CHAPTER 10

Other Related Techniques

Contents

10.1 Introduction 358

10.2 Pharmacophore 359

10.2.1 Concept and definition 359

10.2.2 Background and early days of pharmacophore 360

10.2.3 Methodology of pharmacophore mapping 361

10.2.3.1 Diverse conformation generation 361

10.2.3.2 Generation of 3D pharmacophore 362

10.2.3.3 Assessment of the quality of pharmacophore hypotheses 363

10.2.3.4 Validation of the pharmacophore model 365

10.2.4 Types of pharmacophore 366

10.2.4.1 Ligand-based pharmacophore modeling 366

10.2.4.2 Structure-based pharmacophore modeling 368

10.2.5 Application of pharmacophore models 372

10.2.5.1 Pharmacophore model—based VS 372

10.2.5.2 Pharmacophore-based de novo *design 373*

10.2.6 Advantages and limitations of pharmacophore 374

10.2.7 Software tools for pharmacophore analysis 375

10.3 Structure-Based Design—Docking 375

10.3.1 Concept and definition of docking 375

10.3.2 Definition of fundamental terms of docking 379

10.3.3 Essential requirements of docking 381

10.3.4 Categorization of docking 382

10.3.4.1 Receptor/protein flexibility 383

10.3.4.2 Ligand sampling and flexibility 385

10.3.4.3 Docking scoring functions 387

10.3.5 Basic steps of docking 388

10.3.6 Challenges and required improvements in docking studies 389

10.3.7 Applications of docking 392

10.3.8 Docking software tools 393

10.4 Combination of Structure- and Ligand-Based Design Tools 396

10.4.1 Comparative binding energy analysis 396

10.4.1.1 The concept of comparative binding energy 396

10.4.1.2 The methodology of COMBINE 397

10.4.1.3 Importance and advantages of COMBINE 399

10.4.1.4 Drawbacks and required improvements 399

10.4.1.5 Applications of COMBINE 399

10.4.1.6 Software for COMBINE 400

Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment.
ISBN: 978-0-12-801505-6, DOI: http://dx.doi.org/10.1016/B978-0-12-801505-6.00010-7

10.4.2 Comparative residue interaction analysis 400
10.4.2.1 Concept of CoRIA *400*
10.4.2.2 Methodology of CoRIA *401*
10.4.2.3 Variants of CoRIA *401*
10.4.2.4 Importance and application of CoRIA *402*
10.4.2.5 Drawback of CoRIA *402*
10.4.2.6 Future perspective of CoRIA *403*
10.5 *In silico* Screening of Chemical Libraries: VS 403
10.5.1 Concept 403
10.5.2 Workflow and types of VS 403
10.5.2.1 Selection of chemical libraries/databases *404*
10.5.2.2 Preprocessing of chemical libraries *404*
10.5.2.3 Filtering of druglike molecules *404*
10.5.2.4 Screening *404*
10.5.2.5 Hit selection to new chemical entity generation *405*
10.5.3 Successful application of VS: A few case studies 406
10.5.4 Advantages of VS 406
10.5.4.1 Cost-effective *413*
10.5.4.2 Time-saving *413*
10.5.4.3 Labor-efficient *413*
10.5.4.4 Sensible alternative *413*
10.5.5 Pitfalls 413
10.5.6 Databases for the VS 416
10.6 Overview and Conclusions 417
References 421

10.1 INTRODUCTION

Computer-assisted tools in drug design and discovery, with an incredible modernization of computational resources, are very much appreciated throughout the world. Both ligand- and structure-based approaches are increasingly being used for the design of small lead and druglike molecules with anticipated multitarget activities [1]. Various ligand-based methods have been developed for effective and comprehensive application in virtual screening (VS), *de novo* design, and lead optimization. Pharmacophore has become one of the major ligand-based tools in computational chemistry for the drug research and development process [2]. Again, molecular recognitions, including enzyme—substrate, drug—protein, drug—nucleic acid, protein—nucleic acid, and protein—protein interactions, play significant roles in many biological responses. As a consequence, identification of the binding mode and affinity of the drug molecule is crucial to understanding the underlying mechanism of action in the respective therapeutic response. In this perspective, structure-based drug design is always a front-runner among all the available drug design approaches. Molecular docking is one of the largely acclaimed structure-based approaches, widely used for the study of molecular recognition, which aims to predict the binding mode and binding affinity of a complex formed by two or more constituent molecules with known structures [3].

There are a handful of novel techniques invented in the last decade employing the combined information computed from receptors and ligands. These tools can be defined as a combination of structure- and ligand-based design tools in the evolution of drug discovery techniques. Undoubtedly, methods like comparative binding energy analysis (COMBINE) [4] and comparative residue interaction analysis (CoRIA) [5] are the front-runners in the abovementioned approach with encouraging successful applications in drug discovery.

In silico screening is generally defined as VS, which is used rationally to select compounds for biological in vitro/in vivo testing from chemical libraries and databases of hundreds of thousands of compounds [6]. The VS approach is used for computationally prioritizing drug candidate molecules for future synthesis by using certain filters. The filters may be created by employing knowledge about the protein target (in structure-based VS) or known bioactive ligands (in ligand-based VS). These computational methods are powerful tools, as they supply a straightforward way to estimate the properties of the molecules and establish them as probable drug candidates from a huge number of compounds in no time in a cost-effective way. A combination of bioinformatics and chemoinformatics is crucial to the success of VS of chemical libraries, which is an alternative and complementary approach to high-throughput screening (HTS) in the lead discovery process [7]. Simply stated, the VS attempts to improve the probability of identifying bioactive molecules by maximizing the true positive rate—that is, by ranking the truly active molecules as high as possible.

10.2 PHARMACOPHORE

10.2.1 Concept and definition

One of the most promising *in silico* concepts of computer-aided drug design (CADD) is that of the pharmacophore. The term *pharmacophore* was first coined by Paul Ehrlich in the early 1900s, but it was Monty Kier [8,9] who introduced the physical chemical concept of pharmacophore in a series of papers published between 1967 and 1971. The pharmacophore technique in modern drug discovery is extremely useful as an interface between the medicinal chemistry and computational chemistry, both in VS and library design for efficient hit discovery, as well as in the optimization of lead compounds to final drug candidates. Recent research has focused on the practice of parallel screening using pharmacophore models for bioactivity profiling and early-stage risk assessment of probable adverse effects and toxicity due to interaction of drug candidates with antitargets.

The hypothesis of pharmacophore is based on that the molecular recognition of a biological target by a class of compounds can be explained by a set of common features that interact with a set of complementary sites on the biological target [10]. Along with the features, their three-dimensional (3D) relationship with each of the features is another crucial component of the pharmacophore concept. It is closely linked to the widely used principle of bioisosterism, which can be adopted by medicinal chemists while designing bioactive compound series.

The pharmacophore can be simply defined by the following, as stated in the International Union of Pure and Applied Chemistry (IUPAC) definition of the term given in Wermuth et al. [11]:

> *A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response.*
>
> *A pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds toward their target structure.*
>
> *A pharmacophore can be considered as the largest common denominator shared by a set of active molecules. This definition discards a misuse often found in the medicinal chemistry literature, which consists of naming as pharmacophores simple chemical functionalities such as guanidines, sulfonamides, or dihydroimidazoles (formerly imidazolines), or typical structural skeletons such as flavones, phenothiazines, prostaglandins, or steroids.*
>
> *A pharmacophore is defined by pharmacophoric descriptors, including H-bonding, hydrophobic, and electrostatic interaction sites, defined by atoms, ring centers, and virtual points.*

The pharmacophore describes the essential steric and electronic, function-determining points necessary for an optimal interaction with a relevant pharmacological target. It can also be thought of as a template, a partial description of a molecule where certain blanks need to be filled. The types of ligand molecules and the size and diversity of the data set have a great impact on the resulting pharmacophore model. Although a pharmacophore model signifies the key interactions between a ligand and its biological target, neither the structure of the target nor its identity is required to construct a handy pharmacophore model. As a consequence, pharmacophore approaches are often considered to be vital when the accessible information is very restricted. For example, when one knows nothing more than the structures of active ligands, a pharmacophore is the answer.

A simple hypothetical example is illustrated to define the common pharmacophores of three well-known compounds (namely, epinephrine, norepinephrine, and isoprenaline) in Figure 10.1.

10.2.2 Background and early days of pharmacophore

Introducing the term *pharmacophore* in the year 1909, Ehrlich [12], nicknamed the "father of drug discovery," defined it as "a molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*) biological activity." Although the first definition of the term was credited to Ehrlich, it was Kier who introduced the physical chemical concept in the late 1960s and early 1970s when describing common molecular features of ligands of important central nervous system receptors. This was labeled as "muscarinic pharmacophore" by Kier [8,9].

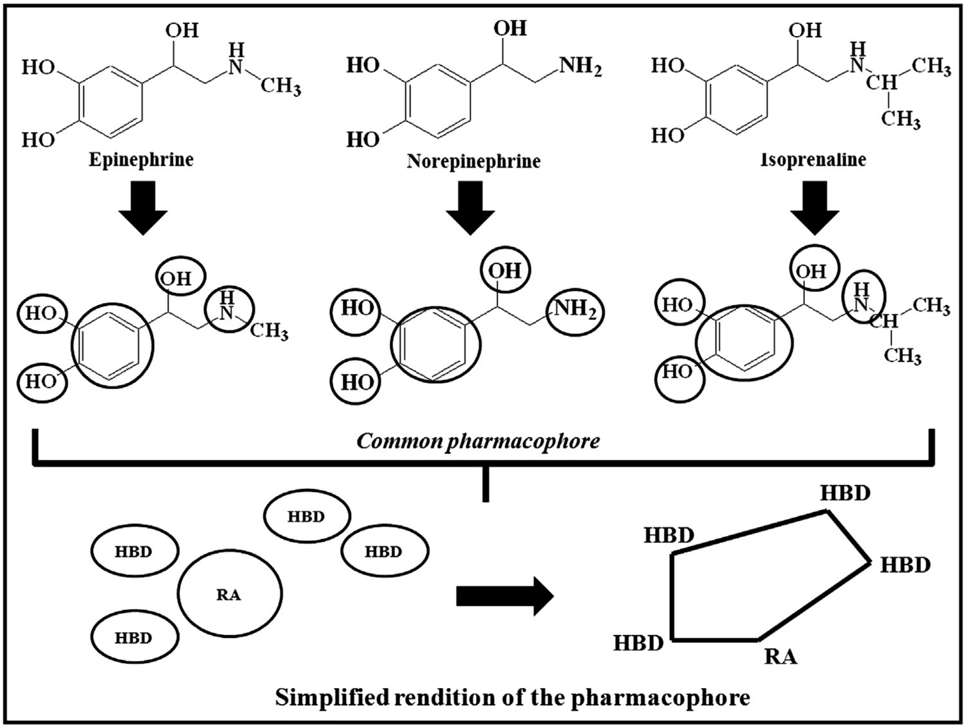


Figure 10.1 Depiction of common pharmacophoric features of three well-known compounds: epinephrine, norepinephrine, and isoprenaline.

In the past, pharmacophore models were mainly worked out manually, assisted through the use of simple interactive molecular graphics visualization programs. Later, the growing complexities of molecular structures required refined computer programs for the determination and use of pharmacophore models. In the evolution of computational chemistry, the fundamental perception of a pharmacophore model as a simple geometric depiction of the key molecular interactions remains unchanged. With the advances in computational chemistry in the past 20 years, a variety of automated tools for pharmacophore modeling and applications emerged. A considerable number of studies have been carried out since the development of the pharmacophore approach [13]. Pharmacophore approaches have been used comprehensively in VS, *de novo* design, as well as in lead optimization and multitarget drug design [14].

10.2.3 Methodology of pharmacophore mapping

10.2.3.1 Diverse conformation generation

Conformational expansion is the most critical step, since the goal is not only to have the most representative coverage of the conformational space of a molecule, but also to have either the bioactive conformation as part of the set of generated conformations or at least a cluster of conformations that are close enough to the bioactive conformation. This conformational search can be divided into four categories: (i) systematic search in the torsional space, (ii) clustering (if wanted or needed), (iii) stochastic methods, such as Monte Carlo (MC), sampling, and Poling, and (iv) molecular dynamics [15]. Commonly employed conformational search methods are BEST, FAST, and conformer algorithms based on energy screening and recursive buildup (CAESAR) [16], all of which generate conformations that provide broad coverage of the accessible conformational space. The FAST conformation generation method searches conformations only in the torsion space and takes less time. The BEST method provides a complete and improved coverage of conformational space by performing a rigorous energy minimization and optimizing the conformations in both torsional and Cartesian space using the Poling algorithm. CAESAR is based on a divide-and-conquer and recursive conformation approach. This approach is also combined in cases of local rotational symmetry so that conformation duplicates due to topological symmetry in a systematic search can be efficiently eliminated.

10.2.3.2 Generation of 3D pharmacophore

The next step is three-dimensional (3D) pharmacophore generation, where Hypogen and HipHop are the two most commonly used algorithms [17,18]. Predictive 3D pharmacophores are generated in three phases: a constructive, a subtractive, and an optimization phase, as follows:

Constructive phase: HipHop is intended to derive common feature hypothesis-based pharmacophore models using information from a set of active compounds. HipHop does not require the selection of a template; rather, each molecule is treated as a template in turn. Different configurations of chemical features are identified in the template molecule using a pruned exhaustive search, which starts with small sets of features and then extends until no larger configuration is found. Next, each configuration is compared with the remaining molecules to identify configurations that are common to all molecules. The resulting pharmacophores are ranked using a combination of how well the molecules in the training set map onto the pharmacophore model. In HipHop, the user can define how many molecules must map completely or partially to a pharmacophore configuration. Again, HypoGen [18] is an algorithm that uses the activity values of the small compounds in the training set to generate hypotheses to build 3D pharmacophore models. HypoGen identifies all allowable pharmacophores consisting of up to five features among the two most active compounds and investigates the remaining active compounds in the list.

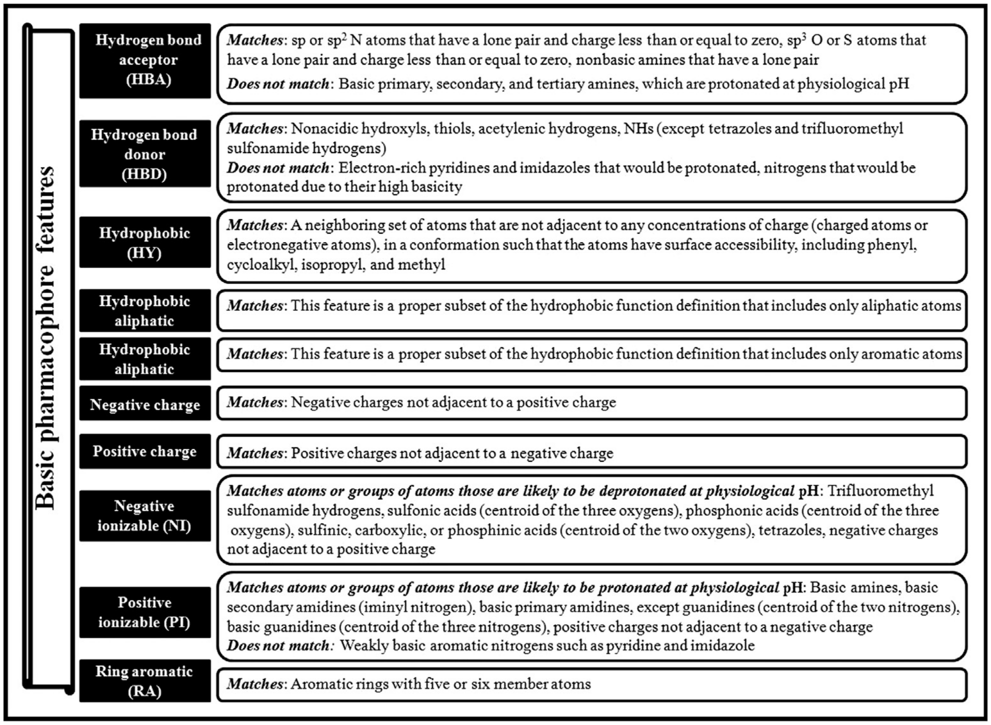
Subtractive phase: This phase deals with pharmacophores that were created in the constructive phase and removes pharmacophores from the data structure that are not likely to be useful.

Optimization phase: The optimization phase is performed using the simulated annealing algorithm. A maximum of 10 hypotheses are generated for each run. HypoGen develops models with different pharmacophore features: (i) hydrogen-bond acceptor (HBA); (ii) hydrogen-bond donor (HBD); (iii) hydrophobic (HYD), HYDROPHOBIC (aliphatic) and HYDROPHOBIC (aromatic); (iv) negative charge (NEG CHARGE); (v) negative ionizable (NI); (vi) positive charge (POS CHARGE); (vii) positive ionizable (PI); and (viii) ring aromatic (RA). The hypotheses generated are analyzed in terms of their correlation coefficients and the cost function values.

The basic pharmacophore features are illustrated in Figure 10.2. Pharmacophore models are usually labeled based on the number of features. For example, pharmacophore models consisting of three and four features are termed as three-point pharmacophore and four-point pharmacophore, respectively. A simple graphical representation is shown in Figure 10.3.

10.2.3.3 Assessment of the quality of pharmacophore hypotheses

The *HypoGen* module performs a fixed cost calculation that represents the simple model that fits all the data, and a null cost calculation that assumes that there is no



Basic pharmacophore features

Hydrogen bond acceptor (HBA)	***Matches***: sp or sp^2 N atoms that have a lone pair and charge less than or equal to zero, sp^3 O or S atoms that have a lone pair and charge less than or equal to zero, nonbasic amines that have a lone pair ***Does not match***: Basic primary, secondary, and tertiary amines, which are protonated at physiological pH
Hydrogen bond donor (HBD)	***Matches***: Nonacidic hydroxyls, thiols, acetylenic hydrogens, NHs (except tetrazoles and trifluoromethyl sulfonamide hydrogens) ***Does not match***: Electron-rich pyridines and imidazoles that would be protonated, nitrogens that would be protonated due to their high basicity
Hydrophobic (HY)	***Matches***: A neighboring set of atoms that are not adjacent to any concentrations of charge (charged atoms or electronegative atoms), in a conformation such that the atoms have surface accessibility, including phenyl, cycloalkyl, isopropyl, and methyl
Hydrophobic aliphatic	***Matches***: This feature is a proper subset of the hydrophobic function definition that includes only aliphatic atoms
Hydrophobic aliphatic	***Matches***: This feature is a proper subset of the hydrophobic function definition that includes only aromatic atoms
Negative charge	***Matches***: Negative charges not adjacent to a positive charge
Positive charge	***Matches***: Positive charges not adjacent to a negative charge
Negative ionizable (NI)	***Matches atoms or groups of atoms those are likely to be deprotonated at physiological* pH**: Trifluoromethyl sulfonamide hydrogens, sulfonic acids (centroid of the three oxygens), phosphonic acids (centroid of the three oxygens), sulfinic, carboxylic, or phosphinic acids (centroid of the two oxygens), tetrazoles, negative charges not adjacent to a positive charge
Positive ionizable (PI)	***Matches atoms or groups of atoms those are likely to be protonated at physiological* pH**: Basic amines, basic secondary amidines (iminyl nitrogen), basic primary amidines, except guanidines (centroid of the two nitrogens), basic guanidines (centroid of the three nitrogens), positive charges not adjacent to a negative charge ***Does not match:*** Weakly basic aromatic nitrogens such as pyridine and imidazole
Ring aromatic (RA)	***Matches***: Aromatic rings with five or six member atoms

Figure 10.2 Basic pharmacophore features and their definitions.

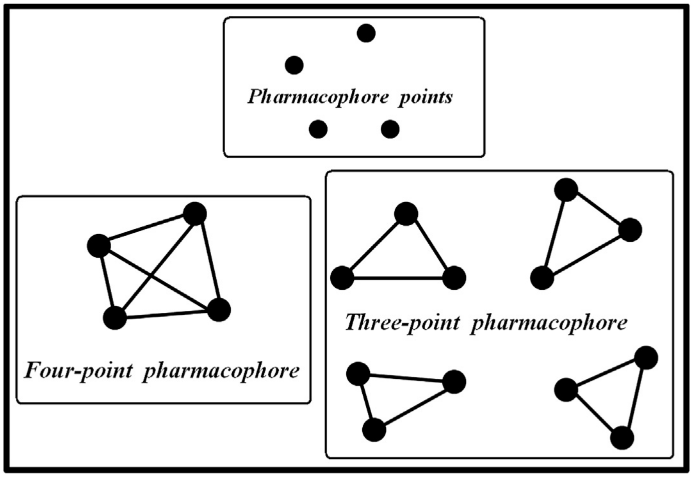


Figure 10.3 Point-based pharmacophore concepts.

relationship in the data set and that the experimental activities are normally distributed about their average value. A small range of the total hypothesis cost obtained for each of the hypotheses indicates homogeneity of the corresponding hypothesis, and the training set selected for the purpose of pharmacophore generation is adequate. Again, values of total cost close to those of fixed cost indicate the fact that the hypotheses generated are statistically robust [19,20]. The total cost of a hypothesis is calculated as per Eq. (10.1):

$$\text{Cost} = eE + wW + cC \tag{10.1}$$

where e, w, and c are the coefficients associated with the error (E), weight (W), and configuration (C) components, respectively. The other two important costs involved are the fixed cost and null cost. The fixed cost represents the simplest model that perfectly fits the data and is calculated by Eq. (10.2):

$$\text{Fixed cost} = eE(x = 0) + wW(x = 0) + cC \tag{10.2}$$

where x is the deviation from the expected values of weight and error. The null cost is the cost of a pharmacophore when the activity data of every molecule in the training set is the average value of all activities in the set and the pharmacophore has no features. Therefore, the contribution from the weight or configuration component does not apply. The null cost is calculated as per Eq. (10.3):

$$\text{Null cost} = eE(\chi_{\text{est}} = \overline{\chi}) \tag{10.3}$$

where χ_{est} is the averaged scaled activity of the training set molecules. It has been suggested that the differences between cost of the generated hypothesis and the null hypothesis should be as large as possible; a value of 40–60 bits difference may indicate that it has a 75–90% chance of representing a true correlation in the data set used. The total cost of any hypothesis should be toward the value of fixed cost to represent any meaningful model. Two other very important output parameters are the configuration cost and the error cost. Any value of configuration cost higher than 17 may indicate that the correlation from any generated pharmacophore is most likely due to chance. The error cost increases as the value of the root mean square (RMS) increases. The RMS deviations (RMSDs) represent the quality of the correlation between the estimated and the actual activity data.

10.2.3.4 Validation of the pharmacophore model

The pharmacophore models selected based on the acceptable correlation coefficient (R) and cost analysis, should be validated in three subsequent steps: (i) Fischer's randomization test, (ii) test set prediction, and (iii) Güner–Henry (GH) scoring method.

Fischer's randomization test: First, cross-validation is performed and statistical significance of the structure–activity correlation is estimated by randomizing the data using the Fischer's randomization test [20]. This is done by scrambling the activity data of the training set molecules and assigning them new values, followed by the generation of pharmacophore hypotheses using the same features and parameters as those used to develop the original pharmacophore hypothesis. The original hypothesis is considered to be generated by mere chance if the randomized data set results in the generation of a pharmacophore with better or nearly equal correlation compared to the original one.

Test set prediction: The purpose of the pharmacophore hypothesis generation is not only to predict the activity of the training set compounds [21], but also to predict the activities of external molecules. With the objective of verifying whether the pharmacophore is able to predict the activity of test set molecules in agreement with the experimentally determined value, the activities of the test set molecules are estimated based on the mapping of the test set molecules to the developed pharmacophore model. The conformers are generated for the test set molecules based on the method that is used during the conformer generation of the training set, and they are mapped using the corresponding pharmacophore models. Thus, the predictive capacity of the models is judged based on the predictive R^2 values (R^2_{pred} with a threshold value of 0.5) or classification-based methods (such as sensitivity, specificity, precision, and accuracy). The test set should cover similar structural diversity as the training set in order to establish the broadness of the pharmacophore predictability.

GH scoring: The GH scoring method is employed following test set validation to evaluate the quality of the pharmacophore models [22–24]. The GH score can be successfully applied to quantify model selectivity precision of hits and the recall of actives from a directory of useful decoys (DUD) data set [25] consisting of known actives and inactives. The DUD is a publicly available database for free use, generated based on the observation that physical characteristics of the decoy background can be used for the classification of different compounds. The DUD can be downloaded from http://dud.docking.org.

The method involves evaluation of the following: the percent yield of actives in a database (%*Y*, recall), the percent ratio of actives in the hit list (%*A*, precision), the enrichment factor *E*, and the GH score. The GH score ranges from 0 to 1, where a value of 1 signifies the ideal model. The following are the metrics used for analyzing hit lists by a pharmacophore model—based database search:

$$\%A = \frac{\mathrm{Ha}}{A} \times 100 \tag{10.4}$$

$$\%Y = \frac{\mathrm{Ha}}{\mathrm{Ht}} \times 100 \tag{10.5}$$

$$E = \frac{\mathrm{Ha}/\mathrm{Ht}}{A/D} \tag{10.6}$$

$$\mathrm{GH} = \left[\frac{\mathrm{Ha}(3A + \mathrm{Ht})}{4\mathrm{Ht}A}\right]\left(1 - \frac{\mathrm{Ht} - \mathrm{Ha}}{D - A}\right) \tag{10.7}$$

In these equations, %*A* is the percentage of known active compounds retrieved from the database (precision); Ha is the number of actives in the hit list (true positives); *A* is the number of active compounds in the database; %*Y* is the percentage of known actives in the hit list (recall); Ht is the number of hits retrieved; *D* is the number of compounds in the database; and *E* is the enrichment of the concentration of actives by the model relative to random screening without any pharmacophoric approach.

The basic steps of pharmacophore formalism are represented in Figure 10.4.

10.2.4 Types of pharmacophore

A pharmacophore model can be generated in two ways. The first method is ligand-based modeling, where a set of active molecules are superimposed and common chemical features are extracted that are necessary for their bioactivity; the second is structure-based modeling performed by probing possible interaction points between the macromolecular target and ligands.

10.2.4.1 Ligand-based pharmacophore modeling

Ligand-based pharmacophore (LBP) modeling has become an important computational tool for assisting drug discovery in the case of nonavailability of a

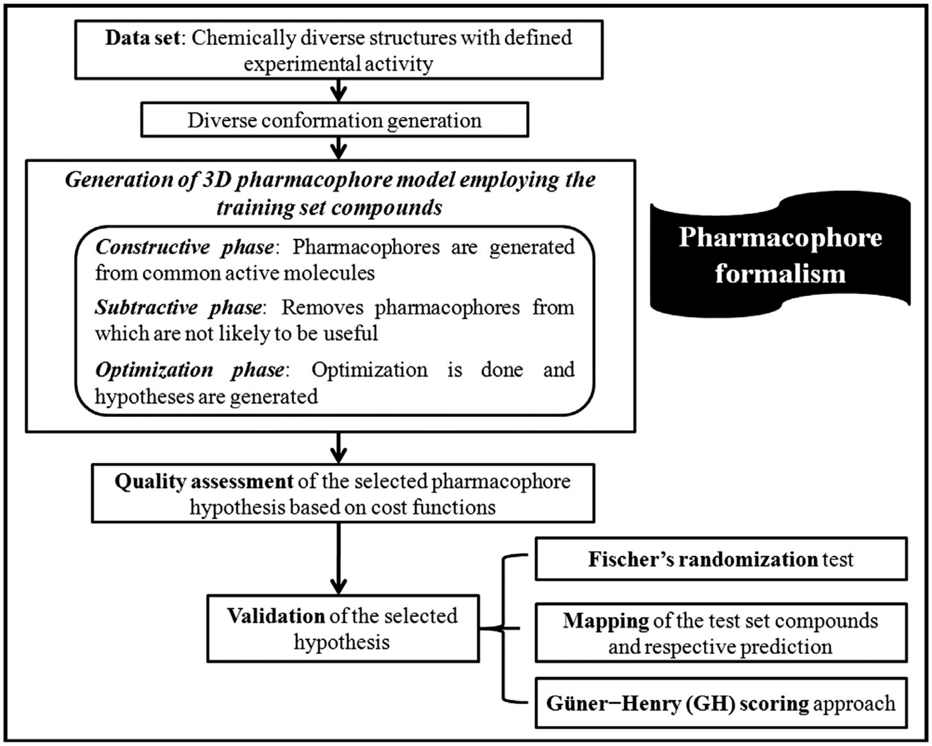

Data set: Chemically diverse structures with defined experimental activity
Diverse conformation generation
Generation of 3D pharmacophore model employing the training set compounds
Constructive phase: Pharmacophores are generated from common active molecules
Subtractive phase: Removes pharmacophores from which are not likely to be useful
Optimization phase: Optimization is done and hypotheses are generated
Pharmacophore formalism
Quality assessment of the selected pharmacophore hypothesis based on cost functions
Validation of the selected hypothesis
Fischer's randomization test
Mapping of the test set compounds and respective prediction
Güner–Henry (GH) scoring approach

Figure 10.4 Fundamental steps of pharmacophore formalism.

macromolecular target structure [26,27]. The LBP is usually carried out by extracting common chemical features from the 3D structures of a known set of ligands representative of fundamental interactions between the ligands and a specific macromolecular target. In the case of LBP modeling, pharmacophore generation from multiple ligands involves two major steps: First, creation of the conformational space for each ligand in the training set to represent conformational flexibility of the ligands and to align the multiple ligands in the training set, and second, determination of the essential common chemical features to build the pharmacophore model. The conformational analysis of ligands and performing molecular alignment are the key techniques as well as the main complexities in any LBP modeling.

A few challenges still exist in spite of the great advances of LBP modeling:

a. The first problem, and one of the most serious, is the modeling of ligand flexibility. Presently, two strategies are utilized to deal with this problem. The first is the preenumerating method, in which multiple conformations for each ligand are precomputed and saved in a database [28]. The advantage of this approach is lower computing cost for conducting molecular alignment at the expense of a possible need for a mass storage capacity. The second approach is the on-the-fly method, in which the conformation analysis is carried out in the pharmacophore modeling process [28]. This approach does not need mass storage but might need higher central processing unit (CPU) time for conducting meticulous optimization. It has been demonstrated that the preenumerating method outperforms the on-the-fly calculation approach [29]. Recently, a considerable number of advanced algorithms [14] have been established to sample the conformational spaces of small molecules, which are listed in Table 10.1.

Most importantly, a good conformation generator should ensure the following conditions: (i) proficiently generating all the putative bound conformations that small molecules adopt when they interact with macromolecules, (ii) keeping the list of low-energy conformations as short as possible to avoid the combinational explosion problem, and (iii) being less time consuming for the conformational calculations.

b. Molecular alignment is the second issue of concern in LBP modeling. The alignment methods can be classified into two approaches in terms of their elementary nature: point-based and property-based [29]. The points of the point-based method can be further discriminated as atoms, fragments, or chemical features [30]. In the point-based algorithms, pairs of atoms, fragments, or chemical feature points are usually superimposed employing a least-squares fitting. The major disadvantage of this approach is the requirement for predefined anchor points because the generation of these points can become problematic in the case of different ligands. Consequently, the property-based algorithms utilize molecular field descriptors, generally represented by sets of Gaussian functions, to generate alignments. Recently, new alignment methods have been developed, including stochastic proximity embedding [31], atomic property fields [32], fuzzy pattern recognition [33], and grid-based interaction energies [34].

c. The third challenge lies in the appropriate selection of training set compounds. Although this problem is simple and nontechnical, but it often puzzles researchers nonetheless. The type of ligand molecules, the size of the data set, and its chemical diversity largely affect considerably the final generated pharmacophore model [28].

10.2.4.2 Structure-based pharmacophore modeling

Structure-based pharmacophore (SBP) modeling is directly dependent on the 3D structures of macromolecular targets or macromolecule—ligand complexes. As the number of experimentally determined 3D structure of targets has grown to a very large number, SBP methods have attracted significant interest in the last decade. The approach is considered as the complementary one to the docking procedures, providing the same level of information as well as less demanding with respect to required computational resources. The protocol of SBP modeling involves analyzing the

Table 10.1 Various conformational sampling methods

Conformational sampling method	Characteristics
3DGEN	1. An algorithm for exhaustive generation of 3D isomers proceeding from molecular topology 2. A systematic approach 3. Based on a combinatorial process
Balloon	1. A stochastic algorithm 2. A multiobjective GA is employed 3. Removing conformational duplicates 4. Can effectively produce low-energy conformers, which are geometrically distinct from each other
CAESAR	1. A systematic search method 2. Based on a divide-and-conquer and recursive conformer buildup approach 3. Avoids conformer duplicates due to topological symmetry 4. Capable of reproducing the receptor-bound conformation
CONAN	1. A fragment-based, buildup approach combined with the rule-based method 2. Intersection strategy is used for conformational analysis
ConfGen	1. A systematic approach based on divide-and-conquer strategy 2. A rule-based approach is incorporated 3. Intended for the high-throughput generation of 3D databases
Conformation import workflow in the molecular operating environment (MOE)	1. A systematic approach with the use of divide-and-conquer strategy and combined with rule-based method 2. Each fragment is subject to a stochastic search algorithm, expected to locate most its low-energy conformers 3. A high-throughput conformer generator for library preparation
Corina	1. Fast approach 2. Straightforward performance, reasonable execution time, simplicity, and applicability to building large, 3D chemical inventories 3. Often gives a larger average RMSD to the bioactive conformation
Cyndi	1. Nondeterministic method 2. Based on a multiobjective genetic algorithm (MOGA) 3. Efficient, particularly when reproducing a bioactive conformation

(*Continued*)

Table 10.1 (Continued)

Conformational sampling method	Characteristics
Directed tweak	1. Originally for 3D query searches 2. A torsional space minimizer 3. Involving the use of analytical derivatives, and is fast as a result 4. Allowing 3D flexible searching on an interactive time scale
Genetic algorithm	1. Nondeterministic method 2. Suitable for the superimposition of sets of flexible molecules
MED-3DMC	1. Nondeterministic method 2. A Metropolis MC algorithm based on a SMARTS mapping of the rotational bonds and the MMFF94 VDW energy term is used 3. Capable of sampling the conformational space with a small average RMSD to the bioactive conformation.
MIMUMBA	1. A rule-based method 2. The revised version is in conjunction with the OMEGA approach 3. The rules can be extracted from statistical observations from a training portion of the CSD
Molecular dynamics simulation method	1. Can create reasonable conformers 2. Time consuming 3. Depends on temperature and simulation time 4. Is of special interest when dealing with molecular conformations in solution
Monte Carlo	1. A stochastic method 2. Solvation effects can be studied in an explicit solvent 3. Do not guarantee that the conformational space has been explored exhaustively; in particular, the output of a search may depend on the starting conformation 4. Efficient at the beginning of a search, but efficiency degrades as the search proceeds
OMEGA	1. A rule-based method (heuristic) combined with divide-and-conquer strategy 2. A high-throughput conformer generator for library preparation
Poling restraints	1. A stochastic method 2. Promoting conformational variation 3. Avoiding analogous conformers 4. Covering most of the pharmacophore space with significantly fewer conformers

(Continued)

Table 10.1 (Continued)

Conformational sampling method	Characteristics
Self-organization method	1. A distance geometry (DG) approach 2. Conformations generated are consistent with a set of geometric constraints, which include interatomic distance bounds and chiral volumes derived from the molecular connectivity table 3. Tending to produce relatively compact conformers
Systematic torsional grid method	1. A systematic search method 2. Uses a recursive tree search algorithm 3. Can generate all conformations of polypeptides which satisfy experimental NMR restraints 4. Time consuming
WIZARD	1. A rule-based method (heuristic) 2. Expert system techniques are adopted

complementary chemical features of the active site and their spatial relationships, and developing a pharmacophore model assembly with selected features [2].

SBP modeling can be further classified into two subclasses: macromolecule—ligand—complex based and macromolecule (without ligand) based. The macromolecule—ligand—complex-based approach is suitable in identifying the ligand binding site of the macromolecular target and determining the key interaction points between ligands and the target protein. LigandScout [35] is an excellent software program that incorporates the macromolecule—ligand—complex-based scheme. Programs like Pocket v.2 [36] and GBPM [37] are based on the same approach. The major limitation of this process is the requirement for the 3D structure of the macromolecule—ligand complex. As a consequence, it cannot be applied to cases when no ligands targeting the binding site of interest are known. This can be solved by the macromolecule-based approach. The SBP method implemented in the Discovery Studio software [18] is a typical example of a macromolecule-based approach [38].

The most commonly encountered difficulty for SBP modeling is the identification of too many chemical features for a specific binding site of the macromolecular target. A pharmacophore model consisting of too many chemical features (e.g., more than seven) is not appropriate for practical applications (e.g., 3D database screening). Therefore, it is always important to pick a restricted number of chemical features (usually three to seven) to create a reliable pharmacophore hypothesis. One more significant drawback is that the obtained pharmacophore hypothesis cannot replicate the quantitative structure—activity relationship (QSAR) because the model is generated based just on a single macromolecule—ligand complex or a single macromolecule.

10.2.5 Application of pharmacophore models

Enrichment in the pharmacophore techniques in the last two decades has made the approach one of the most significant tools in drug discovery. In spite of the advances in key techniques of pharmacophore modeling, there is space for additional improvement to derive more precise and best possible pharmacophore models, which include better handling of ligand flexibility, proficient molecular alignment algorithms, and more precise model optimization. Along with the pharmacophore-based VS and *de novo* design, the applications of pharmacophore have been extended to lead optimization [39], multitarget drug design [40], activity profiling [41], and target identification [42]. Application of the pharmacophore technique is demonstrated in a schematic way in Figure 10.5.

10.2.5.1 Pharmacophore model–based VS

Pharmacophore models can be used for querying the 3D chemical database to search for potential ligands; this process is termed *pharmacophore-based VS*. In the case of the pharmacophore-based VS approach, a pharmacophore hypothesis is taken as a

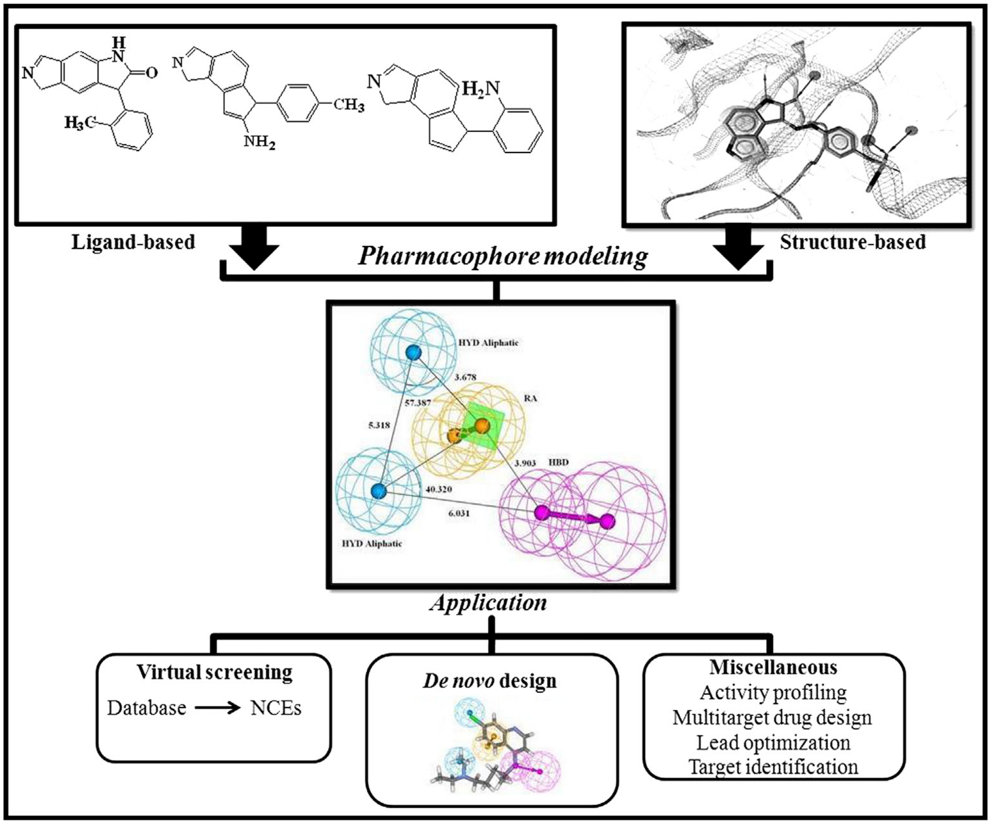


Figure 10.5 Diverse applications of the pharmacophore technique.

template. The intention behind the screening is actually to discover such hits that have chemical features similar to those of the template. Sometimes these hits might be related to known active compounds, but few have completely novel scaffolds. The screening process involves two major difficulties: handling the conformational flexibility of small molecules and pharmacophore pattern identification.

The flexibility of small molecules is handled either by preenumerating multiple conformations for each molecule or conformational sampling at search time. Pharmacophore pattern identification, usually known as *substructure searching*, is performed to check whether a query pharmacophore is present in a given conformer of a molecule. The commonly used approaches for substructure searching are Ullmann [43], the backtracking algorithm [44], and the Generic Match Algorithm (GMA) [45].

The most challenging problem for pharmacophore-based VS is that few percentages of the virtual hits are really bioactive. In simpler words, the screening results produce a higher false-positive rate, a higher false-negative rate, or both. Many factors like the quality and composition of the pharmacophore model and the macromolecular target information can contribute to this problem. The most probable factors are as follows:

a. The most critical one is the development of a robust and reliable pharmacophore hypothesis. Addressing this issue requires an inclusive validation and optimization of the pharmacophore model.

b. Different molecules can be retrieved in VS from different hypotheses of a single pharmacophore model, which is probably an important reason for the higher false-positive/false-negative rates in some studies.

c. The flexibility of target macromolecule in pharmacophore approaches is handled by introducing a tolerance radius for each pharmacophoric feature, which is unlikely to entirely account for macromolecular flexibility in some cases. Recent attempts [46] to integrate molecular dynamics simulation (MDS) into pharmacophore modeling have recommended that the pharmacophore models generated from MDS trajectories explain the considerably enhanced representation of the flexibility of pharmacophores.

d. The steric restriction by the macromolecular target which is not adequately considered in pharmacophore models, although it is partially accounted for by the consideration of excluded volumes. In most of the cases, interactions between a ligand and a protein are distance-sensitive, particularly the short-range interactions, such as the electrostatic interaction, which a pharmacophore model is tricky to account for. As a consequence, the combination of pharmacophore-based and docking-based VS can be considered as an efficient approach for VS.

10.2.5.2 Pharmacophore-based de novo *design*

Another vital application of pharmacophore is *de novo* design of ligands. In the case of pharmacophore-based VS, the obtained compounds are generally existing chemicals that might be patent protected. On the contrary, the *de novo* design approach can be used to generate entirely novel candidate structures that match to the requirements of a given pharmacophore. The first pharmacophore-based *de novo* design program is NEWLEAD [47]. It uses a set of disconnected molecular fragments that are consistent with a pharmacophore model as input. The selected sets of disconnected pharmacophore fragments are subsequently connected by using various linkers (such as atoms, chains, or ring moieties).

The limitation with NEWLEAD is that it can only handle cases in which the pharmacophore features are functional groups (not typical chemical features). The additional inadequacy of the NEWLEAD program is that the sterically illicit region of the binding site is not considered. As a result, the compounds created by the NEWLEAD program might be tricky to chemically synthesize. There are programs like LUDI [10] and BUILDER [48] that can also be used to amalgamate identification of SBP with *de novo* design. Both programs require knowledge of the 3D structures of the macromolecular targets.

More recently, a program called PhDD (a pharmacophore-based *de novo* design method of druglike molecules) has been designed by Huang et al. [49], to overcome the limitations of the present pharmacophore-based, *de novo* design software tools. PhDD can involuntarily create druglike compounds that satisfy the necessities of an input pharmacophore hypothesis. The pharmacophore used in PhDD can be consisted of a set of abstract chemical features and excluded volumes which are the sterically forbidden region of the binding site. In the case of PhDD, it first generates a set of new molecules that entirely conform to the requirements of the given pharmacophore model. Thereafter, a series of evaluation to the generated molecules are carried out, including the assessments of drug-likeness, bioactivity, and synthetic convenience.

10.2.6 Advantages and limitations of pharmacophore

Like any other approach, pharmacophore has both advantages and disadvantages. The major advantages and limitations are as follows:

Advantages

- Pharmacophore models can be used for VS on a large database.
- There is no need to know the binding site of the ligands in the macromolecular target protein, although this is true only for LBP modeling.
- It can be used for the design, optimization of drugs, and scaffolds hopping.
- It can conceptually be obtained even for 2D structural representation.
- This approach is comprehensive and editable. By adding or omitting chemical feature constraints, information can be easily traced to its source.

Limitations

- 2D pharmacophore is faster but less accurate than 3D pharmacophore.

- A pharmacophore is based only on the ligand structure and conformation. No interactions with the proteins are integrated. It is interesting to point out that in this case, SBP modeling can be used to solve the problem.
- It is sensitive to physicochemical features.

10.2.7 Software tools for pharmacophore analysis

Pharmacophore modeling is extensively used because of its immense accessibility through commercial software packages. Also, there is a freely available web server called PharmaGist (http://bioinfo3d.cs.tau.ac.il/PharmaGist/) for detecting a pharmacophore from a group of ligands known to bind to a particular target. A complete list of different commercialized and freely available software and program modules [19,35–38]) used for pharmacophore modeling is given in Table 10.2.

10.3 STRUCTURE-BASED DESIGN—DOCKING

10.3.1 Concept and definition of docking

Molecular docking is the study of how two or more molecular structures (e.g., drug and enzyme or protein) fit together [50]. In a simple definition, docking is a molecular modeling technique that is used to predict how a protein (enzyme) interacts with small molecules (ligands). The ability of a protein (enzyme) and nucleic acid to interact with small molecules to form a supramolecular complex plays a major role in the dynamics of the protein, which may enhance or inhibit its biological function. The behavior of small molecules in the binding pockets of target proteins can be described by molecular docking. The method aims to identify correct poses of ligands in the binding pocket of a protein and to predict the affinity between the ligand and the protein. Based on the types of ligand, docking can be classified as

- Protein–small molecule (ligand) docking
- Protein–nucleic acid docking
- Protein–protein docking

Protein–small molecule (ligand) docking represents a simpler end of the complexity spectrum, and there are many available programs that perform particularly well in predicting molecules that may potentially inhibit proteins. Protein–protein docking is typically much more complex. The reason is that proteins are flexible and their conformational space is quite vast.

Docking can be performed by placing the rigid molecules or fragments into the protein's active site using different approaches like clique-searching, geometric hashing, or pose clustering. The performance of docking depends on the search algorithm [e.g., MC methods, genetic algorithms (GAs), fragment-based methods, Tabu searches, distance geometry methods, and the scoring functions like force field (FF)

Table 10.2 Software and programs for pharmacophore modeling

Ligand-based methods

Software	Conformational analysis algorithm	Molecular alignment	Significant characteristics	Remarks
ALADDIN	N/A*	N/A*	Design and pharmacophore generation from geometric, steric, and substructure searching of 3D structures	Not commercialized
Apex-3D	Preenumerating method	Feature-based method	An expert system developed to represent, elucidate, and utilize knowledge on structure—activity relationships	Catalyst (Biovia, http://accelrys.com/)
APOLLO	Preenumerating method	Feature-based method	Identifying from a set of ligands their interaction points belonging to the receptor site and creating a pseudoreceptor	Not commercialized
CLEW	N/A*	Feature-based method	Utilizing the machine-learning method and geometrical fitting to develop the pharmacophore	Not commercialized
DANTE	N/A*	N/A*	Inferring pharmacophores automatically from structure—activity data, which include information about the shape of the binding cavity	Not commercialized
DISCO	Preenumerating method by Concord and Confort via the Sybyl interface	Bron—Kerbosh clique-detection algorithm	Considering 3D conformations of compounds as sets of interpoint distances	Integrated into the Sybyl interface, which is available from Tripos Inc. (www.tripos.com)
GALAHAD	Both preenumerating method and on-the-fly	Atom-based method	A more sophisticated GA is used for pharmacophore modeling	Integrated into the Sybyl interface, which is available from Tripos Inc. (www.tripos.com)

(*Continued*)

Table 10.2 (Continued)

Ligand-based methods

Software	Conformational analysis algorithm	Molecular alignment	Significant characteristics	Remarks
GAMMA	On-the-fly	Atom-based method	The conformational search and the pattern identification are performed simultaneously by utilizing the GA technique	Not commercialized
GASP	On-the-fly	Atom-based method	A flexible GA is used for pharmacophore identification	Integrated into the Sybyl interface, which is available from Tripos Inc. (www.tripos.com)
HipHop	Preenumerating method by the Poling algorithm	Feature-based method	Identifying common features by a pruned exhaustive search (qualitative model)	Discovery Studio (Biovia, http://accelrys.com/)
HypoGen	Preenumerating method by the Poling algorithm	Feature-based method	Designed to correlate structure and activity (quantitative model)	Discovery Studio (Biovia, http://accelrys.com/)
HypoRefine	Preenumerating method by the Poling algorithm	Feature-based method	An extension to the HypoGen Exclusion volumes are involved	Discovery Studio (Biovia, http://accelrys.com/)
MOE	Preenumerating method ranging from molecular dynamics to stochastic methods and systematic search	Property-based algorithm	A pharmacophore is defined manually by applying schemes using a Pharmacophore Query Editor	Chemical Computing Group, Inc. (www.chemcomp.com)
MPHIL	On-the-fly	Atom-based method (rigid)	Based on clique detection and GA	Not commercialized
PharmaGist	On-the-fly	Feature-based method	A webserver for LBP detection	http://bioinfo3d.cs.tau.ac.il/PharmaGist
PHASE	Preenumerating method by Schrödinger's ConfGen technology	Feature-based method (called *sites*)	Very flexible and user friendly. SMARTS pattern matching is used for feature location. Excluded volumes are included.	Schrödinger Inc. (www.schrodinger.com)

(*Continued*)

Table 10.2 (Continued)

Ligand-based methods

Software	Conformational analysis algorithm	Molecular alignment	Significant characteristics	Remarks
RAPID	Preenumerating method	Atom-based method	A rigid alignment based on mapping triangles of 3D atom coordinates	Not commercialized
SCAMPI	On the fly	Feature-based method	Can handle large heterogeneous data sets	Not commercialized
XED	Preenumerating method	Molecular field-based method	Using field points to describe the VDW and electrostatic potential that surround molecules	Marketed by Cresset Biomolecular (http://www.cresset-group.com/)

Structure-based methods

Software	Molecular alignment	Significant characteristics	Remarks
GBPM	Complex-based	Based on logical and clustering operations with 3D maps computed by the GRID program on structurally known molecular complexes. Particularly suitable for identifying protein—protein interaction areas.	Not commercialized
LigandScout	Complex-based	Incorporating a complete definition of 3D chemical features. Pharmacophoric feature points-based pattern-matching alignment algorithm is used. Intuitive and easy to use.	Marketed by Inte:Ligand (www.inteligand.com/ligandscout/)
Pocket v.2	Complex-based	Capable to generate a pharmacophore model with a rational number of features when one complex structure is available.	Not commercialized
SBP	Apoprotein-based	Directly converting LUDI interaction maps within the protein binding site into Catalyst pharmacophoric features.	Discovery Studio (Biovia, http://accelrys.com/)

N/A*: Not applicable or the exact information is not available.

methods and empirical free energy scoring functions]. The first step of docking is the generation of composition of all possible conformations and orientations of the protein paired with the ligand. The second step is that the scoring function takes input and returns a number indicating favorable interaction [51].

To identify the active site of the protein, first, selection of the required X-ray cocrystallized structure from the protein data bank (PDB) is performed, and then extracting the bound ligand, one can optimize the protein active site of interest. But the process of identification of the active site in a protein is critical when the bound ligand is absent in the crystal structure. In that case, one has to do the following procedures:

a. One can perform comprehensive literature review of the source papers (from which the X-ray crystal structure has been included in PDB) to identify the active site of residues.

b. If any established drug giving the same pharmacological action of interest is available for the protein, then the active sites for this drug should be identified. In the initial phase of analysis, one can try these residues as active binding sites for the test ligands.

c. Every docking software program usually has a particular algorithm to identify the active site of the protein by allowing binding of the ligand in different parts of the protein and exploring the best possible binding position of the ligands with the protein.

10.3.2 Definition of fundamental terms of docking

To understand the docking study better, one needs to know the basic terms related with the docking study. The most commonly used terms connected with docking studies are defined next. All the discussed terms are graphically represented in Figure 10.6:

Receptor: A *receptor* is a protein molecule or a polymeric structure in or on a cell that distinctively recognizes and binds a molecule (ligand) acting as a molecular messenger. When such ligands bind to a receptor, they cause some kind of cellular response.

Ligand: A *ligand* is the complementary partner molecule that binds to the receptor for effective bimolecular response. Ligands are most often small drug molecules, neurotransmitters, hormones, lymphokines, lectins, and antigens, but they could also be another biopolymer or macromolecule (in the case of protein—protein docking).

Docking: *Docking* is a molecular modeling technique designed to find the proper fit between a ligand and its binding site (receptor).

Dock pose: A ligand molecule can bind with a receptor in a multiple positions, conformations, and orientations. Each such docking mode is called a *dock pose.*

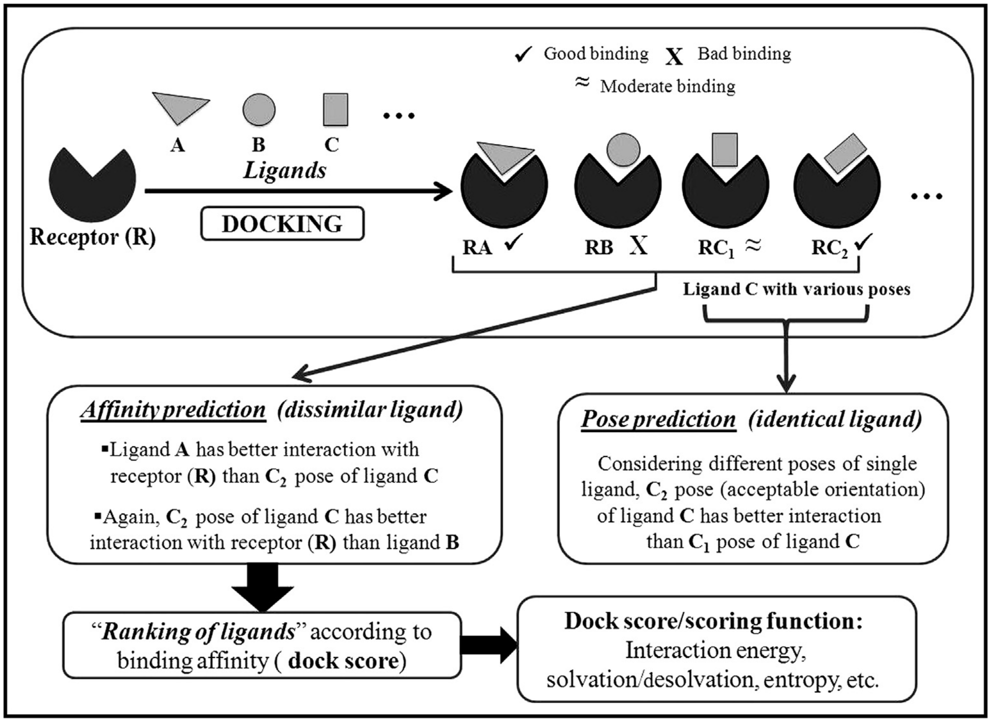

✓ Good binding X Bad binding
≈ Moderate binding
A B C ...
Ligands
DOCKING
Receptor (R)
RA ✓ RB X RC₁ ≈ RC₂ ✓ ...
Ligand C with various poses
Affinity prediction (dissimilar ligand)
▪Ligand A has better interaction with receptor (R) than C₂ pose of ligand C
▪Again, C₂ pose of ligand C has better interaction with receptor (R) than ligand B
Pose prediction (identical ligand)
Considering different poses of single ligand, C₂ pose (acceptable orientation) of ligand C has better interaction than C₁ pose of ligand C
"Ranking of ligands" according to binding affinity (dock score)
Dock score/scoring function:
Interaction energy, solvation/desolvation, entropy, etc.

Figure 10.6 Graphical representation of commonly used terms in docking studies.

Binding mode: *Binding mode* is the orientation of the ligand relative to the receptor, as well as the conformation of the ligand and receptor when they are bound to each other.

Dock score: The process of evaluating a particular pose by counting the number of favorable intermolecular interactions such as hydrogen bonds and hydrophobic contacts. In order to recognize the energetically most favorable pose, each pose is evaluated based on its compatibility to the target in terms of shape and properties such as electrostatics and generate corresponding dock score. A good dock score for a given ligand signifies that it is potentially a good binder.

Ranking: *Ranking* is the process of classifying which ligands are most likely to interact favorably to a particular receptor based on the predicted free energy of binding. After completion of docking, all ligands are consequently ranked by their respective dock scores (i.e., their predicted affinities). This rank-ordered list is then employed for further synthesis and biological investigation only for those compounds that are predicted to be most active.

Pose prediction: *Pose prediction* can be defined as searching for the accurate binding mode of a ligand, which is typically carried out by performing a number of trials and keeping those poses that are energetically best. It involves finding the correct orientation and the correct conformation of the docked ligand due to their flexible nature.

Scoring or affinity prediction: Affinity prediction or scoring functions are applied to the energetically best pose or *n* number of best poses found for each ligand, and comparing the affinity scores for different ligands give their relative rank ordering. [52].

Scoring functions are generally divided into two main groups. One main group comprises knowledge-based scoring functions that are derived using statistics for the observed interatomic contact frequencies, distances, or both in a large database of crystal structures of protein—ligand complexes. The other group contains scoring schemes based on physical interaction terms [53]. These so-called energy component methods are based on the assumption that the change in free energy upon binding of a ligand to its target can be decomposed into a sum of individual contributions:

$$\Delta G_{\text{bind}} = \Delta G_{\text{int}} + \Delta G_{\text{solv}} + \Delta G_{\text{conf}} + \Delta G_{\text{motion}} \tag{10.8}$$

The terms defined for the main energetic contributions to the binding event are as follows: specific ligand—receptor interactions (ΔG_{int}), the interactions of ligand and receptor with solvent (ΔG_{solv}), the conformational changes in the ligand and the receptor (ΔG_{conf}), and the motions in the protein and the ligand during the complex formation (ΔG_{motion}).

10.3.3 Essential requirements of docking

1. *Receptor crystal structures*: To execute the docking study, it is essential to have the receptor structures of interest. The structure of the receptor can be determined by experimental techniques such as X-ray crystallography or nuclear magnetic resonance (NMR), and can be easily downloaded from the PDB (http://www.rcsb.org/pdb/home/home.do). The quality of the receptor structure plays a crucial role in the success of docking studies. In general, the higher the resolution (preferably <2 Å) of the employed crystal structure, the better the observed docking results are. Another important criterion for examining the quality of a receptor structure is Debye—Waller factor (DWF), or *B*-factor; or the temperature factor, which is used to describe the attenuation of X-ray scattering or coherent neutron scattering caused by thermal motion. It signifies the relative vibrational motion of different parts of the protein. Atoms with low *B*-factors belong to a part of the structure that is well ordered, and atoms with large *B*-factors generally belong to part of the structure that is very flexible. As a consequence, it is important to ensure that the *B*-factors of the atoms in the binding site region are logical, as high values imply that their coordinates are less reliable. Identification of the bound-ligand in the cocrystal structure and the knowledge about its interaction

with the corresponding protein's amino acid residues are very important before starting a docking study.

2. *Receptor homology modeling and threading techniques*: On the contrary, if the X-ray crystal structure of the protein is not available, one can opt for the protein structure prediction techniques. In that case, most commonly applied techniques are "threading" and "homology modeling" [54,55]. In the case of threading or the fold recognition technique, an estimation is made whether a given amino acid sequence is compatible with one of the ligands in a database. On the other hand, homology or comparative modeling relies on a correlation or homology between the sequence of the target protein and at least one known structure. Correct homology models can be generated, provided that the sequence identity of a given target sequence is >50% to a known structure template. Modest homology model building efforts could potentially create receptor structures for entire target families. Importantly, homology modeling is a comparatively economical method for generating a diversity of receptor conformations using either single-template or multiple-template structures enhancing the understanding of selectivity.
3. *A set of ligands of interest*: Once the 3D structure of the protein of interest has been attained from either experiments (X-ray crystal structure) or predictions (receptor from homology modeling), the docking study can be performed using ligands of interest employing a multiplicity of docking techniques. If the function of the protein is unknown, it may be vital to search its structure for hypothetical binding sites. These binding sites can be explored for the binding of selected ligands or they can be compared with known binding sites. An analysis of the binding site characteristics and the interactions with a given ligand can lead to important insights for the design of novel ligands or the docking of assumed ligand molecules [56].

10.3.4 Categorization of docking

As discussed earlier in Section 10.3.1, docking can be categorized into three main classes: (i) protein−ligand docking, (ii) protein−nucleic acid docking, and (iii) protein−protein docking. Among these, protein−ligand docking is a common research area because of its importance to structure-based drug design [3]. Again, the protein−ligand docking can be classified in the following manner: (a) rigid-body docking, where both the receptor and ligand are treated as rigid; (b) flexible ligand docking, where the receptor is held rigid, but the ligand is treated as flexible; and (c) flexible docking, where both receptor and ligand flexibility is considered. Thus far, the most commonly used docking algorithms use the rigid receptor/flexible ligand model. Here, we have categorized the protein−ligand docking in terms of the three most important aspects: (i) protein flexibility, (ii) ligand sampling, and (iii) scoring function, as illustrated in Figure 10.7.

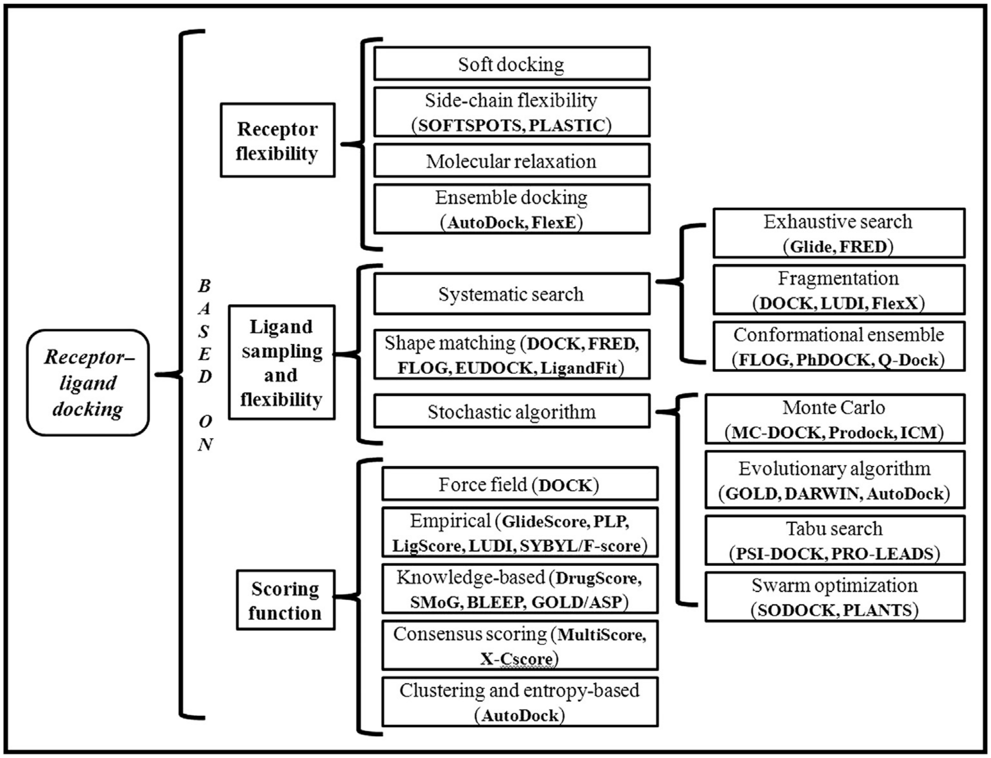

Receptor–ligand docking
BASED ON
Receptor flexibility
Soft docking
Side-chain flexibility (SOFTSPOTS, PLASTIC)
Molecular relaxation
Ensemble docking (AutoDock, FlexE)
Ligand sampling and flexibility
Systematic search
Shape matching (DOCK, FRED, FLOG, EUDOCK, LigandFit)
Stochastic algorithm
Exhaustive search (Glide, FRED)
Fragmentation (DOCK, LUDI, FlexX)
Conformational ensemble (FLOG, PhDOCK, Q-Dock)
Monte Carlo (MC-DOCK, Prodock, ICM)
Evolutionary algorithm (GOLD, DARWIN, AutoDock)
Tabu search (PSI-DOCK, PRO-LEADS)
Swarm optimization (SODOCK, PLANTS)
Scoring function
Force field (DOCK)
Empirical (GlideScore, PLP, LigScore, LUDI, SYBYL/F-score)
Knowledge-based (DrugScore, SMoG, BLEEP, GOLD/ASP)
Consensus scoring (MultiScore, X-Cscore)
Clustering and entropy-based (AutoDock)

Figure 10.7 Categorization of protein–ligand docking.

10.3.4.1 Receptor/protein flexibility

Ligand binding usually induces protein conformational changes (ranging from local rearrangements of side-chains to large domain motions) or induced fit (Figure 10.8) upon ligand binding in order to maximize energetically favorable interactions with the ligand [57]. The algorithm behind the most induced-fit mechanisms is hydrophobic interaction or hydrophobic collapse of the receptor around the bound ligand [58]. Due to the large size and many degrees of freedom of proteins, their flexibility is one of the most challenging issues in molecular docking. There are varying degrees of receptor flexibility. The degree of flexibility that one could incorporate in a given experiment is directly proportional to computational complexity and cost. The protein flexibility can be grouped into four major categories: (i) soft docking, (ii) side-chain flexibility, (iii) molecular relaxation, and (iv) protein ensemble docking [59].

10.3.4.1.1 Soft docking

Soft docking is the simplest approach, which considers protein flexibility in absolute terms. Soft docking algorithms attempt to allow flexibility of the receptor and ligand

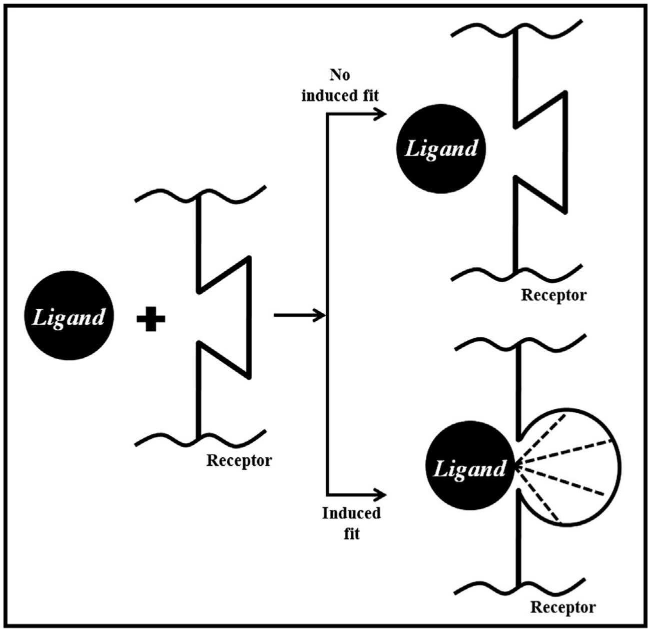


Figure 10.8 Graphical representation of induced-fit docking.

structures by using a relaxed representation of the molecular surface. The method allows for a small degree of overlap between the ligand and the protein through the use of additional energy terms—usually interatomic van der Waals (VDW)—in the empirical scoring function [60]. The advantages of soft docking are its computational competence and easiness for implementation. It is important to remember that soft docking can account for only minute conformational changes. The soft docking concept proposed by Jiang and Kim describes the molecular surface and volume as a "cube representation" [61]. This cube representation implies implicit conformational changes by way of size/shape complementarity, close packing, and, most important, liberal steric overlap.

10.3.4.1.2 Side-chain flexibility

Allowing active site side-chain flexibility is another way to provide receptor flexibility, in which backbones are kept fixed and side-chain conformations are sampled. The method originally proposed by Leach [62] uses pregenerated side-chain rotamer libraries that subsequently are subjected to optimization during a ligand docking procedure via the dead-end elimination algorithm. The optimized ligand/side-chain orientations are then scored in order to rank the lowest energy combination of side-chain and ligand conformers. Since the invention of this approach, researchers have proposed many improved techniques to incorporate continuous or discrete side-chain flexibility in ligand docking [63,64].

10.3.4.1.3 Molecular relaxation

The molecular relaxation method accounts for protein flexibility by first using rigid-body docking to place the ligand into the binding site and then relaxing the protein backbone and side-chain atoms nearby. Initially, the rigid-body docking allows atomic clashes between the protein and the placed ligand conformations in order to consider the protein conformational changes. Thereafter, the formed complexes are relaxed or minimized by MC, MDS, or other methods [65]. The MDS calculate the time-dependent behavior of a molecular system, which provides detailed information on the fluctuations and conformational changes of proteins and nucleic acids. These methods are now regularly used to examine the structure, dynamics, and thermodynamics of biological molecules and their complexes.

The advantage of the molecular relaxation method is the addition of certain backbone flexibility in addition to the side-chain conformational changes. However, compared to the side-chain flexibility methods, the relaxation method is more demanding on the scoring function because it involves not only the side-chain movement, but also the more challenging task of backbone sampling. One of the significant drawbacks of this approach is that it is time consuming.

10.3.4.1.4 Docking of multiple protein structures/ensemble docking

Ensemble docking, which has gained considerable attention as a method of incorporating protein flexibility, utilizes an ensemble of protein structures to represent different possible conformational changes [66]. Commonly, the full receptor ensembles are generated by MDS, MC simulation, or homology modeling approaches. The ensembles can be generated experimentally from NMR solution structure determination or multiple X-ray crystal structures. Strict comparisons have revealed that there is a considerable overlap of dynamic information between theoretically derived molecular dynamics ensembles and experimentally derived NMR ensembles [67]. The first ensemble study was done by Knegtel et al. [68], in which an averaged energy grid was constructed by combining the energy grids generated from each experimentally determined protein structures using a weighting scheme, followed by standard ligand docking. Generally, the ensemble docking algorithm is not used for generating new protein structures; instead, it is used for selecting the induced-fit structure from a given protein ensemble.

10.3.4.2 Ligand sampling and flexibility

Ligand sampling is one of the most basic components in protein–ligand docking. Given a protein target, the sampling algorithm generates possible ligand orientations or conformations (poses) around the selected binding site of the protein. It is interesting to point out that the binding site can be the experimentally determined active site, a dimer interface, or another site of interest. Without any doubt, ligand sampling and its flexibility are the significant areas in protein—ligand docking research. There are three types of ligand-sampling algorithms: shape matching, systematic search, and stochastic algorithms, all of which are discussed in the next sections.

10.3.4.2.1 Shape matching

The shape matching approach is one of the common sampling algorithms that is employed in the initial stages of the docking or in the earlier step of other, more advanced ligand sampling methods. The ligand is placed using the criterion that the molecular surface of the placed ligand must harmonize the molecular surface of the binding site on the protein. Generally, three translational and three rotational degrees of freedom of the ligand are allocated for many possible ligand-binding orientations. Therefore, how the placed ligand gets bound in the protein site with a good shape complementarity is the major goal of the shape matching algorithm. It is important to remember that the conformation of the ligand is normally fixed during shape matching [69]. The major advantage of shape matching is its computational efficiency.

10.3.4.2.2 Systematic search

Systematic search algorithms are usually employed for flexible ligand docking, which create all the probable ligand binding conformations by exploring all degrees of freedom of the ligand. The systematic search method can be divided into three subclasses:

a. *Exhaustive search*: The most uncomplicated systematic algorithms are exhaustive search methods, in which flexible ligand docking is performed by systematically rotating all possible rotatable bonds of the ligand at a given interval. In spite of its sampling totality for ligand conformations, the number of the choices can be huge due to an increase in the number of rotatable bonds. As a consequence, to make the docking process realistic, geometric and chemical constraints are normally applied to the initial screening of ligand poses, and the filtered ligand conformations are further subject to the more precise refinement and optimization measures.

b. *Fragmentation approach*: The basic idea behind this approach is that the ligand is first divided into a number of fragments. Then, the ligand-binding conformation is grown by placing one fragment at a time in the binding site or by docking all the fragments into the binding site and linking them covalently.

c. *Conformational ensemble*: In the conformational ensemble methods [69], ligand flexibility is achieved by rigidly docking an ensemble of pregenerated ligand conformations with other programs (e.g., OMEGA). Then, ligand-binding modes from different docking runs are collected and ranked according to their binding energy scores.

10.3.4.2.3 Stochastic algorithms

The fundamental algorithm behind the stochastic approach is that ligand-binding orientations and conformations are sampled by making random changes to the ligand at each step in the conformational space and the translational and rotational space of the ligand, respectively. The random change will be accepted or rejected according to a probabilistic criterion. The stochastic algorithms can be classified into four different categories [70]:

a. *MC methods*: The probability to allow a random change is determined by employing the Boltzmann probability function.

b. *Evolutionary algorithms*: These involve a search for the right ligand-binding mode based on the idea from the evolutionary process in biological systems.

c. *Tabu search methods*: The probability of approval relies on the explored areas in the conformational space of the ligand. The random change will be rejected if the RMSD between the present ligand-binding conformation and any of the formerly recorded solutions is less than a cutoff; otherwise, the random change will be accepted.

d. *Swarm optimization method*: This particular algorithm tries to determine the best possible solution in a search space by modeling swarm intelligence. Movements of a ligand mode through the search space are directed by the information of the best positions of its neighbors.

10.3.4.3 Docking scoring functions

The fundamental element behind determining the accuracy of a protein—ligand docking algorithm is the generated scoring function during the docking study [71]. Swiftness and precision are the two essential aspects of any scoring function. An ideal scoring function would be both computationally proficient and consistent. Numerous scoring functions have been developed since the introduction of docking studies. The scoring functions are broadly grouped into five basic categories according to their methods of derivation.

10.3.4.3.1 FF scoring functions

FF scoring functions [72] rely on the partitioning of the ligand-binding energy into individual interaction terms such as VDW energies, electrostatic energies, and bond stretching/bending/torsional energies, employing a set of derived FF parameters such as the AMBER [73] or CHARMM [74] FFs. The major challenges in FF scoring functions are accounting for the solvent effect and accounting for the entropic effect.

10.3.4.3.2 Empirical scoring functions

The binding energy score of a complex is calculated by adding up a set of weighted empirical energy terms (such as VDW energy, electrostatic energy, hydrogen-bonding energy, desolvation term, entropy term, and hydrophobicity term) in empirical scoring functions.

Compared to the FF scoring functions, the empirical scoring functions are usually much more computationally proficient due to their simple energy terms. It is interesting to point out here that the general applicability of an empirical scoring function relies on the training set due to the fact that it fits known binding affinities of its training set.

10.3.4.3.3 Knowledge-based scoring functions

Knowledge-based scoring functions result from the structural information in experimentally determined protein—ligand complexes [75]. The theory beneath the knowledge-based scoring functions is the potential of mean force, which is defined by the inverse Boltzmann relation. This scoring function maintains a good balance between accuracy and speed. The difficulty for this scoring function is the calculation for the aforementioned reference state. It can be classified into three categories based on the methods of computation: (a) traditional atom-randomized reference state, (b) corrected reference state, and (c) circumventing the reference state.

10.3.4.3.4 Consensus scoring

Consensus scoring is not a typical scoring function; rather, it is a technique involved in protein—ligand docking [76]. It advances the probability of finding an accurate solution by amalgamating the scoring information from multiple scoring functions in anticipation of eliminating the inaccuracies of the individual scoring functions. As a consequence, the main difficulty in consensus scoring is how to create the combination rule for each score so that the true binders can be discriminated from others according to the consensus rule.

10.3.4.3.5 Clustering- and entropy-based scoring methods

To enhance the performances of scoring functions, there is another new technique called the *clustering-based scoring method*, which includes the entropic effects by dividing generated ligand-binding modes into different clusters [77]. The entropic contribution in each cluster is calculated by the configurational space covered by the ligand poses or the number of ligand poses in the cluster. One disadvantage of clustering-based scoring methods is that its performance relies on the ligand sampling protocol, which is highly dependent on the docking program.

10.3.5 Basic steps of docking

Fundamentally, docking is a three-step process irrespective of software and docking algorithms [78]. The steps are as follows:

a. *Ligand preparation*: The first step is to prepare the ligands. In this process, all the duplicate structures should be removed, and options for ionization change, tautomer, isomer generation, and 3D generator must be set in the working software platform for the respective ligands.

b. *Protein preparation*: Hydrogen atoms should be added and the protein must be minimized using software-specific FF, followed by the removal of water molecules except in the active site. The protein should be adjusted by fixing any serious errors like incomplete residues near the active site. The charges and atom types for any metal atoms should be set properly, if needed. If there are bonds to metal ions, the bonds should be deleted, followed by adjusting the formal charges of the atoms that were attached to the metal, as well as the metal itself. The protein molecule, thus prepared, is the total receptor ready for docking.

c. *Ligand—protein docking*: After ensuring that protein and ligands are in the correct form for docking, in a few cases the receptor grid files are generated using a grid—receptor generation program for grid-based docking. The grid box is generally generated at the centroid of the ligand bound to the active site of the receptor. In other cases, active pockets of the protein are identified to dock the prepared ligand in those identified pockets. Initially, all the molecules of the data set should be docked into the active site of the protein and the interaction energies between each ligand and the receptor can be calculated. The obtained results are then needed to be compared with those of the bound ligand of the crystallized protein structure in order to assess whether the molecules fit into the specified active site of the receptor or not. A set number of ligand poses should be saved for each conformation of the ligand. A predefined number of docking poses thus saved for each conformation of the compound can be ranked according to their dock score function, and then their interaction with the receptor can be analyzed. From the docking studies, the receptor—ligand interactions are correlated with the biological activity of the data set compounds. The structural validation of the docking procedure is done by extracting the cocrystallized ligand from the active site of the receptor and redocking it to the receptor to ensure that it binds to the same active site and interacts with the same amino acid residues as before. The basic steps are schematically illustrated in Figure 10.9.

10.3.6 Challenges and required improvements in docking studies

A significant amount of work has been performed to devise superior docking programs and scoring functions over the past years. However, there is still room for improvement. This section presents some of the primary challenges and the required improvements that will advance the performance of docking and scoring [79].

Challenges:

a. *Water molecules in protein*: Water molecules often play a significant role in protein—ligand interaction. If water-mediated interactions during docking is ignored, the estimated interaction energy of a given ligand conformation may be too low. On the contrary, if one holds water molecules present in the crystal

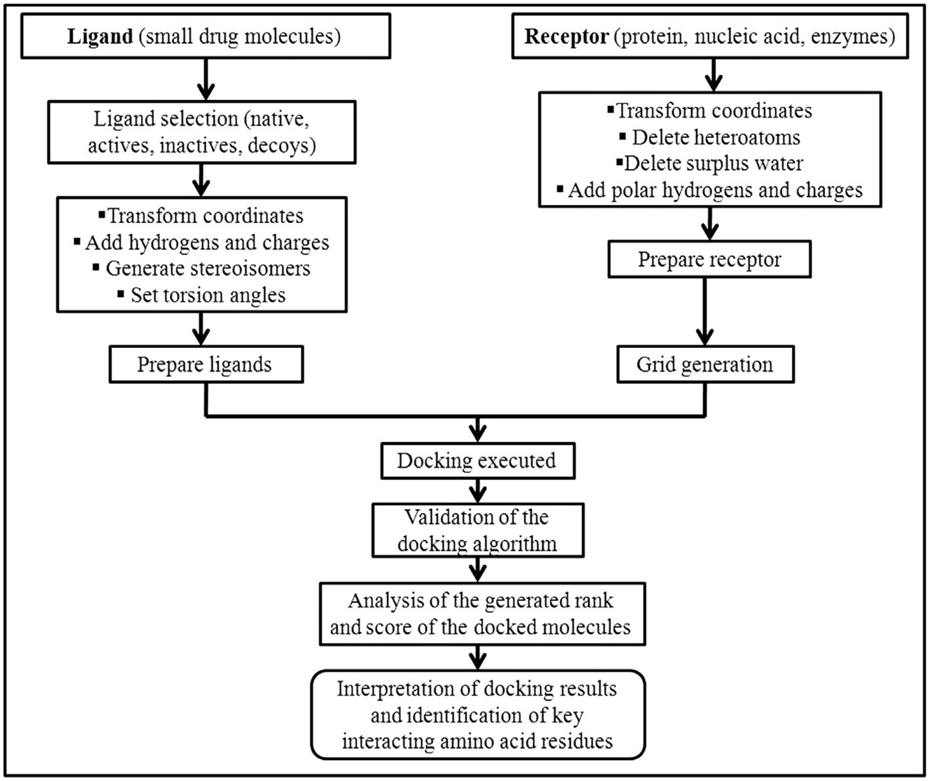

Ligand (small drug molecules)
Ligand selection (native, actives, inactives, decoys)
▪Transform coordinates
▪Add hydrogens and charges
▪Generate stereoisomers
▪Set torsion angles
Prepare ligands
Receptor (protein, nucleic acid, enzymes)
▪Transform coordinates
▪Delete heteroatoms
▪Delete surplus water
▪Add polar hydrogens and charges
Prepare receptor
Grid generation
Docking executed
Validation of the docking algorithm
Analysis of the generated rank and score of the docked molecules
Interpretation of docking results and identification of key interacting amino acid residues

Figure 10.9 Basic steps of the docking formalism.

protein structure, then the binding pose and affinity of a ligand will not be reliable. Thus, it is very difficult to treat the water molecules effectively. To perform reliable and acceptable docking, one first need to recognize probable positions for water molecules where they could interact with the protein and ligand, and subsequently, one must be capable to predict whether a water molecule is indeed present at that position.

b. *Tautomers and protomers*: Another significant challenge with docking is consideration for the various tautomeric and protomeric states that the molecules can adopt. Most of the time, molecules such as acids or amines are stored in their neutral forms. As they are ionized under physiological conditions, it is essential to ionize them prior to docking. Although ionization is easy to attain, but the problem of tautomer generation is already much more difficult, as other questions will arise, such as: Which tautomer should one use? Should one use more than one or all possible tautomers for a given molecule? Not only tautomers, but also different ionization states of ligands provide real challenges in docking.

c. *Docking into flexible receptors*: One of the most challenging problems in docking is dealing with flexible receptors. Numerous examples have become known where the same protein adopts different conformations depending on which ligand it binds to [80]. In order to deal with the trouble of flexible receptors in docking, several approaches have been proposed: (1) letting the receptor or parts of it move during docking; (2) docking the compounds into numerous different conformations of the same receptor and aggregating the results; and (3) docking into averaged receptor representations. In a few cases, more than one of these methods are used based on the requirements in question.

Required improvements:

d. *Multiple active-site corrections (MASC)*: A possible way of improving docking results is the application of MASC, a simple statistical correction [81]. The scoring functions prefer certain ligand types or characteristics, such as large or hydrophobic ligands. As a consequence, some ligands are predicted to be good binders regardless of whether these ligands will bind to specific active sites. Therefore, MASC has been introduced, which can be interpreted either as a statistical measure of ligand specificity or as a correction for ligand-related bias in the scoring function. In order to calculate the MASC scores, each ligand is docked into a number of unrelated binding sites of different binding site characteristics. The corrected score (or MASC score) S'_{ij} for ligand molecule i in binding site j is calculated as follows:

$$S'_{ij} = \frac{(S_{ij} - \mu_i)}{\sigma_i} \tag{10.9}$$

where S_{ij} is the uncorrected score for the ligand, and μ_i and σ_i represent the mean and standard deviation of the scores for the ligand molecule i across the different binding sites. Thus, the MASC score S'_{ij} represents a measure of specificity of molecule i for binding site j compared to the other binding sites.

e. *Docking with constraints*: By introducing a constraint during docking, it is feasible to control the way the poses are generated and the ones that are preferentially set aside. For example, in the case of the DockIt program [82], one can apply distance constraints between the ligand and protein that are consequently utilized during pose generation via a distance geometry approach.

f. *Postprocessing*: There are two approaches of postprocessing that can be employed in the case of a docking study: (i) applying postdock filters and (ii) using tailor-made rescoring functions. The postdock filters are theoretically simple and may correspond to certain geometric criteria, like the existence of certain interactions (e.g., a hydrogen bond with a selected residue or a polar interaction) or the filling of a specified pocket in the active site. Again, all scoring functions may exhibit biased behavior with certain compound classes or functional groups. To diminish the impact of this difficulty and to decrease the statistical noise, composite scoring, or rescoring methods have been introduced [83]. Rather than using a single scoring function, several scoring functions are merged such that in order to be classified as a potential binder, a molecule has to be scored well by a number of different scoring functions. Another way of postprocessing is to use the docking results as input to develop a Bayesian model with the aim of reducing the numbers of false positives and false negatives [84].

10.3.7 Applications of docking

The docking technology is successfully applied at multiple stages of the drug design and discovery process for three main purposes: (1) predicting the binding mode of a known active ligand, (2) identifying new ligands using VS, and (3) predicting the binding affinities of allied compounds from a known active series. The prediction of a ligand-binding mode in a protein active site has been the most successful area. In the broader perspective, the major specific applications of docking are listed here to get a proper dimension of the use of docking studies in the drug discovery process:

- The determination of the lowest free-energy structures for the receptor—ligand complex
- Calculation of the differential binding of a ligand to two different macromolecular receptors
- Study of the geometry of a particular ligand—receptor complex.
- Searching of a database and ranking of hits for lead generation and optimization for future drug candidate.
- To propose the modification of lead molecules to optimize potency or other properties.
- Library design and data bank generation.
- Screening for the side effects that can be caused by interactions with proteins, like proteases and cytochrome P450, can be done.
- It is also possible to check the specificity of a potential drug against homologous proteins through docking.
- Docking is also a widely used tool in predicting protein—protein interactions.
- Docking can create knowledge of the molecular association, which aids in understanding a variety of pathways taking place in the living system.
- To reveal possible potential pharmacological targets.
- Docking-based virtual HTS is less expensive than normal HTS and faster than conventional screening.

The docking study has a huge role not only in lead drug identification process, but also in search of potential target identification for different diseases [79]. A representative list of marketed or clinical trial drugs employing structure-based drug design—docking study is given in Table 10.3. For a more elaborate illustration, please see Chapter 11.

Table 10.3 Representative examples of marketed drugs employing the structure-based drug design—docking study

Generic name	Manufacturer	Inhibit/Target
AG85, ag337, ag331	Agouron	Thymidylate synthase
Aliskiren	Novartis	Renin inhibitors
Amprenavir	GlaxoSmithKline	HIV protease
Boceprevir	Schering—Plough	Protease inhibitor used for treating hepatitis caused by hepatitis C virus (HCV)
Captopril	Bristol Myers—Squibb	Reversible inhibitor of angiotensin-converting enzyme (ACE)
Dorzolamid	Merck Sharp and Dohme	Carbonic anhydrase (hypercapnic ventilatory failure)
ERα and ERβ	Information not available	Estradiol (E2) analogs
Indinavir	Merck	HIV protease
Inverase	Hoffman La Roche	HIV protease
LY-517717	Lilly/Protherics	Inhibitors of factor Xa serine protease
Nelfinavir	Hoffman La Roche	HIV protease
Nolatrexed dihydrochloride	Agouron	Thymidylate synthase (TS)
Norvir	Abbot	HIV protease
NVP-AUY922	Novartis	Heat shock protein 90 (HSP90)
Raltitrexed	AstraZeneca	Thymidalate
Raltegravir	Merck	HIV integrase
Rupintrivir	Agouron	Irreversible inhibitors of human rhinovirus (HRV) 3C protease
Saquinavir	Hoffman La Roche	HIV protease
TMI-005	—	Dual inhibitor of tumor necrosis factor-α (TNFα) converting enzyme (TACE) and matrix metalloproteinases (MMPs)
Zanamivir	Gilead Sciences	Neuraminidase inhibitor

10.3.8 Docking software tools

A large number of docking programs and search algorithms have been reported since the invention of docking [79]. Although the basic steps of docking processes are more or less identical, these docking programs vary fundamentally with respect to the docking algorithm, ligand search strategy, and scoring function techniques. A list of popular docking software programs is given in Table 10.4.

Table 10.4 Available software tools for the docking study

Software	Algorithm and remarks
AutoDock	AutoDock is a suite of automated docking tools capable of predicting how small molecules, such as substrates or drug candidates, bind to a receptor of a known 3D structure. The Lamarckian GA is used as the algorithm. Website: http://autodock.scripps.edu/
Discovery Studio	The conformational search of the ligand poses is performed by the MC trial method. Preprocessing of ligands is performed using the ligand fit program with selecting one of the energy grid out of three energy grids (PLP1, Dreiding, and CFF) available in Discovery Studio. The docking poses saved for each conformation of the compound are ranked according to their dock scores based on LigScore1, LigScore2, PLP1, PLP2, Jain, and PMF function. Website: http://accelrys.com/
DOCK	DOCK is a program that can examine possible binding orientations of protein–protein and protein–DNA complexes. It can be used to search databases of molecular structures for compounds that act as enzyme inhibitors or bind to target receptors. The shape matching (sphere images) algorithm is employed here. Website: http://www.cmpharm.ucsf.edu/kuntz/dock.html
DOT	Daughter Of Turnip (DOT) is a program for docking macromolecules to other molecules of any size. It can predict binding modes of small molecule–protein complexes. The intermolecular energies for all configurations generated by this search are calculated as the sum of electrostatic and VDW energies. Website: http://www.sdsc.edu/CCMS/DOT/
FADE and PADRE	Fast Atomic Density Evaluator (FADE) and Pairwise Atomic Density Reverse Engineering (PADRE) programs are designed to aid in the molecular modeling of proteins. In particular, the programs can rapidly elucidate features of interest such as crevices, grooves, and protrusions. The topographical information produced by FADE and PADRE can help researchers easily pinpoint the most prominent features of a protein, regions that are likely to participate in interactions with other molecules. In addition, it provides shape descriptors to aid in analyzing single molecules.
FlexiDock	FlexiDock is a commercial software performs flexible docking of ligands into receptor binding sites. Website: http://www.tripos.com/software/fdock.html
FlexX	Incremental construction algorithm is employed in FlexX. The FlexX predicts the geometry of the protein–ligand complex and estimates the binding affinity. The two main applications of FlexX are complex prediction and VS. Complex prediction is used, when one have a protein and a small molecule binding to it but no structure of the protein–ligand complex is available. Website: http://www.biosolveit.de/flexx/

(*Continued*)

Table 10.4 (Continued)

Software	Algorithm and remarks
FRED	The shape matching (Gaussian functions) algorithm is employed in the Fast Rigid Exhaustive Docking (FRED) software.
FTDock	Fourier Transform Docking (FTDock) is a free program that performs rigid-body docking on two biomolecules in order to predict their correct binding geometry.
Glide	Glide is a fast and accurate docking program that addresses a number of problems, ranging from fast database screening to highly accurate docking. The descriptor matching/MC is the principal algorithm of Glide. The hierarchical filters in Glide ensure a fast and efficient reduction of large data sets to the few drug candidates that bind best with the target. Website: http://www.schrodinger.com/Glide/
GOLD	GOLD is a GA-based method for ligand protein docking. GOLD accounts for receptor flexibility through side-chain flexibility and, most important, ensemble docking. Website: http://www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx
GRAMM	Global Range Molecular Matching (GRAMM) is a free program for protein docking. To predict the structure of a complex, it requires only the atomic coordinates of the two molecules (no information of the binding sites is needed). The molecular pairs may be two proteins, a protein and a smaller compound, two transmembrane helices, etc. The program performs an exhaustive 6D search through the relative translations and rotations of the molecules. Website: http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/
Hammerhead	Hammerhead is suitable for screening large databases of flexible molecules by binding to a protein of known structure. The approach is completely automated, from the elucidation of protein binding sites, through the docking of molecules, to the final selection of compounds.
HINT	HINT is a software package that utilizes experimental solvent partitioning data as a basis for an empirical molecular interaction model. The program calculates empirical atom-based hydropathic parameters that, in a sense, encode all significant intermolecular and intramolecular noncovalent interactions implicated in drug binding or protein folding.
Liaison	Liaison is a commercial program for fast estimation of free energy of binding between a receptor and a ligand. The free energy of binding can be approximated by an equation in which only the free and bound states of the ligand are calculated. The method combines high-level molecular mechanics calculations with experimental data to build a scoring function for the evaluation of ligand—receptor binding free energies.
LigandFit	The shape matching (moments of inertia) algorithm is employed.

(*Continued*)

Table 10.4 (Continued)

Software	Algorithm and remarks
MOE	MOE is a fast and accurate docking program. The dock poses were ranked according to the GBVI/WSA binding free-energy calculation and minimized using MMFF94x within a rigid receptor.
Molegro Virtual Docker	Molegro Virtual Docker is an integrated platform for predicting protein—ligand interactions. Molegro Virtual Docker handles all aspects of the docking process, from preparation of the molecules to determination of the potential binding sites of the target protein, and prediction of the binding modes of the ligands.
QSite	QSite is a mixed-mode QM/MM program for highly accurate energy calculations of protein—ligand interactions in the active site. The program is specifically designed for proteins and allows a number of different QM/MM boundaries for residues in the active site. QSite uses the power and speed of Jaguar to perform the quantum mechanical part of the calculations and OPLS-AA to perform the molecular mechanical part of the calculations.
Situs	Situs is a program package for the docking of protein crystal structures to single-molecule, low-resolution maps from electron microscopy or small-angle X-ray scattering.
SLIDE	Descriptor matching algorithm is employed in SLIDE.
SuperStar	SuperStar is a program for generating maps of interaction sites in proteins using experimental information about intermolecular interactions. The generated interaction maps are therefore fully knowledge-based. SuperStar retrieves its data from IsoStar, CCDC interaction database. IsoStar contains information about nonbonded interactions from both the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB).

10.4 COMBINATION OF STRUCTURE- AND LIGAND-BASED DESIGN TOOLS

In recent years, there has been increasing attention paid to developing new methods employing the combined information generated from receptors and ligands. Most of the common present-day and potential future approaches are discussed in this section.

10.4.1 Comparative binding energy analysis

10.4.1.1 The concept of comparative binding energy

Comparative binding energy (COMBINE) analysis is a method of developing a system-specific expression to compute binding free energy using the 3D structures of receptor—ligand complexes [4]. This technique is based upon the hypothesis that the free energy of binding can be correlated with a subset of energy components calculated from the structures of receptors and ligands in bound and unbound forms [4,85]. Computation of binding free energies is very challenging due to the need to sample conformational space effectively in order to compute entropic contributions. Empirical scoring functions, which are fast to calculate, have been derived to approximate binding free energy using a single structure of a receptor–ligand complex [86]. If some experimental binding data are accessible for a set of related complexes, then this information can be used to derive a target-specific scoring function. This algorithm is taken in the COMBINE analysis in which the binding free energy (ΔG) or inhibition constant (K_i) or other related properties are correlated with a subset of weighted interaction energy components determined from the structures of energy minimized receptor–ligand complexes. The receptor binding free energy (ΔG) of a ligand can be expressed as

$$\Delta G = \sum_{i=1}^{n} \omega_i \, \Delta u_i^{\text{rep}} + C \tag{10.10}$$

The n terms Δu_i^{rep} of the ligand–receptor binding energy ΔU are selected, and the coefficients ω_i and constant C are determined by the statistical analysis. ΔU is calculated for representative conformations of the ligand–receptor complexes and the unbound ligands and the receptor using a molecular mechanics FF. The ligands are divided into n_{l} fragments, and the receptor into n_{r} regions (e.g., amino acid residues), and thus

$$\begin{aligned}\Delta U = {} & \sum_{i=1}^{n_{\text{l}}}\sum_{j=1}^{n_{\text{r}}} u_{ij}^{\text{VDW}} + \sum_{i=1}^{n_{\text{l}}}\sum_{j=1}^{n_{\text{r}}} u_{ij}^{\text{ELE}} + \\ & \sum_{i=1}^{n_{\text{l}}} \Delta u_i^{B,L} + \sum_{i=1}^{n_{\text{l}}} \Delta u_i^{A,L} + \sum_{i=1}^{n_{\text{l}}} \Delta u_i^{T,L} + \sum_{i<i'}^{n_{\text{l}}} \Delta u_{ii'}^{\text{NB},L} + \\ & \sum_{j=1}^{n_{\text{r}}} \Delta u_j^{B,R} + \sum_{j=1}^{n_{\text{r}}} \Delta u_j^{A,R} + \sum_{j=1}^{n_{\text{r}}} \Delta u_j^{T,R} + \sum_{j<j'}^{n_{\text{r}}} \Delta u_j^{\text{NB},R}\end{aligned} \tag{11.10}$$

The first two terms on the right side of the equation describe the intermolecular interaction energies between each fragment i of the ligand and each region j of the receptor. The next four terms describe changes in the bonded (bond, angle, and torsion) and the nonbonded (a combination of Lennard–Jones and electrostatic) energies of the ligand fragments upon binding to the receptor, and the last four terms account for changes in the bonded and nonbonded energies of the receptor regions upon binding of the ligand.

10.4.1.2 The methodology of COMBINE

To derive the COMBINE model, fundamentally three steps are to be followed: namely, modeling the molecules and their complexes, measuring the interaction energies between ligands and the receptor, and finally, performing chemometric analysis to derive the regression equation [4]. The methodology for the COMBINE analysis is outlined schematically in Figure 10.10.

a. *Molecular modeling*: To develop the COMBINE models, the ligands should be divided into fragments, and then the same number of fragments must be allocated to all the compounds, adding dummy fragments to the ligands lacking the correct number. The 3D models of the ligand–receptor complexes and the unbound receptor and ligands can be derived with a standard molecular mechanics program. Different regression equations can be produced by using the following factors:
- Different starting conformations of the receptor
- The inclusion of positional restraints on parts of the receptor
- Different convergence criteria during energy minimization
- Different ways of treating the solute–solvent interface
- The dielectric environment

b. *Measurement of the interaction energies*: The objective of this step is the computation of the nonbonded (VDW and electrostatic) interaction energies between each residue of the receptor and every fragment of the ligand, using a molecular mechanics FF. Along with the interaction energies, the energies between all pairs of residues/fragments for the complexes and for the free ligands and receptor on the basis of the distance-based dielectric constant should be computed as well. Finally, a matrix will be formed, with columns representing the energy components and rows

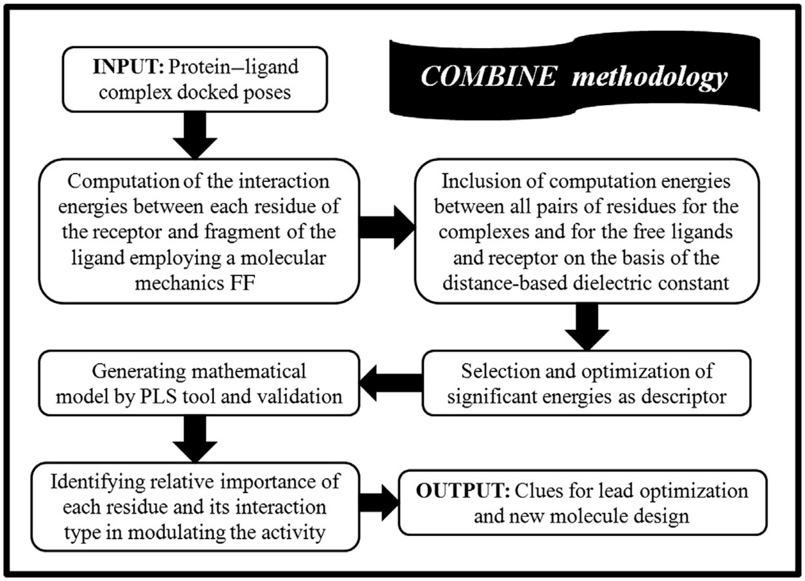

INPUT: Protein–ligand complex docked poses
COMBINE methodology
Computation of the interaction energies between each residue of the receptor and fragment of the ligand employing a molecular mechanics FF
Inclusion of computation energies between all pairs of residues for the complexes and for the free ligands and receptor on the basis of the distance-based dielectric constant
Selection and optimization of significant energies as descriptor
Generating mathematical model by PLS tool and validation
Identifying relative importance of each residue and its interaction type in modulating the activity
OUTPUT: Clues for lead optimization and new molecule design

Figure 10.10 The methodology of the COMBINE analysis.

representing each compound in the set. A final column containing experimental activities or inhibitory activities or binding affinities is then added to the matrix as the dependent variable for model development.

c. *Chemometric analysis*: After the completion of steps 1 and 2, significant descriptors should be retained and others must be eliminated from the study matrix. Due to the large number of variables and their intercorrelation nature, partial least squares (PLS) is the technique of choice for deriving the QSAR model that can quantify the most important energy interactions in terms of activity prediction [87,88].

10.4.1.3 Importance and advantages of COMBINE

Comparing the COMBINE method with the calculation of binding energies via classical molecular mechanics, the advantages of ligand–receptor interaction energies to statistical analysis are as follows [4]:

a. The noise due to inaccuracy in the potential energy functions and molecular models can be reduced.

b. Mechanistically important interaction terms can be identified.

c. Compared to more traditional QSAR analysis, this approach can be anticipated to be more predictive, as it incorporates more physically relevant information about the energies of the ligand–receptor interaction.

d. It helps in the screening of lead compounds based on the required properties that interact favorably with the key residues.

10.4.1.4 Drawbacks and required improvements

COMBINE experiences the intrinsic errors implicated in the computation of the interaction energies between ligand and macromolecular complexes like all other interaction energy-based 3D-QSAR methods. The predictability of the method can be improved by making advances in several aspects, like the description of the electrostatic term, the addition of appropriate descriptors for solvation and entropic effects, and the optimization of the methodology, such as the choice of ligand fragment definitions and the details of the variable selection protocol [4].

10.4.1.5 Applications of COMBINE

COMBINE analysis was originally developed to study the interactions of one target protein with a set of related ligands. It has been established in recent times that the approach can be applied tactfully to a wide range of complexes, including enzyme–substrate and inhibitor complexes [89], protein–protein/peptide complexes [90], and protein–DNA complexes [91]. It has also been employed to examine binding to more than one target protein receptor.

10.4.1.6 Software for COMBINE

SCOPE: To make the COMBINE method more user friendly, the method has been implemented in the structure-based compound optimization, prioritization, and evolution (SCOPE) module of VLifeMDS, which uses this approach to derive a 3D-QSAR between the experimental biological activities and the calculated ligand interaction energy terms [92]. First, to execute COMBINE analysis, each of the ligands against a particular target has to be docked into its target. It requires a training set of docked and optimized ligand—receptor complexes, and the unbound ligands and receptor for which intermolecular and intramolecular interaction energies are calculated. The calculated descriptors are then correlated with the experimental activity of the studied compounds to develop a QSAR model. Finally, the interpretation of the developed mathematical equation can enlighten the important ligand—receptor interactions for future drug designing and development process.

gCOMBINE: gCOMBINE is an user-friendly tool for performing COMBINE analysis in drug design research programs. It is a graphical user interface (GUI) written in Java with the purpose of performing COMBINE analysis on a set of ligand—receptor complexes with the intention of deriving highly informative QSAR models [93]. The objective of this method is to generate the ligand—receptor interaction energies into a series of variables, explore the origins of the variance within the set employing principal component analysis (PCA), and then allocate weights to the chosen ligand—residue interactions by using PLS analysis to correlate with the experimental activities or binding affinities. The major advantages of using a GUI are that it allows plenty of interactivity and provides multiple plots representing the energy descriptors entering the analysis, scores, loadings, experimental versus predicted regression lines, and the evolution of classical validation parameters. Using the GUI, one can carry out numerous added tasks, such as possible truncation of positive interaction energy values and generation of ready-made PDB files containing information related to the importance of the activity of individual protein residues. This information can be aptly displayed and color-coded using a molecular graphics program like PyMOL.

10.4.2 Comparative residue interaction analysis

10.4.2.1 Concept of CoRIA

The CoRIA analysis is a relatively recent innovation in the field of QSAR studies. It is a 3D-QSAR approach, which uses the descriptors that describe the thermodynamic events involved in ligand binding to the receptor to explore both the qualitative and quantitative facets of the ligand—receptor recognition process. The main emphasis of CoRIA is to calculate and analyze the receptor—ligand complex and thereafter predict the binding affinity of the complex [5]. The binding free-energy difference (ΔG_{bind})

between the free and bound states of the receptor and ligand ($\Delta G_{\text{complex}} - \Delta G_{\text{uncomplexed}}$) is related to the binding constant (K_d) of the ligand to the receptor and can be expressed as an additive interaction of different events using the classical binding free energy equation [94]:

$$\Delta G_{\text{bind}} = \Delta G_{\text{solv}} + \Delta G_{\text{conf}} + \Delta G_{\text{inter}} + \Delta G_{\text{motion}} \tag{10.12}$$

That is, the total free energy of binding (ΔG_{bind}) is an additive interaction of solvation of ligand (ΔG_{solv}), which is the difference between the unbound (e.g., cellular) and bound states, conformational changes that occur in the receptor and ligand (ΔG_{conf}), specific interactions between the ligand and receptor as a consequence of their proximity (ΔG_{inter}), and the motion in the receptor and ligand once they are close to each other (ΔG_{motion}).

10.4.2.2 Methodology of CoRIA

The first step of CoRIA is the calculation of the binding energies in the form of nonbonded interaction energies (like VDW and Coulombic), which describe thermodynamic events involved in ligand binding to the active site of the receptor. Thereafter, employing a genetic version of the PLS technique (namely, G/PLS), these calculated energies should be correlated with the biological activities of molecules, along with the other physiochemical variables like molar refractivity, surface area, molecular volume, Jurs descriptors, and strain energy [5,95,96]. Further, validation has to be performed for the developed CoRIA models based on various validation metrics to ensure the acceptability of the developed models. The methodology of the CoRIA is schematically presented in Figure 10.11.

10.4.2.3 Variants of CoRIA

In recent years, to deal with the problems of peptide QSAR, CoRIA methodology has gone through several advanced modifications. Two newly developed variants of CoRIA are [5,95]:

a. *reverse*-CoRIA (*r*CoRIA): When the peptide (ligand) is fragmented into individual amino acids, and the interaction energies (VDW, Coulombic, and hydrophobic interactions) of each amino acid in the peptide with *the total receptor* is calculated, the technique is known as *rCoRIA*.

b. *mixed*-CoRIA (*m*CoRIA): When the interaction energies of each amino acid in the peptide with the *individual active site residues in the receptor* is calculated, the approach is defined as *mCoRIA*.

For both approaches, along with the interaction energies, other thermodynamic descriptors (like free energy of solvation, entropy loss on binding, strain energy, and solvent assessable surface area) are also included as independent variables, which are correlated to the biological activity using a G/PLS technique like general CoRIA.

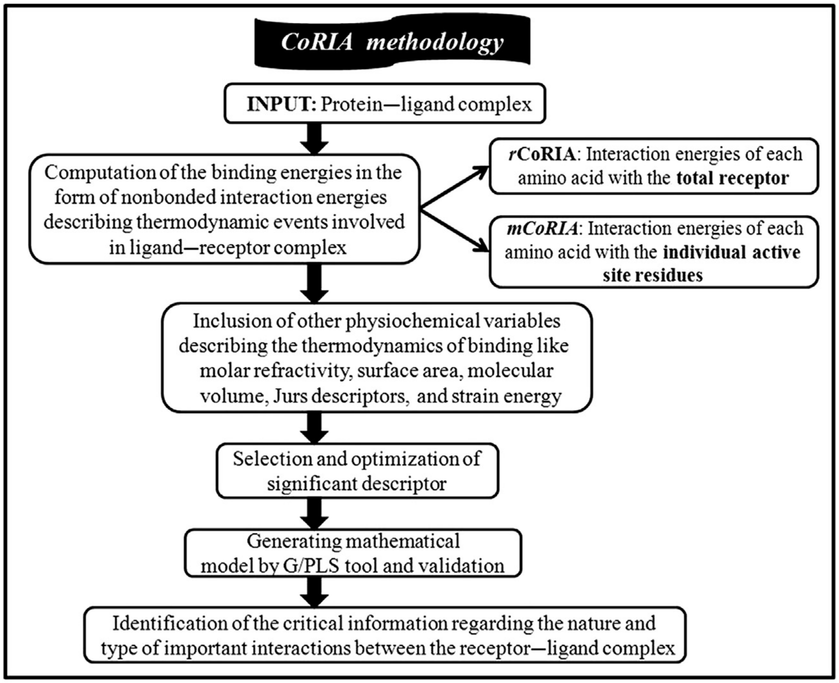

CoRIA methodology
INPUT: Protein—ligand complex
Computation of the binding energies in the form of nonbonded interaction energies describing thermodynamic events involved in ligand—receptor complex
rCoRIA: Interaction energies of each amino acid with the total receptor
mCoRIA: Interaction energies of each amino acid with the individual active site residues
Inclusion of other physiochemical variables describing the thermodynamics of binding like molar refractivity, surface area, molecular volume, Jurs descriptors, and strain energy
Selection and optimization of significant descriptor
Generating mathematical model by G/PLS tool and validation
Identification of the critical information regarding the nature and type of important interactions between the receptor—ligand complex

Figure 10.11 The methodology of the CoRIA analysis.

10.4.2.4 Importance and application of CoRIA

The most significant importance of the CoRIA methodology is that it is capable of extracting critical information regarding the nature and type of important interactions at the level of both the receptor and the ligand. The generated rich source of information can be directly employed in the design of new molecules and drug targets [87]. The approaches have the ability to forecast modifications in both the ligand and the receptor, provided that structures of some ligand—receptor complexes are available. The CoRIA approach can be used to identify crucial interactions of the inhibitors with the enzyme at the residue level, which can be profitably exploited in optimizing the inhibitory activity of ligands. Furthermore, it can be used to guide point mutation studies—yet another advantage.

10.4.2.5 Drawback of CoRIA

The major drawback of CoRIA is that it cannot be applied with small organic molecules. This is because unlike peptides, there is no rational or unanimously established protocol for fragmenting small molecules [87].

10.4.2.6 Future perspective of CoRIA

The algorithm of this methodology can be further improved in the near future by considering the following points:

- Solvation of the entire ligand—protein complexes
- Extensive conformational sampling by molecular dynamics
- Inclusion of other important interactions like hydrogen bonding

10.5 *IN SILICO* SCREENING OF CHEMICAL LIBRARIES: VS

10.5.1 Concept

VS is a technique to identify novel hits (i.e., bioactive molecules) from large chemical libraries through computational means by applying knowledge about the protein target (structure-based VS) or known bioactive ligands (ligand-based VS) [97]. The ligand-based approaches utilize structure—activity data from a set of known actives in order to identify drug candidates for experimental evaluation. The ligand-based methods include approaches like similarity and substructure searching, QSAR, pharmacophore-based search, and 3D shape matching [98,99]. Apparently, structure-based VS mainly employ the docking approach, where the 3D structure of the biological target protein or receptor is used to dock the candidate molecules and rank them based on their predicted binding affinity (docking score). These techniques, like ePharmacophore and protein—ligand fingerprints, also can be used under structure-based VS. It is important to mention that based on the requirements of the researchers, one can use ligand- and structure-based approaches one by one, as a layered screening technique, or both approaches concurrently.

The VS technology has emerged as a response to the pressure from the combinatorial/HTS community. VS can be considered as the mining of chemical spaces with the aim to identify molecules that possess a desired property [100]. The VS approach is highly dependent on the quantity and quality of available data and the predictability of the underlying algorithm. As a consequence, there is no universal guideline or workflow for the VS approaches, and the researcher has to apply his computational knowledge and experience to find the active drug candidate from the sea of drug databases and chemical libraries applying the best possible source of tools as per his requirements.

10.5.2 Workflow and types of VS

The experimental efforts to carry out the biological screening of billions of compounds are still considerably high, and therefore, CADD approaches have become attractive alternatives. One has to remember that the workflows employed in the VS are not universal. The workflow is solely dependent on the researchers' needs, the diversity of chemical library, and available sources for sensible and practical VS. Here, we have tried to describe a general, commonly employed workflow.

10.5.2.1 Selection of chemical libraries/databases

The first criterion of any VS approach is the selection of the required chemical library. Taking the requirements into consideration, the researcher has to select the chemical library from the available large pool of public and commercial databases. The database may cover a particular class of compounds (structural or pharmacological) or diverse classes of molecules. A significant amount of information regarding various types of chemical libraries has been provided in Section 10.5.6.

10.5.2.2 Preprocessing of chemical libraries

After selection of the required database, one has to perform the preprocessing of the chemical library by removing the duplicate structures, tautomers, counter ions, and protonated ones.

10.5.2.3 Filtering of druglike molecules

In the next step, to filter the druglike molecules from the preprocessed chemical library, different druglike filters need to be employed:

a. *Lipinski's rule of five*: It is a well-known rule of thumb of encoding a simple profile for the permeability of orally available drugs. The filter demonstrates that poor absorption or permeation are more likely to occur when (i) molecular weight (MW) is over 500, (ii) calculated octanol/water partition coefficient (logP) is over 5, (iii) presence of more than 5 HBDs, and (iv) presence of more than 10 HBAs [101]. With the exception of logP, all other criteria are additive and can be accurately computed for screening of virtual libraries. However, Lipinski's rule of five fails to distinguish between drugs and nondrugs, rather serves as a method to predict compounds with poor absorption or permeability. One has to remember that antibiotics fall outside the scope of this rule.

b. *ADMET filter*: In addition to the Lipinski's filter, ADMET filters [102] can be employed for filtering. To get the early information regarding absorption, distribution, metabolism, excretion (ADME), and toxicity data (ADMET data), the ADMET filter screen is very useful. The late stage failure of the molecules in the clinical trials is primarily attributed to their inability to meet the necessary pharmacokinetic profile. Accurate prediction of ADMET properties enables to eliminate unwanted molecules and aids the lead optimization process.

10.5.2.4 Screening

The ultimate screening step of the VS of the filtered druglike compounds is based on two fundamental approaches; namely, a ligand-based approach and a receptor-based approach [103].

a. *Ligand-based approach*: In this approach, molecules with physical and chemical properties similar to those of the known ligands are identified using QSAR models, pharmacophore-based search, substructure search, and 2D and 3D atomic property-based search approaches. The ligand-based approach is possible without protein information and can be employed for scaffold hopping. Again, since this technique is biased by the properties of known ligands, it limits the diversity of the hits generated.

b. *Receptor-based approach*: The approach uses techniques like protein—ligand docking, different scoring functions, and active-site-directed SBPs for the molecular recognition between a ligand and a target protein to select chemical entities that bind to the active sites of biologically relevant targets with known 3D structures. The major advantages of this approach are the following: It is possible to carry out this process without ligand information, the entire capability of the protein pocket is taken into account, prediction of binding modes is possible, scaffold hopping and profiling without any bias toward existing ligands can be done.

c. *Combination of ligand- and receptor-based approach*: As there is no universal method for the VS, one can use ligand- and structure-based methods separately (e.g., pharmacophore and docking one by one as a two-layer technique), or can employ the combined ligand- and structure-based methods like COMBINE and CoRIA. The COMBINE and CoRIA approaches are reliable. as they consider the ligand and receptor information as well as information regarding their binding complexes.

d. *Machine learning techniques*: Apart from the ligand- and receptor-based approaches, machine learning techniques like support vector machine (SVM) and binary kernel discrimination (BKD) can be tactfully applied in a few cases of VS. The SVM predicts the bioactivity by representing the lead in *n*-dimensional real space using molecular descriptors and fingerprint technology, where *n* represents the number of features or attributes. The SVM approach is based on the fuzzy logic fingerprint. The BKD is a recently developed computational approach. In BKD, the molecule is represented as 2D fragment bit-string. It consists of three components. First, the structural representation section; second, the similarity searching section using different coefficients; and third, the section with different weighting schemes for lead compounds.

10.5.2.5 Hit selection to new chemical entity generation

Once hits are selected from the final screening process, one has to synthesize or purchase the hits for further study. The selected hits have to go through different in vitro/in vivo bioassays for final confirmation of their pharmacological actions. Compounds showing encouraging pharmacological activity are considered as the leading ones for further preclinical and clinical studies to establish them as the final drug candidates. A schematic illustration of various steps of the VS is presented in Figure 10.12.

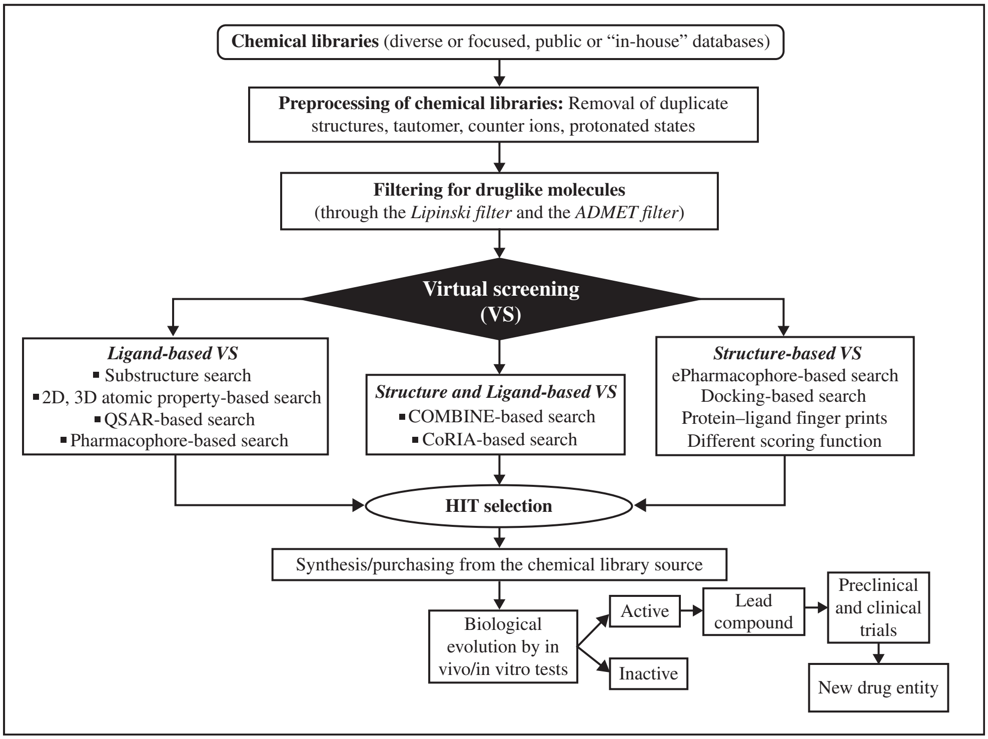

Chemical libraries (diverse or focused, public or "in-house" databases)
Preprocessing of chemical libraries: Removal of duplicate structures, tautomer, counter ions, protonated states
Filtering for druglike molecules (through the Lipinski filter and the ADMET filter)
Virtual screening (VS)
Ligand-based VS
▪ Substructure search
▪ 2D, 3D atomic property-based search
▪ QSAR-based search
▪ Pharmacophore-based search
Structure and Ligand-based VS
▪ COMBINE-based search
▪ CoRIA-based search
Structure-based VS
ePharmacophore-based search
Docking-based search
Protein–ligand finger prints
Different scoring function
HIT selection
Synthesis/purchasing from the chemical library source
Biological evolution by in vivo/in vitro tests
Active
Inactive
Lead compound
Preclinical and clinical trials
New drug entity

Figure 10.12 Fundamental steps of the VS approach.

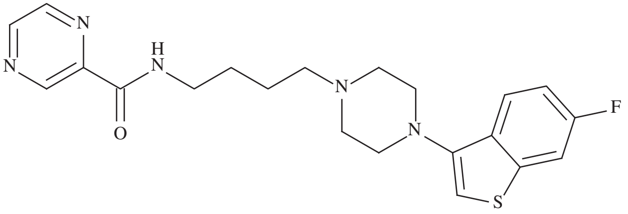
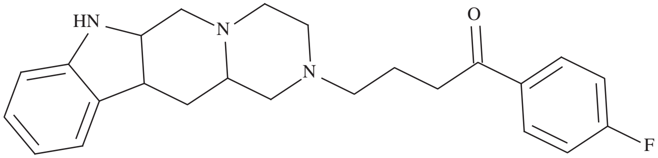
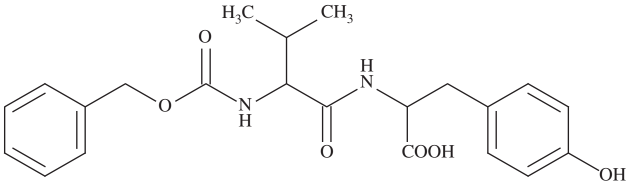
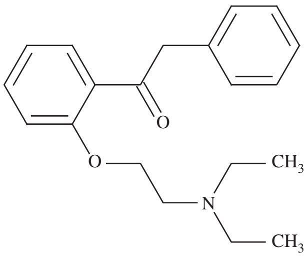
10.5.3 Successful application of VS: A few case studies

Employing VS approaches, many drugs have been obtained that are already on the market, and a few others are in the different stages of clinical trials. Liebeschuetz et al. [104] used library design- and structure-based VS to develop inhibitors of factor Xa serine protease, an important target in the blood coagulation cascade. Sharma et al. [105] carried out VS to find the neuraminidase inhibitors (potential targets for swine flu), and two of the metabolites (Hesperidin and Narirutin) were predicted to be more potent than the existing drugs (Oseltamivir). Dahlgren et al. [106] developed salicylidene acylhydrazides as inhibitors of type III secretion (T3S) in the gram-negative pathogen *Yersinia pseudotuberculosis* from a set of 4416 virtual compounds employing three QSAR models. As the studies are so numerous, we have made a representative list of successful applications of VS-based [107] drug discovery in Table 10.5.

10.5.4 Advantages of VS

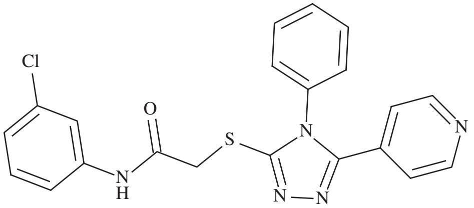
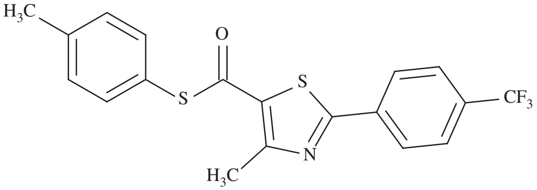
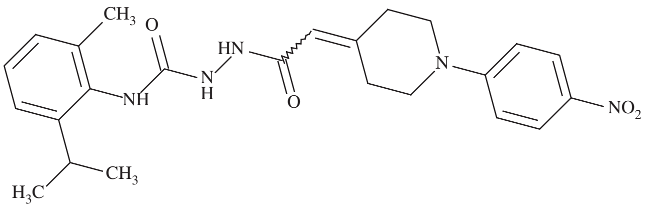
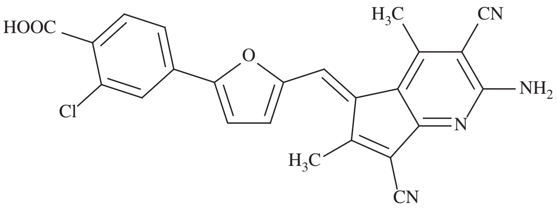
Application of the VS techniques increases the chance of successful drug discovery by many times. Without any hesitation, we can say that the VS has emerged as a reliable,

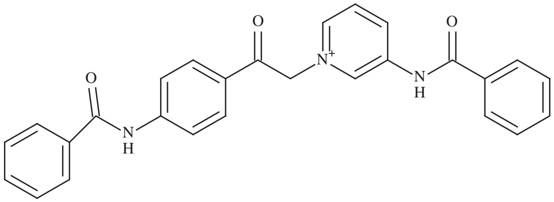
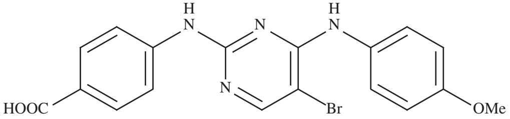
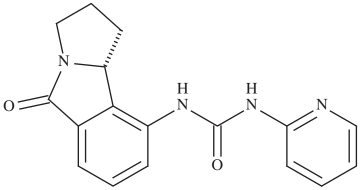
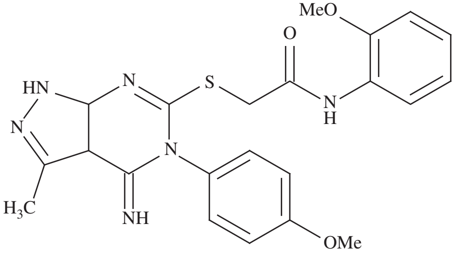
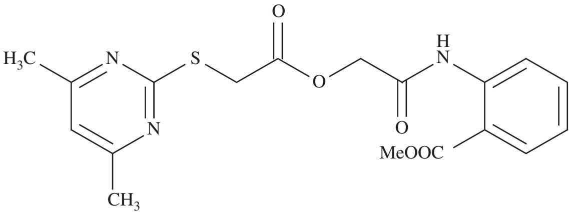
Table 10.5 Representative case studies of successful application of VS

Target			Database	Methods employed	Structure of the most active hit
Receptor	G protein −coupled	*α_{1A} adrenergic*	Aventis in-house compound and MDDR	Pharmacophore and docking	 α_{1A} adrenergic receptor antagonist
		Dopamine D3	NCI	Pharmacophore and docking	 Dopamine D3 receptor antagonist
		Endothelin A	Maybridge database	Pharmacophore	 Endothelin A (ET_A) receptor antagonist
		Muscarinic M3	Astra Charnwood in-house compound repository	Pharmacophore	 Muscarinic M3 receptor antagonist

(*Continued*)

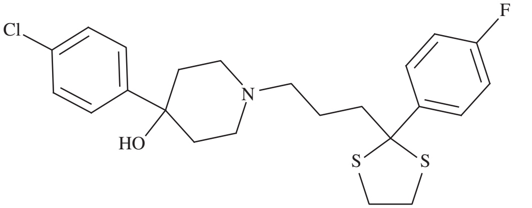
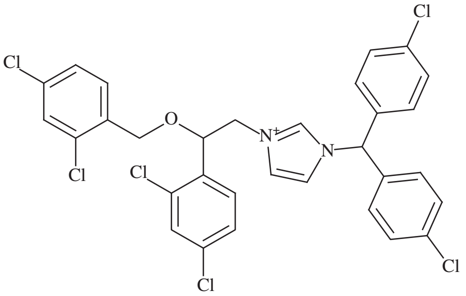
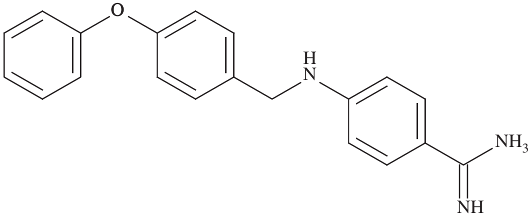
Table 10.5 (Continued)

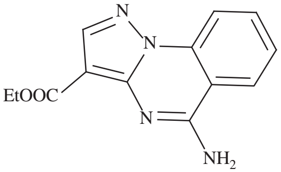
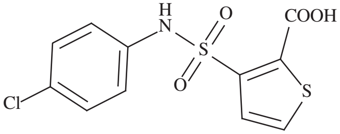
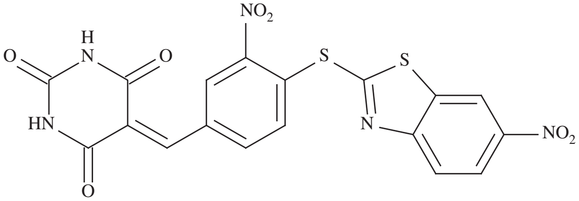
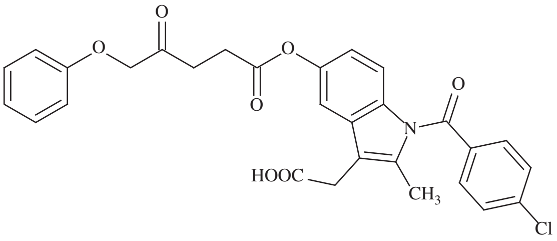
Target			Database	Methods employed	Structure of the most active hit
		Neurokinin-1 (NK$_1$)	826,952 compounds Merging various databases	Pharmacophore and docking	 Neurokinin NK$_1$ antagonist
	Nuclear receptors	*Retinoic acid receptor*	ACD	Docked into the retinoic acid receptor (RAR) binding site	 Retinoic acid receptor α antagonist
		Thyroid hormone receptor	ACD	Docking	 Thyroid hormone receptor antagonist
Enzymes	Kinase	*Akt 1 (protein kinase Bα, PKBα)*	ChemBridge	Flexible docking and employing different scoring functions	 Akt 1 inhibitor

		Bcr-Abl tyrosine kinase	ChemDiv	Lipinski filter and docking	 Bcr-Abl Tyrosine kinase inhibitor
		Checkpoint kinase 1 (Chk-1)	AstraZeneca in-house compound	Pharmacophore and flexible docking	 Chk-1 inhibitor
		Cyclin-dependent kinase 4 (Cdk4)	ACD	*De novo* design program LEGEND was combined with the program SEEDS to extract relevant scaffolds	 Cdk4 inhibitor
		p56 Lymphoid T cell tyrosine kinase (Lck)	3D database of 2 million commercial compounds	Docking	 Lck inhibitor
	Proteases	*Falcipain-2*	ChemBridge	Lipinski and ADMET filters, homology modeling along with docking	 Falcipain inhibitor

(*Continued*)

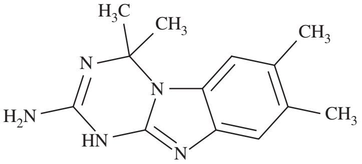
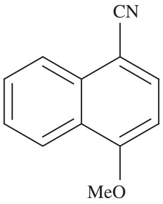
Table 10.5 (Continued)

Target			Database	Methods employed	Structure of the most active hit
		HIV protease	Cambridge	Docking	 HIV protease inhibitor
		SARS CoV 3C-Like proteinase	ACD, MDDR, and NCI	Homology modeling, docking and molecular dynamics,	 SARS CoV 3C-like proteinase inhibitor
		Thrombin	5300 commercial compounds	Docking and *de novo* design	 Thrombin inhibitor

	Hydrolases	*Adenylyl cyclase (edema factor and CyaA)*	ACD	Docking	 Edema factor (EF) adenylyl cyclase inhibitor
		AmpC β-lactamase	ACD	Docking	 mpC β-lactamase noncovalent inhibitor
		Protein tyrosine phosphatase 1B	Pharmacia, the in-house compound	Docking	 Protein tyrosine phosphatase 1B (PTP1B) inhibitor
	Oxidases/ reductases	*Aldose reductase*	ADAM and EVE docking program	Docking	 Aldose reductase inhibitor

(*Continued*)

Table 10.5 (Continued)

Target			Database	Methods employed	Structure of the most active hit
		Dihydrofolate reductase	ACD	Docking	 *Plasmodium falciparum* DHFR inhibitor
		Inosine 5′-monophosphate dehydrogenase Inhibitors (IMPDH)	In-house reagent inventory system	Docking and different scoring functions	 Inosine 5′-monophosphate dehydrogenase (IMPDH) inhibitor

cost-effective, and time-saving technique for the discovery of lead compounds [108]. The main advantages of this method compared to laboratory experiments are described in the next sections.

10.5.4.1 Cost-effective

As no compounds have to be purchased externally or synthesized by a chemist at the initial stages, VS is one of the most cost effective of the drug discovery processes.

10.5.4.2 Time-saving

Synthesis can take an extremely long time, especially in the case of large databases with millions of chemical compounds. But employing computational tools, the VS approach is always efficient in drug discovery.

10.5.4.3 Labor-efficient

Synthesis and bioassays always involve a great amount of human strength, and the chance of getting false positives is always present, even after spending a lot of physical and mental labor. Although undeniably VS also has a chance of resulting in false positives, but it is always labor-efficient in drug development.

10.5.4.4 Sensible alternative

It is possible to investigate compounds that have not been synthesized yet; and conducting HTS experiments is costly, time-consuming, and laborious for large numbers of chemicals. As a result, VS is always a rational option to minimize the initial number of compounds before using HTS methods.

10.5.5 Pitfalls

While applying the VS technique, the researcher must face many difficulties, such as finding the best possible balance between efficiency and precision when evaluating a particular algorithm, determining which method achieves better results and in what situations, and defining whether there is any universal method or workflow for VS. Considering the altitude of settings, parameters, and data sets, researchers have to explore a large number of *ifs* and *buts* during execution of VS. There are many known limitations (as well as still-unreported ones) of VS techniques. The probable pitfalls are discussed in the following section, along with possible ways to resolve them [109]. The pitfalls can be classified into four categories: (a) erroneous assumptions and expectations, (b) data design and content, (c) choice of software, and (d) conformational sampling, as well as ligand and target flexibility. A schematic representation has been shown in Figure 10.13.

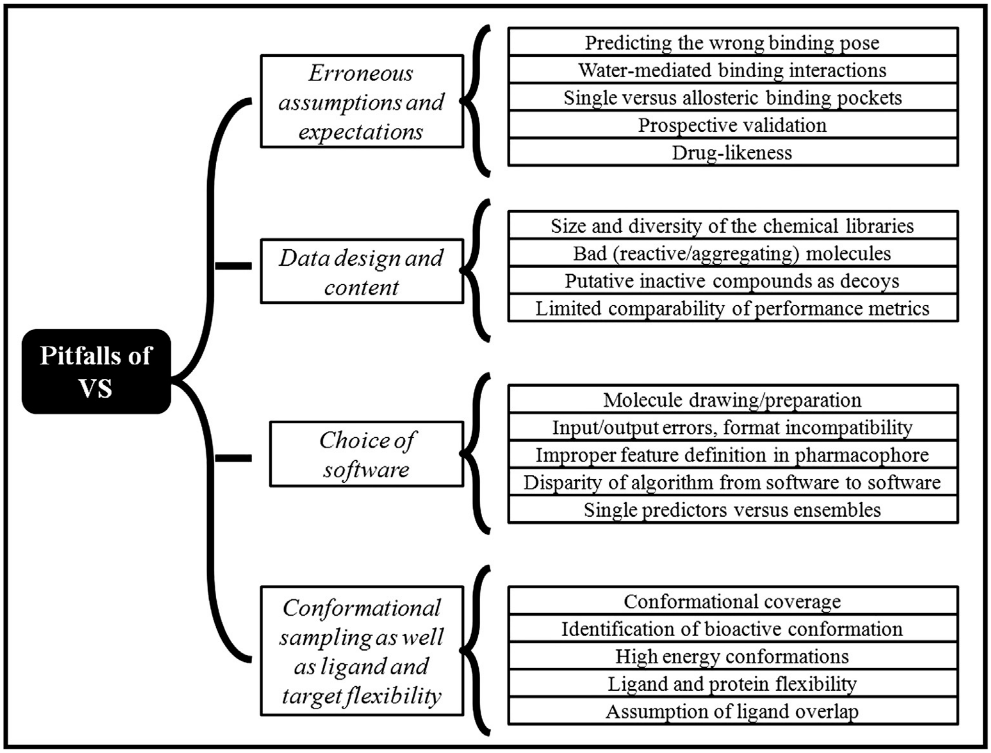

Pitfalls of VS
Erroneous assumptions and expectations
Predicting the wrong binding pose
Water-mediated binding interactions
Single versus allosteric binding pockets
Prospective validation
Drug-likeness
Data design and content
Size and diversity of the chemical libraries
Bad (reactive/aggregating) molecules
Putative inactive compounds as decoys
Limited comparability of performance metrics
Choice of software
Molecule drawing/preparation
Input/output errors, format incompatibility
Improper feature definition in pharmacophore
Disparity of algorithm from software to software
Single predictors versus ensembles
Conformational sampling as well as ligand and target flexibility
Conformational coverage
Identification of bioactive conformation
High energy conformations
Ligand and protein flexibility
Assumption of ligand overlap

Figure 10.13 Major pitfalls of the VS approach.

a. *Erroneous assumptions and expectations*:

- *Predicting the wrong binding pose*: There are a few cases where even though the predicted docking binding poses are wrong, docking screening can accidentally generate high scores to many hits.
- *Water-mediated binding interactions*: In many docking studies, hydrogen bonds between ligand and protein are formed by water, which is often visible in the crystal structure of the complex. Those water-mediated hydrogen bonds can be taken into account in a structure-based VS study, but it is very difficult to predict the exact number, position, and orientation of these interactions.
- *Single- versus allosteric-binding pockets*: Both structure- and ligand-based VS approaches have intrinsic limitations, in that they are incapable of identifying bioactive ligands for the binding pockets, which are not explicitly docked against or implicitly represented in the training set. Again, the unknown binding site of a ligand complicates the problem of properly assessing hit rates in the VS experiments.
- *Prospective validation*: The VS is habitually performed on data sets with known actives, but often only putatively inactive molecules. As a result, in many cases,

a good number of inactive molecules are absent to identify the true inactives after the VS approach.

- *Druglikeness*: The majority of the VS approaches are based on the screening of druglike compounds by employing Lipinski's rule of five as the preliminary screening steps of the VS. One should remember that the rule applies only to oral bioavailability, and that many bioactivity classes such as antibiotics fall outside the scope of this rule. Hence, VS protocols are generally applied and validated on a relatively small fraction of chemical space, and their performance may change drastically from one database to another.

b. *Data design and content*:

- *Size and diversity of the chemical libraries*: In many cases, the employed libraries in the VS are either too small or contain too many closely related analogs, or often both. A data set that lacks sufficient chemical diversity is never an ideal choice for VS.
- *Experimental errors and inappropriate bioassays*: A large pool of data sets is often assembled from different sources, where different bioassay procedures and detection techniques have been used. As a consequence, there is a huge risk of experimental errors and inappropriate assays from molecule to molecule.
- *Bad (reactive/aggregating) molecules*: The data set provided for VS often includes molecules that contain chemically reactive groups or other undesirable functionalities that may interfere with the HTS detection techniques.
- *Putative inactive compounds as decoys*: Experimentally confirmed inactive compounds are helpful as negative controls because only few of them should appear in the hit list when a reliable VS protocol is employed. However, many of the decoys used in VS benchmark studies are only putatively inactive; hence, some assumed true negatives may actually be positives.

c. *Choice of software*:

- *Molecule drawing/preparation*: Adding implicit hydrogen atoms, handling the ionization states of the molecule, and assigning the correct charges at the initial stages of VS screening can easily be forgotten in many cases, where the final result will totally mislead the scientific community.
- *Input/output errors, format incompatibility*: Various errors are introduced when interconverting different molecular formats from one software to another. There is a high possibility of getting distorted information, like changes in atomic coordinates, chirality, hybridization, and protonation states of the employed compounds.
- *Improper feature definitions in the pharmacophore*: Incorrect feature definitions can be detrimental to the outcome of VS. In pharmacophore queries, the definition of pharmacophore features needs to be applied with caution. For example, it is known that nitrogen and oxygen atoms in the same heterocycle (such as

an oxazole) do not both behave as HBAs simultaneously. In the majority of cases, the acceptor in the oxazole ring is the nitrogen.

- *Disparity of algorithm from software to software*: Various software tools apply different algorithms for a particular job. For example, in the case of docking for energy minimization, different forms of FFs are applied from one software to another. Therefore, there is a high probability of getting different hits and making the VS process highly dependent on the use of particular software.
- *Single predictors versus ensembles*: Multiple statistical tools (both free and commercial software and descriptors) are available to perform VS studies. Each module captures different characteristics of molecular similarity. As a consequence, it is always difficult to identify a preferred tool/software/descriptor; therefore, it is often necessary to account for several approaches rather than one.

d. *Conformational sampling as well as ligand and target flexibility*:

- *Conformational coverage*: One of the main challenges in 3D-VS is generating a convenient set of conformations to cover the molecule's conformational space effectively.
- *High-energy conformations*: One has to remember that good conformational coverage is very important, and on the contrary, high-energy or physically unrealistic conformations can be detrimental to VS. Few conformational sampling approaches do not utilize energy minimization to refine and properly rank the resulting geometries. Therefore, the resulting list could contain many false positives.
- *Ligand and protein flexibility*: A common practice in many 3D database search systems is to set a limit on the number of conformations stored for each molecule. The number of conformations accessible to a molecule depends greatly on its size and flexibility. Of course, it is not only the ligands that are flexible; it is the biological targets as well. Protein flexibility is probably the most unexploited aspect of VS.
- *Assumption of ligand overlap*: In 3D shape-based VS, most programs attempt to maximize the overlap between the query and the database molecules. Indeed, different ligands may occupy different regions in the same protein, even in the same binding site, and the overlap between them in 3D space can be much less than assumed by a shape-based VS tool, resulting in more false negatives.

10.5.6 Databases for the VS

Chemically diverse libraries are particularly attractive for identifying novel scaffolds for new or relatively unexplored targets, such as those resulting from diversity-oriented synthesis. One needs to remember that the database library must fit the purpose of the experiment before its selection for screening. A large number of databases are publicly available and the number is increasing day by day. Recent initiatives requiring greater use of *in silico* technologies have called for transparency and development of strong database information that is available to the public at no cost. Electronic information on chemical structure, pharmacological activity, and specificity against known molecular targets can serve a wide variety of purposes in the field of VS. Table 10.6 summarizes a list of most current public and commercial chemical databases that are commonly screened in practice. The scientific community should take initiatives to develop more databases for public and administration use in the near future.

10.6 OVERVIEW AND CONCLUSIONS

The LBP approach and structure-based molecular docking play promising roles in the identification and optimization of leads in modern drug discovery. Pharmacophore and docking-based approaches, employed both alone or concurrently in VS, lead to a much higher hit rate than traditional screening methods (e.g., HTS). In a complete, structurally diverse data set, pharmacophore gives immense confidence about the best probable features that are solely responsible for particular pharmacological activity. On the contrary, the docking method provides an opportunity for the designing of active compounds considering the binding aspects of the ligand with the amino acid residues in the respective receptors. Methods like CoRIA and COMBINE provide a blend of ligand- and structure-based drug design at once, where the best possible interactions of the ligand—receptor complex can be identified. These methods are capable of extracting critical information regarding the nature and type of important interactions at the level of both the receptor and the ligand.

VS approaches have been vigorously implemented by pharmaceutical industries with the intent to obtain as many potential compounds as possible, and with the hope of a greater chance of finding hits from the available large pool of chemical libraries. Many successful examples have been demonstrated recently in the field of computer-aided VS for lead identification. There appears to be no universal method to execute these studies, as each biological target system is unique. Although one cannot ignore the intrinsic restrictions of VS, it remains one of the best possible options to explore a large chemical space, in terms of cost effectiveness and commitment of time and material needed. With the development of new docking methodologies, ligand-based screening techniques and machine-learning tools, the VS techniques are capable of giving better hit prediction rates, and undoubtedly, these will play the front-runner role in drug design in the near future either as a complementary approach to HTS or as a stand-alone approach. One has to remember that technologies are available that need to be employed in the right way and in the right direction to identify novel chemical substances with the scientific use of VS techniques. However, it must be emphasized that VS is not intended to replace the actual experimental approaches. As a matter of fact, the VS and experimental methods are highly complementary to each other.

Table 10.6 Commonly used chemical databases for the VS approach

Compound database	Availability	No. of compounds[a]	Website
ACD	Commercial	3,870,000	http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html
Asinex	Commercial	550,000	http://www.asinex.com
Binding DB	Public	284,206 small ligands with 648 915 binding data, for 5662 protein targets	http://www.bindingdb.org
Chem ID	Public	3,88,000	http://chem.sis.nlm.nih.gov/chemidplus/
ChemBank	Public	800,000	http://chembank.broadinstitute.org
ChEMBL db	Public	658,075 differing bioactive compounds and 8091 targets	https://www.ebi.ac.uk/chembldb/
ChemBridge	Commercial	700,000	http://www.chembridge.com
ChemDiv	Commercial	1.5 million	http://www.chemdiv.com
Chemical Entities of Biological Interest (ChEBI)	Public	584,456	http://www.ebi.ac.uk/chebi/init.do
ChemMine	Public	6,200,000	http://bioweb.ucr.edu/ChemMineV2/
ChemNavigator	Commercial	55.3 million	http://www.chemnavigator.com
ChemSpider	Public	26 million	http://www.chemspider.com
Chimiotheque nationale	Public	44,817 compounds	http://chimiotheque-nationale.enscm.fr/index.php
CoCoCo	Public	6,957,134 molecules	http://cococo.unimore.it/tiki-index.php
Desmond Absolute Solvation Free Energies Set	Public	239	http://www.schrodinger.com/Desmond/Absolute-Solvation-Free-Energies-Set
Developmental Therapeutics Program (DTP)	Public	4,73,965	http://dtp.nci.nih.gov/
DNP	Public	40,000	http://dnp.chemnetbase.com/intro/index.jsp
DUD	Commercial	2950	http://dud.docking.org/
DUD.E	Commercial	22,886	http://dude.docking.org/

DrugBank	Public	7739 drugs	http://www.drugbank.ca
e-Drug3D	Public	1632	http://chemoinfo.ipmc.cnrs.fr/MOLDB/index.html
Enamine	Commercial	1.7 million	http://www.enamine.net
GLIDA	Public	G protein—coupled receptors (GPCRs) related Chemical Genomics database, Over 200	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/index.php
Glide Fragment Library	Commercial	441	http://www.schrodinger.com/Glide/Fragment-Library
Glide Ligand Decoys Set	Commercial	1000	http://www.schrodinger.com/Glide/Ligand-Decoys-Set
GLL	Commercial	25,145	http://cavasotto-lab.net/Databases/GDD/
GVK BIO	Commercial	Focused libraries with target inhibitor	http://www.gvkbio.com/informatics.html
HerbMedPro	Commercial	246	http://www.herbmed.org/
i:lib diverse	Commercial	Druglike fragment set for combinatorial library generation	http://www.inteligand.com/
Interbioscreen	Public	440,000 synthetic and 47,000 natural	http://www.ibscreen.com/index.htm
KKB	Public	>1.54M	http://www.eidogen.com/kinasekb.php
Maybridge	Commercial	56,000	http://www.maybridge.com
Mcule	Commercial	–	https://mcule.com/
MDDR	Commercial	150,000	http://accelrys.com/products/databases/bioactivity/mddr.html
MMsINC	Public	–	http://mms.dsfarm.unipd.it/MMsINC/search/
MORE	Commercial	9.7 million	https://itunes.apple.com/us/app/mobile-reagents-universal/id417616789
Mother of All Databases (MOAD)	Public	14,720 ligand—protein complexes, 4782 structures with binding data, 7064 ligands	http://www.bindingmoad.org
NCI	Public	140,000 million	http://dtp.nci.nih.gov/index.html
NRDBSM	Public	17,000	http://www.scfbio-iitd.res.in/software/nrdbsm/index.jsp

(*Continued*)

Table 10.6 (Continued)

Compound database	Availability	No. of compounds[a]	Website
PDB bind	Commercial	3214 ligand – protein complexes	http://www.pdbbind.org/
PubChem	Public	49,875,000	http://pubchem.ncbi.nlm.nih.gov
Specs	Commercial	240,000	http://www.specs.net
SPRESI^{web}	Commercial	5.68 million	http://www.spresi.com/
Super Drug Database (SDD)	Public	2,396 compounds with 1,08,198 conformers	http://bioinf.charite.de/superdrug/
TCM	Public	32,000	http://tcm.cmu.edu.tw
Therapeutic Target Database	Commercial	1906 targets, 5124 drugs	http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp
U.S. Food and Drug Administration (FDA) database	Public	Drugs@FDA includes most of the drug products approved since 1939	http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
WOMBAT	Commercial	331,872 molecules, 1966 targets	http://www.sunsetmolecular.com
ZINC	Public	13 million	http://zinc.docking.org
ZINClick	Public	16 million	http://www.symech.it/index.asp?catID=31&lang=en

[a] These are approximate numbers; -no exact information is available.

REFERENCES

[1] Schneider G. *De novo* design—hop(p)Ing against hope. Drug Discov Today Technol 2013;10:e453–60.
[2] Langer T. Pharmacophores in drug research. Mol Inf 2010;29(6–7):470–5.
[3] Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotech 2009;20:429–36.
[4] Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. J Med Chem 1995;38(14):2681–91.
[5] Datar PA, Khedkar SA, Malde AK, Coutinho EC. Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. J Comput Aided Mol Des 2006;20:343–60.
[6] Tropsha A. Integrated chemo and bioinformatics approaches to virtual screening. In: Tropsha A, Varnek A, editors. Chemoinformatics approaches to virtual screening. London: RSC Publishing; 2008. pp. 295–325.
[7] Oprea TI. Virtual screening in lead discovery: a viewpoint. Molecules 2002;7:51–62.
[8] Kier LB. Molecular orbital calculation of preferred conformations of acetylcholine, muscarine, and muscarone. Mol Pharmacol 1967;3:487–94.
[9] Kier LB. MO Theory in drug research. New York, NY: Academic Press; 1971.
[10] Wermuth CG. Pharmacophores: historical perspective and viewpoint from a medicinal chemist. In: Langer T, Hoffmann RD, editors. Pharmacophores and pharmacophore searches. Weinheim: Wiley-VCH; 2006. pp. 3–13.
[11] Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1997). Pure Appl Chem 1998;70(5):1129–43.
[12] Ehrlich P. Ueber den jetzigen Stand der Chemotherapie. Ber Dtsch Chem Ges 1909;42:17–47.
[13] Leach AR, Gillet VJ, Lewis RA, Taylor R. Three-dimensional pharmacophore methods in drug discovery. J Med Chem 2010;53(2):539–58.
[14] Yang S-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov Today 2010;15(11–12):444–50.
[15] Smellie A, Teig S, Towbin P. Poling: promoting conformational variation. J Comput Chem 1995; 16:171–87.
[16] Kristam R, Gillet VJ, Lewis RA, Thorner D. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. J Chem Inf Model 2005;45(2):461–76.
[17] Sutter J, Guner OF, Hoffman R, Li H, Waldman M. In: Guner OF, editor. Pharmacophore perception, development, and use in drug design. La Jolla, CA: International University Line; 2000.
[18] Accelrys Inc. Discovery studio 2.1. San Diego, CA: Accelrys Inc.; 2010.
[19] Li H, Sutter J, Hoffmann RD. In: Güner OF, editor. Pharmacophore perception, development, and use in drug design. La Jolla, CA: International University Line; 2000.
[20] Debnath AK. Generation of predictive pharmacophore models for CCR5 antagonists: study with piperidine- and piperazine-based compounds as a new class of HIV-1 entry inhibitors. J Med Chem 2003;46(21):4501–15.
[21] Ekins S, Bravi G, Binkley S, Gillespie JS, Ring BJ, Wikel JH, et al. Drug Metab Dispos 2000; 28:994.
[22] Güner OF, Henry DR. Metric for analyzing hit lists and pharmacophores. In: Güner OF, editor. Pharmacophore perception, development, and use in drug design, IUL biotechnology series. La Jolla, CA: International University Line; 2000. pp. 191–212.
[23] Güner OF, Waldman M, Hoffmann RD, Kim JH. Strategies for database mining and pharmacophore development, 1st. In: Güner OF, editor. Pharmacophore perception, development, and use in drug design, IUL biotechnology series. La Jolla, CA: International University Line; 2000. pp. 213–36.
[24] Clement OO, Freeman CM, Hartmann RW, Handratta VD, Vasaitis TS, Brodie AM, et al. Three dimensional pharmacophore modeling of human CYP17 inhibitors. Potential agents for prostate cancer therapy. J Med Chem 2003;46:2345–51.

[25] Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem 2006;49:6789–801.

[26] Willett P, Clark RD. GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. J Comput-Aided Mol Des 2006;20(9):567–87.

[27] Jones G, Willett P, Glen RC. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. J Comput-Aided Mol Des 1995;9(6):532–49.

[28] Poptodorov K, Luu T, Hoffmann RD. In: Langer T, Hoffmann RD, editors. Methods and principles in medicinal chemistry, pharmacophores and pharmacophores searches, vol. 2. Weinheim, Germany: Wiley-VCH; 2006.

[29] Wolber G, Seidel T, Bendix F, Langer T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today 2008;13(1–2):23–9.

[30] Dror O, Shulman-Peleg A, Nussinov R, Wolfson H. Predicting molecular interactions in silico. I. An updated guide to pharmacophore identification and its applications to drug design. Front Med Chem 2006;3:551–84.

[31] Bandyopadhyay D, Agrafiotis DK. A self-organizing algorithm for molecular alignment and pharmacophore development. J Comput Chem 2008;29:965–82.

[32] Totrov M. Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. Chem Biol Drug Des 2008;71:15–27.

[33] Nettles JH, et al. Flexible 3D pharmacophores as descriptors of dynamic biological space. J Mol Graph Model 2007;26:622–33.

[34] Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. J Chem Inf Model 2007;47:279–94.

[35] Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein bound ligands and their use as virtual screening filters. J Chem Inf Model 2005;45(1):160–9.

[36] Chen J, Lai LH. Pocket v.2: further developments on receptor-based pharmacophore modeling. J Chem Inf Model 2006;46:2684–91.

[37] Ortuso F, Langer T, Alcaro S. GBPM: GRID based pharmacophore model. Concept and application studies to protein–protein recognition. Bioinformatics 2006;22(12):1449–55.

[38] SBP is now incorporated into Discovery Studio, available from Accelrys Inc., San Diego, CA.

[39] Brenk R, Klebe G. "Hot spot" analysis of protein-binding sites as a prerequisite for structure-based virtual screening and lead optimization. In: Langer T, Hoffmann RD, editors. Pharmacophores and pharmacophore searches. Weinheim: Wiley-VCH; 2006. pp. 171–92.

[40] Wei D, Jiang X, Zhou L, Chen J, Chen Z, He C, et al. Discovery of multi-target inhibitors by combining molecular docking with common pharmacophore matching. J Med Chem 2008;51(24):7882–8.

[41] Steindl TM, Schuster D, Laggner C, Langer T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. J Chem Inf Model 2006;46(5):2146–57.

[42] Rollinger JM, Hornick A, Langer T, Stuppner H, Prast H. Acetylcholinesterase inhibitory activity of scopolin and scopoletin discovered by virtual screening of natural products. J Med Chem 2004;47(25):6248–54.

[43] Ullmann JR. An algorithm for subgraph isomorphism. J ACM 1976;23:31–42.

[44] Barnard JM. Substructure searching methods: old and new. J Chem Inf Comput Sci 1993;33:532–8.

[45] Xu J. GMA: a generic match algorithm for structural homomorphism, isomorphism, maximal common substructure match and its applications. J Chem Inf Comput Sci 1996;36:25–34.

[46] Giménez-Oya V, Villacañas O, Fernàndez-Busquets X, Rubio-Martinez J, Imperial S. Mimicking direct protein–protein and solvent mediated interactions in the CDP-methylerythritol kinase homodimer: a pharmacophore-directed virtual screening approach. J Mol Model 2009;15(8):997–1007.

[47] Tschinke V, Cohen N. The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. J Med Chem 1993;36:3863–70.

[48] Roe DC, Kuntz I. BUILDER v.2: improving the chemistry of a *de novo* design strategy. J Comput Aided Mol Des 1995;9:269–82.

[49] Huang Q, et al. PhDD: a new pharmacophore-based *de novo* design method of drug-like molecules combined with assessment of synthetic accessibility. J Mol Graph Model 2010;28(8):775–87.

[50] Kirkpatrick P. Virtual screening: gliding to success. Nat Rev Drug Disc 2004;3:299.

[51] Ewing JAT, Kuntz ID. Critical evaluation of search algorithms for automated molecular docking and database screening. J Comput Chem 1997;18:1175.

[52] Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed 2002;41(15):2644–76.

[53] Gohlke H, Kleb G. Statistical potentials and scoring functions applied to protein–ligand binding. Curr Opin Struct Biol 2001;11(2):231–5.

[54] Peitsch MC, Schwede T, Diemand A, Guex N. In: Jiang T, Xu Y, Zhang MQ, editors. Current topics in computational molecular biology. Cambridge, MA: MIT Press; 2002. pp. 449–66.

[55] Zimmer R, Lengauer T. In: Lengauer T, editor. Bioinformatics—from genomes to drugs. New York, NY: Wiley-VCH; 2002. pp. 237–313.

[56] Bitetti-Putzer R, Joseph-McCarthy D, Hogle JM, Karplus M. Functional group placement in protein binding sites: a comparison of GRID and MCSS. J Comput Aided Mol Des 2001;15(10): 935–60.

[57] Leulliot N, Varani G. Current topics in RNA–protein recognition: control of specificity and biological function through induced fit and conformational capture. Biochemistry 2001;40:7947–56.

[58] Davis AM, Teague SJ. Hydrogen bonding, hydrophobic interactions and failure of the rigid receptor hypothesis. Angew Chem Int Ed Engl 1999;38:736–49.

[59] Totrov M, Abagyan R. Flexible ligand docking to multiple receptor conformations: a practical alternative. Curr Opin Struct Biol 2008;18:178–84.

[60] Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft docking and multiple receptor conformations in virtual screening. J Med Chem 2004;47:5076–84.

[61] Jiang F, Kim SH. Soft docking: matching of molecular surface cubes. J Mol Biol 1991;219:79–102.

[62] Leach AR. Ligand docking to proteins with discrete side-chain flexibility. J Mol Biol 1994;235: 345–56.

[63] Meiler J, Baker D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. Proteins 2006;65:538–84.

[64] Nabuurs SB, Wagener M, de Vlieg J. A flexible approach to induced fit docking. J Med Chem 2007;50:6507–18.

[65] Davis IW, Baker D. ROSETTALIGAND docking with full ligand and receptor flexibility. J Mol Biol 2009;385:381–92.

[66] Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, et al. Target flexibility: an emerging consideration in drug discovery and design. J Med Chem 2008;51:6237–55.

[67] Abseher R, Horstink L, Hilbers CW, Nilges M. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. Proteins 1998;31:370–82.

[68] Knegtel RM, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures. J Mol Biol 1997;266:424–40.

[69] Lorber DM, Shoichet BK. Hierarchical docking of databases of multiple ligand conformations. Curr Top Med Chem 2005;5:739–49.

[70] Huang S-Y, Zou X. Advances and challenges in protein–ligand docking. Int J Mol Sci 2010; 11:3016–34.

[71] Jain AN. Scoring functions for protein–ligand docking. Curr Protein Pept Sci 2006;7:407–20.

[72] Huang N, Kalyanaraman C, Irwin JJ, Jacobson MP. Molecular mechanics methods for predicting protein–ligand binding. J Chem Inf Model 2006;46:243–53.

[73] Weiner PK, Kollman PA. AMBER—assisted model building with energy refinement. A general program for modeling molecules and their interactions. J Comput Chem 1981;2:287–303.

[74] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.

[75] Verkhivker G, Appelt K, Freer ST, Villafranca JE. Empirical free energy calculations of ligand–protein crystallographic complexes. I. Knowledge-based ligand–protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. Protein Eng 1995;8:677–91.

[76] Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42:5100–9.

[77] Lee J, Seok C. A statistical rescoring scheme for protein–ligand docking: consideration of entropic effect. Proteins 2008;70:1074–83.

[78] Venkatesan SK, Shukla AK, Dubey VK. Molecular docking studies of selected tricyclic and quinone derivatives on trypanothione reductase of *Leishmania infantum*. J Comput Chem 2010;31(13):2463.

[79] Kroemer RT. Structure-based drug design: docking and scoring. Curr Protein Pept Sci 2007;8:312–28.

[80] Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2003;2 (7):527–41.

[81] Vigers GPA, Rizzi JP. Multiple active site corrections for docking and virtual screening. J Med Chem 2004;47(1):80–9.

[82] DockIt: Metaphorics, Aliso Viejo, CA, <http://www.metaphorics.com/products/dockit>.

[83] Terp GE, Johansen BN, Christensen IT, Jørgensen FS. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein–ligand binding affinities. J Med Chem 2001;44(14):2333–43.

[84] Klon AE, Glick M, Davies JW. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. J Chem Inf Comput Sci 2004;44:2216–24.

[85] Wade RC, Ortiz AR, Gago F. Comparative binding energy analysis. Persp Drug Discov Des 1998; 11:19–34.

[86] Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol 2008;153 (S1):S7–26.

[87] Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in drug design—a review. Curr Top Med Chem 2010;10(1):95–115.

[88] Lushington GH, Guo JX, Wang JL. Whither combine? New opportunities for receptor-based QSAR. Curr Med Chem 2007;14(17):1863–77.

[89] Kmunicek J, Hynkova K, Jedlicka T, Nagata Y, Negri A, Gago F, et al. Quantitative analysis of substrate specificity of haloalkane dehalogenase LinB from *Sphingomonas paucimobilis* UT26. Biochemistry 2005;44:3390–401.

[90] Wang T, Tomic S, Gabdoulline RR, Wade RC. How optimal are the binding energetics of barnase and barstar? Biophys J 2004;87:1618–30.

[91] Tomic S, Nilsson L, Wade RC. Nuclear receptor–DNA binding specificity: a COMBINE and Free-Wilson QSAR analysis. J Med Chem 2000;43:1780–92.

[92] VLife MDS. 3.5 is a software of VLife Sciences Technologies Private Limited, 2007–2008, <http://www.vlifesciences.com>.

[93] Gil-Redondo R, Klett J, Gago F, Morreale A. gCOMBINE: a graphical user interface to perform structure-based comparative binding energy (COMBINE) analysis. Proteins 2010;78(1):162–72.

[94] Vedani A, Briem H, Dobler M, Dollinger K, McMasters DR. Multiple conformation and protonation-state representation in 4D-QSAR. J Med Chem 2000;43:4416–27.

[95] Verma J, Khedkar VM, Prabhu AS, Khedkar SA, Malde AK, Coutinho EC. A comprehensive analysis of the thermodynamic events involved in ligand–receptor binding using CoRIA and its variants. J Comput Aided Mol Des 2008;22(2):91–104.

[96] Dhaked DK, Verma J, Saran A, Coutinho EC. Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). J Mol Model 2009;15(3):233–45.

[97] Dror O, Shulman-Peleg A, Nussinov R, Wolfson HJ. Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications to drug design. Curr Med Chem 2004; 11:71–90.

[98] Jahn A, Hinselmann G, Fechner N, Zell A. Optimal assignment methods for ligand-based virtual screening. J Cheminform 2009;1:14.

[99] Villoutreix BO, Renault N, Lagorce D, Sperandio O, Montes M, Miteva MA. Free resources to assist structure-based virtual ligand screening experiments. Curr Protein Pept Sci 2007;8:381−411.

[100] Fox S, Farr-Jones S, Yund MA. High throughput screening for drug discovery: continually transitioning into new technology. J Biomol Screen 1999;4:183−6.

[101] 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(1−3):3−25.

[102] QikProp, version 3.4, Schrödinger, LLC, New York, NY; 2011.

[103] Wilson GL, Lill MA. Integrating structure-based and ligand-based approaches for computational drug design. Future Med Chem 2011;3:735−50.

[104] Liebeschuetz JW, Jones SD, Morgan PJ, Murray CW, Rimmer A, Roscoe JM, et al. PRO_SELECT: combining structure-based drug design and array-based chemistry for rapid lead discovery. 2. The development of a series of highly potent and selective factor Xa inhibitors. J Med Chem 2002;45:1221−32.

[105] Sharma A, Tendulkar AV, Wangikar PP. Drug discovery against H1N1 virus (influenza A virus) via computational virtual screening approach. Med Chem Res 2011;20(9):1445−9.

[106] Dahlgren MK, Zetterström CE, Gylfe Å, Linusson A, Elofsson M. Statistical molecular design of a focused salicylidene acylhydrazide library and multivariate QSAR of inhibition of type III secretion in the Gram-negative bacterium Yersinia. Bioorg Med Chem 2010;18(7):2686−703.

[107] Kubinyi H. Success stories of computer-aided design. In: Ekins S, editor. Computer applications in pharmaceutical research and development. New York: John Wiley & Sons; 2006. pp. 377−424.

[108] Schneider G, Böhm H. Virtual screening and fast automated docking methods: combinatorial chemistry. Drug Discov Today 2002;7:64−70.

[109] Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T, et al. Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 2012;52:867−81.