 nM while previously the best molecule was 50 nM (**upper right**).

Fingerprint	0	0	1	1	0	1	0	0	0	1	1	1	1	0	0	0	$\Sigma = 35$
Hologram	0	0	2	5	0	9	0	0	0	1	3	8	7	0	0	0	

FIGURE 2.6 An example of a fingerprint and a hologram. The chemical structure has 35 fragments. In the case of the hashed fingerprint, the representation is purely binary, whereas in the case of the hologram, the bins contain information about the number of fragments hashed into each bin and its contribution (positive, negative, or otherwise to biological activity).

screening [37], and more effective 2D and/or similarity searches, than a structural key approach [39]. The package used [14], however, provides the capabilities of each approach.

Conventional QSARs can identify critical relationships between the properties—the geometric and the chemical characteristics of a molecular system of interest. In QSAR, the measured bioactivity of a set of compounds is correlated with structural descriptors to determine trends and predict the bioactivity of related, untested compounds. HQSAR (like combined QSAR and CoMFA) works by identifying substructural features in sets of molecules that are relevant to biological activity. A key advantage of HQSAR is the relative simplicity of the required input, which in the training phase consists of chemical (2D) structures and their experimental activity (i.e., relative binding affinity, LD_{50} or IC_{50}).

The underlying principle of HQSAR is simply stated: since the structure of the molecule is encoded within its 2D fingerprint and that structure is a key determinant of all molecular properties including biological activity, then it should be possible

to predict the activity of a molecule from its fingerprint, which implicitly encodes the structure. Molecular fingerprints (Fig. 2.6) are strings of 1's and 0's (binary) that indicate the presence or absence of a particular fragment (e.g., a carbonyl, hydroxyl, or halide). The fingerprint is therefore a shorthand list of the fragments present in a structure. A *molecular hologram* extends this concept to incorporate both *branched fragments* and *counts* of the fragments that are present. These additions are important as the inclusion of branched fragments helps distinguish hybridization states, and the counting function differentiates between compounds with (for example) one, two, or more carbonyls. Thus, a hologram is a list of integers, not a bit string, and is therefore a fingerprint. In summary, a molecular hologram contains all possible molecular fragments within a molecule, including overlapping fragments, and maintains a count of the number of times each unique fragment occurs. This process of incorporating information about each fragment, and each of its constituent sub-fragments, implicitly encodes 3D structural information. These fragments are arranged into (or hashed) a linear array of numbers (the hologram, Fig. 2.6). Note that the hashing algorithm guarantees that a particular fragment will always hash to the same numerical value but does not guarantee that different fragments will always have unique hash values. If different fragments, all important for determining biological activity, should hash to the same bin positions in the hologram, it will not be possible to distinguish their effects using a partial least squares analysis (or any other approach for that matter). This phenomenon is called fragment collision and results in poor models with little predictive validity. Using holograms of different lengths (L) can prevent fragments that are responsible for biological activity from being hashed to the same bin.

When the various lengths of holograms are chosen at even intervals, the chances of resolving bad collisions are reduced. For example, if fragment collisions occur with the length set at 400, these collisions will not be resolved at length 200 (nor 100, 50, or 25 as these are all factors of 400). Since the 400 bin hologram is simply folded over to produce the 200 bin hologram, the colliding fragments will still end up in the same bin. For this reason, the default values for the hologram lengths are all prime numbers. This reduces the chances of seeing the same bad collisions at the various lengths. It is wise to use all available lengths so that the best model can be generated in a single HQSAR run.

Despite the fundamental simplicity of HQSAR, there are many options and much terminology connected with the application of the method, only a portion of which is described here. Each molecule in the training set is broken down into structural fragments whose size falls within a specified range; here we will denote the lower limit for the number of atoms in a fragment as M , and the upper limit as N . The parameters M and N are user-defined (typically with $M = 4-7$ and $N = 7-9$), as is the length (L) of the molecular hologram into which the fragments are hashed. Given the length L of the hologram, a fragment with a particular chemical structure always hashes to the same well-defined bin. However, the hashing function can be adjusted to ignore certain molecular features; for example, the user controls whether or not fragments with the same chemical structure but opposite chirality hash into the same or different hologram bins.

The calculation of a set of molecular holograms for a training set of structures yields a data matrix of dimension $n \times L$, where n is the number of compounds in the dataset and L is the length of the molecular hologram. The partial least squares

technique is then employed to generate a statistical model that relates the descriptor variables (occupancy numbers of the bins in the hologram) to an observable property, for example, the biological activity expressed as $-\log IC_{50}$. The selection of the hologram length leading to the “best” HQSAR is based on the PLS analysis that gives either the highest cross-validated q^2 or the lowest standard error associated with the cross-validation analysis. The predictive power of the model is determined by using statistical cross-validation, the default approach being the leave-one-out (LOO) method.

An important role of a QSAR model, besides predicting the activities of untested molecules, is to provide hints about what molecular fragments may be important contributors to activity. This information is of great value to the synthetic chemist. Such information can be combined with knowledge of the maximal common structure (MCS) between compounds to get ideas for the synthesis of new molecules that might lead to suitable drug candidates. Clearly, the HQSAR method, which focuses directly on elements of chemical structure (rather than derived properties such as dipole moment or hydrophobicity), is well suited to proposing changes in chemical structure that are likely to improve activity.

The contributions of individual atoms to activity can be assessed through their representation in the hologram hash bins and can be used to color code the atoms of each molecule. Atoms that contribute positively to activity are colored toward the blue end of the spectrum; atoms that contribute negatively are colored toward the red end, and atoms that do not significantly affect activity (usually constituting the backbone or skeleton of the molecule) are colored white.

An important feature of HQSAR is the capability to recognize the maximal substructure common to the training compounds and to remove this substructure from consideration when constructing holograms. It is usually the case that a series of active molecules will share a significant common framework, and the predictive power of the method is enhanced by excluding portions of the molecule that are shared across the training set, and thus cannot be correlated with activity. Using this maximum common substructure (MCS) feature also allows the user to exclude in a consistent way molecules that lie too far outside the predictive space of the model. Moreover, identifying the MCS allows the user to better rationalize the relationship between structure and activity, and provides a template for proposing new compounds with activity that is likely to be accurately predicted by the HQSAR model. An MCS must contain at minimum seven connected atoms and is calculated purely by the molecules that constitute the training dataset; it is possible for good models to be derived without an MCS, but this is of less use when establishing and developing biological theories and of course for a centralized chemical synthesis route.

Once a HQSAR model has been calculated, the next step is to validate the model quality; the preferred method being the leave-one-out method (this is covered in detail in the next section). After model validation, proposed synthetics or other candidate theory molecules may be tested using the model. Often a search of chemical databases for molecules similar to those in the training set is conducted. Predictions of their activities, based on the model, are organized with the predicted activities in descending order. In a typical application, the database is first pre-screened using a fast measure of similarity, such as the Tanimoto index, with a user-selected value for the minimum similarity cutoff (the default threshold for this index is 0.85). The results of the Tanimoto screen, which is effectively a 2D similarity

search, is carried out, and the hits found from the search are entered onto a spreadsheet. The activities of the compounds are calculated from the HQSAR model. Compounds are sorted according to their predicted measured variable (usually IC_{50}). These findings can be presented in simple bar charts with test compound versus predicted variable. Using the Tanimoto coefficient is not absolutely essential, but can aid in reducing the size of larger databases. Frequently a database including only those molecules proposed by medicinal chemists is used.

Successful HQSAR requires a minimum of five molecules in the training set, but preferably more than ten. These should be structurally different molecules, but with some similarity. The model should not be used to predict the activity of compounds that are structurally very different from those in the training set.

HQSAR Example Vacuolar ATPase is an ATP-dependent proton pump that is important in many biological processes by acidifying specialized cell compartments. Of particular interest is the implication of V-ATPase in bone resorption by osteoclasts. Excessive bone resorption by osteoclasts and failure of osteoblasts laying down new bone are indicative of osteoporosis. The structural defects that accumulate and weaken the bones are primarily due to demineralization of the bone (removal of Ca^{2+} and PO^{3-}). Approximately one in three postmenopausal women and one in twelve older men suffer from osteoporosis. Currently, treatment of osteoporosis is via two methods—bisphosphonates and selective estrogen receptor modulators, the latter not being the favored treatment for men. Current therapies are useful in two respects. First, the bisphosphonates rapidly increase bone mineral density by forming a dead end complex; hence, bone resorption is impaired. Second, estrogenic molecules have the beneficial effect of mimicking endogenous hormones promoting remineralization, but increase in bone density is poor compared to bisphosphonates. Estrogens do offer a considerable advantage in that bone integrity is maintained, whereas with bisphosphonates microfractures tend to accumulate. To date, V-ATPase inhibitors are not suitable because of poor selectivity toward the osteoclast form of the enzyme and because of their more potent effect on kidney, liver, and neural forms. Structures and IC_{50} data from literature sources incorporating medicinal chemistry to improve selectivity [40–45] were used to aid design of new V-ATPase inhibitors. The data was organized in a manner so HQSAR could specifically assess and augment compound activity against desired V-ATPase species (osteoclast), yet decrease activity against undesired V-ATPases (liver and kidney). HQSAR models with high predictive quality were produced and tested, and a selection of proposed compounds were synthesized and evaluated *in silico*.

All compounds from the source papers were entered into a SYBYL database and transferred to a spreadsheet manually and checked for absolute stereochemistry. While constructing HQSAR models is a straightforward process, producing a model that will be robust and have high predictive value requires care in the selection and processing of activity data. The quality of models is assessed using *cross-validation*. In this approach, subsets of the training data are selected at random, and removed. An HQSAR model is constructed using the observations that remain and is used to predict the activities of the points removed. There are different protocols for carrying out this procedure; in the “leave-one-out” technique, every element of the training set is removed, and its activity is predicted based on all the remaining observations. Whatever the protocol, the measure of the quality of the model is the

cross-validated r^2 , denoted q^2 , which measures the error when the points predicted are excluded when constructing the model. Cross-validation is critical not only for HQSAR but for any method that potentially uses a large number of descriptors, and where there is high probability of finding “random” variation in the attributes of the training molecules that will closely match any pattern of activity presented.

In the present case, the quality of the models (as measured by both q^2 and standard error) was improved by two means: normalizing IC_{50} data relative to one compound of the set, or predicting differences between activity against different targets to produce a selectivity index (e.g., activity against chicken osteoclast cOc minus activity measured using bovine chromaffin granules bCG). An ideal set of compounds would contain IC_{50} values that cover a wide range of activity (10–15 orders of magnitude or more) and also a diverse collection of functional groups, so that the resulting model is valid over a large chemical space. In particular, compounds that include large, complex groups (e.g., an Fmoc group, see Compound 4 in [40]) will be poorly predicted if underrepresented in the training set.

With the goal of the study to create, detect, and evaluate new compounds selective for osteoclast V-ATPase inhibition, some HQSAR results (Figs. 2.7–2.9) illustrate a few the best models produced. These were measured by achieving a cross-validated $q^2 > 0.5$ and a standard deviation < 1 between the training dataset (experimentally derived measured variable) and the predicted dataset (HQSAR model-derived measured variable) (Table 2.2). Within the figures the y-axis scale was calculated relative to bafilomycin A_1 ; in essence, this compound serves as a reference value for the HQSAR algorithm, to which all other compounds were compared (i.e., bafilomycin has an inhibitory selective index of 1 for chicken osteoclast

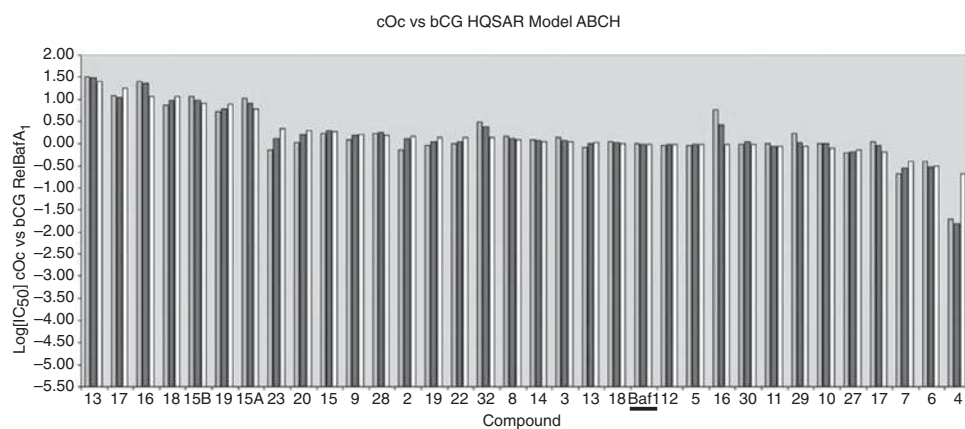


FIGURE 2.7 Selection of HQSAR results I. Actual data (**light gray bars**) $\log IC_{50}$ cOc – $\log IC_{50}$ bCG of a compound measured relative to bafilomycin A_1 V-ATPase phosphate assay from original data [40, 44], versus predicted $\log IC_{50}$ according to HQSAR (**dark gray bars**), versus the cross-validated HQSAR prediction (**white bars**). The graph is sorted in decreasing order of the cross-validated prediction relative to bafilomycin A_1 specificity. The graph suggests a strong correlation between the actual measured values of cOc/bCG selectivity and the predicted values (small difference between actual and predicted data). Hence, the model is likely to be of strong predictive value. To summarize, any compound left of bafilomycin A_1 (**Baf1**) has an increased specificity for cOc; any compound to the right has a decreased specificity for cOc.

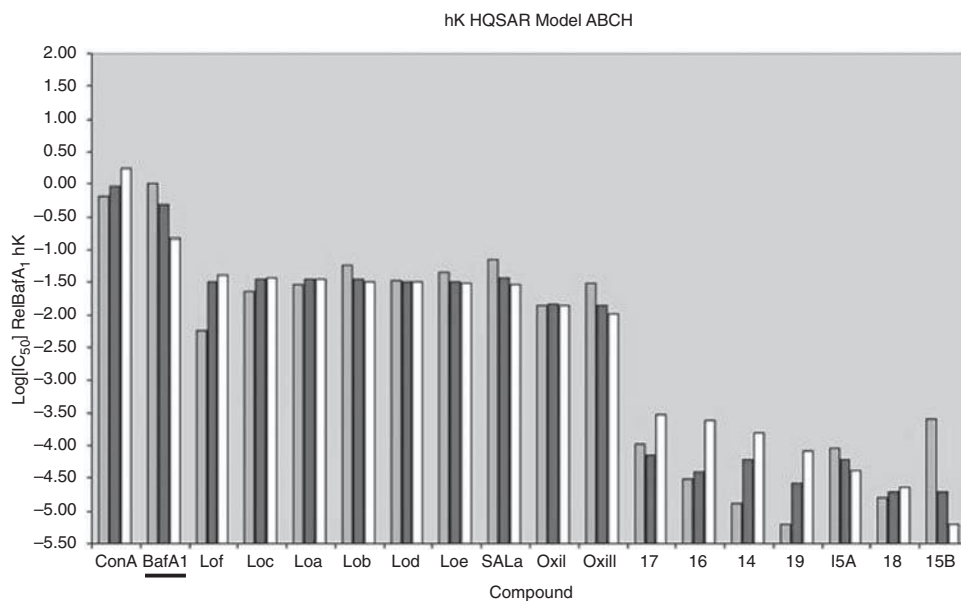


FIGURE 2.8 Selection of HQSAR results II. Actual data (**light gray bars**) $\log IC_{50}$ hK of a compound measured relative to bafilomycin A_1 V-ATPase phosphate assay (**BafA1**) from original data [41, 44], versus predicted according to HQSAR (**dark gray bars**), versus the cross-validated HQSAR prediction (**white bars**). The graph is sorted in decreasing order of the cross-validated prediction and has a relatively strong predictive power. This data could provide a reasonable model for predicting the action of potential lead compounds against hK toxicity.

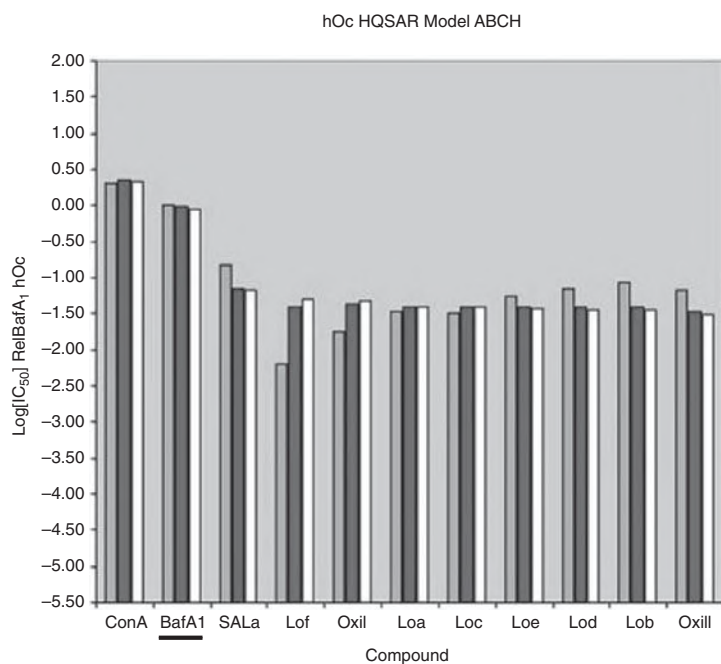


FIGURE 2.9 Selection of HQSAR results III. Actual data (**light gray bars**) $\log IC_{50}$ hOc of a compound measured relative to bafilomycin A_1 V-ATPase phosphate assay (**BafA1**) from original data [41], versus predicted according to HQSAR (**dark gray bars**), versus the cross-validated HQSAR prediction (**white bars**). Despite the small size of this group, the HQSAR model was of high quality and may provide an indication of the activity of the potential lead compounds on hOc.

TABLE 2.2 HQSAR Models^a

Parameter	Model	ApF	q^2	STD_ERR	CVSTD_ERR	Length	Components	MCS	Pred- r^2
hOc	ABC	4 to 7	0.672	0.384	0.437	401	1	N	D2S
hOc	ABCH	4 to 7	0.708	0.365	0.412	401	1	N	D2S
hOc	ACDo	4 to 7	0.606	0.413	0.478	401	1	N	D2S
hOc	ACHDo	4 to 7	0.645	0.395	0.454	401	1	N	D2S
hK	ABC	4 to 7	0.857	0.516	0.681	199	2	N	D2S
hK	ABCH	4 to 7	0.831	0.472	0.766	97	3	N	D2S
hK	ACDo	4 to 7	0.890	0.469	0.598	61	2	N	D2S
hK	ACHDo	4 to 7	0.968	0.189	0.295	59	3	N	D2S
cOc/bCG	ABC	5 to 8	0.819	0.148	0.279	59	5	N	0.7371
cOc/bCG	ABCH	5 to 8	0.766	0.130	0.323	83	6	N	0.8109
cOc/bCG	ABCh	5 to 8	0.817	0.124	0.286	97	6	N	0.8219
cOc/bCG	ABCHCh	5 to 8	0.719	0.115	0.355	353	6	N	0.8204
cOc/bCG	ACChDo	5 to 8	0.775	0.121	0.311	83	5	N	0.8829
cOc/bCG	ACDo	5 to 8	0.766	0.158	0.317	97	5	N	0.8335
cOc	ABC	5 to 8	0.413	0.572	0.904	53	5	N	D2S
cOc	ABCh	5 to 8	0.449	0.370	0.890	199	6	N	D2S
cOc	ACChDo	5 to 8	0.490	0.381	0.857	353	6	N	0.3572
cOc	ACDo	5 to 8	0.461	0.452	0.881	61	6	N	D2S
bCG	ABC	4 to 7	0.577	0.634	0.895	59	5	N	D2S
bCG	ABCH	4 to 7	0.525	0.650	0.948	401	5	N	D2S
bCG	ABCh	4 to 7	0.589	0.563	0.882	199	5	N	0.3111
bCG	ACChDo	4 to 7	0.633	0.360	0.849	257	6	N	0.5615
bCG	ACDo	5 to 8	0.568	0.513	0.921	307	6	N	D2S

^aThe HQSAR model accuracy was assessed by linear regression. The predictive power of the HQSAR model was calculated for each compound dataset (Pred- r^2), where some compounds were omitted in the HQSAR model construction. D2S = Dataset too Small for HQSAR model validation. An $r^2 > 0.35-0.4$ is good, with a value of 1.0 being perfect. The Parameter column is the dataset used and constitutes the dependent variable (i.e., IC₅₀ relative to baflomycin A₁), the Model column specifies the flags used. The Atoms per Fragment (ApF) column indicates the optimum number of atoms per fragment to generate the best q^2 value. STD_ERR = Standard Error, and CVSTD_ERR = Cross-Validated Standard Error. Length is the optimal holographic length; in each case all available hologram lengths were used (53, 59, 61, 71, 83, 97, 151, 199, 257, 307, 353, 401). Components are the fragments that were calculated to be potentially of greatest interest to modulate activity of the ligand (the larger the number, the better and more informative the model). The maximal common structure (MCS) was only effective when comparing the plecomacrolides.

over bovine chromaffin granules). The cOc versus bCG compares the selectivity ratio of bafilomycin A₁ against the two biological targets (computed as the difference of log₁₀ values), and then subtraction of the result from itself. The value is 0 because

$$\text{bafilomycin A}_1 \text{ selectivity (on cOc/bCG)} - \text{bafilomycin A}_1 \text{ selectivity (on cOc/bCG)} = 0$$

For compounds with greater selectivity toward cOc, the value will be >0 (*positive*) and those with less selectivity to cOc, the value will be <0 (*negative*).

The data (Fig. 2.7) classically demonstrates a good quality HQSAR model (relative to bafilomycin A₁); Indole 3 (**13**) was the most selective experimentally derived compound for cOc versus bCG (38-fold). The predicted value from the HQSAR model also indicated **13** as the most selective because it had the greatest magnitude. Likewise, the most nonselective compound for cOc over bCG was concanamycin derivative 4 (**4**), which had correspondingly the greatest negative magnitude. The cross-validated values indicate the predicted activity of omitted compounds and can improve the quality of the model (by making q^2 closer to 1, and standard deviation closer to 0), but in these results the cross-validation impaired the quality. The quality of HQSAR models can be assessed directly by leaving out bioactive molecules that have known IC₅₀ data. It is advisable to select molecules over the majority of the measured data range to ensure a fair test. An HQSAR model is constructed as usual with these selected molecules omitted from the training set. The HQSAR model was then used to evaluate the predicted IC₅₀ of these omitted molecules. Plotting the known IC₅₀ versus the predicted IC₅₀ values should produce a graph of $y = x$, with an ideal graph having the points plotted close to the $y = x$ line. Applying linear regression evaluates the quality of the predicted activities against their known values and hence is a direct measure of the quality of an HQSAR model(s). A linear regression score (r^2) close to 1 is ideal with 0 being the worst possible model. If the predicted r^2 value is close to 1, it adds confidence to the HQSAR model (Table 2.2), although datasets can be too small to apply such a technique and one must rely on the HQSAR model alone (Fig. 2.10). When evaluating proposed new chemical inhibitors (Figs. 2.11–2.13), researchers prefer a predictive r^2 value of around 0.35–0.40 (or greater) when considering candidates for chemical synthesis or chemical modification.

HQSAR has been used for investigations into several other ligand classes [46–52], but while computational output is exciting and valid, links with *wet* experiments and subsequent success are pressing issues and hence such techniques often evade academic interest.

2.6.2 QSAR

In the last ten years, QSAR has become an important tool that is used by nearly every pharmaceutical, agrochemical, and biotechnology company to increase the efficiency of lead discovery. The value of the QSAR approach is that it may be used either in the absence of detailed receptor site knowledge (i.e., binding site structural information) or in conjunction with such information if it is available. A QSAR model is a multivariate mathematical relationship between a set of physicochemical

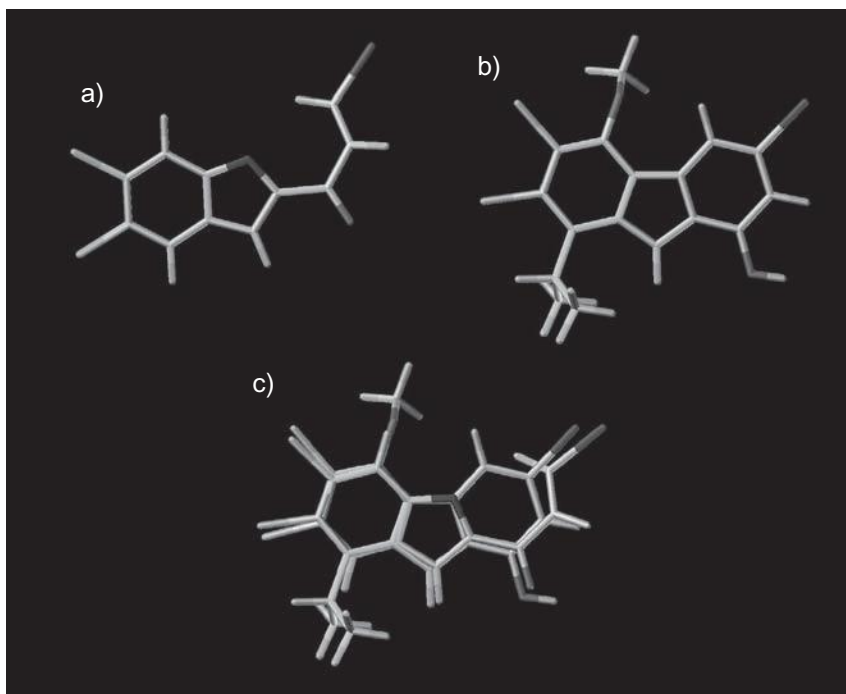


FIGURE 2.10 Two small molecules that at first glance look very different (**a** and **b**). When superimposed (**c**), the fused ring structure of the second molecule (**b**) does not appear too different from the nonfused ring structure (**a**). Only by careful alignment by eye can overlaps such as these be accomplished. Slowly, computers are improving in performing these tasks.

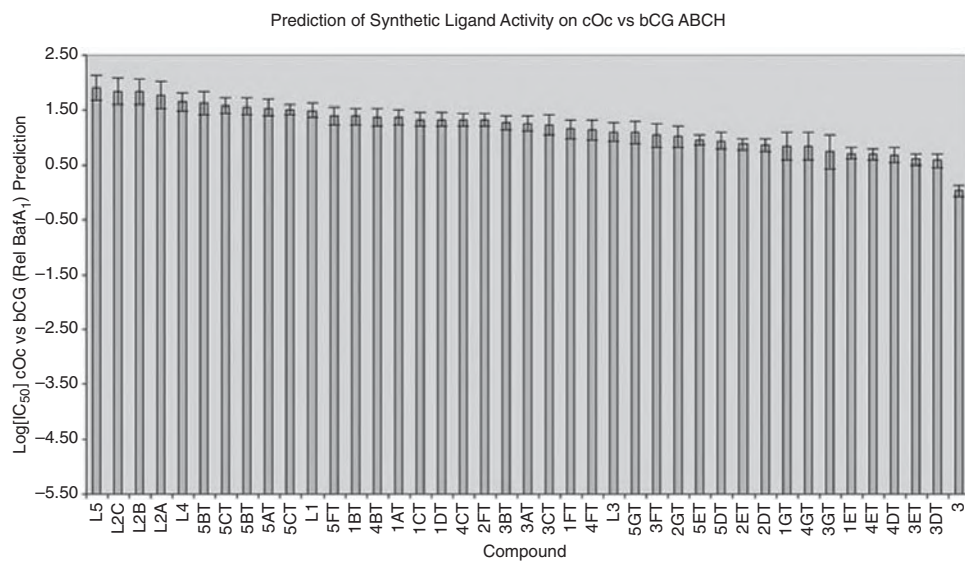


FIGURE 2.11 Application of HQSAR model predictions for *in silico* ligand database assessment I. Evaluation of the database of proposed synthetic ligands for selective action on cOc over bCG V-ATPase. Database compounds of significant interest lie to the left of the graph.

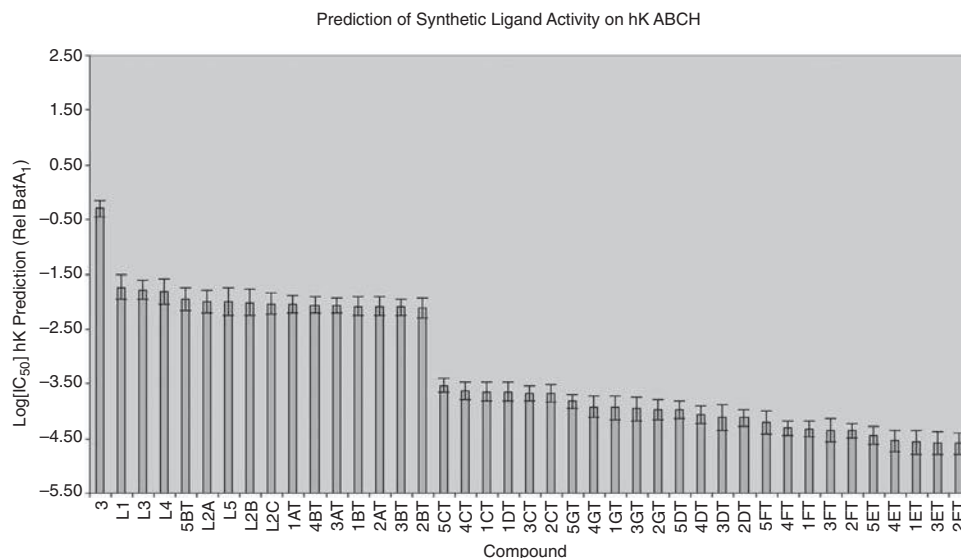


FIGURE 2.12 Application of HQSAR model predictions for *in silico* ligand database assessment II. Evaluation of the database of proposed synthetic ligands sorted by IC₅₀ for hK V-ATPase phosphate assay. Database compounds of significant interest lie to the right of the graph.

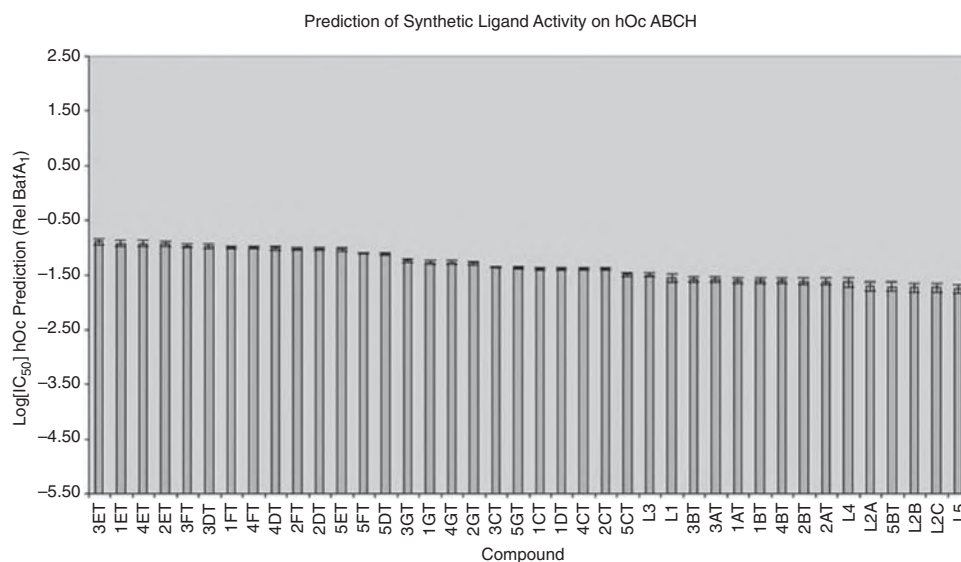


FIGURE 2.13 Application of HQSAR model predictions for *in silico* ligand database assessment III. Evaluation of the database of proposed synthetic ligands sorted by IC₅₀ for hOc V-ATPase phosphate assay. Database compounds of significant interest lie to the left of the graph.

properties (*descriptors*) and a property of the system being studied, such as the biological activity, solubility, or mechanical behavior. QSARs correlate with cogenetic series of compounds, affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological activities, either with certain structural features (e.g., Free Wilson analysis) or with atomic, group, or molecular properties, such as lipophilicity, polarizability, and electronic and steric properties (e.g., Hansch analysis). QSAR models have proved to be reliable tools for speeding up lead discovery [53–55] and have had an important place in molecular informatics throughout the world for at least two decades.

QSAR, like HQSAR, is based on measuring molecular similarity, but in the case of “classical” QSAR that similarity is based on characteristics, such as charge distribution and hydrophobicity, that derive from but are clearly secondary to molecular structure. While it is possible to rely exclusively on descriptors derived from experimental data, it is often the case that the molecular descriptors used in a QSAR study are directly computed from chemical structure; for example, a number of accurate algorithms are available to estimate $\log P$ from molecular structure [56], and likewise the electrostatic properties of molecules can be calculated using *ab initio* quantum chemical methods [57–62]. So, even if molecular structure is not immediately related to function by a QSAR model, it lies close in the background of these studies. In fact, this has become increasingly the case as the chemical libraries being used have increased in size, and it has become less likely that experimental data will be uniformly available for all the compounds being considered.

The elements of a QSAR model are the training data, consisting of selected physicochemical properties, a method for generating a predictive regression model from the training data, and a set of validation data (which are usually removed from the initial training set at the beginning). A variety of mathematical techniques are available to generate predictive models, which map a set of values for the molecular descriptors to a predicted activity. The simplest is classical multivariate linear regression, but this approach becomes less satisfactory as the number of descriptors increases, and it becomes likely that some of the descriptors will be highly correlated. More satisfactory approaches include principal components analysis (PCA) [63] and partial least squares (PLS) [64], which despite some differences in mathematical approach have the same function of automatically generating “aggregate” variables that are linear combinations of descriptors and building regression models based on these. Models constructed using PCA or PLS are less complex and more robust than those based on straightforward regression. They provide the added benefit of highlighting just those descriptors that are important for explaining the activity to be predicted, and effectively dropping those that are uncorrelated with the activity to be explained. Cross-validation is critical, especially when many descriptors are in use, and is performed just as was described in the context of HQSAR (Section 2.6.1).

The goal of QSAR is to derive a function that relates to biological activity with some parameter(s) describing a feature of the molecule. Analyzing the correlation between biological activity and molecular parameters for a series of molecules that have already been tested forms the QSAR relationship. The concept is based on the assumption that the difference in the physical and chemical properties of molecules, whether experimentally measured or computed, accounts for the difference in their observed biological or chemical properties. Thus, in general, the QSAR method

deals with identifying and describing important structural features of molecules that are relevant to explaining variations in biological or chemical properties. The QSAR indicates the descriptors that are most statistically significant in determining the property, and studies can be focused on the molecular characteristics that those descriptors represent. QSARs thus help to make maximum use of data, whether that data is from experiment, simulation, or a database search.

QSAR is often carried out in the context of a *congeneric series*, a collection of molecules that share a common framework, but with significant variation of attached functional groups. Such a series comprises “variations on a theme,” and it is this scenario that is most likely to lead to a robust, useful predictive model. While this might be seen as a drawback for classical QSAR, in that only molecules of a specified class are considered, the approach is very much compatible with most drug design methodologies where one begins with a framework that is amenable to variation using well-understood synthetic routes. The goal in this case is to develop a QSAR that rationalizes changes in activity with respect to structure, and to use this as a guide in proposing additional modifications that will further improve activity.

The “series” (a library of molecules with a shared characteristic) will consist of a common molecular framework with variance only in physicochemical descriptors such as hydrophobic constants, electronic parameters, or individual atoms. These physicochemical adaptations explain why individual molecules in a series have different biological activities. Relationships between activity and physicochemical characteristics can then be postulated, and the postulated relationship can be tested by generation of new compounds with predicted properties. The methods used to establish the equation that best describes the relationship between the property and the descriptors include regression techniques, PCA, and genetic algorithms. The expressed QSAR takes the form of either a search query or a predictive model, which can then be used to select new molecules with the specified activity from a database, or to predict the activity of individual molecules of interest. QSAR thus provides invaluable knowledge of which interactions are important to activity. This understanding provides the basis to formulate new active compounds that possess better overall therapeutic profiles, for example, compounds with increased functionality or that are more orally active. In fact, any biological or chemical activity that can be measured or derived from measurements can potentially be used as input for QSAR.

It should also be remembered that *in vitro* activity data only produce a QSAR that selects molecules that satisfy the *tested biological assay*, and do not necessarily indicate *in vivo* activity. The quality of the assay is therefore important. It has become increasingly important to conduct QSAR using more than one measured activity variable, such as adding a bioavailability measure to the activity data. The addition of *in vivo* data is difficult, as often many variables may exist that affect activity (primarily metabolism and excretion that are not tailored for), but this should be considered as an important development of QSAR.

2.6.3 Superimposition

One of the most compelling similarity techniques is not automated, can be time consuming, but is always done to get the point across clearly and effectively to a

reader or critic. Usually after several techniques have been employed to narrow down the best matches to a particular query molecule, potential candidates can be compared for similarity by 2D superimposition. The coordinates of one molecule (the query) is taken and mapped onto another molecule (the candidate) to assess how similar the two molecules are. However, the drawback is that so far only we humans can deduce where to appropriately overlay them (Fig. 2.10). Therefore, it is hard to implement superimposition until there is a short list of molecules: a list of thousands is not appropriate for evaluation by this method. Ultimately, this technique convinces medicinal chemists that the candidate molecule is indeed similar to another, and they will commence synthesis (providing there are not too many chiral centers, *E/Z* isomers, or complex heterocyclic ring structures).

2.6.4 Program Suites

UNITY is a program suite offered by Tripos [14], which contains a combination of techniques and methods to represent small molecules and identify potential molecules for biological trials via similarity. Another such suite available for purchase is CATALYST [65] offered by the commercial vendor Accelrys [66].

Identifying Molecules Via Behavioral Similarities The principles governing the biological activity of many chemical entities used as drugs are usually due to particular groups and core structure within the molecule. These physicochemical descriptors are usually unique to a particular class of molecule. For instance, a hydrogen bond donor (or acceptor) could interact favorably with the biological target and be crucial for activity; often coordinated water molecules are critical in such cases (e.g., estrogens 1ERR [67]). Maximizing contact between the molecule and the target via van der Waals interactions is the driving force for ligand binding (i.e., a good fit), releasing water used in solvation and increasing the entropy of the system. To employ UNITY, the user needs strong biological understanding of the active ligands known and, if possible, should have an available protein structure.

Before searching large databases, it is imperative to design a robust representation of the parts of the molecule responsible for the drug class activity. This simplified representation is called a *pharmacophore*; it incorporates the features of as many drugs within the class as possible (deemed crucial for activity). If the design is of good quality, the pharmacophore will return the molecules known to be active and thus validate itself. Once the pharmacophore is designed suitably and holds well by returning other molecules within the drug class, the next step can be taken.

Large databases of molecules are available at cost and also without charge (Section 2.5.2) and are an excellent resource to find if there are compounds already similar to the pharmacophore. The goal is to find subtly different molecules that contain the groups and structural features essential for activity. Even if the molecule is not ideal, it may be available and amenable to medicinal chemistry.

In UNITY the procured molecular databases must be processed into a recognizable format for use by the program and often requires expertise. One must be exceptionally careful when building the database and when relying on prepared databases that the compounds are indeed represented correctly. CORINA [16] and STERGEN [17] are among the best tools for this particular job, but naming compounds uniquely is the most imperative and difficult task to accomplish. On con-

struction of UNITY databases, unique names for each molecule are responsible for most of the observed errors reported back. There are more sinister examples of errors such as the interpretation of molecules by SYBYL itself. One needs to be very vigilant with molecules having double bonds and conjugated aromatic and cyclized compounds. It is often the case that subtle changes can occur and the user is oblivious to these mistakes. A particularly good example is Raloxifene from the WDI database being interpreted incorrectly. Many PDB files contain mistakes and can also be interpreted incorrectly. When the database is constructed, it is important to note that 2D fingerprints and 2D macroscreens can be included in the database build. Often incorporating this into the UNITY database design increases the build time dramatically; but on the upside, all the searches will be considerably faster and nonmatching molecules are rapidly evaluated in a prescreen first.

These shortcomings aside, UNITY is still a powerful and reproducible means to search for similar molecules with many extra features, including the possibility of using key amino acids interacting with the ligand for structure-based drug design. What is sadly lacking is a score of how well the hits from the database fit the pharmacophore to describe the quality of the hit itself. This is rectified using an inbuilt feature to obtain Tanimoto coefficients for the returned hits: pitching these molecules against the best known drug molecules and implementing a cutoff to similarity.

Designing New ACE Inhibitors Angiotensin converting enzyme (ACE) inhibitors are an excellent means to demonstrate pharmacophore design, not to mention their significant economic and medicinal importance. Heart disease affects approximately 65+ million Americans. The costs incurred via healthcare bills, lost revenue, and sick days are a serious nationwide medical and financial concern. The prime contributor to heart disease is hypertension (high blood pressure); 95% of cases are diagnosed as “essential hypertension.” Blood pressure higher than 140/90 mmHg is considered seriously hypertensive and requires treatment.

ACE inhibitors work by preventing the final cleavage step of the rennin-angiotensin system; at this step angiotensin I is converted to angiotensin II by the enzyme ACE. Angiotensin II is the most potent endogenous vasoconstrictor known and increases peripheral resistance to blood flow by reducing the bore of arterioles. Inhibition of ACE would logically follow with a reduction in blood pressure and reduce the strain on the heart and circulatory system.

Recently, the crystal structure of lisinopril (Zestril by AstraZeneca) has been solved (PDB file 1O86) [68] and is considerably useful since potential molecules can be evaluated via docking into the active site of the receptor. Several ACE inhibitors combined with known IC_{50} data are available [69], as are pharmacophore designs [70, 71]. To highlight the use of such information and data, we demonstrate how UNITY can be used by incorporating pharmacophore designs for database screening. Our example is somewhat simpler than that presented in the literature for clarity and ease of understanding.

The pharmacophore designed is more simplistic but serves an important purpose in orienting the mind to appreciate that pharmacophores do not need to be overly complex to be effective. As each feature is added and the pharmacophore is built up, it adds further complexity to the screening process, unlike a macroscreen or

fingerprint comparison that is rapid; a spatial or distance constraint between atoms requires a lot more computational time. Hence, a pharmacophore with one defined feature is a 1-point pharmacophore and can be identified very rapidly. Two defined features constitute a 2-point pharmacophore, and so on; typically, a pharmacophore has between 4 and 6 points, with some being very complex indeed [71]. As more points are added, the complexity of the calculations increases exponentially, so when screening, simple pharmacophores followed by a more complex one will save time in investigations. To screen a library of approximately 6 million molecules, it took approximately two weeks (in computer time) to create and screen the UNITY databases with incorporation of 2D macroscreens via fingerprints. This was six weeks in real time to manipulate and screen and modify the residual hit list data. Figure 2.14 shows the simple pharmacophore used incorporating the fragments conserved between all the ACE inhibitors [69], the three-carbon chain, and nitrogen in the five-membered ring. The features of this 2-point pharmacophore (Fig. 2.14a) were rapidly identified by the 2D Macro function and the matches were easily recognized by 1D comparisons. The 3-point pharmacophore (Fig. 2.14b), however, includes distance and spatial constraints, consequently requiring a lot more computational time. Thus, in screening a large library one would almost certainly conduct the search using the 2-point pharmacophore followed by the second. When testing the pharmacophores sequentially, all but one of the 23 known ACE inhibitors were returned, thus validating the pharmacophore design. The pharmacophores were used to screen an NCI database containing all possible stereoisomers of each entry and the residual dataset was organized using the Tanimoto coefficient. The top 4000

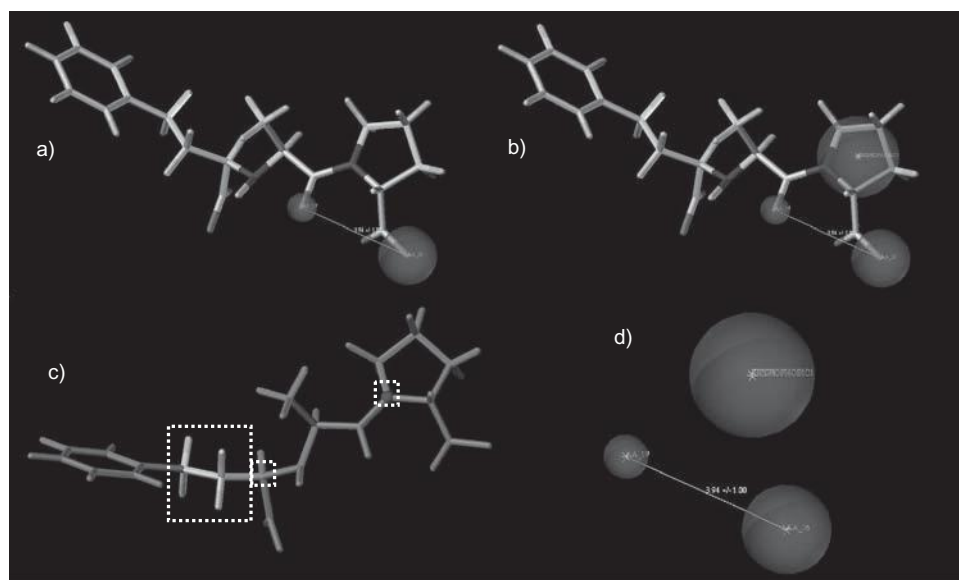


FIGURE 2.14 Three pharmacophores were used in a similarity search: (a) 2 points + distant constraint, (b) 3 points + distant constraint, (c) fragment-based pharmacophore with four hydrogens and 3 carbons (large box) and one nitrogen (small box); (d) what the computer sees for (b).

hits against lisinopril were docked into the PDB file 1O86 with the GOLD algorithm [72]. A comprehensive study of this investigation is currently under way.

UNITY Summary UNITY demonstrates that robust design of a molecular template as a pharmacophore can prove insightful for identifying similar molecules. Although investment is required in software and expertise, it is not a particularly difficult technique to become adept with. Probably the biggest concern is maintenance and networking issues requiring specialized staff capable of system administration. This example merely scratches the surface of what UNITY can do: it is most relevant in identifying similar molecules.

2.6.5 Comparative Molecular Field Analysis (CoMFA) and Related Approaches

Comparative molecular field analysis (CoMFA) [73, 74], a proprietary method developed and marketed by Tripos, Inc. [14], is one of the most popular approaches to constructing quantitative models that link structure and activity [75–78]. Unlike conventional QSAR approaches, which use descriptors that describe the entire molecule (dipole moment, log P , topological indices, etc.), CoMFA is a 3D method that correlates a measured activity (e.g., log IC_{50}) with variations in electrostatic and steric properties for an aligned series of molecules. Since CoMFA identifies regions of the aligned structures that are important for determining activity, it allows the synthetic chemist to focus on sites of an existing molecular framework where variations are likely to be most productive. In favorable circumstances, a CoMFA can be thought of as an inverse model of the target receptor, since positive variations in activity require changes in the steric and electrostatic properties of the ligand that are compatible with (complementary to) those of the host receptor site. A CoMFA model can be generated to rationalize the activities of a series of ligands already characterized, and can then be used in predictive mode to estimate the activities of new molecules being proposed. CoMFA is one of a number of techniques that implicitly measure molecular shape by superimposing a regular grid on a collection of molecules.

The prerequisite for CoMFA is an aligned training set of ligands with known activities. This condition is most easily met by a congeneric series, since the shared molecular scaffold provides an immediate means to align the members of the set. In situations where a common framework is not available, CoMFA still requires that the molecules being analyzed can be superimposed in a meaningful way. One approach to this is *molecular field-fit*, where each molecule is positioned on a grid so as to maximize overlap with a preexisting set of field values (steric or electrostatic) computed on the grid vertices [79]. This approach can be used to construct a progressive alignment, where one or more seed molecules are used to generate a field on the grid, and successive molecules are aligned to the existing grid. As molecules are added to the alignment, their fields can be added to the vertices, potentially reinforcing the alignment field. Other methods that rely on a molecular field representation to carry out alignment include SEAL [80], FLUFF [81], and FIGO [82] (which combines a minimization procedure with 3D descriptors generated by the program GRID [83]).

A simple and effective approach for collections of structurally dissimilar molecules is to align them using their principal moments of inertia [84]—a procedure

that aligns molecules based on their distributions of mass. Yet another approach, applicable when a structure is available for the target protein receptor, is to dock the molecules into the common receptor site, thus aligning them [85–87]. At first this might seem an ideal approach, since the molecules will be aligned with biologically relevant conformations; however, owing to inevitable perturbations in the conformations and positions of the docked ligands, even when the molecules exhibit very similar docking modes, this procedure tends to produce CoMFA models that are inferior to those generated using a receptor-free alignment [88]; but on occasion improvements are evident [89].

Once an alignment has been constructed using an appropriate set of training molecules, the next step is to embed the aligned set in a cubical grid and to compute the interactions of a probe atom with the aligned molecules. The default probe is a carbon atom (with van der Waals parameters appropriate for a tetrahedral carbon) with a charge of +1. The probe is positioned at each vertex of the grid, and its steric (van der Waals 6–12) and electrostatic (Coulomb's law) interactions are computed separately with each of the molecules in the alignment. (It should be pointed out that the choice of probe is flexible, and a number of novel modifications have been introduced, including probes that measure hydrophobicity through the HINT potential [90–92], and use of an orientable water molecule as a probe, to measure local hydrogen-bonding potential [93].) The grid spacing is an important parameter in determining the performance of the resulting CoMFA model, and there is also some influence due to the position and orientation of the aligned molecules with respect to the grid. The importance of these factors can be evaluated by adjusting the grid spacing as well as the orientation of the aligned molecules. In addition, a grid “focusing” procedure is available, which can be used to automatically refine the grid by identifying those grid vertices on which changes in potential are highly correlated with activity [73, 74, 94]; those portions of the grid most important for creating a robust predictive model can then be refined by subdivision, effectively increasing the density of vertices in the regions of the alignment with greatest predictive importance.

The final step in the construction of a CoMFA model is the application of partial least squares (PLS) regression, a method already discussed, to generate a quantitative relationship between the field values computed on the grid and measured activity (although the technique SOMFA shows new and promising advantages [95]). In CoMFA, each vertex of the grid gives rise to at least one descriptor (a steric or electrostatic field value for each molecule in the alignment). Since these typically number in the hundreds or thousands, there is the danger that the PLS procedure will pick out vertices with potential values that correlate well with activity merely by chance. In fact, it is almost always the case that CoMFA will produce a model that predicts the activities of the training molecules with a very favorable r^2 . To ensure that the model is robust and applicable outside the training set, it is an absolute requirement to apply cross-validation. This is usually accomplished using the leave-one-out approach discussed previously (although more sophisticated techniques are available). As a rule of thumb, CoMFA models with a cross-validated r^2 (q^2) of 0.5 or better are considered to be useful as predictive models. As with all established techniques, CoMFA is subject to evaluation [96] and improvement [95, 97].

2.6.6 Shape Signatures

Shape Signatures is a new method for detecting molecular similarity recently developed in our laboratory [98]. The fundamental motivation of the method is to generate compact descriptors that capture the features of molecular shape and polarity while avoiding details of chemical structure. By maintaining a focus on shape, our approach emphasizes those characteristics of molecules important for biological activity, making it much easier to scan chemical libraries for compounds that may be both bioactive and of novel structure.

The key feature of Shape Signatures is our approach for exploring and encoding shape and polarity information. We have adapted the method of ray tracing, widely used in computer animation and presentation graphics, as a probe of molecular geometry. To do this, we initiate a ray in the interior of a molecule, which is bounded by a triangulated representation of the solvent-accessible molecular surface (Fig. 2.15a), and allow the ray to propagate by the laws of optical reflection (Fig. 2.15b,c). Probability distributions are derived from the ray, and it is these distributions, stored as histograms, which we denote as Shape Signatures. The signatures are independent of the orientation of the molecule and can be compared very quickly using simple metrics below:

$$L_1^{1D} = \sum_i |H_i^1 - H_i^2|$$

$$L_1^{2D} = \sum_i \sum_j |H_{i,j}^1 - H_{i,j}^2|$$

The simplest signature is the probability distribution of ray-trace segment lengths, where a segment is the portion of the trace between two successive reflections. We call this a one-dimensional (1D) signature, as the domain of the distribution is one dimensional (Fig. 2.15e). Shape Signatures of higher dimension can easily be generated by combining ray-trace segment length with properties measured on the molecular surface; for example, by computing the joint probability distribution for observing a particular sum of segment lengths on either side of a reflection point, combined with the molecular electrostatic potential (MEP) measured at the reflection, we produce a “2D-MEP” signature with two-dimensional domain (length + electric potential), which encodes both shape and polarity information (Fig. 2.15f).

Although not as well developed as the ligand-based approach just described, it is also possible to apply the Shape Signatures method in *receptor-based* mode. Here, the ray-trace operation is carried out in the volume exterior to a protein receptor site, with Shape Signature histograms accumulated in the same way as in the ligand-centered approach. In this case, a match between the signatures for a potential ligand and the receptor indicates shape complementarity between the small molecule and the shape of the receptor-site volume. While harder to apply than the ligand-based method (primarily because of ambiguities in defining the binding-site volume), the receptor-based approach offers the exciting prospect of scanning chemical libraries for potential bioactive compounds on the basis of shape, without the bias of using specific chemical structure queries, and without the computational expense of database docking.

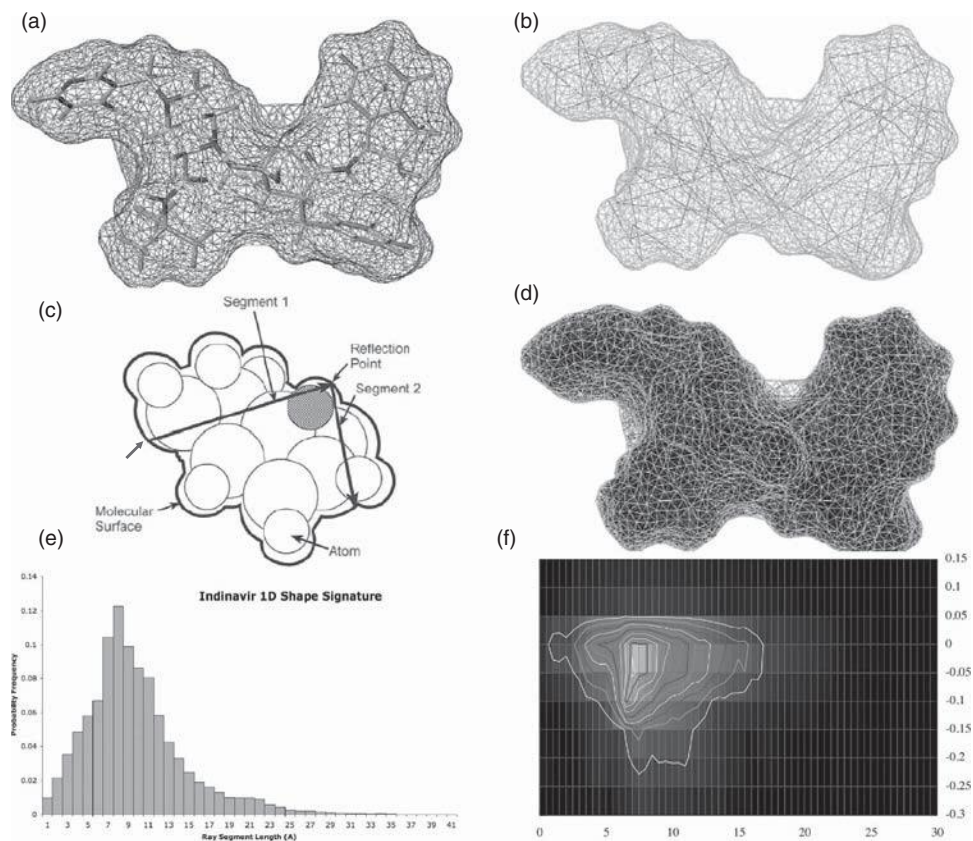


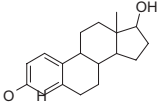
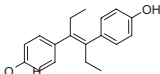
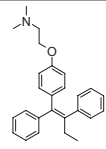
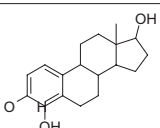
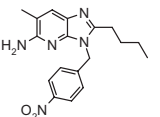
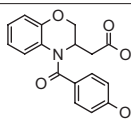
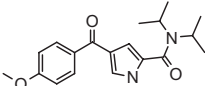
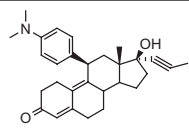
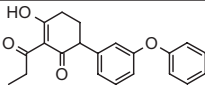
FIGURE 2.15 Process of Shape Signature generation. (a) The structure of Indinavir enclosed in the triangulated solvent-accessible surface is generated using SMART [41]. (b) Propagation of a ray trace around the inside of the triangulated molecular surface. Propagation of the ray trace around the Indinavir molecule with (c) 100 ray-trace segments and (d) 10,000 ray-trace segments is shown. On completion of a Shape Signature, the generated trace (d) is illustrated by (e), a histogram denoting the probability distribution (ordinate) of ray-trace segment lengths (abscissa), a “1D Shape Signature.” A Shape Signature trace defined by ray-trace segment lengths and a mean electrostatic potential (MEP) is shown by (f) a contour plot, a “2D Shape Signature.” It is these data stored as text files that are compared when performing a database screen.

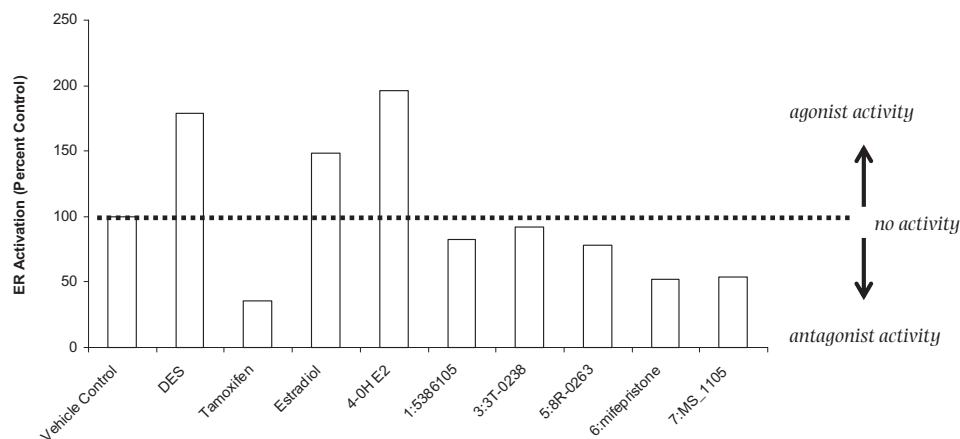
Although the method is in its infancy, Shape Signatures similarity comparisons have already been shown to be very effective for biological application [28–30, 98–100]. A current area of intensive research in our laboratory is to develop further ways of applying Shape Signatures for multiple conformers and stereoisomers, and to further develop the receptor-based strategy along with the existing ligand-based approach. It is clear that it is not unreasonable to expect to be able to carry out well over a thousand million comparisons a day on a single processor (Table 2.3)—numbers that are not easily attainable by other techniques. The method is trivial to

TABLE 2.3 Comparison Between Speeds for Database Construction and Screening

Technique	Number of Molecules/Time Unit					CPU + RAM Computer Specifications
	Minute	Hour	Day	Month	Year	
<i>Database Construction</i>						
Shape Signatures (Ray Tracing)						
	5.6	3.3×10^2	8.0×10^3	2.4×10^5	2.9×10^6	1 × 3.5 GHz Intel Pentium 4, 2 GB
	1.1×10^4	2.1×10^4	5.1×10^5	1.5×10^7	1.9×10^8	32 × Opteron 2.6 GHz, 0.5 GB
Raptor Model (Generation)						
	4.5	2.7×10^2	6.5×10^3	1.9×10^5	2.4×10^6	1 × 1.8 GHz Pentium 4
<i>Database Screening</i>						
Shape Signatures (Screening)						
	1×10^6	6.0×10^7	1.4×10^9	4.3×10^{10}	5.3×10^{11}	1 × 3.5 GHz Intel Pentium 4, 2 GB
	6.4×10^7	3.8×10^9	9.2×10^{10}	2.8×10^{12}	3.4×10^{13}	32 × Opteron 2.6 GHz, 0.5 GB
ROCS (Rapid Overlay of Chemical Structures)						
	7.6×10^2	4.6×10^5	1.1×10^7	3.3×10^8	4.0×10^9	1 × 3.5 GHz Intel Pentium 4, 2 GB
UNITY (Pharmacophore)						
3pt	81.6	4.9×10^3	1.2×10^5	3.5×10^6	4.3×10^7	1 × RG14000 MIPS
4pt	68.9	4.1×10^3	9.9×10^4	3.0×10^6	3.6×10^7	SGI 1 × 500 mHz, 1 GB
5pt	56.6	3.4×10^3	8.2×10^4	2.4×10^6	3.0×10^7	
Docking GOLD 2.2						
	0.6	34.7	8.3×10^2	2.5×10^4	3.0×10^5	2 × 3.0 GHz Intel Xeon, 4 GB
Raptor (Molecule Evaluation)						
	8.0	4.8×10^2	1.2×10	3.5×10^5	4.2×10^6	1 × 1.8 GHz Pentium 4

parallelize: both database generation and comparison—hence submission of queries via the Internet—could be implemented and accomplished in the near future. The data returned dramatically reduces the initial database size, a vital part of any successful CADD strategy [101]. Furthermore, Shape Signatures also achieves the second important quality of CADD: enrichment of molecules from the same class as the query [28–30, 98, 99]. Initial successes with Shape Signatures and estrogenic molecules (Fig. 2.16) were supported with experimental findings [99]. A potential novel molecule involved in analgesia has been discovered using Shape Signatures and offers considerable benefits over current compounds in this class [100]. Many groups [102–104] have previously used similar techniques for virtual-spatial recognition, but they have not been used in biological applications (with the exception of Ankerst and colleagues [105]). The Shape Signatures method is distinct from these

Controls		
 Estradiol (Control: Potent Agonist)	 DES (Control: Potent Agonist)	 Tamoxifen (Control: Potent Antagonist)
Test Compounds		
 4-OH E2	 1:5386105	 3:3T-0238
 5:8R-0263	 6:MS_9579 (mifepristone)	 7:MS_1105



other previous attempts that use histogram comparisons for recognition of objects [102–104] or for lead discovery [105]. The Shape Signatures technique is rotationally invariant, meaning that it does not require, nor depend on, the orientation of the ligand or receptor site to obtain a reproducible histogram profile [103, 107].

As an example of the application of the Shape Signatures method, the query molecule WHI-P131 [106, 107], a tyrosine kinase inhibitor of interest as a therapeutic against several diseases (including leukemia and amyotrophic lateral sclerosis), is shown along with the top Shape Signatures hits located in the NCI database (Table 2.4) (were clarified by superimposition, see Section 2.6.3). Note the clear shape similarity between query and hits that nonetheless differ significantly in details of chemical structure. Finding these matches by other CADD methods would entail either generating a large set of structural queries based on the molecule of interest, or generating a pharmacophore model for the inhibitor. In either case, there

FIGURE 2.16 Examples of molecules selected by Shape Signatures demonstrating estrogenic activity on human estrogen receptor (ER) based on known controls. Three known estrogenic compounds (**top**)—estradiol, diethylbesterol, and tamoxifen—are all known to interact with human ER control. Using these as query molecules to screen an in-house database (1.2×10^6 molecules), the returned molecules (**middle**) were tested via assay [48]. Taking each of the selected “hits” from the search in turn, a 25 μM sample was tested using the NR peptide ER α ELISA kit (Active Motif, Carlsbad, CA) according to manufacturer’s instructions. 17 β -estradiol and tamoxifen were included as part of the kit. Briefly, a precoated 96-well plate was supplied with an optimized peptide containing the consensus binding motif of ER α coactivator SRC-1. Each compound was incubated for 1 hour with MCF-7 nuclear extract and the coactivator peptide in each well. The ligand-activated ER α was first detected using primary antibody specific for ER α and further with HRP-conjugated secondary antibody. The ER competitive binding assays used a gel filtration displacement assay for estrogen receptor alpha (ER α) and was employed to assess competitive binding by selected ER antagonists among the hit compounds. Estrogen receptor binding assays were conducted in duplicate in 50mM Tris-HCl, pH 7.5, 1 mM EDTA, 20% glycerol, and 1 mM DTT buffer. Radioligands that were used included [6,7- ^3H]estradiol (specific activity 44 Ci/mmol, Amersham Biosciences). Binding assays were conducted on ice in a volume of 1 mL with 10 ng of purified full-length ER α and 25 nM ^3H -estradiol in final concentration; 10 μM 17 α -estradiol was used to define specific binding. Following a 1-hour incubation, assays were terminated by filtration through Whatman GF/B filters. Filters were soaked in Ecoscint liquid scintillation mixture (National Diagnostics, Somerville, NJ) and filter bound radioactivity was counted using a Beckman LS 1071 counter. (**Bottom**) One molecule was strongly agonistic (4-OH E2), while two were strongly antagonistic (6:mifepristone and 7:MS_1105). The other three molecules were only marginally antagonistic.

would be the presumption of specialized knowledge to construct the queries actually used to scan the database, and a small omission in constructing that query could mean missing important hits. In contrast, the ligand-based Shape Signatures search is very easy to carry out; the investigator need only present the single compound of interest as the query.

There is one potential drawback of Shape Signatures that must be addressed up front. Because the method collapses a great deal of chemical space onto very compact descriptors, it is inevitable that a Shape Signatures search will turn up false

TABLE 2.4 Shape Signatures Search of the NCI Database with WHI-P131

Rank	Molecule	Score
1D		
1	WHI-P131	0.000000
2	NCI_HIT1	0.050280
3	NCI-HIT2	0.081440
2D		
1	WHI-P131	0.000000
2	NCI_HIT2	0.255917
3	NCI_HIT1	0.273748

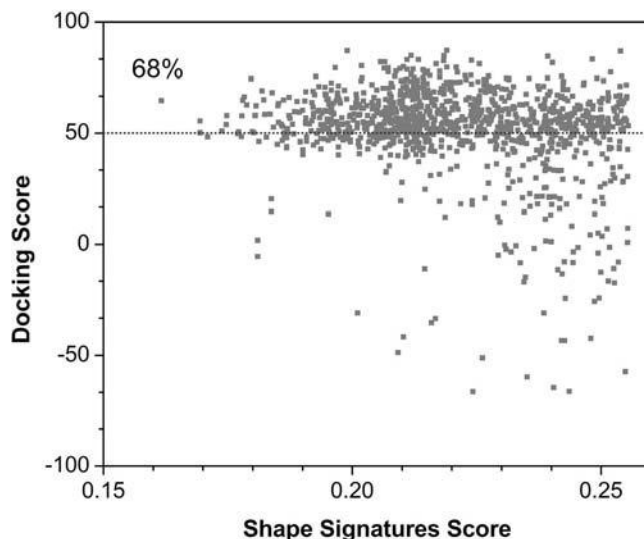


FIGURE 2.17 Evaluation of Shape Signatures with the GOLD algorithm. The plot indicates Shape Signatures score (abscissa) versus GOLD (ordinate) with the dashed line at 50 indicating a good GOLD docking hit.

positives in the hit list. To test how significant this issue might be, an independent assessment of the quality of the hits produced by a Shape Signatures search with the ACE inhibitor Enalapril was established. First, docking of known ACE inhibitors that have measured IC_{50} data [69] into ACE (PDB code 1O86) [68] was carried out using GOLD [72], producing a significant correlation between inhibitor pIC_{50} and the GOLD score, with scores for the known actives ranging from 50 to 87. In the second phase of our study, we identified the top 250 hits produced by a Shape Signatures search for the single query, Enalapril, conducted on an in-house database of 423 drugs and the NCI database [10] of over 250,000 compounds. The hit compounds were extracted from the database and docked into the active site of ACE using GOLD. Not only did our search, conducted using a single query, immediately identify 11 of the 20 well-characterized ACE inhibitors, but the fitness scores by GOLD evaluation of all 250 hits gave >75 for 4%, and >50 for 70% of the hits (Fig. 2.17). Assuming that the correlation between the experimental IC_{50} and the GOLD score holds, the Shape Signatures method is shown to readily identify known actives and interesting lead molecules, and with a modest proportion of false positives.

2.7 CONCLUSION

In this chapter we have merely scratched the surface of available methods to measure molecular similarity, focusing on methods with which we have experience and omitting any number of popular methods. In particular, there are a large and steadily growing body of 3D methods that are similar in spirit to CoMFA, but which nonetheless incorporate variations and refinements. In this context we must mention

approaches that use descriptors generated using GRID [83] and first-cousins to CoMFA, such as COMSIA [108]. We hope that this chapter has provided an introduction to the spirit of molecular similarity methods and will serve as a foundation for further exploration by the interested reader.

ACKNOWLEDGMENTS

We kindly acknowledge Dr. Zhiwei Liu at the University of the Sciences in Philadelphia for her efforts in providing us with Fig. 2.16 for this chapter. We also extend gracious thanks to Ching Y. Wang and her supervisor, Professor William Welsh, at the University of Medicine and Dentistry of New Jersey for their efforts in constructing Fig. 2.15.

REFERENCES

1. Center for Drug Evaluation and Research, 2004, Report to the Nation: Improving Public Health Through Human Drugs. (<http://www.fda.gov/cder/reports/rtn/2004/rtn2004.htm>).
2. Mullin R. Tufts report anticipates upturn; cites drugmakers' actions to accelerate drug development. *Chem Eng News* 2006;84:9.
3. Newman DJ, et al. Natural products as sources of new drugs over the period 1981–2002. *J Nat Prod.* 2003;66:1022–1037.
4. Laatsch H. *AntiBase 2007: The Natural Product Identifier*. Hoboken, NJ: John Wiley & Sons; 2007.
5. Pauli A. *AmicBase 2005*. Hoboken, NJ: John Wiley & Sons; 2005.
6. Thomson Scientific World Drug Index. <http://www.scientific.thomson.com/products/wdi> (accessed July 2006).
7. GlaxoSmithKline, Plc. <http://www.gsk.com> (accessed July 2006).
8. Pfizer, Inc. <http://www.pfizer.com> (accessed July 2006).
9. Irwin JJ, Shoichet BK. ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;45:177–182. (<http://zinc.docking.org>).
10. NCI Database. <http://cactus.nci.nih.gov/ncidb2/download.html> (accessed July 2006).
11. ChemBank Database (1.1 million entries). <http://chembank.broad.harvard.edu/>.
12. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36.
13. Ash S, et al. SYBYL Line Notation (SLN): a versatile language for chemical structure representation. *J Chem Inf Comput Sci* 1997;37:71–79.
14. Tripos, Inc. <http://www.tripos.com> (accessed July 2006).
15. MarvinSketch. <http://www.chemaxon.com/marvin> (accessed July 2006).
16. CORINA. <http://www.mol-net.de/software/corina/index.html> (accessed July 2006).
17. STERGEN. <http://www.mol-net.de/software/stergen/index.html> (accessed July 2006).
18. TAUTOMER. <http://www.mol-net.de/software/tautomer/index.html> (accessed July 2006).
19. Halgren TA. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J Comp Chem* 1995;17:553–586.

20. Gasteiger J, Marsili M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett* 1978;34:3181–3184.
21. Gasteiger J, Saller H. Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept. *Angew Chem Int Ed Engl* 1985;24:687–689.
22. Dalby A, et al. Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited. *J Chem Inf Comput Sci.* 1992;32: 244–255.
23. Hammett LP. *Physical Organic Chemistry*, 1st and 2nd eds. New York: McGraw Hill; 1940, 1970. Hammond GS. A correlation of reaction rates. *J Am Chem Soc* 1955;77:334–338.
24. Hansch C, et al. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Hoboken, NJ: Wiley; 1979. Leo A, et al. Partition coefficients and their uses. *Chem Rev* 1971;71:525–554.
25. ROTATE. <http://www.mol-net.de/software/rotate/index.html> (accessed July 2006).
26. Example of energy minimization algorithm. <http://www.chemsoc.org/exemplarchem/entries/pkirby/exemchem/mmintro.html> (accessed July 2006).
27. Weininger D, inventor; Daylight Chemical Information Systems, Inc, assignee. Method and apparatus for designing molecules with desired properties by evolving successive populations. US patent US5434796, 1995.
28. Zauhar RJ, et al. Shape Signatures: a new approach to ligand- and receptor-based molecular design. *J Med Chem.* 2003;46:5674–5690.
29. Nagarajan K, et al. Enrichment of ligands for the serotonin receptor using the Shape Signatures approach. *J Chem Inf Model* 2005;45:49–57.
30. Meek PJ, et al. Shape Signatures: speeding up computer aided drug discovery. *Drug Discov Today* 2006;11:895–904.
31. Nordling E, Homan E. Generalization of a targeted library design protocol: application to 5-HT7 receptor ligands. *J Chem Inf Comput Sci* 2004;44:2207–2215.
32. Milton RC, et al. Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity. *Science* 1992;256:1403–1404. Erratum, *Science* 1992;257(Jul):147.
33. Thalidomide toxicity information. <http://www.k-faktor.com/thalidomide> (accessed July 2006).
34. CONCORD. http://www.tripos.com/index.php?family=modules,SimplePage,,&page=sybyl_concord (accessed July 2006).
35. Barnard JM, et al. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J Mol Graph Model* 2000;18:452–463.
36. Molecular Operating Environment (MOE). <http://www.chemcomp.com/software.htm> (accessed July 2006).
37. James CA, Weininger D. *Daylight Theory Manual*. Daylight Chemical Information Systems, Inc; 1995.
38. Smith EG, et al. The Wiswesser line-formula chemical notation (WLN). Cheery Hill, NJ: Chemical Information Management; 1975.
39. MACCS-II. San Leandro, CA: MDL Limited; 1992.
40. Dröse S, et al. Semisynthetic derivatives of concanamycin A and C, as inhibitors of V- and P-type ATPases: structure–activity investigations and developments of photoaffinity probes. *Biochemistry* 2001;40:2816–2825.

41. Boyd MR, et al. Discovery of a novel antitumor benzolactone enamide class that selectively inhibits mammalian vacuolar-type (H^+)-atpases. *J Pharmacol Exp Ther* 2001;297:114–120.
42. Farina C, et al. Novel bone antiresorptive agents that selectively inhibit the osteoclast V- H^+ -ATPase. *Farmaco* 2001;56:113–116.
43. Gagliardi S, et al. 5-(5,6-Dichloro-2-indolyl)-2-methoxy-2,4-pentadiamides: novel and selective inhibitors of the vacuolar H^+ -atpase of osteoclasts with bone antiresorptive activity. *J Med Chem* 1998;41:1568–1573.
44. Gagliardi S, et al. Synthesis and structure–activity relationships of bafilomycin A1 derivatives as inhibitors of vacuolar H^+ ATPase. *J Med Chem* 1998;41:1883–1893.
45. Nadler G, et al. (2Z,4E)-5-(5,6-dichloro-2-indolyl)-2-methoxy-N-(1,2,2,6,6-pentamethylpiperidin-4-yl)-2,4-pentadienamides, a novel, potent and selective inhibitor of the osteoclast V-ATPase. *Bioorg Med Chem Lett* 1998;8:3621–3626.
46. Chen D, et al. Holographic QSAR of selected esters. *Chemosphere* 2004;57:1739–1745.
47. Zhang H, et al. CoMFA, CoMSIA, and molecular hologram QSAR studies of novel neuronal nAChRs ligands—open ring analogues of 3-pyridyl ether. *J Chem Inf Model* 2005;45:440–448.
48. Zhu W, et al. QSAR analyses on ginkgolides and their analogues using CoMFA, CoMSIA, and HQSAR. *Bioorg Med Chem* 2005;13:313–322.
49. Castilho MS, et al. Two- and three-dimensional quantitative structure–activity relationships for a series of purine nucleoside phosphorylase inhibitors. *Bioorg Med Chem* 2006;14:516–527.
50. Honorio KM, et al. Hologram quantitative structure–activity relationships for a series of farnesoid X receptor activators. *Bioorg Med Chem Lett* 2005;15:3119–3125.
51. Doddareddy MR, et al. Hologram quantitative structure–activity relationship studies on 5-HT₆ antagonists. *Bioorg Med Chem* 2004;12:3815–3824.
52. Pungpo P, et al. Hologram quantitative structure–activity relationships investigations of non-nucleoside reverse transcriptase inhibitors. *Curr Med Chem* 2003;10:1661–1677.
53. Lohray BB, et al. 3D QSAR studies of N-4-arylacryloylpiperazin-1-yl-phenyl-oxazolidinones: a novel class of antibacterial agents. *Bioorg Med Chem Lett* 2006;16:3817–3823.
54. Desai PV, et al. Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. *J Med Chem* 2006;49:1576–1584.
55. Allan GM, et al. Modification of estrone at the 6, 16, and 17 positions: novel potent inhibitors of 17 β -hydroxysteroid dehydrogenase type 1. *J Med Chem* 2006;49:1325–1345.
56. Eros D, et al. Reliability of log *P* predictions based on calculated molecular descriptors: a critical review. *Curr Med Chem* 2002;9:1819–1829.
57. Goller AH, et al. *In silico* prediction of buffer solubility based on quantum-mechanical and HQSAR- and topology-based descriptors. *J Chem Inf Model* 2006;46:648–658.
58. Ermondi G, et al. A combined *in silico* strategy to describe the variation of some 3D molecular properties of beta-cyclodextrin due to the formation of inclusion complexes. *J Mol Graph Model* 2006;25:296–303.
59. Clark M. Generalized fragment–substructure based property prediction method. *J Chem Inf Model* 2005;45:30–38.
60. Votano JR, et al. New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Mol Divers* 2004;8:379–391.

61. Holm R, Hoest J. Successful *in silico* predicting of intestinal lymphatic transfer. *Int J Pharm* 2004;272:189–193.
62. Lobell M, Sivarajah V. *In silico* prediction of aqueous solubility, human plasma protein binding and volume of distribution of compounds from calculated pK_a and $A \log P_{98}$ values. *Mol Divers* 2003;7:69–87.
63. Introduction of Principal Components Analysis (PCA). <http://www.statsoft.com/textbook/stfacan.html> (accessed July 2006).
64. Introduction of Partial Least Squares (PLS). <http://www.statsoft.com/textbook/stpls.html> (accessed July 2006).
65. CATALYST. <http://www.accelrys.com/products/catalyst> (accessed July 2006).
66. Accelrys, Inc. <http://www.accelrys.com/products> (accessed July 2006).
67. Brzozowski AM, et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 1997;389:753–758.
68. Natesh R, et al. Crystal structure of the human angiotensin-converting enzyme–lisinopril complex. *Nature* 2003;421:551–554.
69. Kamenska V, et al. The COREPA approach to lead generation: an application to ACE-inhibitors. *Eur J Med Chem* 1999;34:687–699.
70. Tzakos AG, Gerathanassis IP. Domain-selective ligand-binding modes and atomic level pharmacophore refinement in angiotensin I converting enzyme (ACE) inhibitors. *Chem Biochem* 2005;6:1089–1103.
71. Sutherland JJ, et al. Pruned receptor surface models and pharmacophores database searching. *J Med Chem* 2004;47:3777–3787.
72. Jones G, et al. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 1997;267:727–748.
73. Cramer RD III, et al. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 1988;110:5959–5967.
74. CoMFA. http://www.tripos.com/index.php?family=modules,SimplePage,,,&page=sybyl_qsar_with_comfa
75. Sheng C, et al. Structure-based optimization of azole antifungal agents by CoMFA, CoMSIA, and molecular docking. *J Med Chem* 2006;49:2512–2525.
76. Menezes IR, et al. Three-dimensional models of non-steroidal ligands: a comparative molecular field analysis. *Steroids* 2006;71:417–428.
77. Demyttenaere-Kovatcheva A, et al. Identification of the structural requirements of the receptor-binding affinity of diphenolic azoles to estrogen receptors alpha and beta by three-dimensional quantitative structure–activity relationship and structure–activity relationship analysis. *J Med Chem* 2005;48:7628–7636.
78. Aguirre G, et al. New potent 5-nitrofuryl derivatives as inhibitors of *Trypanosoma cruzi* growth. 3D-QSAR (CoMFA) studies. *Eur J Med Chem* 2006;41:457–466.
79. Dixon S, et al. QMQSAR: utilization of a semiempirical probe potential in a field-based QSAR method. *J Comput Chem* 2005;26:23–34.
80. Kearsley SK, Smith GM. An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput Methodol* 1990;3:615–633.
81. Korhonen SP, et al. Comparing the performance of FLUFF-BALL to SEAL-CoMFA with a large diverse estrogen data set: from relevant superpositions to solid predictions. *J Chem Inf Model* 2005;45:1874–1883.
82. Melani F, et al. Field interaction and geometrical overlap: a new simplex and experimental design based computational procedure for superposing small ligand molecules. *J Med Chem* 2003;46:1359–1371.

83. Goodford PJA. Computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985;28:849–857. (GRID, version 19; Molecular Discovery Ltd, Mayfair, London, England, 2001.)
84. Collantes ER, et al. Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. *Anal Chem* 1996;68:2038–2043.
85. Gallardo-Godoy A, et al. Sulfur-substituted alpha-alkyl phenethylamines as selective and reversible MAO-A inhibitors: biological activities, CoMFA analysis, and active site modeling. *J Med Chem* 2005;48:2407–2419.
86. Thaimattam R, et al. 3D-QSAR CoMFA, CoMSIA studies on substituted ureas as Raf-1 kinase inhibitors and its confirmation with structure-based studies. *Bioorg Med Chem* 2004;12:6415–6425.
87. Iskander MN, et al. Optimization of a pharmacophore model for 5-HT4 agonists using CoMFA and receptor based alignment. *Eur J Med Chem* 2006;41:16–26.
88. Kamath S, Buolamwini JK. Receptor-guided alignment-based comparative 3D-QSAR studies of benzylidene malonitrile tyrophostins as EGFR and HER-2 kinase inhibitors. *J Med Chem* 2003;46:4657–4668.
89. Datar PA, Coutinho EC. A CoMFA study of COX-2 inhibitors with receptor based alignment. *J Mol Graph Model* 2004;23:239–251.
90. Abraham DJ, Leo AJ. A program has recently become available that provides a functionality long missing from the molecular modeling world. HINT is a program that performs approximations of atom-based hydrophobicity based on the hydrophobic fragment work of Hansch and Leo. *Proteins* 1987;2:130–152.
91. Kellogg GE, et al. HINT—a new method of empirical hydrophobic field calculation for CoMFA. *J Comput Aided Mol Des* 1991;5:545–552.
92. Huey R, et al. Olson grid-based hydrogen bond potentials with improved directionality. *Lett Drug Des Discov* 2004;1:178–183.
93. Kim KH, et al. Use of the hydrogen bond potential function in a comparative molecular field analysis (CoMFA) on a set of benzodiazepines. *J. Comput Aided Mol Des* 1993;7:263–280.
94. Stuti G, et al. CoMFA and CoMSIA studies on a set of benzyl piperazines, piperadines, pyrazinopyridoindoles, pyrazinoisoquinolines and semi rigid analogs of diphenhydramine. *Med Chem Res* 2004;13:746–757.
95. Korhonen SP, et al. Improving the performance of SOMFA by use of standard multivariate methods. *SAR QSAR Environ Res* 2005;16:567–579.
96. Cavalli A, et al. Linking CoMFA and protein homology models of enzyme-inhibitor interactions: an application to non-steroidal aromatase inhibitors. *Bioorg Med Chem* 2000;8:2771–2780.
97. Moitessier N, et al. Combining pharmacophore search, automated docking, and molecular dynamics simulations as a novel strategy for flexible docking. Proof of concept: docking of arginine-glycine-aspartic acid-like compounds into the alphavbeta3 binding site. *J Med Chem* 2004;47:4178–4187.
98. Zauhar RJ, et al. Shape Signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J Med Chem* 2003;46:5674–5690.
99. Wang CY, et al. Identification of previously unrecognized antiestrogenic chemicals using a novel virtual screening approach. *Chem Res Toxicol* 2006;19:1595–1601.
100. Zhang Q, et al. Discovery of novel triazole-based opioid receptor antagonists. *J Med Chem* 2006;49:4044–4047.
101. Zeman SP. Charting chemical space: finding new tools to explore biology. *Nature* 1–3. The 4th Horizon Symposium, Black Point Inn, Maine, USA, 20–22 May 2004.

102. Liu X, et al. *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*; 2003, pp 1–8.
103. Osada R, et al. Matching 3D models with shape distributions. *Shape Model Int* 2001;May:154–166.
104. Ohbuchi R, et al. Shape-similarity search of 3D models by using enhanced shape functions. *Int J Comput Appl Tech* 2005;23:70–85.
105. Ankerst M, et al. Nearest neighbor classification in 3D protein databases. In *Proceedings of the 7th Conference on Intelligent Systems for Molecular Biology (ISMB'99)*; 1999, pp 34–43.
106. Amin HM, et al. Inhibition of JAK3 induces apoptosis and decreases anaplastic lymphoma kinase activity in anaplastic large cell lymphoma. *Oncogene* 2003;22: 5399–5407.
107. Jilek RJ, et al. Lead hopping method based on topomer similarity. *Chem Inf Comput Sci* 2004;44:1221–1227.
108. Klebe G. Comparative molecular similarity indices: CoMSIA. In *Kubinyi H, Folkers G, Martin YC, Eds. 3D QSAR in Drug Design*. London: Kluwer Academic Publishers; 1998, Vol 3, p 87.

3

PROTEIN–PROTEIN INTERACTIONS

KAMALJIT KAUR, DIPANKAR DAS, AND MAVANUR R. SURESH

University of Alberta, Edmonton, Alberta, Canada

Contents

- 3.1 Introduction
- 3.2 Protein–Protein Interactions and Human Pathogenesis
 - 3.2.1 Oncogenesis
 - 3.2.2 Pathogen–Host Interaction
 - 3.2.3 Loss of Normal Protein–Protein Interaction
- 3.3 Screening of Protein–Protein Interaction Inhibitors
 - 3.3.1 Structure–Activity Relationship
 - 3.3.2 Genetic Screening Systems and Phage Display
 - 3.3.3 Yeast Two Hybrid System and Intracellular Antibodies
- 3.4 Inhibitors of Protein–Protein Interactions
 - 3.4.1 Peptide and Peptidomimetic Inhibitors
 - 3.4.2 Small-Molecule Inhibitors
 - 3.4.3 Molecules Containing Porphyrin or Peptidocalixarene Scaffolds
- 3.5 Conclusion
- References

3.1 INTRODUCTION

Living organisms are almost exclusively comprised of four classes of molecules, namely, proteins, nucleic acids, polysaccharides, and lipids. Of these, barring lipids, all other classes can be regarded as macromolecules that are built from a limited number of building blocks or monomers. In the case of proteins, such building blocks are amino acids. Proteins are formed by polymerization of essentially twenty

“standard” amino acids. Yet, the myriad of proteins and their diverse functions, ranging from basic metabolism to structural and reproductive functions, can be astounding and constitute the very basis of life on Earth. For instance, an *Escherichia coli* bacterium contains over 4000 different proteins participating in virtually every life sustaining function of the cell.

The Greek root of the word protein, *proteios*, meaning *of first importance*, identifies the paramount role of this class of macromolecules in eukaryotes. It is perhaps discernible that twenty amino acids can be combined in different manners to yield virtually innumerable proteins, with a variety of functions. It is, however, interesting to note that even a slight alteration of the amino acid sequence can significantly alter the structure and function of a protein molecule. A well known example is the modification of hemoglobin in sickle cell anemia. Normal hemoglobin (HbA) contains a Glu at the sixth position of each β -chain, which is replaced by Val in the sickle cell hemoglobin (HbS). This single substitution causes aggregation of HbS into stiff filaments, leading to the deformation of the red blood cells into elongated “sickled” shapes, and the consequent symptoms of sickle cell anemia.

Being macromolecules, proteins predominantly interact with other molecules, including other proteins, via weak long-range interactions. These are also referred to as nonbonded interactions and are of two types, van der Waals and electrostatic. As two interacting proteins approach very close to each other, stronger specific bonds can be formed between them. For instance, the strongest specific interaction between two proteins occurs in the form of a covalent bond in the so-called disulfide linkage. Hydrogen bonds, less strong than covalent bonds, are more common forms of specific interactions exhibited by proteins. The four common types of interactions observed are listed in Fig. 3.1. A second factor that modifies long-range interactions is the tertiary structure of proteins. Proteins or peptides with defined secondary structural elements like helices, strands, and coils fold into three-dimensional arrangements, which give them a tertiary structure. For example, an α -helix, a β -sheet, and a coil region in the three-dimensional solution structure of an antibacterial peptide are highlighted in Fig. 3.2. Sequences with less than 50 amino acids are generally considered as peptides and more than 50 amino acids are called proteins.

Protein-protein interactions are key to several biological pathways and thus are attractive targets for therapeutic intervention. Such approaches are frequently based on a sound assessment of the strong and weak interactions and the protein’s secondary or tertiary structures. The modulation or disruption of protein-protein inter-

Interactions	
Long-Range/Weak	Short-Range/Strong
Electrostatic (5 kcal/mol)	Covalent bonds (40 kcal/mol)
van der Waals (< 1 kcal/mol)	Hydrogen bonds (3-7 kcal/mol)

FIGURE 3.1 Long-range and short-range interactions between biomolecules. The energy values in parentheses are approximate.

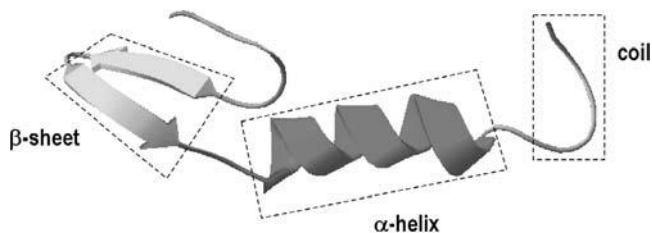


FIGURE 3.2 Tertiary structure of an antibacterial peptide (leucocin A) displaying secondary structure elements, namely, α -helix, β -sheet, and coil. The figure was generated from the 1CW6 Protein Data Bank coordinates [1].

actions has been difficult owing to their large surface areas of interaction. For several extracellular protein complexes, antibodies and other proteins have been identified as successful antagonists. The main rationale for this being that large macromolecules can readily disrupt the interaction between two proteins. From a cursory glance, this appears to be an attractive therapeutic option. However, a closer inspection of the problem indicates that large proteins are generally not orally bioavailable or cell permeable. Thus, these cannot be particularly effective for targeting intracellular protein-protein interactions. In light of this, one might presume that small molecules will provide more attractive therapeutic intervention options. However, such small molecules have been less successful in this regard, since they cannot provide specific recognition needed for a large protein surface. Protein surfaces do not often present deep indentations for small molecules to bind, and affinity is achieved through summing up several weak interactions.

Recent studies report several protein complexes as targets for drug design, with some of these targets amenable to small molecule inhibition [2-7]. Here we review the interactions between some of the important protein-protein pairs, followed by the recent successes in developing peptides, peptidomimetics, or small organic molecules as inhibitors of these interactions. This field is still in its infancy as most of the compounds identified are still in preclinical stages. Recent developments made in this broad field that have pharmaceutical and clinical implications will be discussed in this chapter.

3.2 PROTEIN-PROTEIN INTERACTIONS AND HUMAN PATHOGENESIS

3.2.1 Oncogenesis

Interaction between specific regions in a protein has been found to be essential for all stages of development and homeostasis. Subsequently, many human diseases occur by either loss of essential protein-protein interaction or through the formation of a protein complex at an inappropriate time or location. Several such interactions have been found to be responsible for the onset of oncogenesis and have been well studied [4, 8]. In the following, interactions between some of the well known protein pairs that lead to oncogenesis are discussed.

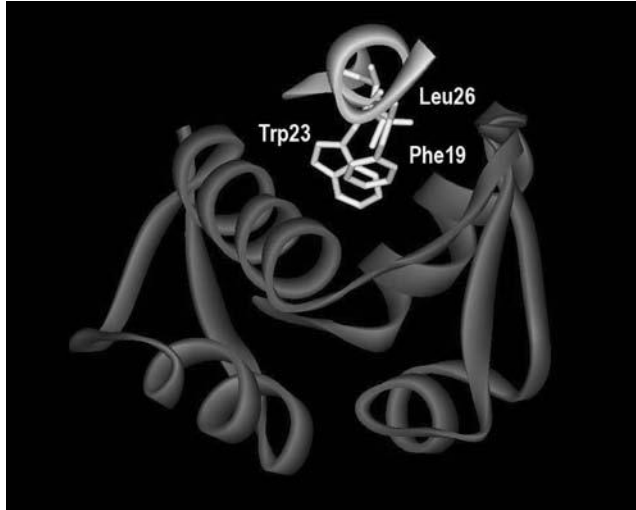


FIGURE 3.3 The ribbon representation of the MDM2-p53 complex (PDB entry 1YCR [10]) displaying the hydrophobic cleft of MDM2 where the p53 peptide binds as an amphipathic α -helix. The hydrophobic side chains of Phe19, Trp23, and Leu26 of p53 inserting deep into the MDM2 cleft are shown.

HDM2-p53 The tumor suppressor protein p53 is involved in the maintenance of the genomic integrity of the cell. It coordinates the cellular response to DNA damage by binding to specific DNA sequences and activating genes responsible for growth arrest or apoptosis. The inactivation of p53 by the binding of the cellular oncoprotein HDM2 (human double minute 2) has been identified as an important step in tumorigenesis [9]. In normal cells, HDM2 and p53 form a negative feedback loop that helps to limit the growth-suppressing activity of p53. The identification of this key negative regulator HDM2 provided a great opportunity to manipulate the levels of the tumor suppressor p53 in cancer cells.

The mouse homolog MDM2 binds the N terminus of p53, thereby interfering with the transcriptional ability of p53. The crystal structure of the amino terminal domain of MDM2 bound to a small region on the N terminus of p53 displayed the specific interaction between a hydrophobic cleft in the MDM2 protein and an amphipathic α -helix of p53 [10]. Several van der Waals contacts augmented by two intermolecular hydrogen bonds were found at the interface. The MDM2 cleft lined with hydrophobic and aromatic amino acids interacts with the hydrophobic face of p53 amphipathic helix. As shown in Fig. 3.3, the side chains of Phe19, Trp23, and Leu26 from the p53 α -helical region nestle deep inside the hydrophobic pocket of MDM2. Since residue Leu22 and Trp23 of p53 are critical for its transcriptional activity, p53 is rendered transcriptionally inactive after binding to MDM2. Inhibitors of the MDM2-p53 interaction have thus been found as attractive targets to gain activity of p53 in tumor cells [11].

Bcl-2/Bcl-X_L-BH3 The proteins in the Bcl-2 family regulate apoptosis by maintaining a fine balance between the pro- and antiapoptotic proteins within the cell

[12–14]. Proapoptotic members of Bcl-2 family such as Bax, Bak, Bid, and Bad, and antiapoptotic members such as Bcl-2 and Bcl-X_L exist as homodimers or mixed heterodimers. The nature of dimerization between these proteins dictates how a cell will respond to an apoptotic signal. Antiapoptotic proteins, Bcl-2, Bcl-X_L, or both, are overexpressed in the majority of human cancers and may play a vital role in cancer development. Therefore, inhibition of Bcl-2/Bcl-X_L activity is gaining recognition for the development of potent therapeutics as anticancer drugs. Several strategies have been employed to target these proteins, including inhibition of expression levels using antisense oligonucleotides and identification of peptide ligands that affect protein–protein association [15, 16].

The antiapoptotic function of Bcl-2 and Bcl-X_L is partially attributed to their ability to heterodimerize with proapoptotic members and antagonize their proapoptotic function. Three regions of the antiapoptotic proteins, namely, the Bcl-2 homology 1 (BH1), BH2, and BH3 binding sites, participate in their death-inhibiting activity and heterodimerization with the proapoptotic protein. However, only the BH3 binding site of the Bcl-2 and Bcl-X_L proteins is found critical for the ability to antagonize apoptosis. Small, truncated peptides derived from the BH3 region of Bak (16mer) and Bad (25mer) have been found necessary and sufficient both for promoting cell death and binding to Bcl-X_L [17, 18]. Furthermore, a synthetic cell permeable BH3 peptide derived from proapoptotic Bad has been shown to induce apoptosis *in vitro* and to have *in vivo* activity in human myeloid leukemia growth in severe combined immunodeficient mice [19].

The NMR solution structures of Bcl-2 and Bcl-X_L alone [20, 21] and Bcl-X_L in complex with Bak or Bad BH3 peptides [17, 18] have provided detailed information about the binding interactions of these proteins to the BH3 peptides (Fig. 3.4). The three-dimensional structures illustrate the formation of a hydrophobic cleft (the BH3 binding site) by the three BH domains of Bcl-2 and Bcl-X_L in which the Bak or Bad BH3 domain binds. The overall binding motif of Bcl-X_L/Bak and Bcl-X_L/Bad

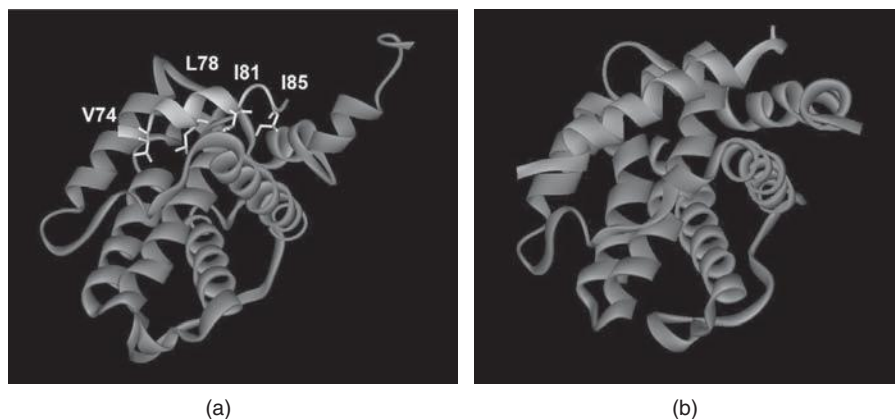


FIGURE 3.4 The ribbon representation of the NMR solution structures of Bcl-X_L in complex with (a) Bak BH3 peptide and (b) Bad BH3 peptide derived from the PDB codes 1BXL [17] and 1G5J [18], respectively. Bak and Bad BH3 peptides bind in the hydrophobic cleft of the Bcl-X_L protein. The hydrophobic side chains Val74, Leu78, Ile81, and Ile85 of Bak BH3 peptide inserting into the Bcl-X_L cleft are highlighted as a stick model.

complex was found to be very similar with only a few differences at the protein-peptide interface. The Bcl-X_L/Bak structure shows that the hydrophobic side chains of the Bak peptide (Val74, Leu78, Ile81, and Ile85) point into the hydrophobic cleft (BH3 binding site) of Bcl-X_L and stabilize complex formation. Several electrostatic interactions between the oppositely charged residues of Bcl-X_L (Glu129, Arg139, and Arg100) and the Bak peptide (Arg76, Asp83, and Asp84, respectively) were also present. Similar interactions were also found between the Bcl-X_L and 25 mer Bad peptide. However, the longer Bad peptide makes additional contact at the two ends of the BH3 binding site, forming a tighter complex (K_d 0.6 nM) as compared to the Bcl-X_L/Bak 16 mer complex (K_d 480 nM). These observations suggest that the BH3 binding pocket of Bcl-X_L as well as Bcl-2 is essential for its antiapoptotic function, and small molecules that bind to the BH3 binding pocket of Bcl-X_L/Bcl-2 can block the interaction between Bcl-X_L/Bcl-2 and proapoptotic proteins such as Bak, Bax, and Bad.

XIAP-Caspase Inhibitors of apoptosis proteins (IAPs) are important but incompletely understood negative regulators of apoptosis. Among other mechanisms, IAPs selectively bind and inhibit caspases-3, -7, and -9, but not caspase-8. Currently, there are eight members of the IAP family. Of these members, X-linked inhibitor of apoptosis protein (XIAP) is upregulated in many cancers and has thus garnered the most attention as a drug discovery target [22].

XIAP is a 57 kDa protein with three zinc-binding baculovirus IAP repeat domains (BIR1-3) and a really interesting new gene (RING)-finger that binds and inhibits caspases with nanomolar affinity (Fig. 3.5). The BIR2 domain inhibits caspases-3 and -7, whereas BIR3 domain inhibits caspase-9. The function of the BIR1 domain has not yet been determined. The RING finger contains an E3 ubiquitin ligase. The proapoptotic protein SMAC is an endogenous human IAP antagonist that binds and inhibits XIAP, thereby releasing caspases and reactivating apoptosis. Structural studies map the interaction between XIAP and SMAC and provide a basis for the development of small molecule XIAP inhibitors. These studies demonstrate that SMAC binds to both the BIR3 and BIR2 domains of XIAP. Crystal structure of BIR3 domain of XIAP in complex with caspase-9 [23] and SMAC [24] illustrates the key interactions between the complexes. The N terminus of the small subunit of caspase-9 binds the same shallow groove on BIR3 as the N terminus of SMAC (Fig. 3.6). The N-terminal 4-7 amino acids of active SMAC are necessary and sufficient for binding the BIR3 pocket of XIAP and preventing XIAP from binding and inhibiting caspase-9. A 4 mer peptide, Ala-Val-Pro-Ile, derived from SMAC binds to XIAP with ~500 nM affinity [25]. These results indicate that small molecules that

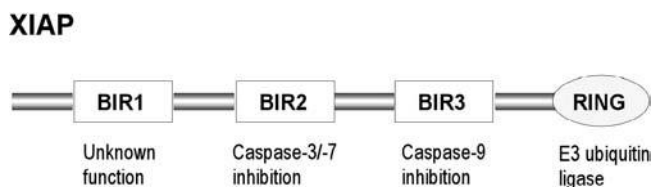


FIGURE 3.5 Schematic representation of XIAP showing different domains and their functions.

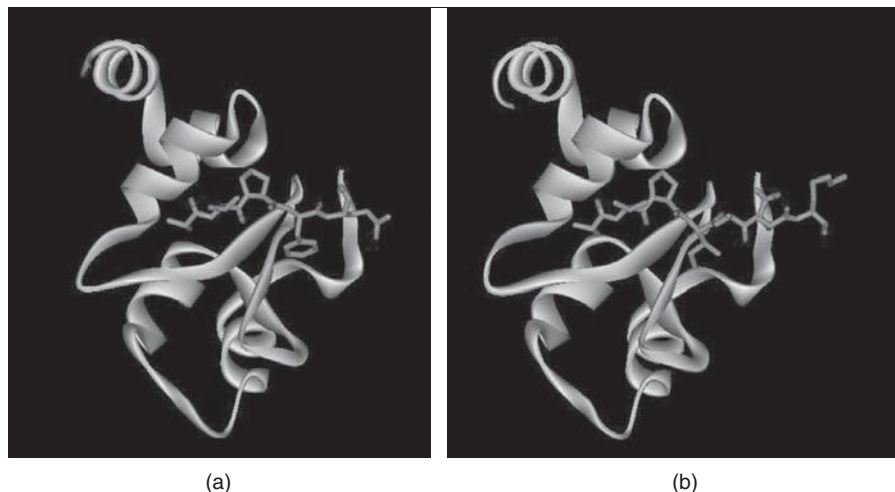


FIGURE 3.6 The three-dimensional structures of XIAP-BIR3 domain in complex with (a) N terminus of the small subunit of caspase-9 (Ala316, Thr317, Pro318, Phe319, and Gln320 shown as stick model) and (b) N terminus of SMAC (Ala1, Val2, Pro3, Ile4, Ala5, Gln6, and Lys7 shown as stick model). Both caspase-9 and SMAC bind in the same pocket of BIR3 as illustrated by the crystal structures of the complexes (PDB accession codes 1NW9 [23] and 1G73 [24], respectively).

mimic the actions of SMAC could be identified, providing an opportunity for a structure-based approach to the design of BIR3 inhibitors.

Stat3 Dimerization Signal transducer and activator of transcription 3 (Stat3) is a cytoplasmic transcription factor that is activated in response to cytokines and growth factors. Upstream regulators of Stat3 constitute JAKs, Src, and EGF receptors and downstream targets include antiapoptotic and cell cycling genes such as Bcl-X_L and cyclin D1. Stat3 is overactivated in a surprisingly large number of cancers including head and neck, breast, brain, prostate, lung, leukemia, multiple myeloma, lymphoma, and pancreas and has therefore been identified as a potential target for cancer drug development [26].

Stat3 is composed of several domains, namely, an oligomerization domain (N terminal), a coiled coil domain, a DNA-binding domain, a linker domain, a Src homology 2 (SH2) domain, a critical tyrosine at position 705 (C-terminal end), and a C-terminal transactivation domain. On activation, JAK kinase-2 phosphorylates the tyrosine residues of its coreceptor, thereby facilitating the binding of Stat3 to specific phosphotyrosine residues of JAK-2 through its SH2 domain. This leads to phosphorylation of Tyr705 on the C terminus of Stat3, followed by Stat3 dimerization by the reciprocal interaction between the SH2 domain of one monomer and the phosphorylated tyrosine of the other. The activated dimers translocate to the nucleus, where they bind to specific DNA sequences and activate gene expression. These observations have pointed out Stat3 inhibition, such as inhibition of JAK-2/Stat3 interaction and Stat3 dimerization, as a novel molecular target for the advancement of a broad anticancer therapy. The X-ray elucidation of three-dimensional

structure of the Stat3 homodimer bound to DNA reveals some details about the interaction between the two monomers and may facilitate the development of Stat3 inhibitors [27].

Rac1-GEF Rho family GTPases, such as Rac1, control signaling pathways that are involved in cell adhesion, cell migration, and other cellular processes. Overexpression or upregulation of Rho GTPases has been discovered in many human tumors, including colon, breast, lung, myeloma, and head and neck squamous-cell carcinoma [28]. They can be activated through specific interaction with guanine nucleotide exchange factor (GEF) proteins that catalyze the exchange of GDP for GTP. One strategy to control tumor spreading is the selective inhibition of Rho GTPase activation by its GEF TrioN or Tiam1. Trio is a large multifunctional domain molecule with the amino-terminal module (TrioN) displaying the Rac1-specific GEF activity. Similarly, Tiam1, the T-cell invasion and metastasis gene product of the Dbl family, is shown to be an active GEF for Rac1. The three-dimensional structures of GEF-Rho protein complexes discern the specific interactions between GEFs and Rho GTPases needed for the signaling specificity mediated by Rho proteins. The cocrystal structure of Rac1/Tiam1 complex [29] shows that a domain of Tiam1, mainly dominated by α -helices, binds a shallow groove of Rac1, suggesting the presence of a small-molecule binding site (Fig. 3.7). A micromolar inhibitor of Rac1/TrioN interaction that selectively inhibits Rac1/Tiam1 and Rac1/TrioN versus related complexes and inhibits Rac1 activation in cells has been reported [30]. This study indicates that inhibition of the Rac1-GEF protein-protein interaction is possible, and such interactions have cellular consequences.

Integrin $\alpha_v\beta_3$ - Fibronectin Another well studied example of cell adhesion proteins is the integrins. Integrins are the cell-surface receptors that act as molecular recognition sites for other proteins (for cell-cell and cell-extracellular matrix inter-



FIGURE 3.7 Crystal structure of Rac1 (left) in complex with the guanine nucleotide exchange region of Tiam1 (right) determined by Worthylake et al. [29] (PDB code 1FOE). The extensive interface of the complex buries over 3000 \AA^2 of primarily hydrophobic surface area.

actions) as well as signaling molecules transferring ligand-binding information to the cytoplasm. Integrins are heterodimeric proteins consisting of α and β subunits and typically have a high molecular mass of ~ 300 kDa. At least, 25 $\alpha\beta$ integrin heterodimers have been reported and six of them are currently being evaluated in clinical trials for cancer [31]. Integrin $\alpha_v\beta_3$ has received particular attention as a potential target for anticancer drug design. The expression of $\alpha_v\beta_3$ is significantly increased on vascular cells in human tumors, but is weakly expressed on normal or quiescent endothelial cells. Since this integrin is relatively limited in its normal distribution, inhibition of its action is considered as an effective means of depriving tumors of nascent blood vessels without involving normal tissues.

Integrin recognition of the extracellular matrix ligands, such as fibronectin, collagen, and vitronectin, relies on the concerted binding of both the α and β subunits to regions of the ligand containing the Arg-Gly-Asp (RGD) sequence [32]. The RGD motif was the first integrin binding motif discovered. Several new motifs have been found since then that bind to a specific class of integrins [33]. Studies on RGD sequence have led to the discovery of cyclic pentapeptide Arg-Gly-Asp-{D-Phe}-{*N*-methyl-Val} or cyclo RGDf{NMe}V that specifically binds and inhibits integrin $\alpha_v\beta_3$ [34].

3.2.2 Pathogen-Host Interaction

Some viruses and bacteria enter eukaryotes by attachment to specific cell-surface receptors or cell-surface receptor-binding proteins. Viruses infect higher eukaryotes to reproduce themselves, whereas bacterial pathogens invade primarily to gain protection against the host immune system. Pathogens have always “enjoyed” invading human cells and have coevolved with their hosts to enable efficient entry, replication, and exit during their infectious cycles. An excellent review by Dimitrov [35] describes in depth the different virus entry mechanisms at the molecular level and opportunities for therapeutic intervention by inhibiting these processes. In this section, interaction between specific proteins of virus or bacteria and the target cell that facilitates pathogen entry into the target cell are discussed.

Papillomavirus E2 Protein Infection by papillomavirus causes benign lesions that can lead to cervical cancer and other tumors [36, 37]. Papillomaviruses are small DNA viruses that infect higher eukaryotes by invading the basal layer of epithelial cells where they replicate successfully. Viral E2 protein has been found essential for replication and survival. E2 protein contains two conserved domains, the C-terminal viral DNA binding domain and the N-terminal transactivation domain that binds the viral E1 protein. Molecules that can bind these two domains of E2, thereby inhibiting the E2/DNA or E1/E2 interaction, are attractive targets for the development of therapeutics to prevent or treat papillomavirus infections. The three-dimensional structure of E1 bound to E2 reveals some important contact points between the complex [38]. The interaction surface, comprised of three helices from the N-terminal domain of E2, buries ~ 940 Å² surface area per protomer on E1-E2 complex formation (Fig. 3.8).

HCV-Envelope Protein 2 Hepatitis C virus (HCV) infection, another important target for antiviral drug design, causes severe medical problems, including chronic hepatitis, cirrhosis, and hepatocellular carcinoma. HCV genome is composed of a



FIGURE 3.8 The X-ray structure of papillomavirus E1 helicase (upper structure) in complex with its molecular partner E2 (lower structure). PDB accession code: 1TUE [38].

single-stranded positive sense RNA of approximately 9600 nucleotides that are translated into a polyprotein precursor of about 3000 amino acids. The HCV polyprotein precursor is processed by host and viral proteases to yield structural and nonstructural proteins, which are essential for replication and assembly of new viral particles. The viral envelope E2 protein initiates the infection by association with specific cell-surface receptor(s). Many groups have demonstrated that the truncated soluble versions of E2 bind specifically to hepatocytes [39, 40]. This glycoprotein is found to interact with CD81, scavenger receptor class B type 1 (SR-B1), and dendritic cell-specific intracellular adhesion molecule 3-grabbing nonintegrin (DC-SIGN). Such findings suggest that these proteins may act as receptors for HCV on the cell surface. Therefore, inhibition of interaction between E2 and the cell-surface receptors, such as CD81, has been identified as a possible target for designing anti-HCV molecules [41, 42].

SARS-Angiotensin Receptor SARS-CoV is a member of the Coronaviridae, a family of positive strand RNA virus that have long been known to cause severe acute respiratory syndrome in many animals and more recently in humans. Similar to other known coronaviruses, SARS-CoV is an enveloped virus containing four structural proteins, namely, the membrane (M), envelope (E) glycoprotein, spike (S) glycoprotein, and nucleocapsid (NP) proteins [43]. The spike protein of SARS-CoV is a large type I glycoprotein and is made up of two domains, the S1 near the N terminus and the S2 near the C terminus. Unlike other coronaviruses, the spike protein of SARS-CoV is not posttranslationally cleaved in virus producing cells. The S1 and S2 domains form the globular head and the stalk of the spike protein and play an important role in specific receptor recognition and cell fusion. The S1 domain mediates receptor association whereas the S2 domain is membrane associated and likely undergoes structural rearrangements. This conformational change initiates the

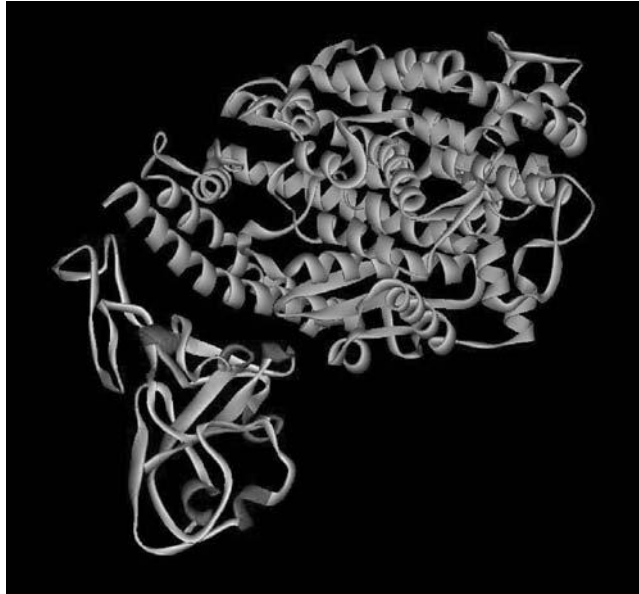


FIGURE 3.9 Crystal structure of SARS-CoV spike protein RBD (lower ribbon structure) in complex with human receptor ACE2 (upper structure). PDB accession code 1AJF [47].

fusion of the virus and host cell membrane, allowing for entry of the virus. The first step in viral infection is the binding of viral proteins to certain host cell receptors. The spike protein of coronavirus is considered as the main site of viral attachment to the host cells. It has been demonstrated that a metallopeptidase, angiotensin converting enzyme 2 (ACE2) isolated from Vero E6 cells, efficiently binds the S1 domain of the SARS-CoV spike protein [44]. A discrete receptor binding domain (RBD) of the spike protein has been defined at residues 318–510 of the S1 domain [45] and this receptor binding domain is the critical determinant of virus receptor interaction and thus of viral host range and tropism. It has been demonstrated that this RBD binds ACE2 with higher affinity than does the full length S1 domain [46]. The crystal structure (Fig. 3.9) of SARS-CoV RBD complexed with ACE2 receptor at 2.9 Å shows that the RBD presents a gentle concave surface, which cradles the N-terminal lobe of the peptidase [47].

Bacterial Fibronectin-Binding Proteins One of the mechanisms by which bacterial pathogens invade cells is by displaying fibronectin-binding proteins (FBPs) on their surface. This approach to internalize into the host cell has been adopted by some pathogenic gram-positive bacteria. FBPs contain tandem arrays of intrinsically disordered repeat sequences that bind fibronectin–integrin complexes. The NMR solution structure of a complex comprising a peptide fragment of a streptococcal fibronectin-binding protein bound to the first two domains of human fibronectin reveals the tandem β -zipper interactions between the two fragments [48]. The tandem β -zipper is created by the β -strand conformation of the repeat sequences of FBP peptide fragment that extends existing antiparallel β -sheets in both directions upon binding to fibronectin. The binding affinity of these complexes is relatively weak but

increases significantly when additional domains of both proteins are present. For example, the binding affinity (K_A) of two FBP repeats to pairs of fibronectin domains is $\sim 10^6 M^{-1}$. The tandem β -zipper interaction is a common phenomenon found in several pathogenic gram-positive bacteria, such as *Staphylococcus aureus*, *Streptococcus pyogenes*, and *Borrelia burgdorferi*, and may prove to be a widespread mechanism for bacterial foray of host cells [49–51]. Molecules that disrupt these β -zipper interactions may prove to be useful therapeutics for bacterial infections.

3.2.3 Loss of Normal Protein-Protein Interaction

Modular protein-protein interactions mediated by the tandem β -zipper have also been observed in eukaryotes [52, 53]. The LIM domains, found only in eukaryotes, are proteins with diverse functions such as transcription factors and protein kinases [54]. They are known to mediate specific protein-protein interactions through their LIM domains. Human genome encodes four LIM-only (LMO) and 12 LIM-homeodomain (LIM-HD) proteins each with a pair of tandem LIM domains at their N terminus (Fig. 3.10a). Three out of four LMO proteins have been implicated in oncogenesis. The LIM domains of all LMO and LIM-HD proteins bind the LIM domain-binding protein, Ldb1, through the 30 residue LIM interaction domain (LID) of this protein. Ldb1 is a ubiquitously expressed protein that contains an N-terminal dimerization domain, LID, and several other binding domains and is an

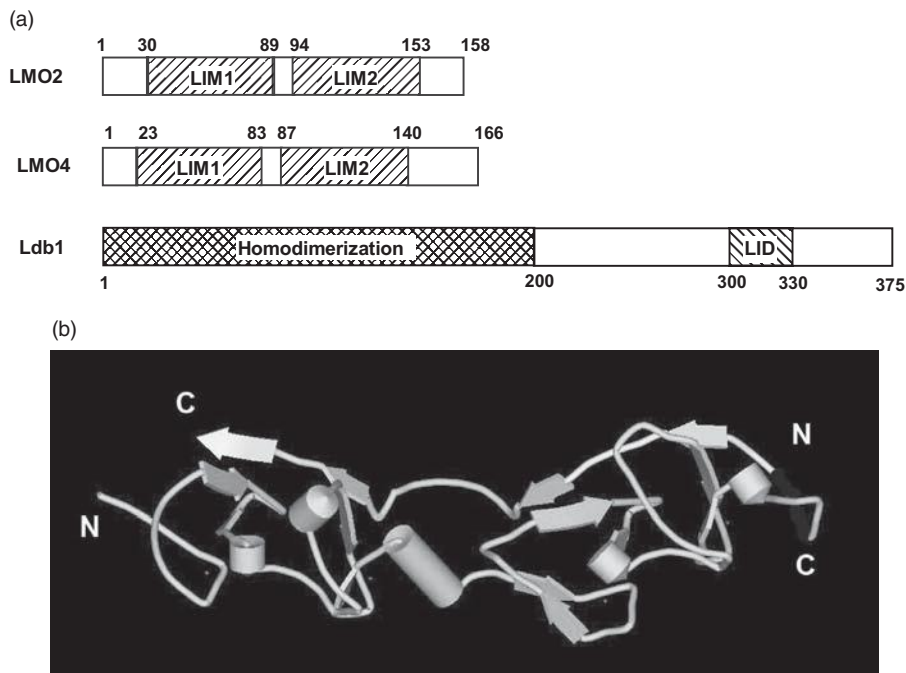


FIGURE 3.10 (a) Schematic representation of different domains of LMO2, LMO4, and Ldb1 proteins. (b) The schematic of LMO4 in complex with the Ldb1-LID domain displaying the “tandem β -zipper” interaction (PDB code 1RUT) [52]. The β -strand conformation of the peptide extends the existing β -structure in the partner protein in a modular fashion.

essential cofactor that plays diverse roles in the development of complex organisms. Since LMO proteins bind the same region of Ldb1 (LID) as the LIM-HD, LMO proteins can regulate the transcriptional activity of LIM-HD by competing for binding to Ldb1. The displacement of endogenous LMO4 by ectopically expressed LMO2, as the normal binding partner for Ldb1, has been directly linked with the overexpression of LMO2 in T cells and the onset of T-cell acute lymphoblastic leukemia in children [55].

The three-dimensional structures of the complexes comprising LIM domains and Ldb1-LID present that the intrinsically disordered Ldb1-LID forms a “tandem β -zipper” upon binding to LIM domains (Fig. 3.10b) [52]. In the complex, the four β -strands of Ldb1-LID extend across one face of LMO4 and remain in continuous contact with the LIM domains, mainly by backbone-backbone hydrogen bonds, burying a total surface area of 3800 \AA^2 . Ldb1 is known to interact with LMO and LMO-HD proteins only and not with any other LIM domain containing proteins. Recent studies [52, 53], highlight the specific features of the interaction between Ldb1 and the LMO2 or LMO4 proteins. Structural, mutagenesis, and yeast two hybrid analysis are used to identify the key binding determinants for the complex formation. The differences in the binding interaction between the two protein complexes suggest that molecules that could bind specifically to LMO2 or LMO4 may have potential uses in the treatment of neoplastic disorders.

The protein complexes discussed above are a few examples of interaction pairs that have been identified as possible drug design targets where the interactions have been mapped at the atomic level. A wide variety of approaches are utilized in the identification of these protein-protein interaction pairs and their inhibitors. Some of the most popular approaches are discussed in the following section.

3.3 SCREENING OF PROTEIN-PROTEIN INTERACTION INHIBITORS

Several approaches have been utilized in the identification of protein-protein interaction inhibitors with the aim of developing therapies for a variety of human diseases [6, 56, 57]. An analysis of current strategies employed for the identification of lead molecules demonstrates that a search for competitors of a known binder is the basis of traditional screening as well as more modern approaches. In the following sections, some of these approaches are described in detail.

3.3.1 Structure-Activity Relationship

A common approach relies on the experimentally determined (NMR or X-ray) structure of the protein complex. In this approach, one attempts to disrupt the interfacial interactions between the two proteins by developing mimics of the interface amino acid residues (peptide fragment) for one of the binding partners. The structure of the interface peptide fragment is modified using computer docking and molecular dynamics simulations to obtain peptidomimetics or small organic molecules [3]. The resulting peptides or peptide analogues present the interacting functional groups in similar spatial orientations as the interface amino acid residues. Peptidomimetic or small molecule inhibitors have higher affinity, better selectivity, and often better pharmacokinetic properties than the parent peptide. A second

approach involves screening of a huge library of compounds using computer docking to find molecules with high affinity toward one of the binding partners of the protein complex. The lead structures identified using the above procedures are further optimized for potency and selectivity by structure-activity relationship (SAR) studies [58]. The SAR methods utilize experimental techniques, like nuclear magnetic resonance (NMR) spectroscopy, or computation methods, such as docking studies, and provide a structural perspective throughout the discovery and optimization of a lead molecule.

3.3.2 Genetic Screening Systems and Phage Display

The new library methodologies, such as phage display, allow generation of a large number of molecules with a fast screening and selection procedure to identify the most interesting lead candidates. Phage display technology has proved to be a very powerful *in vitro* technique for generating libraries containing millions of different peptides, proteins, or small molecules. Using the same technique, these libraries have been screened to identify ligands for peptide receptors, to define epitopes for monoclonal antibodies, to select substrates for enzymes, and to screen cloned antibody repertoires [59, 60].

In the phage display technique, filamentous virus is used as a platform for cloning of a DNA library (a library of genes or gene segments) encoding millions of variants of certain ligands into the phage genome and is fused to the gene encoding the phage coat or tail protein. Upon expression in the *E. coli* host in the presence of helper phage, the fusion protein (e.g., Coat protein-scFv) is incorporated into new phage particles that are assembled in the periplasmic space of the bacterium. Expression of the target gene fusion product and its subsequent incorporation into the mature phage coat results in the ligand being presented on the phage surface, while its genetic material resides within the phage genome. The proteins that are encoded by the library are expressed on the surface of phage and can be selected on the immobilized target molecule by biopanning. This interaction allows selection of high affinity binders for a variety of biomedical applications. Phages that bind the target molecule contain the gene for the protein and have the ability to replicate while nonadherent phages are washed away. This method can be used to efficiently clone genes encoding proteins with particular binding characteristics. In antibody phage display, the Fab or single chain fragment of IgG variable proteins is displayed on phage. This approach for antibody development offers advantages over immunization of animals and hybridoma technology [61]. Phage display can produce antibodies more quickly in a cost effective manner than traditional approaches. Additionally, antibody phage display techniques can potentially isolate antibodies to molecules that are not immunogenic in animals due to tolerance mechanism. Phage selection is not limited to the isolation of antibodies or short peptides. This approach has also been instrumental in studies and manipulation of a variety of other biologically active molecules and their designer variants [62].

3.3.3 Yeast Two Hybrid System and Intracellular Antibodies

Cell-based assays that monitor the intracellular behavior of target molecules, rather than binding or catalytic activity of purified proteins, are also being used in high

throughput screening of protein interaction inhibitors. These assays offer an opportunity to discover entirely new classes of compounds, molecules that act primarily by modulating protein interactions in living cells.

The yeast two hybrid system is a cell-based genetic selection assay that has been successfully used to identify protein-protein interactions *in vivo*. The model originally developed by Fields and Song [63] exploits the fact that transcription factors are comprised of two domains, a DNA binding domain and a transactivation domain. As an example, GAL4 protein of yeast (*Saccharomyces cerevisiae*) is a transcriptional activator required for the expression of genes encoding enzymes of galactose utilization. The native GAL4 protein contains two domains: an N-terminal domain that binds to specific DNA sequences but fails to activate transcription; and the C-terminal acidic domain that is necessary to activate transcription but cannot initiate function without the N-terminal domain. The basic strategy of the two hybrid system involves two proteins of interest that are expressed as two different fusion proteins. One fusion protein, known as the bait, is fused to the DNA binding domain to bind at specific sites upstream of the reporter gene. The second fusion protein, known as prey, is fused to the transactivation domain. If a physical interaction occurs between the two proteins, it brings the GAL4 domain in sufficient proximity to activate the GAL4-dependent transcription of a reporter gene. There will be no expression of the reporter gene if the two proteins do not interact in the intracellular milieu.

The two hybrid system may not be a useful tool for all protein-protein interactions. The limitation of the technique includes where the protein of interest is able to initiate GAL4-dependent transcription. Toxicity of the expressed protein or misfolding of the chimeric protein inside the cell might result in a limited activity or inaccessibility of binding site to the other protein. Furthermore, some protein-protein interactions depend on posttranslational modification (S-S bond, glycosylation, and phosphorylation) that may not appropriately occur in yeast. Two hybrid systems need the fusion protein to be targeted to the yeast nucleus and it might be a disadvantage for extracellular proteins. Weak and transient interactions are often the most interesting in signaling cascades. These are more rapidly detected in the two hybrid system in view of the significant amplification of the reporter gene in this system.

Intracellular antibodies are antibody fragments that are targeted and expressed inside the cells for interaction with cellular target antigens. This strategy can inhibit the regular function or in some cases mediate cell killing following antigen binding. Specific activity of certain intracellular proteins has been blocked by microinjection of full length antibodies [64, 65] or of hybridoma mRNA [66, 67] into the cytoplasm of various cell types. Recent advances in DNA technology and antibody engineering have allowed the development of specific, high affinity antibodies to target antigens. These probes could be targeted intracellularly as unique nontoxic therapeutics. Recombinant antibody reductants that provide many of the essential features of antibodies are suitable forms to be expressed *in vivo* or internalized efficiently inside the cells. The recombinant single chain Fv fragment (scFv) has been the most widely used for intracellular antibodies [68]. Intracellular single domain antibodies have also been isolated from yeast libraries with good antigen binding affinities [69].

The first step of intracellular antibody isolation is the derivation of the V regions of the heavy and light chains of a high affinity monoclonal antibody against a

target antigen. The VH and VL sequences could be amplified by RT-PCR of mRNA isolated from the hybridoma cells, assembled and cloned as a scFv [70]. Alternatively, one of the *in vitro* display systems, such as phage display [60], yeast display [71], or ribosome display [72] techniques, could be employed to generate the scFv libraries from immunized mouse spleen total mRNA and screen scFv libraries with the desired antigen to select specific scFv clones. Intracellular antibody capture technology [73, 74] has also been developed for *in vivo* screening of scFv libraries for a target antigen. This involves *in vitro* biopanning of diverse scFv libraries developed by phage display, followed by *in vivo* screening of antigen-antibody interaction using the yeast two hybrid system. The coding sequences of the antigen are cloned in one of the two hybrid vectors expressing the GAL4 DNA binding domain-antigen fusion protein. The coding sequence of scFv is cloned in the other two hybrid vector resulting in expression of GAL4 activation domain-scFv fusion protein. Yeast cells cotransformed with both the vectors will result in the expression of the reporter gene if the antigen and scFv interact with each other. There will be no reporter gene expression if the antibody fragment does not functionally interact with the antigen inside the yeast cells. By this technology it is possible to select and isolate intracellular antibodies, which could widely interact with the target protein inside the cells to alter or affect the protein function. Such induced intracellular protein-protein interactions could be an efficient pathogen neutralizing strategy for several viral and bacterial diseases [68, 75].

3.4 INHIBITORS OF PROTEIN-PROTEIN INTERACTIONS

As mentioned earlier, a large number of protein interaction complexes are emerging as potential targets for developing therapeutic agents. However, a big portion of these are ruled out at the onset due to the intricacies involved at the interaction site such as innate mutations and the atomic details of the binding site. The binding site may not present particular indentations, or if a pocket is present, its dimensions may be too small, or its geometry may be too shallow. Such features do not support tight binding of a drug-like molecule. Some of the above issues can be handled and a drug can be produced by generating antibodies against the target. In fact, therapeutic antibodies, including chimeric, humanized, and multivalent antibodies, and antibody fragments have been utilized in several instances and comprise over 30% of biopharmaceuticals currently undergoing clinical trials [76]. Several monoclonal antibodies against growth factors or their receptors are found effective in the treatment of solid tumors [77]. Antibodies tend to bind their targets with both high affinity and specificity and therefore block protein-protein interactions efficiently. However, antibodies are incompatible for intracellular targets, encountering problems such as poor delivery due to their relatively large size and lower stability of their disulfide-bonded structure in the reducing environment of the cell. Peptide inhibitors provide a much smaller substitute for *in vitro* inhibition of protein-protein interactions but are often not stable *in vivo* to be successful drugs. More stable variants, such as crosslinked peptides, peptide mimetics, and small molecule inhibitors, may prove to be better blocking agents for both intracellular and extracellular protein-protein interactions.

3.4.1 Peptide and Peptidomimetic Inhibitors

Specific recognition needed for a large protein surface seeks at least $\sim 6 \text{ nm}^2$ area buried at the interface. The unique spatial distribution of the charged, polar, and hydrophobic residues at the interface are deemed important for recognition. Despite being conceptually simple, mimetics of the large interfacial area required for specific recognition remains a challenging endeavor. Nonetheless, steady progress has been made in the discovery of compounds that mimic protein surface and function.

A variety of peptide inhibitors have been reported over the last decade for blocking the MDM2-p53 association [11]. These peptides have been helpful in mapping out the interaction between the two proteins. The key interactions between MDM2 and p53 involve a relatively small area represented by three amino acids, namely, Phe19, Trp23, and Leu26 of p53. Optimization of peptides has led to the discovery of several low nanomolar inhibitors of MDM2 that have recently been reviewed by Fotouhi and Graves [11]. Peptide mimetics have also been explored in order to increase the metabolic and proteolytic stability over α -peptide inhibitors. Schepartz and colleagues [78] targeted the HDM2-p53 interaction with a 14-helical structure made of beta amino acids to display the functional groups of Phe19, Trp23, and Leu26 in the same spatial orientation as found in p53. These synthetic β^3 -peptides exhibited significant helical character in aqueous buffer and one of the oligomers, **1** (Fig. 3.11), selectively inhibited HDM2 interaction with nanomolar affinity. Similarly, Hamilton and associates [79] have utilized terphenyl scaffold to mimic one face of α -helical peptides. Substitution of the three ortho positions of the scaffold projected one side of the molecule analogous to the i , $i + 4$, and $i + 7$ residues of an α -helix. A terphenyl derivative with three hydrophobic side chains, compound **2** (Fig. 3.11), was found to bind specifically at the p53 binding site of HDM2 and exhibited a K_i of 182 nM. More recently, α -helical peptidomimetics with the terphenyl scaffold and more soluble terephthalamide scaffold inhibited the Bak BH3-Bcl-X_L interactions in the low micromolar range [80, 81].

Verdine and co-workers [82] targeted the BID/ Bcl-X_L interaction by synthesizing hydrocarbon stapled helices to mimic the amphipathic α -helix BH3 domain of BID. These molecules, for example, **3** (Fig. 3.11), with constraint helix became proteolytically stable, cell permeable, and bound Bcl-X_L with nanomolar affinity. In a cell-based assay, these compounds induced apoptosis and in an *in vivo* experiment inhibited the growth of human leukemia xenografts. Gellman and co-workers [83] generated chimeric ($\alpha/\beta + \alpha$)-peptides that mimic the α -helical display of BH3 domain of Bak. These peptides are tight binders and therefore potent inhibitors ($K_i = 0.7 \text{ nM}$) of Bak/ Bcl-X_L interaction.

RGD motif present in the extracellular matrix ligands, such as fibronectin, was the first integrin binding motif identified. Since then, development of RGD mimetics that bind selectively to a single integrin has been a subject of intense research [33]. Studies have primarily focused on four integrins— $\alpha_4\beta_1$, $\alpha_5\beta_1$, $\alpha_v\beta_3$, and $\alpha_{IIb}\beta_3$ —that bind RGD containing ligands and are thought to have the most clinical significance. An important RGD containing molecule that emerged out of these efforts is cyclic RGDf{NMe}V pentapeptide [34]. This cyclic peptide specifically inhibits integrin $\alpha_v\beta_3$ with an IC₅₀ of 0.6 nM and is in Phase II clinical trials as an anticancer drug under the name cilengitide. The crystal structure (Fig. 3.12) of the cyclic peptide bound to the extracellular segment of $\alpha_v\beta_3$ integrin in the presence of Mn²⁺ metal

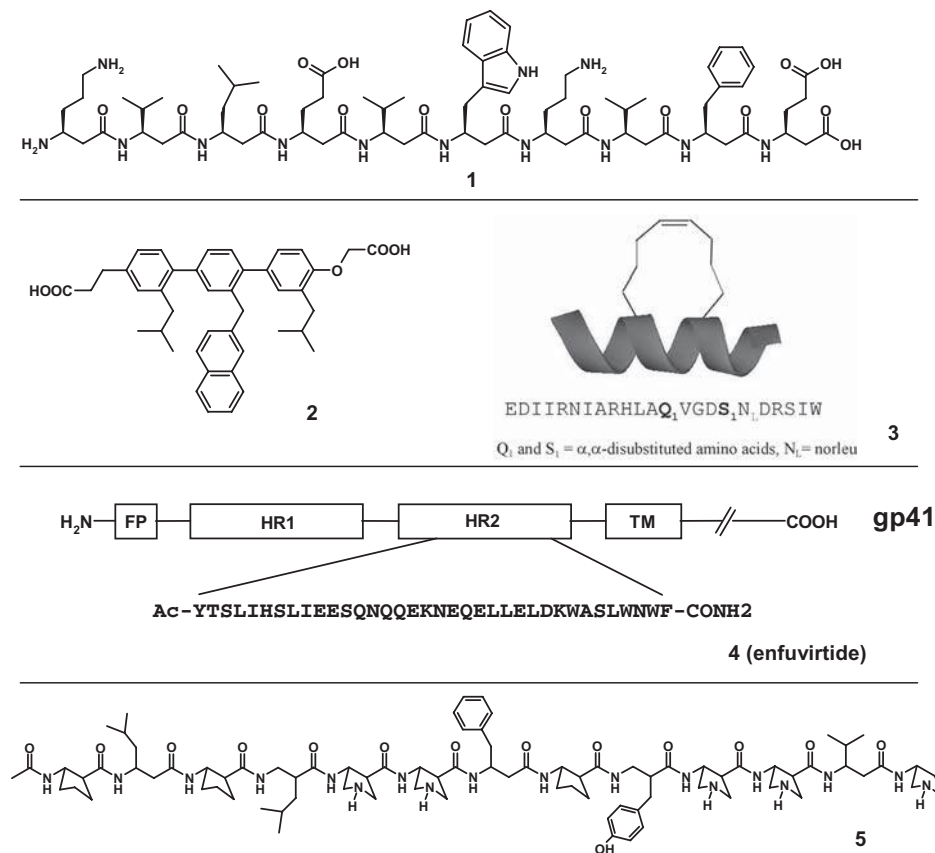


FIGURE 3.11 Structure of peptide or peptidomimetic inhibitors of protein-protein interactions.

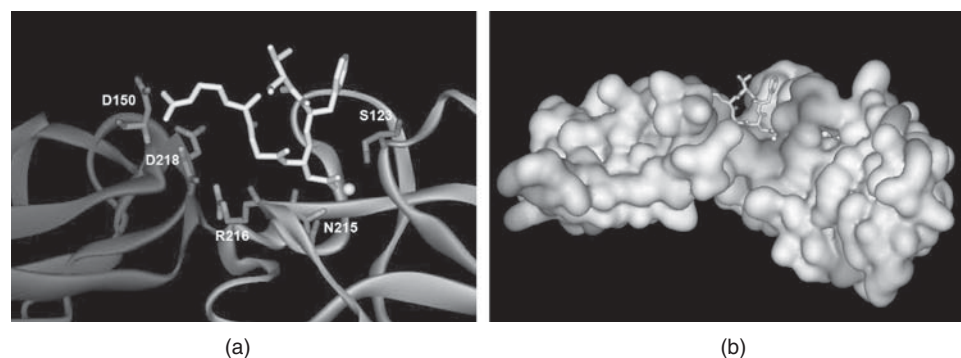


FIGURE 3.12 (a) Crystal structure of cyclo RGDf[NMe]V peptide (stick model) bound in the active site of integrin $\alpha_v\beta_3$ (PDB accession code 1L5G [84]). The Asp side chain carboxyl of the cyclic peptide interacts with one of the Mn^{2+} . The interacting residues (stick model) from integrin $\alpha_v\beta_3$ are derived from both the α and β subunits. (b) Surface representation of the two subunits of integrin $\alpha_v\beta_3$ bound to the cyclic peptide (stick model).

cation reveals important binding interactions in the complex [84]. The peptide binds at the major interface between the α_v and β_3 subunits burying about 45% (355 \AA^2) of its total surface area. The RGD sequence makes the main contact with the integrin subunits. The availability of the complex structure has prompted search of nonpeptide antagonists of $\alpha_v\beta_3$ integrin using detailed computer docking experiments.

Peptides have also been exploited in the design of antiviral agents [85]. A successful example in this case is the HIV antiviral drug Fuzeon (**4**), enfuvirtide, or T-20, Fig. 3.11) [86]. Fuzeon is a 36 amino acid α -peptide that was introduced in 2003 into the clinics as a potent HIV entry inhibitor [87]. HIV entry into the target cells takes place in several steps, beginning with the binding of viral envelope protein gp120 to CD4 receptors on the target cells, followed by the exposure of the buried transmembrane fusion protein gp41 and conformational changes for the assembly of the hexameric six-helix bundle gp41 that allows the fusion to take place. As depicted in Fig. 3.11, the gp41 is made of an N-terminal fusion peptide (FP), two heptad repeat regions HR1 and HR2, and a transmembrane region. Fuzeon (**4**) is a small peptide derived from the HR2 region of gp41 that competitively inhibits the last step of the viral fusion. Peptidomimetics of the HR2 region would serve as good alternate inhibitors of the HIV fusion, perhaps with better pharmacokinetic profile than the α -peptide. Hamilton and co-workers [88] have utilized the terphenyl derivatives to mimic the helical HR2 domain. The most potent molecule with hydrophobic side chains at the ortho position of the three phenyl rings efficiently inhibits HIV-1 infection in a cell fusion assay (IC_{50} 15.7 $\mu\text{g/mL}$). Furthermore, *in vivo* studies of these molecules as anti-HIV agents are in progress.

Nozaki et al. [89] discovered that a small peptide fragment from the milk glycoprotein human lactoferrin is able to block the entry of hepatitis C virus (HCV) particles into the hepatocytes. Virus entry inhibitors act extracellularly by blocking the binding of the virus to the host cell and this process of shielding the virus from attachment to the target cells seems more facile compared to targeting other intracellular sites that require exacting precision. The authors demonstrated that the mechanism of action of this peptide fragment is by binding to the E2 protein of HCV, thereby blocking its entry into the host cell, rather than binding to the host cell-surface receptors. Peptide mimics of this 33 amino acid fragment that can resist proteolysis and are metabolically stable will be of great clinical interest, as there is no vaccine for HCV and current therapeutic strategies yield roughly 40% response rates. English et al. [90] have attempted to construct peptidomimetic entry inhibitors of human cytomegalovirus (HCMV). The authors have prepared and tested several β -peptides, oligomers of β -amino acid, as entry inhibitors of HCMV. The most potent β -peptide, **5** (Fig. 3.11), inhibited HCMV infection in a cell-based assay with an $IC_{50} \approx 30 \mu\text{M}$.

The above examples reveal the prospect of peptides and peptidomimetics as potential therapeutics for various diseases. Currently, there are more than 40 marketed peptides worldwide and about 670 peptides are in either clinical or advanced preclinical phases. Several different classes of peptidomimetics are also entering the preclinical stages. Peptidomimetics like β -peptides, peptoids, and azapeptides are receiving particular attention as these unnatural oligomers, also called foldamers, fold into a conformationally ordered state in solution, like the natural biopolymers. These unnatural oligomers do not have disadvantageous peptide characteristics

and therefore may generate viable pharmaceuticals. They are protease resistant, resistant to metabolism, and may have reduced immunogenicity relative to peptide analogues.

3.4.2 Small-Molecule Inhibitors

A broad range of screening initiatives has helped identify several protein-protein complexes that are amenable to inhibition by small molecules. Several compounds have been identified that help characterize proteins such as MDM2, Bcl-2, and XIAP as drug targets. Additionally, small-molecule antagonists have recently been described for several new targets, including Rac1-Tiam1, β -catenin-T cell factor, and Sur-2-ESX. Several of these small molecule protein-protein inhibitors are virtually at the threshold of becoming therapeutics.

Fotouhi and Graves [11] have reviewed some interesting new scaffolds and leads as MDM2 inhibitors. Among several reported molecules, only a series of compounds termed Nutlins (**6**, Fig. 3.13) possessed *in vivo* activity and therefore drug-like properties [58]. The Nutlins with a core imidazoline mimic the α -helical structure of the p53 backbone. The three aryl rings on the imidazoline are presented in the same space as the side chains of Phe19, Trp23, and Leu26 of p53. Furthermore, it was demonstrated by 2D NMR spectroscopy that they bind to MDM2 at the p53-binding site. The specific interaction of Nutlins with MDM2 indeed translated to the selective growth inhibition of cells containing wild-type p53 ($IC_{50} \approx 1.5 \mu\text{M}$) and showed 10–20-fold selectivity for cells with active versus mutated p53. Compound **6** was well tolerated, orally bioavailable, and inhibited the growth of an MDM2-overexpressing tumor in mice. It achieved a high steady-state concentration during the study, indicating good pharmacokinetic properties.

Bad and Bak proteins bind to Bcl-2 and Bcl-X_L by inserting ~20 residue long α -helical Bad/Bak BH3 peptide into a hydrophobic groove [91]. Isolated BH3-like peptides also bind in this groove, suggesting the existence of a small-molecule binding site. There has been significant progress in developing compounds that bind in this groove on Bcl-2 and/or Bcl-X_L and thereby augment cell death [15, 16]. Recently, several molecules, such as **7** [92], **8** [93], and **9** [94] (Fig. 3.13), have been developed and are moving into clinical trials. GX15-070, a small-molecule inhibitor from Gemin X, is specifically designed to inhibit all of the antiapoptotic members of the Bcl-2 protein family and is the first such small-molecule inhibitor tested in clinical trials. Phase I clinical trials of GX15-070 in patients with refractory solid tumors and lymphomas showed promising results, advancing GX15-070 into Phase II clinical trials.

Small molecules that inhibit the BIR domains of XIAP have been found to be promising candidates for the development of therapeutic XIAP inhibitors [4, 22]. Molecules have been developed to specifically target BIR2 and BIR3 regions of XIAP. This is due to the difference in the mechanism of caspase inhibition by the BIR2 and BIR3 domains and their ability to inhibit different caspases. The structural data available for the interaction between the BIR3 domain of XIAP and caspase-9 suggests that small molecules binding the BIR3 pocket of XIAP could mimic the action of SMAC and inhibit the interaction between XIAP and caspase-9. These structural studies have facilitated several research groups toward the discovery of cell-active ligands for XIAP. Tripeptide inhibitors, such as **10** and **11** (Fig. 3.13) with

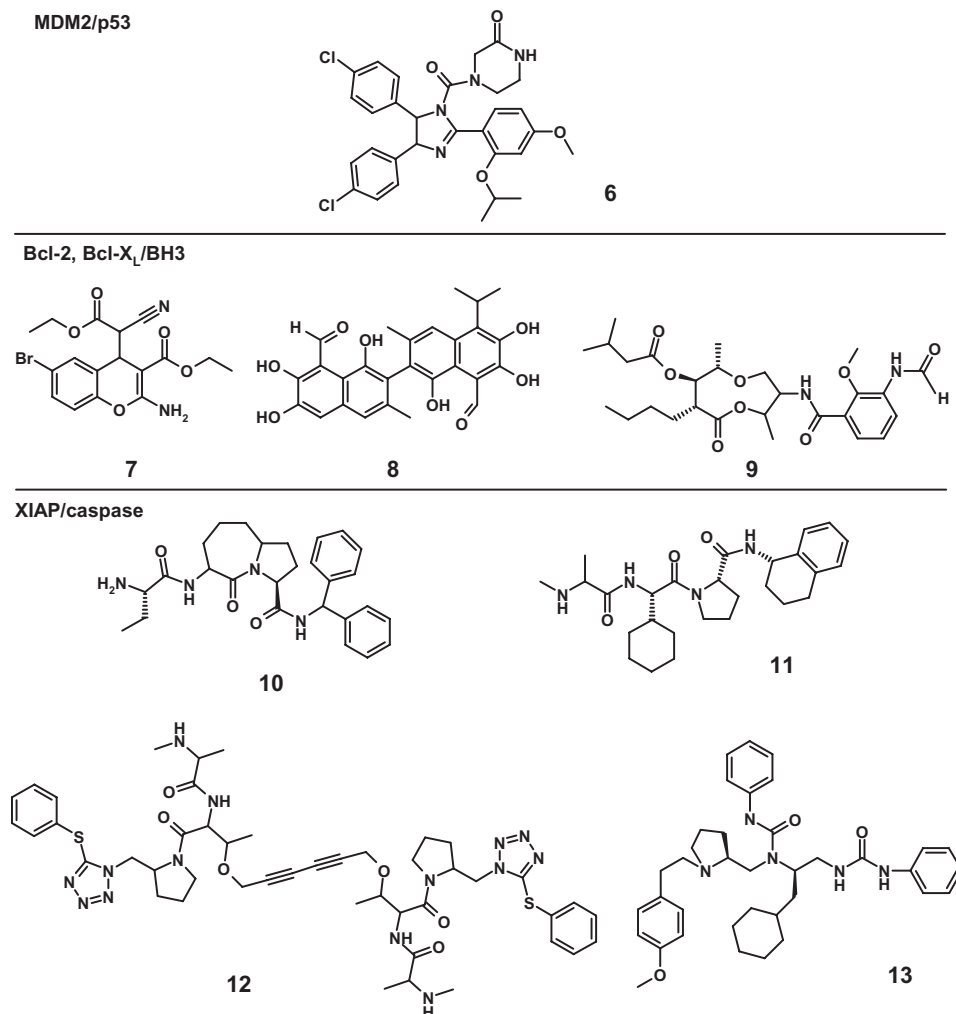


FIGURE 3.13 Small-molecule inhibitors of protein-protein interactions in cancer.

unnatural amino acids were identified and tested for their binding to the BIR3 domain of XIAP by NMR spectroscopy and fluorescent polarization assay [95–97]. These molecules bound to the BIR3 domain at SMAC binding site with nanomolar affinity. Unlike **10**, the BIR3 ligand **11** could activate apoptosis in a caspase-dependent manner in the absence of additional stimuli. Li et al. [98] used computer-aided drug design to mimic SMAC peptide. Lead compound identified included a tetrazoyl thioether moiety and was modified to form a C₂-symmetric diyne. The compound **12** (Fig. 3.13) bound the BIR3 domain of XIAP with an affinity similar to SMAC peptides and also bound cIAP-1 and cIAP-2 in cells. The proposed bivalent binding mechanism of **12** to XIAP resembled the wild-type SMAC, interacting simultaneously with the BIR2 and BIR3 domains of XIAP. SMAC is a dimer and interacts with the BIR2 and BIR3 domains at the same time to inhibit XIAP.

Compounds with polyphenylurea pharmacophore were identified by screening a large combinatorial library for activation of caspase-3 in the presence of XIAP [99]. Compound **13** (Fig. 3.13) was found to be selective for caspases-3 and -7 over caspase-9 and did not inhibit the SMAC-XIAP interaction. The polyphenylurea inhibitors were toxic to a wide spectrum of malignant cell lines and demonstrated preferential toxicity to primary malignant cells over normal cells. In xenograft models, these compounds were found to delay the growth of tumors of the prostate, breast, and colon carcinoma cells without any unpleasant toxicity to the mice. The above examples of XIAP inhibitors clearly suggest additional studies are required to discern the feasibility of small-molecule XIAP inhibitors as potential therapeutics. However, the data already point to XIAP as an interesting target for therapy.

The strategy of inhibiting protein-protein interactions with a small molecule has also been applied to the design of antiviral and antibacterial compounds. White et al. [100] reported a small-molecule inhibitor of papilloma virus E1-E2 dimerization. The inandione inhibitor **14** (Fig. 3.14) binds to the hydrophobic pocket of the E2 protein as shown by the cocrystal structure of the complex (Fig. 3.15) [101]. A second weakly bound inandione molecule was also observed in the crystal structure, suggesting the presence of additional binding region on the E2 protein for exploiting

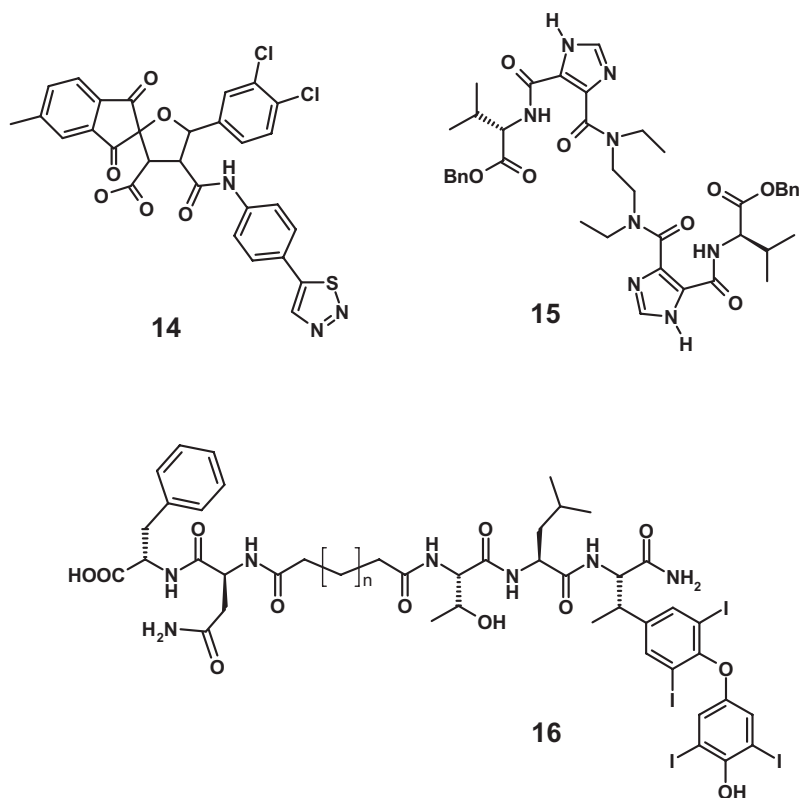


FIGURE 3.14 Small-molecule inhibitors of papilloma virus E1 – E2 heterodimerization (**14**), HCV-E2 – CD81 interaction (**15**), and HIV-1 protease dimerization (**16**).



FIGURE 3.15 The structure of a small-molecule inhibitor, an inandione bound to papilloma virus E2 protein (PDB accession code 1R6N [101]). The inhibitor binding to E2 prevents E1–E2 heterodimerization and disrupts viral replication. A second weakly bound inhibitor (top) suggests that additional functional groups could be added to the side chains of the inhibitor to gain binding affinity.

inhibitor design. Todd and colleagues [41] developed molecules like **15** (Fig. 3.14), with a novel bis-imidazole scaffold, as mimics of helix D of CD81 that reversibly inhibited binding of HCV-E2 to CD81 receptor protein.

Inhibition of HIV-1 protease dimerization is a promising strategy for anti-HIV drug design as opposed to the active-site directed inhibitors. In this regard, Chmielewski and co-workers [102] discovered a nanomolar inhibitor (**16**, $K_i = 71$ nM, Fig. 3.14) of HIV-1 protease dimerization using a focused library approach. More importantly, the potent molecules of this class were equally active against wild-type and a mutant form of the enzyme. The mutant enzyme was resistant to active-site directed inhibitors, suggesting the importance of alternative drug design strategy.

3.4.3 Molecules Containing Porphyrin or Peptidocalixarene Scaffolds

Protein surface recognition by molecular architectures such as porphyrin and calixarene scaffolds has also been utilized in several instances. Hamilton and co-workers [103, 104] have used tetraphenylporphyrin derivatives to recognize the surface of cytochrome-*c* and identified subnanomolar binders. These molecules consist of peripheral anionic groups that bind positively charged Arg and Lys residues present on the cytochrome-*c* surface. The binding of phenylporphyrins induces unfolding of the protein, leading to disruption of tertiary and secondary structure. This denaturation of the protein facilitates proteolytic degradation. Another group used similar

porphyrin-based derivatives for blocking potassium channels. Trauner and colleagues [105] used porphyrins to match the fourfold symmetry of the homotetrameric human $K_v1.3$ potassium channel. Using competitive binding assays, the authors showed that tetraphenylporphyrin derivatives with peripheral cationic groups strongly interact with potassium channel, thereby reducing the current through the channel.

Several synthetic receptors containing calixarene scaffolds have been designed to bind protein surfaces to block protein-protein interactions or the entry of small molecules into the active site of certain enzymes [106–109]. For example, calix[8]arene receptors decorated with basic amino acids competitively inhibit recombinant lung trypsin by binding to the acidic residues at the central junction of the tetrameric protein [107]. These molecules, most likely, bind at the entrance of the active site and block the approach of the substrate. Similarly, peptidocalix[4]arenes have been shown to bind the surface of transglutaminase, inhibiting its activity [108]. However, the competition assays suggest that these molecules bind to the surface of protein other than the enzyme active site, causing a conformational change in the protein or sterically blocking the approach of the substrate. Aachmann et al. [109] have designed β -cyclodextrin that binds to a specific site on the insulin surface via its solvent-exposed aromatic side chain. These studies suggest that porphyrin and calixarene scaffolds are certainly promising candidates for protein surface recognition and further work in this area may lead to novel therapeutic agents.

3.5 CONCLUSION

Over the past decade, protein complexes have become prime targets for therapeutic intervention. This has opened immense opportunities in the treatment of hitherto incurable diseases such as cancer. A large number of protein pairs have been identified as drug targets with reported successful inhibitors, suggesting the possibility of fighting disease in near future. However, with the identification of hundreds of possible drug targets in the “class” of protein-protein interaction complexes, picking targets for inhibition by peptides, peptide mimetics, or small molecules is going to be critical. Many protein pairs have interfaces where a small, linear region of one protein binds into a hydrophobic cleft of the other. Inhibitors for such interfaces have been discovered, using several approaches ranging from screening to structure-based design, that display sufficient potency and cellular activity. Unfortunately, a potent, specifically binding molecule does not necessarily make a good drug. The true potential of these molecules as therapeutics will only be realized following successful clinical trials.

REFERENCES

1. Fregeau Gallagher NL, Sailer M, Niemczura WP, Nakashima TT, Stiles ME, Vederas JC. Three-dimensional structure of leucocin A in trifluoroethanol and dodecylphosphocholine micelles: spatial location of residues critical for biological activity in type IIa bacteriocins from lactic acid bacteria. *Biochemistry* 1997;36:15062–15072.
2. Ryan DP, Matthews JM. Protein-protein interactions in human disease. *Curr Opin Struct Biol* 2005;15:441–446.

3. Zhao L, Chmielewski J. Inhibiting protein–protein interactions using designed molecules. *Curr Opin Struct Biol* 2005;15:31–34.
4. Arkin M. Protein–protein interactions and cancer: small molecules going in for the kill. *Curr Opin Chem Biol* 2005;9:317–324.
5. Fletcher S, Hamilton AD. Protein surface recognition and proteomimetics: mimics of protein surface structure and function. *Curr Opin Chem Biol* 2005;9:632–638.
6. Sillerud LO, Larson RS. Design and structure of peptide and peptidomimetic antagonists of protein–protein interaction. *Curr Protein Pept Sci* 2005;6:151–169.
7. Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, Krog-Jensen C, Butcher S. Emerging classes of protein–protein interaction inhibitors and new tools for their development. *Curr Opin Chem Biol* 2004;8:442–449.
8. Fry DC, Vassilev LT. Targeting protein–protein interactions for cancer therapy. *J Mol Med* 2005;83:955–963.
9. Bottger A, Bottger V, Garcia-Echeverria C, Chene P, Hochkeppel HK, Sampson W, Ang K, Howard SF, Picksley SM, Lane DP. Molecular characterization of the hdm2–p53 interaction. *J Mol Biol* 1997;269:744–756.
10. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* 1996;274:948–953.
11. Fotouhi N, Graves B. Small molecule inhibitors of p53/MDM2 interaction. *Curr Top Med Chem* 2005;5:159–165.
12. Adams JM, Cory S. The Bcl-2 protein family: arbiters of cell survival. *Science* 1998;281:1322–1326.
13. Chao DT, Korsmeyer SJ. BCL-2 family: regulators of cell death. *Annu Rev Immunol* 1998;16:395–419.
14. Cory S, Huang DC, Adams JM. The Bcl-2 family: roles in cell survival and oncogenesis. *Oncogene* 2003;22:8590–8607.
15. O’Neill J, Manion M, Schwartz P, Hockenbery DM. Promises and challenges of targeting Bcl-2 anti-apoptotic proteins for cancer therapy. *Biochim Biophys Acta* 2004;1705:43–51.
16. Wang S, Yang D, Lippman ME. Targeting Bcl-2 and Bcl-XL with nonpeptidic small-molecule antagonists. *Semin Oncol* 2003;30:133–142.
17. Sattler M, Liang H, Nettlesheim D, Meadows RP, Harlan JE, Eberstadt M, et al. Structure of Bcl-xL–Bak peptide complex: recognition between regulators of apoptosis. *Science* 1997;275:983–986.
18. Petros AM, Nettlesheim DG, Wang Y, Olejniczak ET, Meadows RP, Mack J, et al. Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. *Protein Sci* 2000;9:2528–2534.
19. Wang JL, Zhang ZJ, Choksi S, Shan S, Lu Z, Croce CM, Alnemri ES, Korngold R, Huang Z. Cell permeable Bcl-2 binding peptides: a chemical approach to apoptosis induction in tumor cells. *Cancer Res* 2000;60:1498–1502.
20. Petros AM, Medek A, Nettlesheim DG, Kim DH, Yoon HS, Swift K, Matayoshi ED, Oltersdorf T, Fesik SW. Solution structure of the antiapoptotic protein bcl-2. *Proc Natl Acad Sci USA* 2001;98:3012–3017.
21. Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, Yoon HS, et al. X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature* 1996;381:335–341.

22. Schimmer AD, Dalili S, Batey RA, Riedl SJ. Targeting XIAP for the treatment of malignancy. *Cell Death Differ* 2006;13:179–188.
23. Shiozaki EN, Chai J, Rigotti DJ, Riedl SJ, Li P, Srinivasula SM, Alnemri ES, Fairman R, Shi Y. Mechanism of XIAP-mediated inhibition of caspase-9. *Mol Cell* 2003;11:519–527.
24. Web: Drugs used in the treatment of HIV infection. <http://wwwfdagov/oashi/aids/viralshhtml>. 2000.
25. Liu Z, Sun C, Olejniczak ET, Meadows RP, Betz SF, Oost T, Herrmann J, Wu JC, Fesik SW. Structural basis for binding of Smac/DIABLO to the XIAP BIR3 domain. *Nature* 2000;408:1004–1008.
26. Jing N, Tweardy DJ. Targeting Stat3 in cancer therapy. *Anticancer Drugs* 2005;16:601–607.
27. Becker S, Groner B, Muller CW. Three-dimensional structure of the Stat3beta homodimer bound to DNA. *Nature* 1998;394:145–151.
28. Sahai E, Marshall CJ. RHO-GTPases and cancer. *Nat Rev Cancer* 2002;2:133–142.
29. Worthylake DK, Rossman KL, Sondek J. Crystal structure of Rac1 in complex with the guanine nucleotide exchange region of Tiam1. *Nature* 2000;408:682–688.
30. Gao Y, Dickerson JB, Guo F, Zheng J, Zheng Y. Rational design and characterization of a Rac GTPase-specific small molecule inhibitor. *Proc Natl Acad Sci USA* 2004;101:7618–7623.
31. Tucker GC. Integrins: molecular targets in cancer therapy. *Curr Oncol Rep* 2006;8:96–103.
32. Humphries MJ. The molecular basis and specificity of integrin–ligand interactions. *J Cell Sci* 1990;97:585–592.
33. D’Andrea LD, Del Gatto A, Pedone C, Benedetti E. Peptide-based molecules in angiogenesis. *Chem Biol Drug Des* 2006;67:115–126.
34. Dechantsreiter MA, Planker E, Matha B, Lohof E, Holzemann G, Jonczyk A, Goodman SL, Kessler H. N-methylated cyclic RGD peptides as highly active and selective alpha(V)beta(3) integrin antagonists. *J Med Chem* 1999;42:3033–3040.
35. Dimitrov DS. Virus entry: molecular mechanisms and biomedical applications. *Nat Rev Microbiol* 2004;2:109–122.
36. zur Hausen H. Papillomaviruses and cancer: from basic studies to clinical application. *Nat Rev Cancer* 2002;2:342–350.
37. Baseman JG, Koutsky LA. The epidemiology of human papillomavirus infections. *J Clin Virol* 2005;32:S16–S24.
38. Abbate EA, Berger JM, Botchan MR. The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. *Genes Dev* 2004;18:1981–1996.
39. Bartosch B, Vitelli A, Granier C, Goujon C, Dubuisson J, Pascale S, Scarselli E, Cortese R, Nicosia A, Cosset FL. Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor. *J Biol Chem* 2003;278:41624–41630.
40. Cormier EG, Tsamis F, Kajumo F, Durso RJ, Gardner JP, Dragic T. CD81 is an entry coreceptor for hepatitis C virus. *Proc Natl Acad Sci USA* 2004;101:7270–7274.
41. VanCompernelle SE, Wiznycia AV, Rush JR, Dhanasekaran M, Baures PW, Todd SC. Small molecule inhibition of hepatitis C virus E2 binding to CD81. *Virology* 2003;314:371–380.

42. Wagner CE, Mohler ML, Kang GS, Miller DD, Geisert EE, Chang YA, Fleischer EB, Shea KJ. Synthesis of 1-boraadamantaneamine derivatives with selective astrocyte vs C6 glioma antiproliferative activity. A novel class of anti-hepatitis C agents with potential to bind CD81. *J Med Chem* 2003;46:2823–2833.
43. Peiris JS, Guan Y, Yuen KY. Severe acute respiratory syndrome. *Nat Med* 2004;10:S88–S97.
44. Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003;426:450–454.
45. Xiao X, Chakraborti S, Dimitrov AS, Gramatikoff K, Dimitrov DS. The SARS-CoV S glycoprotein: expression and functional characterization. *Biochem Biophys Res Commun* 2003;312:1159–1164.
46. Wong SK, Li W, Moore MJ, Choe H, Farzan M. A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2. *J Biol Chem* 2004;279:3197–3201.
47. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* 2005;309:1864–1868.
48. Schwarz-Linek U, Werner JM, Pickford AR, Gurusiddappa S, Kim JH, Pilka ES, et al. Pathogenic bacteria attach to human fibronectin through a tandem beta-zipper. *Nature* 2003;423:177–181.
49. Pilka ES, Werner JM, Schwarz-Linek U, Pickford AR, Meenan NA, Campbell ID, Potts JR. Structural insight into binding of *Staphylococcus aureus* to human fibronectin. *FEBS Lett* 2006;580:273–277.
50. Schwarz-Linek U, Pilka ES, Pickford AR, Kim JH, Hook M, Campbell ID, Potts JR. High affinity streptococcal binding to human fibronectin requires specific recognition of sequential F1 modules. *J Biol Chem* 2004;279:39017–39025.
51. Raibaud S, Schwarz-Linek U, Kim JH, Jenkins HT, Baines ER, Gurusiddappa S, Hook M, Potts JR. *Borrelia burgdorferi* binds fibronectin through a tandem beta-zipper, a common mechanism of fibronectin binding in staphylococci, streptococci, and spirochetes. *J Biol Chem* 2005;280:18803–18809.
52. Deane JE, Ryan DP, Sunde M, Maher MJ, Guss JM, Visvader JE, Matthews JM. Tandem LIM domains provide synergistic binding in the LMO4:Ldb1 complex. *EMBO J* 2004;23:3589–3598.
53. Ryan DP, Sunde M, Kwan AH, Marianayagam NJ, Nancarrow AL, Vanden Hoven RN, et al. Identification of the key LMO2-binding determinants on Ldb1. *J Mol Biol* 2006;359:66–75.
54. Matthews JM, Visvader JE. LIM-domain-binding protein 1: a multifunctional cofactor that interacts with diverse proteins. *EMBO Rep* 2003;4:1132–1137.
55. Hammond SM, Crable SC, Anderson KP. Negative regulatory elements are present in the human LMO2 oncogene and may contribute to its expression in leukemia. *Leuk Res* 2005;29:89–97.
56. Gadek TR. Strategies and methods in the identification of antagonists of protein–protein interactions. *Biotechniques Suppl* 2003;21–24.
57. Arkin MR, Wells JA. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat Rev Drug Discov* 2004;3:301–317.
58. Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, Filipovic Z, et al. *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* 2004;303:844–848. Epub 2004 Jan 2002.
59. Benhar I. Biotechnological applications of phage and cell display. *Biotechnol Adv* 2001;19:1–33.

60. Clackson T, Hoogenboom HR, Griffiths AD, Winter G. Making antibody fragments using phage display libraries. *Nature* 1991;352:624–628.
61. Griffiths AD, Duncan AR. Strategies for selection of antibodies by phage display. *Curr Opin Biotechnol* 1998;9:102–108.
62. Hoogenboom HR, de Bruine AP, Hufton SE, Hoet RM, Arends JW, Roovers RC. Antibody phage display technology and its applications. *Immunotechnology* 1998;4:1–20.
63. Fields S, Song O. A novel genetic system to detect protein–protein interactions. *Nature* 1989;340:245–246.
64. Graessmann A, Graessmann M, Mueller C. Microinjection of early SV40 DNA fragments and T antigen. *Methods Enzymol* 1980;65:816–825.
65. Morgan DO, Roth RA. Analysis of intracellular protein function by antibody injection. *Immunol Today* 1988;9:84–88.
66. Valle G, Jones EA, Colman A. Anti-ovalbumin monoclonal antibodies interact with their antigen in internal membranes of *Xenopus* oocytes. *Nature* 1982;300:71–74.
67. Burke B, Warren G. Microinjection of mRNA coding for an anti-Golgi antibody inhibits intracellular transport of a viral membrane protein. *Cell* 1984;36:847–856.
68. Lobato MN, Rabbitts TH. Intracellular antibodies as specific reagents for functional ablation: future therapeutic molecules. *Curr Mol Med* 2004;4:519–528.
69. Tanaka T, Lobato MN, Rabbitts TH. Single domain intracellular antibodies: a minimal fragment for direct *in vivo* selection of antigen-specific intrabodies. *J Mol Biol* 2003;331:1109–1120.
70. Das D, Suresh MR. Producing bispecific and bifunctional antibodies. *Methods Mol Med* 2005;109:329–346.
71. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* 1997;15:553–557.
72. Lipovsek D, Pluckthun A. *In-vitro* protein evolution by ribosome display and mRNA display. *J Immunol Methods* 2004;290:51–67.
73. Tse E, Lobato MN, Forster A, Tanaka T, Chung GT, Rabbitts TH. Intracellular antibody capture technology: application to selection of intracellular antibodies recognising the BCR-ABL oncogenic protein. *J Mol Biol* 2002;317:85–94.
74. Visintin M, Settanni G, Maritan A, Graziosi S, Marks JD, Cattaneo A. The intracellular antibody capture technology (IACT): towards a consensus sequence for intracellular antibodies. *J Mol Biol* 2002;317:73–83.
75. Lobato MN, Rabbitts TH. Intracellular antibodies and challenges facing their use as therapeutic agents. *Trends Mol Med* 2003;9:390–396.
76. Hudson PJ, Souriau C. Engineered antibodies. *Nat Med* 2003;9:129–134.
77. Hinoda Y, Sasaki S, Ishida T, Imai K. Monoclonal antibodies as effective therapeutic agents for solid tumors. *Cancer Sci* 2004;95:621–625.
78. Kritzer JA, Lear JD, Hodsdon ME, Schepartz A. Helical beta-peptide inhibitors of the p53–hDM2 interaction. *J Am Chem Soc* 2004;126:9468–9469.
79. Yin H, Lee GI, Park HS, Payne GA, Rodriguez JM, Sebt SM, Hamilton AD. Terphenyl-based helical mimetics that disrupt the p53/HDM2 interaction. *Angew Chem Int Ed Engl* 2005;44:2704–2707.
80. Yin H, Lee GI, Sedey KA, Rodriguez JM, Wang HG, Sebt SM, Hamilton AD. Terephthalamide derivatives as mimetics of helical peptides: disruption of the Bcl-x(L)/Bak interaction. *J Am Chem Soc* 2005;127:5463–5468.

81. Yin H, Lee GI, Sedey KA, Kutzki O, Park HS, Orner BP, Ernst JT, Wang HG, Sebti SM, Hamilton AD. Terphenyl-based Bak BH3 alpha-helical proteomimetics as low-molecular-weight antagonists of Bcl-xL. *J Am Chem Soc* 2005;127:10191–10196.
82. Walensky LD, Kung AL, Escher I, Malia TJ, Barbuto S, Wright RD, Wagner G, Verdine GL, Korsmeyer SJ. Activation of apoptosis *in vivo* by a hydrocarbon-stapled BH3 helix. *Science* 2004;305:1466–1470.
83. Sadowsky JD, Schmitt MA, Lee HS, Umezawa N, Wang S, Tomita Y, Gellman SH. Chimeric (alpha/beta + alpha)-peptide ligands for the BH3-recognition cleft of Bcl-XL: critical role of the molecular scaffold in protein surface recognition. *J Am Chem Soc* 2005;127:11966–11968.
84. Xiong JP, Stehle T, Zhang R, Joachimiak A, Frech M, Goodman SL, Arnaout MA. Crystal structure of the extracellular segment of integrin alpha Vbeta3 in complex with an Arg-Gly-Asp ligand. *Science* 2002;296:151–155. Epub 2002 Mar 2007.
85. Altmeyer R. Virus attachment and entry offer numerous targets for antiviral therapy. *Curr Pharm Des* 2004;10:3701–3712.
86. Moore JP, Doms RW. The entry of entry inhibitors: a fusion of science and medicine. *Proc Natl Acad Sci USA* 2003;100:10598–10602.
87. Web: Drugs used in the treatment of HIV infection. <http://wwwfdagov/oashi/aids/viralshhtml>. Accessed 9 August 2006.
88. Ernst JT, Kutzki O, Debnath AK, Jiang S, Lu H, Hamilton AD. Design of a protein surface antagonist based on alpha-helix mimicry: inhibition of gp41 assembly and viral fusion. *Angew Chem Int Ed Engl* 2002;41:278–281.
89. Nozaki A, Ikeda M, Naganuma A, Nakamura T, Inudoh M, Tanaka K, Kato N. Identification of a lactoferrin-derived peptide possessing binding activity to hepatitis C virus E2 envelope protein. *J Biol Chem* 2003;278:10162–10173.
90. English EP, Chumanov RS, Gellman SH, Compton T. Rational development of beta-peptide inhibitors of human cytomegalovirus entry. *J Biol Chem* 2006;281:2661–2667.
91. Petros AM, Olejniczak ET, Fesik SW. Structural biology of the Bcl-2 family of proteins. *Biochim Biophys Acta* 2004;1644:83–94.
92. Wang JL, Liu D, Zhang ZJ, Shan S, Han X, Srinivasula SM, Croce CM, Alnemri ES, Huang Z. Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells. *Proc Natl Acad Sci USA* 2000;97:7124–7129.
93. Qian J, Voorbach MJ, Huth JR, Coen ML, Zhang H, Ng SC, et al. Discovery of novel inhibitors of Bcl-xL using multiple high-throughput screening platforms. *Anal Biochem* 2004;328:131–138.
94. Tzung SP, Kim KM, Basanez G, Giedt CD, Simon J, Zimmerberg J, Zhang KY, Hockenbery DM. Antimycin A mimics a cell-death-inducing Bcl-2 homology domain 3. *Nat Cell Biol* 2001;3:183–191.
95. Sun H, Nikolovska-Coleska Z, Yang CY, Xu L, Liu M, Tomita Y, et al. Structure-based design of potent, conformationally constrained Smac mimetics. *J Am Chem Soc* 2004;126:16686–16687.
96. Sun H, Nikolovska-Coleska Z, Yang CY, Xu L, Tomita Y, Krajewski K, Roller PP, Wang S. Structure-based design, synthesis, and evaluation of conformationally constrained mimetics of the second mitochondria-derived activator of caspase that target the X-linked inhibitor of apoptosis protein/caspase-9 interaction site. *J Med Chem* 2004;47:4147–4150.
97. Sun H, Nikolovska-Coleska Z, Chen J, Yang CY, Tomita Y, Pan H, Yoshioka Y, Krajewski K, Roller PP, Wang S. Structure-based design, synthesis and biochemical testing of novel and potent Smac peptido-mimetics. *Bioorg Med Chem Lett* 2005;15:793–797.

98. Li L, Thomas RM, Suzuki H, De Brabander JK, Wang X, Harran PG. A small molecule Smac mimic potentiates TRAIL- and TNF α -mediated cell death. *Science* 2004;305:1471–1474.
99. Schimmer AD, Welsh K, Pinilla C, Wang Z, Krajewska M, Bonneau MJ, et al. Small-molecule antagonists of apoptosis suppressor XIAP exhibit broad antitumor activity. *Cancer Cell* 2004;5:25–35.
100. White PW, Titolo S, Brault K, Thauvette L, Pelletier A, Welchner E, et al. Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1–E2 protein interaction. *J Biol Chem* 2003;278:26765–26772.
101. Wang Y, Coulombe R, Cameron DR, Thauvette L, Massariol MJ, Amon LM, et al. Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. *J Biol Chem* 2004;279:6976–6985.
102. Shultz MD, Ham YW, Lee SG, Davis DA, Brown C, Chmielewski J. Small-molecule dimerization inhibitors of wild-type and mutant HIV protease: a focused library approach. *J Am Chem Soc* 2004;126:9886–9887.
103. Aya T, Hamilton AD. Tetrabiphenylporphyrin-based receptors for protein surfaces show sub-nanomolar affinity and enhance unfolding. *Bioorg Med Chem Lett* 2003;13:2651–2654.
104. Groves K, Wilson AJ, Hamilton AD. Catalytic unfolding and proteolysis of cytochrome C induced by synthetic binding agents. *J Am Chem Soc* 2004;126:12833–12842.
105. Gradl SN, Felix JP, Isacoff EY, Garcia ML, Trauner D. Protein surface recognition by rational design: nanomolar ligands for potassium channels. *J Am Chem Soc* 2003;125:12668–12669.
106. Park HS, Lin Q, Hamilton AD. Modulation of protein–protein interactions by synthetic receptors: design of molecules that disrupt serine protease–proteinaceous inhibitor interaction. *Proc Natl Acad Sci USA* 2002;99:5105–5109.
107. Mecca T, Consoli GM, Geraci C, Cunsolo F. Designed calix[8]arene-based ligands for selective trypsin surface recognition. *Bioorg Med Chem* 2004;12:5057–5062.
108. Francese S, Cozzolino A, Caputo I, Esposito C, Martino M, Gaeta C, Troisi F, Neri P. Transglutaminase surface recognition by peptidocalix[4]arene diversomers. *Tetrahedron Lett* 2005;46:1611–1615.
109. Achmann FL, Otzen DE, Larsen KL, Wimmer R. Structural background of cyclodextrin–protein interactions. *Protein Eng* 2003;16:905–912.