CHAPTER TWO

Recent Trends in Quantitative Structure-Activity Relationships

A. TROPSHA
Laboratory for Molecular Modeling
School of Pharmacy
University of North Carolina
Chapel Hill, North Carolina

Contents

1 Introduction, 50
1.1 A Unified Concept of QSAR, 51
1.2 The Taxonomy of QSAR Approaches, 52
2 Multiple Descriptors of Molecular Structure, 54
2.1 Topological Descriptors, 54
2.2 3D Descriptors, 55
3 QSAR Modeling Approaches, 58
3.1 3D-QSAR, 58
3.2 The Descriptor Pharmacophore Concept and Variable Selection QSAR, 60
3.2.1 Linear Models, 61
3.2.2 Nonlinear Models, 62
4 Validation of QSAR Models, 63
4.1 Beware of q^2, 64
4.2 Rational Selection of Training and Test Sets, 64
4.3 Guiding Principles of Safe QSAR, 66
5 QSAR Models as Virtual Screening Tools, 66
5.1 Data Mining and SAR Analysis, 66
5.2 Virtual Screening, 67
5.3 Rational Library Design by use of QSAR, 68
6 Conclusions, 69

Burger's Medicinal Chemistry and Drug Discovery
Sixth Edition, Volume 1: Drug Discovery
Edited by Donald J. Abraham
ISBN 0-471-27090-3 © 2003 John Wiley & Sons, Inc.

1 INTRODUCTION

Quantitative structure-activity relationship (QSAR) methodology was introduced by Hansch et al. in the early 1960s (1, 2). The approach stemmed from linear free-energy relationships in general and the Hammett equation in particular (3). It is based on the assumption that the difference in structural properties accounts for the difference in biological activities of compounds. According to this approach, the structural changes that affect the biological activities of a set of congeners are of three major types: electronic, steric, and hydrophobic (4). These structural properties are often described by Hammett electronic constants (5), Verloop STERIMOL parameters (6), hydrophobic constants (5), to name but a few. The relationship between a biological activity (or chemical property) and the structural parameters is obtained through the use of linear or multiple linear regression (MLR) analysis. The fundamentals and applications of this method in chemistry and biology have been summarized by Hansch and Leo (4) and an account of the most recent developments in this area of traditional QSAR appears in the chapter by Celassie in this series (7). As discussed in that chapter, the history of modern QSAR counts over 40 years of active research in method development and its applications. It is practically impossible to review all, even relatively recent, developments in the field in a single chapter. Several reviews and monographs on QSAR and its applications have been published in recent years (4, 8–12) and the reader is referred to this collection of general references and publications cited therein for additional in-depth information.

One of the most characteristic features of the modern age QSAR as an integral part of drug design and discovery is an unprecedented growth of biomolecular databases, which contain data on chemical structure and, in some cases, biological activity (or other relevant drug properties such as toxicity or mutagenicity) of chemicals. Figure 2.1 illustrates the fast growth of one of such databases, the Chemical Abstract Service (CAS) registry file (13). The

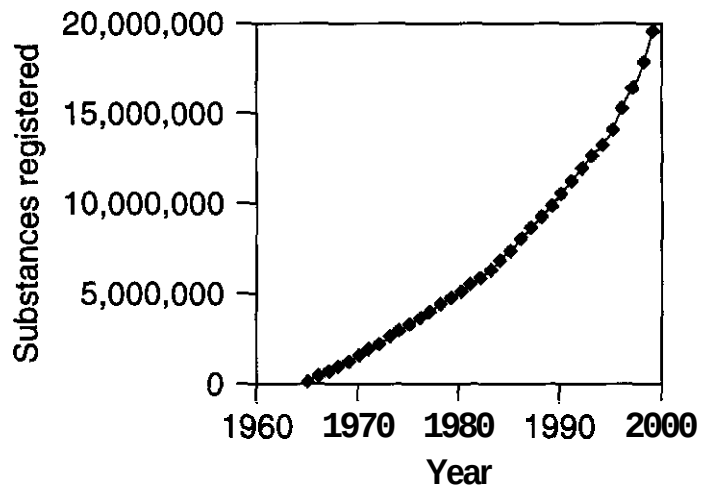


Figure 2.1. Growth in the number of chemical compounds, excluding biopolymers, registered by the Chemical Abstract Service (CAS).

growth has been phenomenal: CAS currently contains more than 39 million compounds, including biological sequences [and it does not include chemical libraries, which literally include billions of compounds (14)]. Naturally, the growth of molecular databases has been concurrent with the acceleration of the drug discovery process. According to an excellent, recent historical account of drug discovery (15), as the result of high throughput screening (HTS) technologies, the amount of raw data points obtained by a large pharmaceutical company per year has increased from approximately 200,000 at the beginning of last decade to around 50 million today. The total number of drugs used worldwide is approximately 80,000, which reportedly act at less than 500 confirmed molecular targets (15). Recent estimates suggest that the number of potential targets lies between 5000 and 10,000, approximately 10-fold greater than the number of targets currently pursued (15). Although traditional QSAR modeling has been typically limited to deal with a maximum of several dozen compounds at a time, rapid generation of large quantities of data requires new methodologies for data analysis. New approaches need to be developed to establish QSAR models for hundreds, if not thousands, of molecules. These new methods should be robust, yet computationally efficient, to compete with the experimental methods of drug discovery, such as combinatorial chemistry and HTS.

This chapter concentrates on recent trends and developments in QSAR methodology, which are characterized by the growing size of the data sets subjected to the QSAR analysis, use of multiple descriptors of chemical structure, application of both linear and, especially, nonlinear optimization algorithms applicable to multidimensional data modeling, growing emphasis on the rigorous model validation, and application of QSAR models as virtual screening tools in database mining and chemical library design. We begin by presenting a unified concept of QSAR, emphasizing common aspects of different QSAR methodologies. We then consider some popular approaches to the derivation of molecular descriptors and optimization algorithms in the context of three important components of any QSAR investigation: model development, model validation, and model utility. We conclude with several remarks on present status and future developments in this exciting research discipline.

1.1 A Unified Concept of QSAR

An inexperienced user or sometimes even an avid practitioner of QSAR could be easily confused by the multitude of methodologies and naming conventions used in QSAR studies. Two-dimensional (2D) and three-dimensional (3D) QSAR, variable selection and artificial neural network methods, comparative molecular field analysis (CoMFA), and binary QSAR present examples of various terms that may appear to describe totally independent approaches, which cannot be even compared to each other. In fact, any QSAR method can be generally defined as the application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) of the form $P_i = \hat{k}(D_1, D_2, \cdots D_n)$, where P_i are biological activities (or other properties of interest) of molecules, $D_1, D_2, \ldots, D_n$ are calculated (or, sometimes, experimentally measured) structural properties (molecular descriptors) of compounds, and $\hat{k}$ is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules. The relationship between values of descriptors D and target properties P can be linear [e.g., multiple linear regression (MLR)

Structure Id	Target Property (EC_{50}, K_i, etc.)	Structural Properties (descriptors)			
Comp. 1	P1	D11	D12	···	Dln
Comp. 2	P2	D21	D22	···	D2n
···	···				
Comp. m	Pm	Dm1	Dm2	···	Dmn

$$\{P\} = \hat{K}\{D\}$$

Figure 2.2. Standard QSAR **table** is a general starting point of any QSAR approach.

as in the Hansch QSAR approach], where target property can be calculated directly from the descriptor values, or nonlinear (such as artificial neural networks or classification QSAR methods), where descriptor values are used in characterizing chemical similarity between molecules, which in turn is used to predict compound activity. In general, each compound can be represented by a point in a multidimensional space, in which descriptors $D_1, D_2, \cdots, D_n$ serve as independent coordinates of the compound. The goal of QSAR modeling is to establish a trend in the descriptor values, which correlates, in a linear or nonlinear fashion, with the trend in biological activity. All QSAR approaches imply, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities. This implies that points representing compounds with similar activities in multidimentional descriptor space should be geometrically close to each other, and vice versa.

Despite formal differences between various methodologies, any QSAR method is based on a QSAR table, which can be generalized, as shown in Fig. 2.2. To initiate a QSAR study, this table must include some identifiers of chemical structures (e.g., company's ID numbers, first column of the table in Fig. 2.2), reliably measured values of biological activity [or any other target property of interest (e.g., solubility, metabolic transformation rate, etc.;

second column)], and calculated values of molecular descriptors in all remaining columns (sometimes, experimentally determined physical properties of compounds can be used as descriptors as well).

The differences in various QSAR methodologies can be understood in terms of types of target property values, types of descriptors, and differences in optimization algorithms used to relate descriptors to the target properties. The target property values can be defined as activity classes [i.e., active or inactive, frequently encoded numerically for the purpose of the subsequent analysis as one (for active) or zero (for inactive)] or as a continuous range of values; the corresponding methods of data analysis are referred to as classification or continuous property QSAR, respectively. Descriptors can be generated from various representations of molecules (e.g., 2D chemical graphs or 3D molecular geometries), giving rise to the terms of 2D- or 3D-QSAR, respectively. Finally, the types of optimization algorithms used in the QSAR model development lead to the definitions of linear versus nonlinear QSAR methods.

In some cases, the types of biological data, the choice of descriptors, and the class of optimization methods are closely related and mutually inclusive. For instance, multiple linear regression can be applied only when a relatively small number of molecular descriptors are used (at least five to six times smaller than the total number of compounds) and the target property is characterized by a continuous range of values. The use of multiple descriptors makes it impossible to use MLR because of a high chance of spurious correlation (16) and requires the use of partial least squares or nonlinear optimization techniques. However, in general, for any given data set a user could choose between various types of descriptors and various optimization schemes, combining them in a practically mix-and-match mode, to arrive at statistically significant QSAR models in a variety of ways. This situation is in essence analogous to molecular mechanics calculations (17), where different force fields and differently derived parameters are developed by different groups, although the common goal is to compute (unique) optimized geometries of molecules from their chemical composition and coordinates of all atoms. Thus, in general, all QSAR models can be universally compared in terms of their statistical significance and, most important, their ability to predict accurately biological activities (or other target properties) of molecules not included in the training set (cf. molecular mechanics, where different methods are ultimately compared by their ability to reproduce experimental molecular geometries). This concept of statistical robustness and the predictive ability as universal characteristics of any QSAR model independent of the particulars of individual approaches should be kept in mind as we consider examples of QSAR tools, their applications, and pitfalls in the subsequent sections of this chapter.

1.2 The Taxonomy of QSAR Approaches

Many different approaches to QSAR have been developed since Hansch's seminal work. As briefly discussed above, the major differences between these methods can be analyzed from two viewpoints: (1) the types of structural parameters that are used to characterize molecular identities, starting from different representation of molecules, from simple chemical formulas to three-dimensional conformations; and (2) the mathematical procedure that is employed to obtain the quantitative relationship between these structural parameters and biological activity.

On the basis of the origin of molecular descriptors used in calculations, QSAR methods can be divided into three groups. One group is based on a relatively small number (usually many times smaller than the number of compounds in a data set) of physicochemical properties and parameters describing, for example, hydrophobic, steric, and electrostatic effects. Usually, these descriptors are used as independent variables in multiple regression approaches (18). In the literature, these methods are typically referred to as Hansch analysis (8). These types of descriptors and corresponding linear optimization methods used in traditional QSAR analyses are discussed extensively in the chapter by Celassie (7) and therefore is not reviewed here.

More recent methods are based on quantitative characteristics of molecular graphs (molecular topological descriptors). Because

molecular graphs or structural formulas are "two-dimensional," these methods are referred to as 2D-QSAR. Most of the 2D-QSAR methods are based on graph theoretical indices, which have been extensively studied by Randíc (19) and Kier and Hall (20–22). They include, for example, molecular connectivity indices (19, 20), molecular shape indices (23, 24), topological (25) and electrotopological state indices (26–29), and atom-pair descriptors (30, 31). Sometimes, topological descriptors are also combined with physicochemical properties of molecules. Although these structural indices represent different aspects of molecular structures, and, what is important for QSAR, different structures provide numerically different values of indices, their physicochemical meaning is frequently unclear. The successful applications of topological indices combined with multiple linear regression (MLR) analysis have been summarized by Kier and Hall (20, 21, 28).

The third group of methods is based on descriptors derived from spatial (three-dimensional) representation of molecular structures. Correspondingly, these methods are referred to as three-dimensional or 3D-QSAR; they have become increasingly popular with the development of fast and accurate computational methods for generating 3D conformations and alignments of chemical structures. The early examples of 3D-QSAR include molecular shape analysis (MSA) (32), distance geometry (33, 34), and Voronoi techniques (35). The first method uses shape descriptors and multiple linear regression analysis, whereas the latter methods apply atomic refractivity as structural descriptors and the solution of mathematical inequalities to obtain the quantitative relationships. These two methods have been applied to the study of structure-activity relationships of many data sets by Hopfinger (e.g., Refs. 36, 37) and Crippen (e.g., Refs. 38, 39), respectively.

Perhaps the most popular example of 3D-QSAR is the comparative molecular field analysis (CoMFA), developed by Cramer et al. (40), which has elegantly combined the power of 3D molecular modeling and partial least-square (PLS) optimization technique (41, 42) and found wide applications in medicinal chemistry and toxicity analysis (see below). Most of 3D-QSAR methods require 3D alignment of all molecules according to a phannacophore model or based on ligand docking to a receptor-binding site. Descriptors in the case of CoMFA (40, 43) and CoMFA-like methods such as COMBINE (44), COMSiA (45), and QsiAR (46) represent electrostatic, steric, and hydrophobic field values (to name but a few examples) in the grid points surrounding molecules.

Finally, QSAR methods can also be classified by the type of the correlation methods used in model development. Linear methods include linear regression or MLR, PLS (41, 42, 47), or principal component regression (PCR), whereas nonlinear methods can be exemplified, for example, by k-Nearest Neighbors (kNN) (48, 49) and artificial neural networks (50) methods. An example of the linear methods is provided by the ADAPT system, which employs topological indices as well as other calculable structural parameters (e.g., steric and quantum mechanical parameters), and the MLR method for QSAR analysis. It has been extensively applied to QSAR/QSPR studies in analytical chemistry, toxicity analysis, and other biological activity prediction (51–54). Parameters derived from various experiments through chemometric methods have also been used in the study of peptide QSAR (55), where PLS analysis was employed. The latter technique has been used almost exclusively in 3D-QSAR, where the number of descriptors characterizing molecular fields may exceed the number of compounds by orders of magnitude.

There has been a great deal of interest, especially more recently, in the use of data mining methods to extract the information from large and/or chemically inhomogeneous data sets. Examples of these methods include pattern recognition (56, 57), automated structure evaluation (58, 59), neural network (60–62), and machine learning (63–65). Recent trends in QSAR studies also include developing optimal QSAR models through variable selection, that is, by selecting a subset of available descriptors in either MLR, PLS, or nonlinear classification or artificial neural networks (ANN) analysis as applied either in 2D- (66–72) or in 3D-QSAR (73). These methods employ either generalized simulated annealing

(67), or genetic algorithms (68), or evolutionary algorithms (69–72) as optimization tools. The effectiveness and convergence of these algorithms are strongly affected by the choice of a fitting function, which drives the optimization process (70–72). It has been demonstrated that optimization combined with variable selection effectively improves QSAR models as compared to those without variable selection. For example, GOLPE (74) was developed through the use of chemometric principles and q^2-GRS (75) was developed on the basis of independent CoMFA analysis of small areas of CoMFA descriptor space, to address the issue of region selection. Both of these methods have been shown to improve QSAR models compared to the original CoMFA technique.

Different QSAR methods have their own strengths and weaknesses. For example, 3D-QSAR methods generally result in the diagrams of important molecular fields that can be easily interpreted in terms of specific steric and electrostatic interactions important for the ligand binding to their receptor. However, the need to align structures in 3D, which is time-consuming and subjective, precludes the use of 3D-QSAR techniques for the analysis of large data sets. On the other hand, 2D-QSAR methods are much faster and more amenable to automation because they require no conformational search and structural alignment. Thus, 2D methods are best suited for the analysis of large numbers of compounds and computational screening of molecular databases; however, the interpretation of the resulting models in familiar chemical terms is frequently difficult, if not impossible.

The generality of the QSAR modeling approach as a drug discovery tool, irrespective of descriptor types or optimization algorithms, can be best demonstrated in the context of inverse QSAR, which can be defined as designing or discovering molecular structures with a desired property on the basis of QSAR models (76–78). In practical terms, inverse QSAR also includes searching for molecules with a desired target property in chemical databases or virtual chemical libraries. These considerations emphasize the universal importance of establishing QSAR model robustness and predictive ability as opposed to concentrating on explanatory power, which has been a characteristic feature of many traditional QSAR approaches.

2 MULTIPLE DESCRIPTORS OF MOLECULAR STRUCTURE

It has been said frequently that there are three keys to the success of any QSAR model building exercise: descriptors, descriptors, and descriptors. Many different molecular representations have been proposed, exemplified by Hansch-type parameters (2), topological indices (19, 79), quantum mechanical descriptors (80), molecular shapes (32, 81), molecular fields (40), atomic counts (82), 2D fragments (83–85), 3D fragments (86–88), molecular eigenvalues (89), molecular multipole moments (90), E-state fields (28), molecular fragment-based hash codes (91, 92), and molecular holograms (93). A recent review by Livingstone provides an excellent survey of various 2D and 3D descriptors, along with some associated diversity and similarity functions (9). Various physicochemical parameters such as the partition coefficient, molar refractivity, and quantum mechanical quantities such as highest occupied molecular orbital (HOMO) and lowest occupied molecular orbital (LUMO) energies have been used to represent molecular identities in early QSAR studies by the use of linear and multiple linear regression. However, these descriptors are not suited for the analysis of large numbers of molecules, either because of the lack of physicochemical parameters for compounds yet to be synthesized or because of the computational expenses required by quantum mechanical methods. Recent years have seen the application of various topological descriptors that are usually derived from either 2D or 3D molecular structural information based on the graph theory or molecular topology (20–22, 94). These descriptors are generated on the basis of the molecular connectivity, 3D molecular topography, and molecular field properties.

2.1 Topological Descriptors

Two widely applied examples of 2D molecular descriptors are molecular connectivity indices

(MCI) and atom-pair (AP) descriptors. Molecular connectivity indices, χ, were first formulated by Randíc (19) and subsequently generalized and extended by Kier and Hall (20–22). The fundamentals and applications of molecular connectivity indices have been thoroughly reviewed (22, 28). A popular MolConnZ software (95) affords the computation of a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, the graph's radius and diameter, Wiener and Platt indices, Shannon and Bonchev-Trinajstic information indices, counts of different vertices, and counts of paths and edges between different kinds of vertices (19, 20, 96–100).

Overall, MolConnZ (95) produces over 400 different descriptors. Most of these descriptors characterize chemical structure, but several depend on the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In a typical QSAR study, only about one-half of all possible MolConnZ descriptors are eventually used, after deleting descriptors with zero value or zero variance. Figure 2.3 provides a summary of these molecular descriptors and presents some algorithms used in their derivation.

The idea of using atom pairs as molecular features in structure-activity studies was first proposed by Carhart et al. (84). AP descriptors are defined by their atom types and topological distance bins. An AP is a substructure defined by two atom types and the shortest path separation (or graph distance) between the atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The general form of an atom-pair descriptor is as follows:

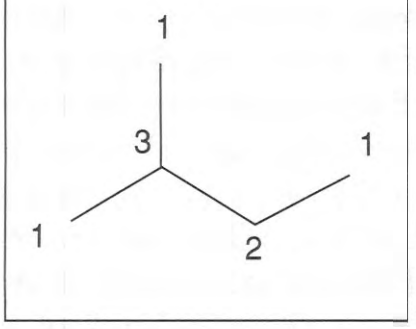
$$\text{atom type } i \text{ —— (distance) —— atom type } j$$

where atom chemical types are typically defined by the user. For example, 15 atom types can be defined by use of the SYBYL mol2 format (101) as follows: (1) negative charge center (NCC); (2) positive charge center (PCC); (3) hydrogen bond acceptor (HA); (4) hydrogen bond donor (HD); (5) aromatic ring center (ARC); (6) nitrogen atoms (N); (7) oxygen atoms (O); (8) sulfur atoms (S); (9) phosphorous atoms (P); (10) fluorine atoms (FL); (11) chlorine, bromine, iodine atoms (HAL); (12) carbon atoms (C); (13) all other elements (OE); (14) triple bond center (TBC); and (15) double bond center (DBC). Apparently, the total number of pairwise combinations of all 15 atom types is 120. Furthermore, distance bins should be defined to discriminate between identical atom pairs separated by different graph distances and therefore representing different molecular substructures. Thus, 15 distance bins can be introduced in the interval between graph distance zero (i.e., zero atoms separating an atom pair) to 14 and greater. Thus, in this format a total of 1800 (120×15) AP descriptors can be generated for any molecular structure. An example of an atom-pair descriptor is shown on Fig. 2.4. Frequently, as applied to particular data sets, many of the theoretically possible AP descriptors have zero value (implying that certain atom types or atom pairs are absent in molecular structures). For instance, in our recent studies of 48 anticonvulsant agents, only 273 descriptors with nonzero value and nonzero variance were' generated (102).

2.2 3D Descriptors

The rapid increase in structural three-dimensional (3D) information of bioorganic molecules (103, 104), coupled with the development of fast methods for 3D structure generation [e.g., CONCORD (105, 106) and CORINA (107)] and alignment [e.g., Active Analog Approach (43, 108)], have led to the development of 3D structural descriptors and associated 3D-QSAR methods. Many 3D-QSAR methods (considered below) make use of so-called molecular field descriptors. To calculate these descriptors, steric and electrostatic fields of all molecules are sampled with a probe atom, usually carbon sp^3 bearing a $+1$ charge, on a rectangular grid that encompasses structurally aligned molecules. The values of both van der Waals and electrostatic interactions between the probe atom and all

Hydrogen - depleted molecular graph and vertex degrees a;



Extended connectivity indices

$$ {}^{n}EC = \sum_{i=1}^{N} {}^{n}b_i = \sum_{i=1}^{N} \sum_{k=1}^{j_i} {}^{n-1}b_k $$

$$ {}^{n}b_i = \sum_{k=1}^{j_i} {}^{n-1}b_k $$ is the sum of vertex degrees connected to vertex i, ${}^{0}b_i = a_i$

Connectivity indices

$$ \sum_{i=1}^{N} (a_i)^f $$

$$ \sum (a_{i_1} a_{i_2})^f $$ **All edges**

$$ \sum (a_{i_1} a_{i_2} \ldots a_{i_v})^f $$ All (n – 1)-edge subgraphs

f = – 0.5	$^{n-}\chi$	Molecular connectivity indices
f = 1	$^{n-1}M_2$	Zagreb group indices
f = 2	$^{n-1}M_1$	Zagreb group indices

Overall connectivity indices

$$ {}^{n}TC_f = \sum_{i=1}^{j_n} \sum_{j=1}^{k_i} {}^{n}b^{i}_{j} $$

$^{n}b^{i}_{j}$ vertex degree of atom j in subgraph k_i of order n,
f = 0, vertex degrees of subgraphs
f = 1, vertex degrees of the whole molecular graph

Figure 2.3. Examples of topological descriptors frequently used in QSAR studies.

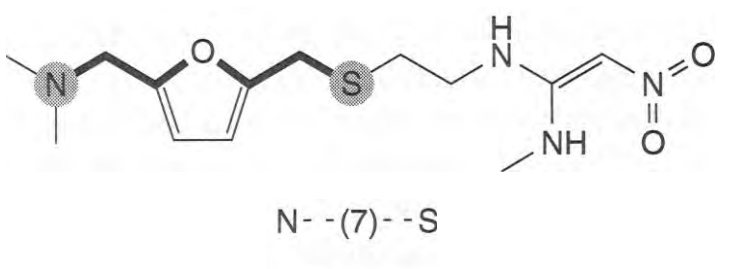
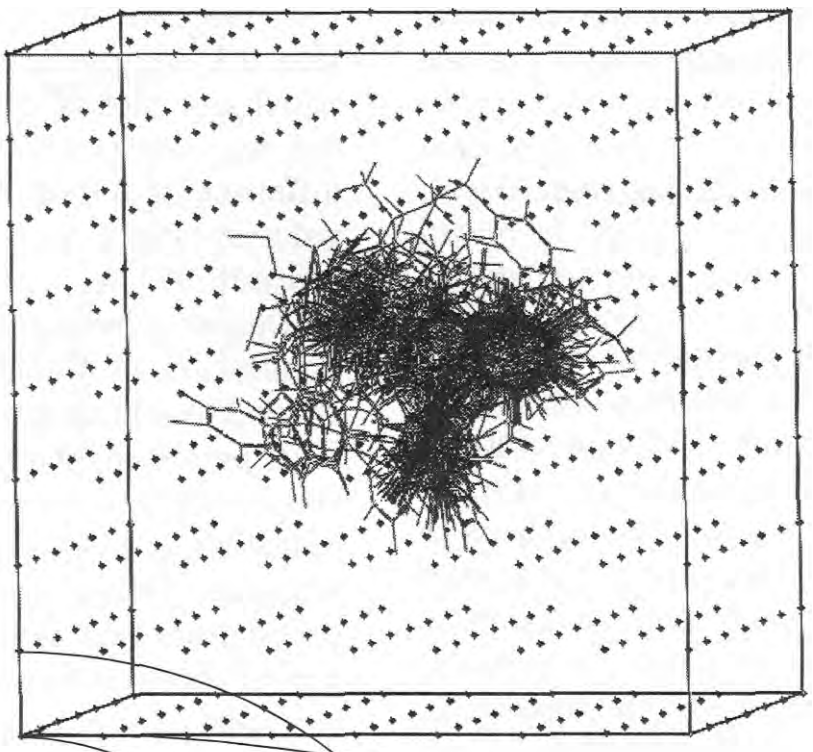


Figure 2.4. Example of an AP descriptor: two atom types, aliphatic nitrogen and aliphatic sulfur, separated by the shortest chemical graph path of seven.

atoms of each molecule are calculated in every lattice point by use of the force field equation described above and entered into the CoMFA QSAR table (Fig. 2.5), which typically contains thousands of columns. Additional molecular field descriptors such as HINT (**Hydropathic INTeraction**) descriptors (109) could improve the CoMFA model. PLS algorithms coupled with leave-one-out (LOO) cross-validation is typically used to arrive at statistically significant CoMFA models.



	Bio Act.	S001	S002 ... S998	E001 ... E998
Cdpl	5.1			
Cdp2	6.8			
...				
...				
...				

Figure 2.5. Process of steric and electrostatic descriptor generation in CoMFA. Note that this process results in a familiar QSAR table (cf. Fig. 2.2). PLS is used as a standard analytical technique in CoMFA.

One of the most attractive features of the CoMFA and CoMFA-like methods is that, because of the nature of molecular field descriptors, these approaches yield models that are relatively easy to interpret in chemical terms. Famous CoMFA contour plots, which are obtained as a result of any successful CoMFA study, tell chemists in rather plain terms how the change in the compounds' size or charge distribution as a result of chemical modification correlate with the binding constant or activity. These observations may immediately suggest to a chemist possible ways to modify molecules to increase their potencies. However; as demonstrated in the next section, these predictions should be taken with caution only after sufficient work has been done to prove the statistical significance and predictive ability of the models.

By analogy with 2D atom-pair descriptors (Fig. 2.4), 3D AP descriptors can also be defined through the use of similar atom types and atom pairs and 3D molecular topography; in this case, a physical distance between atom types is used in place of chemical graph distance. The distance between two "atoms" is measured and then assigned into one or two distance bins. Typically, the width of each distance bin is chosen as 1.0 Å. Because it is also designed to let the adjacent bins have 10% overlap with each other, the actual length of each distance bin is 1.2 Å. Any distance located in the overlap region is assigned to both bins. This "fuzzy distance" concept is adopted to alleviate the possible unfavorable boundary effects of the distance bins. For example, with strict boundary conditions, a distance of 2.05 Å will be assigned only to bin No. 2, but it can be reasonably argued that it is almost as close to the upper half of bin No. 1 as to bin No. 2. With fuzzy boundary conditions, 2.05 Å belongs to both bin No. 1 and bin No. 2, allowing

a possible match to either. All the distances greater than 20 Å are assigned into the last bin.

3 QSAR MODELING APPROACHES

3.1 3D-QSAR

Two original 3D-QSAR methods, CoMFA (40) and GRID (110), were developed almost simultaneously in the mid- to late-1980s (9). Since its introduction, the CoMFA approach has rapidly become one of the most popular methods of QSAR. Over the years, this approach has been applied to a wide variety of receptor and enzyme ligands [many reviews appeared in a recent monograph (10)]. Undoubtedly, the further development of this and related methods is of great importance and interest to many scientists working in the area of rational drug design.

CoMFA methodology is based on the assumption that because, in most cases, the drug-receptor interactions are noncovalent, the changes in the biological activities or binding affinities of sample compounds correlate with changes in the steric and electrostatic fields of these molecules. In a standard CoMFA procedure, all molecules under investigation are first structurally aligned, and the steric and electrostatic fields around them are sampled with probe atoms, usually sp^3 carbon with a +1 charge, on a rectangular grid that encompasses aligned molecules. The results of the field evaluation in every grid point for every molecule in the data set are placed in the CoMFA QSAR table, which therefore contains thousands of columns (Fig. 2.5). The analysis of this table by the means of standard multiple regression is practically impossible; however, the application of special multivariate statistical analysis routines, such as PLS analysis and LOO cross-validation ensures the statistical significance of the final CoMFA equation. The outcome from this procedure is a cross-validated correlation coefficient R^2 (q^2), which is calculated according to the formula

$$q^2 = 1 - \frac{\Sigma (y_i - \hat{y}_i)^2}{\Sigma (y_i - \bar{y})^2}, \qquad (2.1)$$

where y_i, $\hat{y}_i$, and $\bar{y}$ are the actual, estimated, and averaged (over the entire data set) activities, respectively. The summations in Equation 2.1 are performed over all compounds, which are used to build a model for the training set. The statistical meaning of the q^2 is different from that of the conventional r^2: a q^2 value greater than 0.3 is often considered significant (111).

Despite obviously successful and growing application of CoMFA in molecular design, several problems intrinsic to this methodology have persisted. Studies revealed that CoMFA results can be extremely sensitive to a number of factors, such as alignment rules, overall orientation, lattice placement, step size, and probe atom type (40, 75, 112–114). The problem of three-dimensional alignment has been the most notorious among others. Even with the development of automated or semiautomated alignment protocols such as the Active Analog Approach (108, 115) or DISCO (116) and the opportunity to use, in some cases, the structural information about the target receptor (112, 117) to align molecules, in general there is no standard recipe as to how to align all molecules under consideration in a unique and unambiguous fashion. A QSAR analysis of 60 acetylcholinesterase inhibitors (117) is particularly illustrative with respect to this point. In that study, the combination of structure-based alignment and CoMFA was employed to obtain a QSAR model for 60 chemically diverse inhibitors of acetylcholinesterase (AChE). The great structural diversity of the AChE inhibitors, ranging from choline to decamethonium, made it practically impossible to structurally align all the inhibitors in any unbiased way and generate a unique three-dimensional pharmacophore. X-ray crystallographic analysis of AChE from *Torpedo californica* (EC 3.1.1.7) (118), followed by X-ray determination of the complexes of the enzyme with three structurally diverse inhibitors, tacrine, edrophonium, and decamethonium (119), provided crucial information with respect to the orientation of these inhibitors in the active site of the enzyme. The crystallographic data indicated that each of the three inhibitors had a unique binding orientation in the active site of the enzyme (Fig. 2.6). Their natural structural alignment would probably never have been predicted by any of the existing automated algorithms for ligand align-

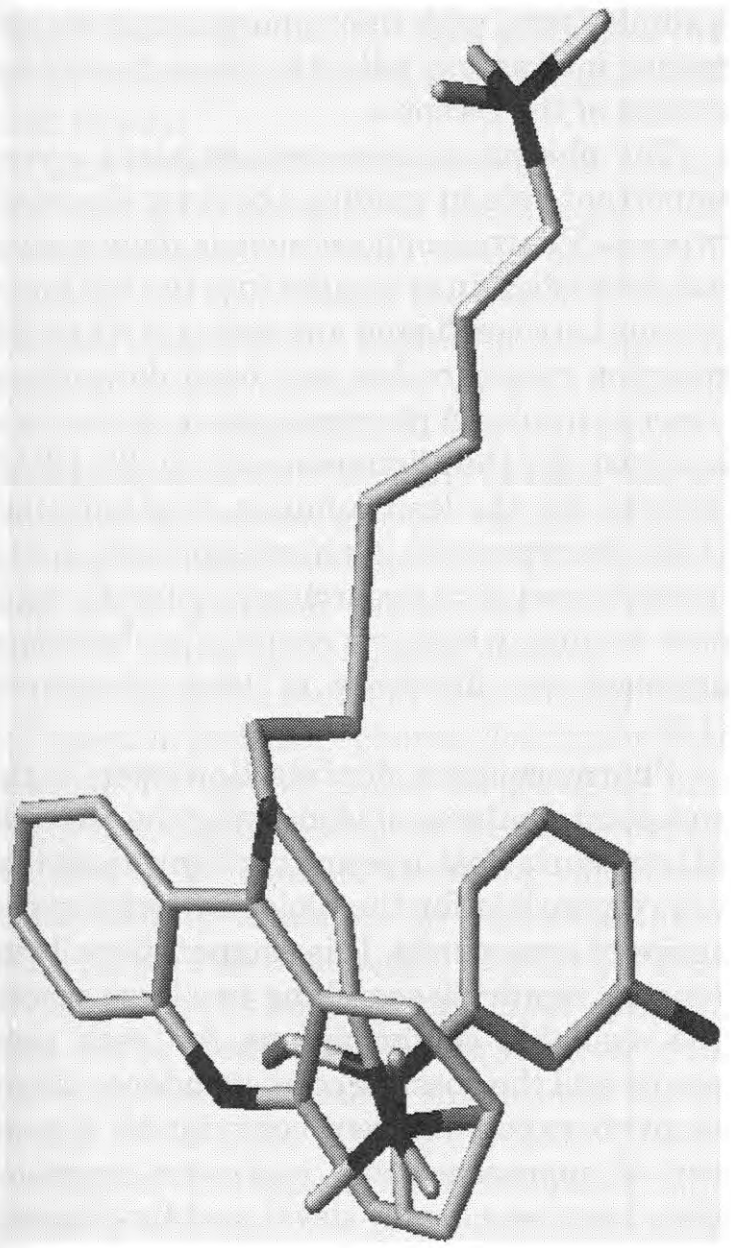


Figure 2.6. Superposition of three inhibitors of AChE in the active site of the enzyme based on crystallographic structures of enzyme-inhibitor complexes. Obviously, no common pharmacophore can be found for these molecules.

ment or even by the researcher's imagination on the basis of the ligand chemical structure alone. This consideration demonstrates the general difficulty of generating a unique and meaningful alignment in 3D-QSAR studies that leads to interpretable and predictive models.

The 3D alignment problem is the main source of ambiguity in obtaining and analyzing CoMFA results, especially in the case of structurally diverse compounds. However, it was also shown that, even if the structural alignment is fixed, the resulting q^2 value could also be sensitive to the orientation of rigidly aligned molecules on the user terminals (75), which can be explained as follows.

The grid orientation in CoMFA is fixed in the coordinate system of the computer; thus, every time when the orientation of the molecular aggregate is changed, the size of the grid may change but not its orientation. The orientation of the assembled molecules therefore affects the placement of probe atoms, which, in turn, influences the field sampling process. This leads to the variability of the q^2 values, mostly attributable to the reasons outlined earlier. The effect of variability of q^2 as a function of molecular aggregate orientation was more pronounced in the case of structurally diverse molecules (e.g., cephalotaxine esters and 5-HT_3 receptor ligands) than in the case of much less structurally diverse molecules (e.g., HIV protease inhibitors) (75). This effect may be attributed to the fact that the pattern of probe atom placement with respect to the aligned molecules changes more dramatically when one changes the orientation of more structurally diverse molecules than it does when the data set is composed of structurally similar molecules.

In the conventional CoMFA implementation, the steric and electrostatic fields, which theoretically form a continuum, are sampled on a fairly coarse grid. As a result, these fields are represented inadequately, and the results are not strictly reproducible. Intuitively, decreasing the grid spacing may increase the adequacy of sampling, as was suggested by Cramer et al. (120). Indeed, it was shown that decreasing the grid spacing from 2.0 to 1.0 Å minimized the fluctuation in the observed q^2 values (75). Most probably, the reason for this phenomenon is that the decrease in grid spacing increases the number of probe atoms, which in turn should raise the probability of placing the probe atoms in a region where the steric and electrostatic field changes can be best correlated with biological activity. However, as was noticed by Cramer et al. (120), the increase in the number of probe atoms also increases the noise in PLS analysis and leads to a less statistically significant q^2 (121).

An important feature of conventional CoMFA routine is that it assumes equal sampling and *a priori* equal importance of all lattice points for PLS analysis, whereas the final CoMFA result actually emphasizes the limited areas of three-dimensional space as important

for biological activity. Indeed, the deficiencies of conventional CoMFA routine mentioned earlier may be effectively dealt with by eliminating from the analyses those areas of three-dimensional space where changes in steric and electrostatic fields do not correlate with changes in biological activity. The q^2-GRS routine was devised (75) to eliminate those areas from the analysis based on the (low) value of the q^2 obtained for such regions individually. The major feature of this routine is that it optimizes the region selection for the final PLS analysis. In this regard, it is intellectually analogous to the GOLPE approach (74).

3D-QSAR remains an active area of research and method development. Several recent approaches such as COMSiA (45), QSiAR (46), and GRIND (122) address the most notorious CoMFA problems dealing with the grid artifacts. However, it should be kept in mind that 3D-QSAR modeling is a difficult process. It is reasonably successful when underlying molecules are relatively rigid and similar, so that the identification of the 3D pharmacophore is straightforward. With the increased complexity and flexibility of molecules and a possibility of multiple mechanisms of binding with the receptor, the derivation of unambiguous pharmacophore and unique alignment is sometimes practically impossible (as shown above in the case of AchE inhibitors), and extreme care is important in trying to obtain reproducible and validated QSAR models.

3.2 The Descriptor Pharmacophore Concept and Variable Selection QSAR

The termpharmacophore, introduced by Ehrlich in the early 1900s (123), was originally referred to the molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*) activity. Nowadays, this term has almost the opposite meaning as applied to three-dimensional (3D) molecular structure. A 3D pharmacophore is defined as a collection of particular chemical features (functional groups) and their spatial arrangement, which define pharmacological specificity of a series of compounds (124). The pharmacophore concept assumes that structurally diverse molecules bind to their receptor site in a similar way, with their pharmacophoric elements interacting with the same functional groups of the receptor.

The pharmacophore concept plays a very important role in guiding the drug discovery process. Pharmacophore models help medicinal chemists gain an insight into the key interactions between ligand and receptor when the receptor structure has not been determined experimentally. A pharmacophore can be used as a basis for the alignment rules in 3D-QSAR analysis for the lead compound optimization (125). Furthermore, a pharmacophore can be directly used as the search query for 3D database mining, which is a common and efficient approach for discovery of lead compounds (126).

Pharmacophore identification refers to the computational way of identifying the essential 3D structural features and configurations that are responsible for the biological activity of a series of compounds. It is computationally intensive, requiring searching two huge spaces: the available conformations for each compound and the possible correspondence (alignment) between different compounds. A number of approaches and computer programs have been specifically developed for pharmacophore identification including, for example, Active Analog Approach, AAA (108, 127, 128), Ensemble distance geometry (129), DISCO (116), Chem-X (130), Catalyst/Hypo (131, 132), Catalyst/HipHop (133, 134), and Apex-3D (135).

An obvious parallel can be established between the identification of descriptors contributing the most to the correlation with biological activity, and search for pharmacophoric elements, which are mainly responsible for the specificity of drug action. Indeed, individual pharmacophoric elements are typically identified in the course of experimental structure-activity studies. Considering molecules as a collection of substructures, pharmacophoric elements can also be viewed as specific chemical features selected from all chemical fragments present in a molecular data set. Thus, the selection of specific pharmacophoric features responsible for biological activity is directly analogous to the selection of specific chemical descriptors contributing to the most explanatory QSAR model. Frequently, the

QSAR modeling that involves descriptor (feature) selection is referred to as variable selection QSAR.

This consideration emphasizes the analogy between pharmacophore identification and variable selection QSAR. On the basis of this analogy, we now expand the notion of chemical pharmacophore to that of the more general descriptor pharmacophore. We shall define descriptor pharmacophore as a special subset of molecular descriptors (of any nature, not only chemical functional groups) optimized in the process of variable selection QSAR, to achieve the most significant correlation between descriptor values and biological activity.

Similar to the common areas of application of chemical pharmacophores, descriptor pharmacophores can be applied for database mining. First, a preconstructed QSAR model can be used as a means of screening compounds from existing databases (or virtual libraries) for high predicted biological activity. Alternatively, variables selected by QSAR optimization can be used for similarity searches to improve the performance of the rational library design or database mining methods. The advantage of this approach for database mining is that it affords not only the compound selection but also the quantitative prediction of their activity.

3.2.1 **Linear Models.** Variable selection approaches can be applied in combination with both linear and nonlinear optimization algorithms. Exhaustive analysis of all possible combinations of descriptor subsets to find a specific subset of variables that affords the best correlation with the target property is practically impossible because of the combinatorial nature of this problem. Thus, stochastic sampling approaches such as genetic or evolutionary algorithms (GA or EA) or simulated annealing (SA) are employed. To illustrate one such application we shall consider the GA-PLS method, which was implemented as follows (136).

Step 1. Multiple descriptors such as molecular connectivity indices or atom pair descriptors (cf. Section 2.1) are generated initially for every compound in a data set.

Step 2. An initial population of 100 different random combinations of subsets of these descriptors (parents) is generated as follows. Each parent is described by a string of random binary numbers (i.e., one or zero), with the length (total number of digits) equal to the total number of descriptors selected for each data set. The value of one in each string implied that the corresponding descriptor is included for the parent, and the value of zero implies that the descriptor is excluded.

Step 3. For every random combination of descriptors (i.e., every parent), a QSAR equation is generated for the training data set by use of the PLS algorithm (41). Thus, for each parent a q^2 value is obtained, and some function of q^2 is used as a fitness function to guide GA.

Step 4. Two parents are selected randomly and subjected to a crossover (i.e., the exchange of the equal length substrings), which produces two offspring. Each offspring is subjected to a random single-point mutation, that is, a randomly selected one (or zero) is changed to zero (or one) and the fitness of each offspring is evaluated as described above (cf. Step 3).

Step 5. If the resulting offspring are characterized by a higher value of the fitness function, then they replaced parents; otherwise, the parents are kept.

Step 6. Steps *3–5* are repeated until a predefined convergence criterion is achieved. For the convergence criterion one can use the difference between the maximum and minimum values of the fitness function. Calculations are terminated when this difference falls below a certain threshold (e.g., 0.02).

In summary, each parent in this method represents a QSAR equation with randomly chosen variables, and the purpose of the calculation is to evolve from the initial population of the QSAR equations to the population with the highest average value of the fitness function. In the course of the GA-PLS process, the initial number of members of the population (100) is maintained while the average value of the fitness function for the whole population converges to a high number. The best model is characterized by the highest value of the fitness function as well as by specific descriptor selection (descriptor pharmacophore) that affords such a model.

3.2.2 Nonlinear Models. Most of the QSAR approaches assume the existence of a linear relationship between a biological activity and molecular descriptors. However, the fast collection of structural and biological data, as a consequence of the recent development of combinatorial chemistry and high throughput screening technologies, has challenged traditional QSAR techniques. First, 3D methods may be computationally too expensive for the analysis of a large volume of data'; and in some cases, an automated and unambiguous alignment of molecular structures is not achievable. Second, although existing 2D techniques are computationally efficient, the assumption of linearity in the SAR may not hold true, especially when a large number of structurally diverse molecules are included in the analysis.

These considerations provide an impetus for the development of fast, nonlinear, variable selection QSAR methods that can avoid the aforementioned problems of linear QSAR. Several nonlinear QSAR methods have been proposed in recent years. Most of these methods are based on either artificial neural network (ANN) (50, 61, 137–142) or machine learning techniques (65,143–145). Given that optimization of many parameters is involved in these techniques, the speed of the analysis is relatively slow. More recently, Hirst reported a simple and fast nonlinear QSAR method (146), in which the activity surface was generated from the activities of training set compounds based on some predefined mathematical function.

For illustration. we shall consider here one of the nonlinear variable selection methods that adopts a k-Nearest Neighbor (kNN) principle to QSAR [kNN-QSAR (49)]. Formally, this method implements the active analog principle that lies in the foundation of the modern medicinal chemistry. The kNN-QSAR method employs multiple topological (2D) or topographical (3D) descriptors of chemical structures and predicts biological activity of any compound as the average activity of k most similar molecules. This method can be used to analyze the structure-activity relationships (SAR) of a large number of compounds where a nonlinear SAR may predominate.

In principle, the kNN technique is a conceptually simple, nonlinear approach to pattern-recognition problems (147). In this method, an unknown pattern is classified according to the majority of the class labels of its k nearest neighbors of the training set in the descriptor space. Many variations of the kNN method have been proposed in the past and new and fast algorithms have continued to appear in recent years (148, 149). The applications of the kNN principle in chemistry have been summarized by Strouf (150). In the area of biology, Raymer et al. have successfully applied a kNN pattern-recognition technique with simultaneous feature selection and classification in the analysis of water distribution in protein structures (151). In the area of QSPR, Basak et al. have applied this principle, combined with principal component analysis and graph theoretical indices, in the estimation of physicochemical properties of organic compounds (152–155).

The assumptions underlying the kNN-QSAR method are as follows. First, structurally similar compounds should have similar biological activities, and the activity of a compound can be predicted (or estimated) simply as the average of the activities of similar compounds. Second, the perception of structural similarity is relative and should always be considered in the context of a particular biological target. Given that the physicochemical characteristics of the receptor-binding site vary from one target to another, the structural features that can best explain the observed biological similarities between compounds are different for different biological endpoints. These critical structural features can be defined as the descriptor pharmacophore (DP) for the underlying biological activity. Thus, one of the tasks of building a kNN-QSAR model is to identify the best DP. This is achieved by the "bioactivity-driven" variable selection, that is, by selecting a subset of molecular descriptors that afford a highly predictive kNN-QSAR model. Because the number of all possible combinations of descriptors is huge, an exhaustive search of these combinations is not possible. Thus, a stochastic optimization algorithm (i.e., simulated annealing) has been adopted for an efficient sampling of the combinatorial space. Fig-

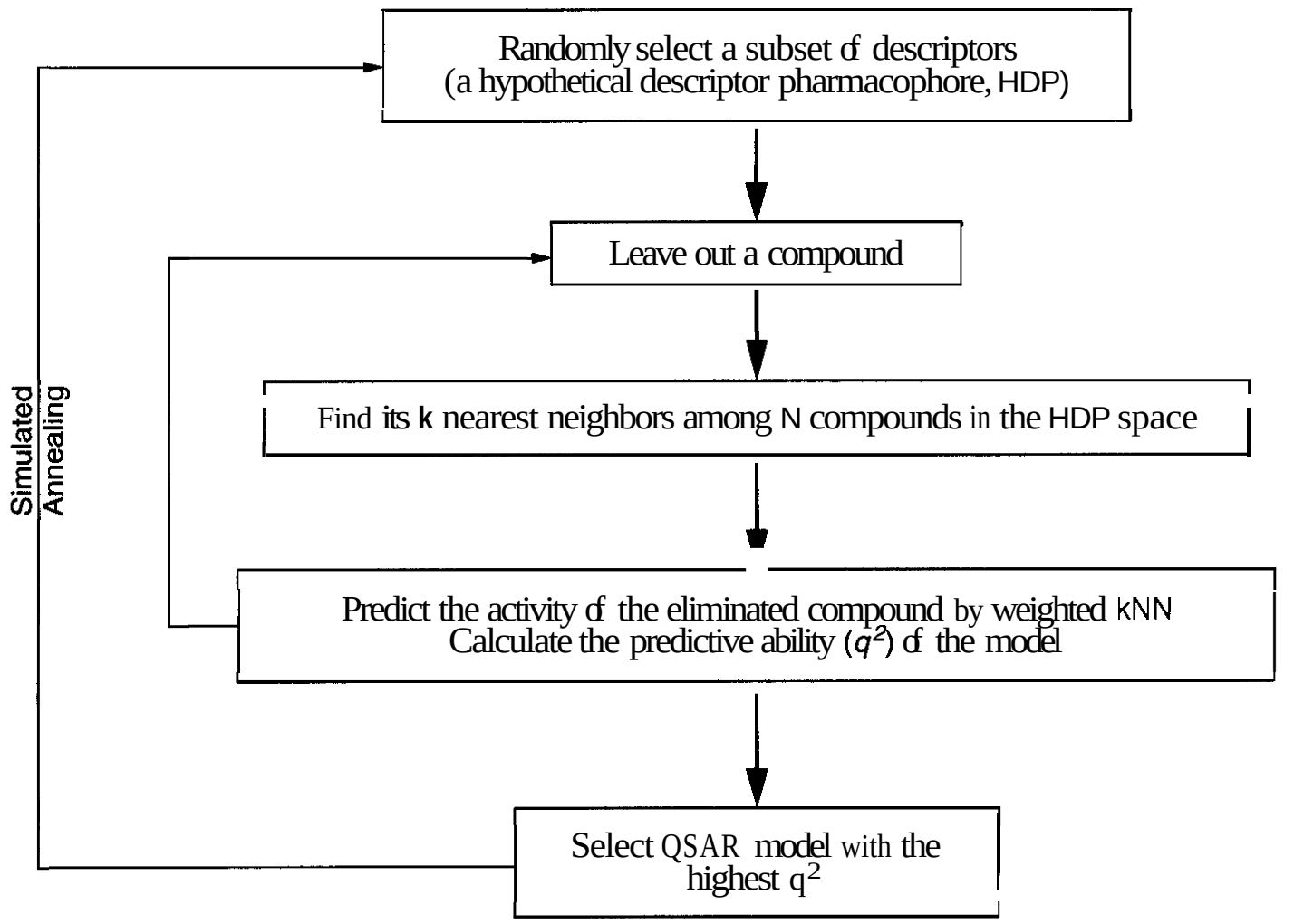


Figure 2.7. Flowchart of the kNN method (49).

ure 2.7 shows the overall flowchart of the kNN-QSAR method, which involves the following steps.

1. Select a subset of n descriptors randomly (n is a number between 1 and the total number of available descriptors) as a hypothetical descriptor pharmacophore (HDP).
2. Validate this HDP by a standard cross-validation procedure, which generates the cross-validated R^2 (or q^2) value for the kNN-QSAR model built by use of this HDP. The standard leave-one-out procedure has been implemented as follows: (*i*) Eliminate a compound from the training set. (*ii*) Calculate the activity of the eliminated compound, which is treated as an unknown, as the average activity of the k most similar compounds found in the remaining molecules (k is set to 1 initially). The similarities between compounds are calculated using only the selected descriptors (i.e., the current trial HDP) instead of the whole set of descriptors. (*iii*) Repeat this procedure until every compound in the training set has been eliminated and predicted once. (*iv*) Calculate the cross-validated R^2 (or q^2) value (cf. Equation 2.1). (*v*) Repeat calculations for $k = 2, 3, 4, \ldots,$ n. The upper limit of k is the total number of compounds in the data set; however, the best value is found empirically between 1 and 5. The k that leads to the best q^2 value is chosen for the current kNN-QSAR model.
3. Repeat steps 1 and 2, the procedure of generating trial HTPs and calculating corresponding q^2 values. The goal is to find the best HTP that maximizes the q^2 value of the corresponding kNN-QSAR model. This process is driven by a generalized simulated annealing by use of q^2 as the objective function.

4 VALIDATION OF QSAR MODELS

One of the most important characteristics of QSAR models is their predictive power. The latter can be defined as the ability of a model to predict accurately the target property (e.g., biological activity) of compounds that were not used for model development. The typical problem of QSAR modeling is that at the time of

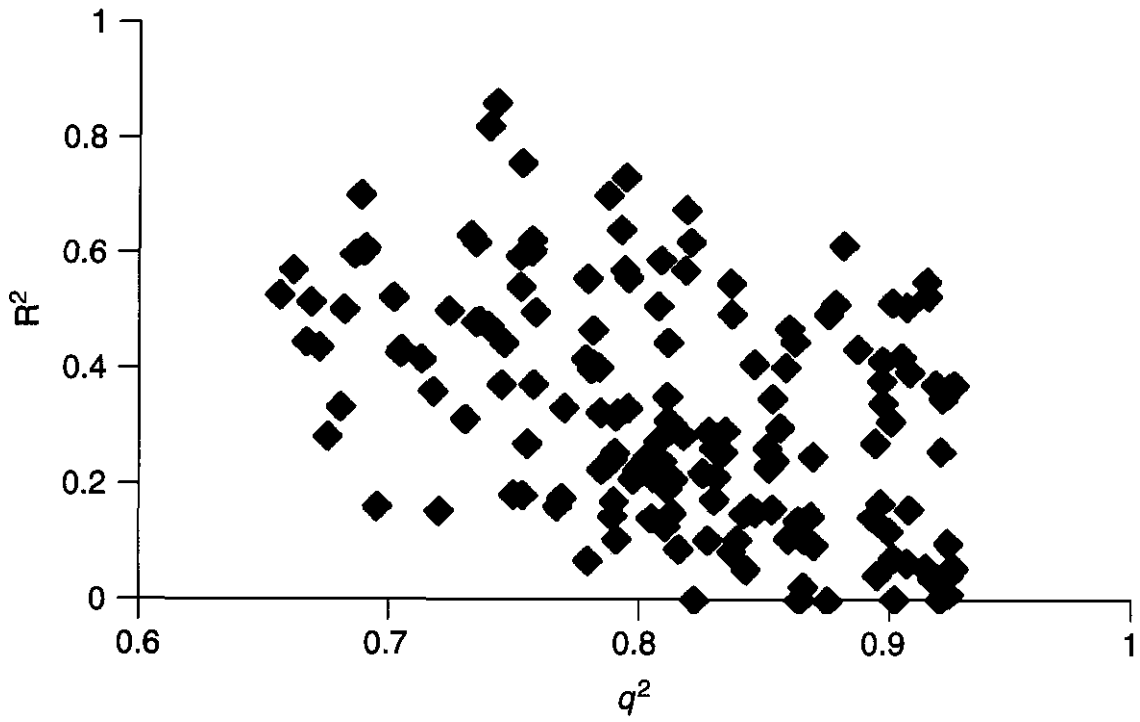


Figure 2.8. Beware of q^2! External R^2 (for the test set) presents no correlation with the "predictive" LOO $q2$ (for the training set). (Adopted from Ref. 163.)

model development a researcher has, essentially, only training set molecules, so predictive ability can be characterized only by statistical characteristics of the training set model, and not by true external validation. Recent research demonstrates that external validation must be made, indeed, a mandatory part of model development. This goal can be achieved by a division of an experimental SAR data set into the training and test sets, which are used for model development and validation, respectively.

It has been shown that the more independent variables are involved in MLR QSAR analysis, the higher the probability of a chance correlation between predicted and observed activities, even if only a small portion of variables is included in the final QSAR equation (16). This conclusion is true not only for MLR QSAR, but also for any QSAR approach when the number of variables (descriptors) is comparable to or higher than the number of compounds in a data set. Thus, model validation is one of the most important aspects of QSAR analysis.

4.1 Beware of q^2

To validate a QSAR model, most of researchers apply the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedures. The outcome from this procedure is a cross-validated correlation coefficient R^2 (q^2) (Equation 2.1). Frequently, q^2 is used as a criterion of both robustness and predictive ability of the model. Many authors consider high q^2 (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of the QSAR model. They do not test the models for their ability to predict the activity of compounds of an external test set (i.e., compounds that have not been used in the QSAR model development). There are several examples of recent publications, in which the authors claim that their models have high predictive ability without validating them by use of an external test set (156–160). Some authors validate their models by the use of only one or two compounds that were not used in QSAR model development (161, 162) and still claim that their models are highly predictive. In contrast with such expectations, it has been shown that if a test set with known values of biological activities is available for prediction, there exists no correlation between LOO cross-validated q^2 and correlation coefficient R^2 between the predicted and observed activities for the test set [Fig. 2.8; (46, 163)].

4.2 Rational Selection of Training and Test Sets

As discussed earlier, to obtain a reliable (validated) QSAR model, an available data set

should be divided into the training and test sets. Ideally, this division must be performed such that points representing both training and test set are distributed within the whole descriptor space occupied by the entire data set, and each point of the test set is close to at least one point of the training set. This approach ensures that the similarity principle can be employed for the activity prediction of the test set. Unfortunately, as we shall see below, this condition cannot always be satisfied.

Many authors use external test sets for validation of QSAR models, but do not provide any rationale as to how and why certain compounds were chosen for the test set (164,165). One of the most widely used methods for dividing a data set into training and test sets is a mere random selection (166, 167). Some authors assign whole structural subgroups of molecules to the training set or the test set (168,169). Another frequently used approach is based on the activity sampling. The whole range of activities is divided into bins, and compounds belonging to each bin are randomly (or in some regular way) assigned to the training set or test set (170,171). These methods (166,170,171) cannot guarantee that the training set compounds represent the entire descriptor space of the original data set, and that each compound point of the test set is close to at least one point of the training set.

In several publications, the division of a data set into training and test sets is performed by use of the Kohonen's Self-Organizing Map (SOM) (172). Representative points falling into the same areas of the SOM are randomly selected for the training and test sets (173, 174). SOM preserves the closeness between points (points that are close to each other in the multidimensional descriptor space are close to each other on the map). Therefore, it is anticipated that the training and test sets must be scattered within the whole area occupied by representative points in the original descriptor space, and that each point of the test set is close to at least one point of the training set. The drawback of this method is that the quantitative methods of prediction use exact values of distances between representative points; because SOM is a nonlinear projection method, the distances between points in the map are distorted.

The division of a data set into the training and test sets can be performed by the use of various clustering techniques. In Burden and Winkler (175) and Burden et al. (176) the *K*-means clustering algorithm (177) was used, and from each cluster one compound for the training set was randomly selected. In Potter and Matter (178), to select a representative subset from a data set, hierarchical clustering and the maximum dissimilarity method (179–181) were used. The authors showed that both methods choose representative subsets of compounds much better than the random selection. Compounds selected through use of the maximum dissimilarity method were used as training sets in 3D-QSAR studies, with all remaining compounds composing the test set. In Wu et al. (166) the Kennard-Stone (182–184) method, which is similar to the maximum dissimilarity method, was applied to the classification of NIR spectra and QSAR analysis. The drawbacks of clustering methods are that different clusters contain different numbers of points and have different densities of representative points. Therefore, the closeness of each point of the test set to at least one point of the training set is not guaranteed. The maximum dissimilarity and Kennard-Stone methods guarantee that the points of the training set are distributed more or less evenly within the whole area occupied by representative points, and the condition of closeness of the test set points to the training set points is satisfied. The maximum distance between training and test set points in these methods does not exceed the radius of the probe sphere.

To select a representative subset of samples from the whole data set, factorial designs (185, 186) and D-optimal designs (187) were used (166, 173, 188). Factorial designs presume that different sample properties (such as substituent groups at certain positions) are divided into groups. The training set includes one representative for each combination of properties. For a diverse data set this approach is impractical, and fractional factorial designs are used, in which only a part of all combinations is included into the training set. Generally, this approach does not guarantee the closeness of the test set points to the training set points in the descriptor space. *D*-optimal design algorithms select samples that

maximize the $|\mathbf{X}'\mathbf{X}|$ determinant, where **X** is the information (variance-covariance)matrix of independent variables (descriptors) (189, 190). The points maximizing the $|\mathbf{X}'\mathbf{X}|$ determinant are spanned across the whole area occupied by representative points. They can be used as a training set, and the points not selected then are used as the test set (166,173).

In Wu et al. (166) four methods of sample selection (random, SOM, Kennard-Stone design, and D-optimal design) were compared. The best models were built when Kennard-Stone and D-optimal designs were used. SOM was better than random selection, and *D*-optimal design was slightly better than the random selection.

4.3 Guiding Principles of Safe QSAR

A widely used approach to establish the model robustness is so-called y-randomization (randomization of response, i.e., in our case, activities) (191). It consists of repeating the calculation procedure with randomized activities and subsequent probability assessment of the resultant statistics. Frequently, it is used along with cross-validation. It is expected that models obtained for the data set with randomized activity should have low values of q^2; otherwise, the original model should be considered insignificant. We suggest that the y-randomization test is a mandatory component of model validation.

Several authors have suggested that the only way to estimate the true predictive power of a QSAR model is to compare the predicted and observed activities of an (sufficiently large) external test set of compounds that were not used in the model development (46, 163, 192–194). To estimate the predictive power of a QSAR model, we recommended use of the following statistical characteristics of the test set (163): (*i*) correlation coefficient R between the predicted and observed activities; (*ii*) coefficients of determination (195) (predicted vs. observed activities R_0^2, and observed vs. predicted activities $R_0'^2$); (*iii*) slopes k and k' of the regression lines through the origin. We consider a QSAR model predictive, if the following conditions are satisfied (163):

$$q^2 > 0.5 \qquad (2.2)$$

$$R^2 > 0.6 \qquad (2.3)$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \qquad (2.4)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \qquad (2.5)$$

The lack of the correlation between q^2 and R^2 was noted in Kubinyi et al. (46), Novellino et al. (192), Norinder (193), and in our recent publication (163), where we demonstrated that all of the above-mentioned criteria are necessary to adequately assess the predictive ability of a QSAR model. We suggest (163) that the external test set must contain at least five compounds, representing the whole range of both descriptor and activities of compounds included into the training set.

5 QSAR MODELS AS VIRTUAL SCREENING TOOLS

5.1 Data Mining and SAR Analysis

Data mining has been of interest to researchers in machine learning, pattern recognition, artificial intelligence, database statistics, and so forth for many years, and widely applied in science, business, and government. Now, chemoinformatitians have also started to plunge into this field because of the increased quantity of data in the drug discovery process. Data mining can be defined as the process of discovering valid, novel, understandable, and potentially useful patterns in data (196, 197). Data mining is an interactive and iterative, multiple-step process, involving the decisions made by the user. It may include data collection, data cleaning, data engineering, algorithm engineering, algorithm running, result evaluation, and knowledge utilization (198, 199).

Data mining methods can be generally divided into two types, unsupervised and supervised. Whereas unsupervised methods seek informative patterns, which directly display the interesting relationship among the data, supervised methods discoverpredictive patterns, which can be used later to predict one or more attributes from the rest.

A wide variety of supervised data mining methods have been applied for analyzing

structure-activity data sets, besides the traditional linear regression methods. Most of them are nonlinear and nonparametric and need no statistical assumptions to apply them. Decision tree and rule induction methods, such as ID3 (200), CART (201), and FIRM (202–204) usually use univariate splits to generate a model in the form of a tree or propositional logic. The inferred model is easy to comprehend, but the approximation power may be significantly restricted by a particular tree or rule representation. Inductive logic programmingmethods, such as GOLEM (64) and PROGOL (65), are designed to induce a model from the more flexible representation of first-order predicate logic. However, this generality comes at the price of significant computational demands. Nonlinear regression and classification methods, such as various neural networks (60–62), train a model by fitting linear and nonlinear combinations of basis functions to the combinations of the input variables. They may be powerful in terms of approximation, but they are statistically poorly characterized, slow (205), and difficult to interpret in chemical terms. Example-based methods, such as nearest-neighbor methods (147), use representative examples from the database as an approximate model and predicate new samples on the basis of the properties of the most similar examples in the model. They are asymptotically powerful for approximating properties, but also difficult to interpret. Furthermore, their performance is strongly dependent on a well-defined distance metric to evaluate distances between data points.

Data mining of chemical databases is still at its very early stage. Nevertheless, as a result of the data explosion in pharmaceutical industry, it is expected that data mining techniques will play an increasingly important role in the drug discovery process. Future studies may include, for example, the definition of chemical space, the validation of various algorithms (206), and the representation of extremely large virtual databases (207).

5.2 Virtual Screening

Although combinatorial chemistry and HTS have offered medicinal chemists a much broader range of possibilities for lead discovery and optimization, the number of chemical compounds that can be reasonably synthesized, which is sometimes called "virtual chemistry space," is still far beyond today's capability of chemical synthesis and biological assay. Therefore, medicinal chemists continue to face the same problem as before: Which compounds should be chosen for the next round of synthesis and testing? For chemoinformatitians, the task is to develop and utilize various computer programs to evaluate a very large number of chemical compounds and recommend the most promising ones for bench medicinal chemists. This process can be called virtual screening (208) or chemical database searching. A large number of computational methods exist for virtual screening, but which one is chosen will depend on the information available and the task at hand in practice.

A substructure search will typically be undertaken if a lead compound has been found. The search query will retrieve all the structures in a database that contain the substructures present in the lead compound that are believed to be important for activity (209). According to graph theory, it is equivalent to searching a series of topological graphs for the existence of a subgraph isomorphism with a specified query graph. Subgraph isomorphism is an NP-complete problem (210), which means that for it, there are no algorithms whose worst-case time requirements do not rise exponentially with the size of the input. However, various backtracking algorithms (211–213) and partitioning algorithms (214–217) have been developed since the 1950s, to reduce the average time required for chemical substructure searching. Today, almost all the chemical database software includes the function of substructure searching.

A similarity search provides a way forward by retrieving the structures that are similar, but not identical, to a lead compound (94). Therefore, it overcomes some limitations of substructure search, for example, not requiring specific knowledge about the substructures responsible for activity, and being able to rank the output structures according to the overall similarity. The search query usually involves a set of descriptors that collectively specify the whole structure of the lead compound. This set of descriptors is compared with the corresponding set of descriptors for

each compound in the database, and then a measure of similarity is calculated between them. There are a wide variety of molecular descriptors for similarity searching (cf. Section 2). Not a single set of molecular descriptors has been found as the best choice in all the cases. The present trend in descriptor selection is to use combined descriptors with many different types. The similarity coefficients that are often used for measuring the similarity between two structures includes Manhattan distance, Euclidean distance, Soergel distance, Tanimoto coefficient, Dice coefficient, Cosine coefficient, and so forth (218), and again no clear-cut winner has been found among them (219). Virtual screening based on QSAR models can serve as a powerful approach to the design of targeted chemical libraries, as illustrated in the following section.

5.3 Rational Library Design by use of QSAR

As discussed earlier, combinatorial chemical synthesis and high throughput screening have significantly increased the speed of the drug discovery process (220–222). However, it remains impossible to synthesize all of the library compounds in a reasonably short period of time. For instance, 3000^3 (2.7×10^{10}) compounds can be synthesized from a molecular scaffold with three different substitution positions when each of the positions has 3000 different substituents. If a chemist could synthesize 1000 compounds per week, 27 million weeks (–0.5 million years) would be required to synthesize all these compounds. Furthermore, many of these compounds can be structurally similar to each other, thus making redundant the chemical information contained in the library. There is a need for rational library design (i.e., rational selection of a subset of available building blocks for combinatorial chemical synthesis), so that a maximum amount of information can be obtained while a minimum number of compounds are synthesized and tested. Similarly, there is a closely related task in computational database mining, that is, rational selection of a subset of compounds from commercially available or proprietary databases for biological testing.

Thus, in many practical cases, the exhaustive synthesis and evaluation of combinatorial libraries is prohibitively expensive, time-consuming, or redundant (223). Modern rational approaches to the design of combinatorial libraries have been explored in a recent monograph (224). Theoretical analysis of available experimental information about the biological target or pharmacological compounds capable of interacting with the target can significantly enhance the rational design of targeted chemical libraries. In many cases, the number of compounds with known biological activity is sufficiently large to develop viable QSAR models for such data sets. These models can be used as a means of selecting virtual library compounds (or actual compounds from existing databases) with (high) predicted biological activity. Alternatively, if a variable selection method has been employed in developing a QSAR model, the use of only selected variables can improve the performance of the rational library design or database mining methods on the basis of the similarity to a probe. This procedure of use of only selected variables in a similarity search in the descriptor space is analogous to more traditional use of conventional chemical pharmacophores in database mining.

QSAR models can be employed for rational design of targeted chemical libraries and database mining by predicting biologically active structures in virtual or actual chemical libraries (225–227). To illustrate this approach, we consider the design of a pentapeptide combinatorial library with the bradykinin activity by use of a QSAR model derived for a small bradykinin peptide data set. Figure 2.9 shows the schematic diagram illustrating the targeted pentapeptide combinatorial library design by use of the FOCUS-2D method (225, 226). The algorithm includes the description, evaluation, and optimization steps.

To identify potentially active compounds in the virtual library, FOCUS-2D employs stochastic optimization methods such as SA (228, 229) and GA (230–232). The latter algorithm was used for targeted pentapeptide library design as follows. Initially, a population of 100 peptides is randomly generated and encoded by use of topological indices or amino acid-dependent physicochemical descriptors. The fitness of each peptide is evaluated by its biological activity predicted from a preconstructed QSAR equation (see below). Two par-

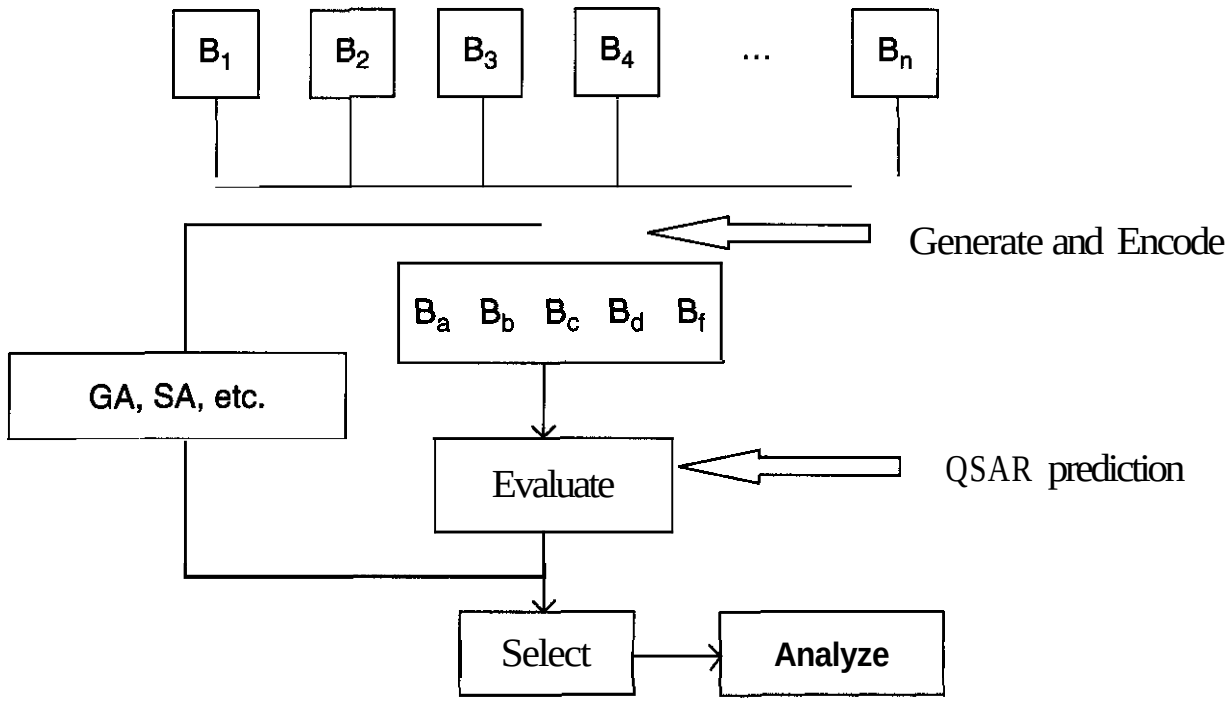


Figure 2.9. Flowchart of the library design approach by FOCUS-2D.

ent peptides are chosen by use of the roulette wheel selection method (i.e., high fitting parents are more likely to be selected). Two offspring peptides are generated by a crossover (i.e., two randomly chosen peptides exchange their fragments) and mutations (i.e., a randomly chosen amino acid in an offspring is changed to any of 19 remaining amino acids). The fitness of the offspring peptides is then evaluated and compared with that of the parent peptides, and the two lowest scoring peptides are eliminated. This process is repeated for 2000 times to evolve the population.

Design of a Targeted Library with Bradykinin (BK) Potentiating Activity. The results obtained with the FOCUS-2D and a QSAR-based prediction are shown in Figure 2.10. The position-dependent frequency distributions of amino acids in the highest scoring pentapepeptides are shown before (Fig. 2.10a) and after (Fig. 2.10b,c) FOCUS-2D. To evaluate the efficiency of stochastic sampling, the entire pentapeptide library (which includes as many as 3.2 million molecules) was also generated and subjected to evaluation by use of the same QSAR model, and the results are shown in Fig. 2.10c. Apparently, the results after FOCUS-2D and the exhaustive search were very similar to each other. FOCUS-2D selected the following amino acids: E, I, K, L, M, Q, R, V, and W. Interestingly, these selected amino acids included most of those found in the two experimentally most active pentapeptides, VEWAK and VKWAP (excluded from the training set for the QSAR model development). Furthermore, the actual spatial positions of these amino acids were correctly identified: the first and fourth positions for V; the second and fifth positions for E; the third position for W; and the second and fifth positions for K. More detailed analysis of these results (cf. Fig. 2.10b,c) may suggest which residues should be preferably chosen for each position in the pentapeptide to achieve a limited size library with high predicted bradykinin activity.

6 CONCLUSIONS

In this chapter, we have reviewed recent and developing trends in the field of QSAR. We have provided common terminology and presented a unified concept of the QSAR approach. We have emphasized that, regardless of the origin of molecular descriptors, any QSAR modeling exercise starts from constructing a two-dimensional data array (Fig. 2.2), which lists molecular IDs, values of the target (or dependent) property of each compound, and values of descriptors (independent variables) for each compound. We have considered various protocols employed by QSAR practitioners to develop quantitative models of biological activity by the use of chemical descriptors and linear or nonlinear optimiza-

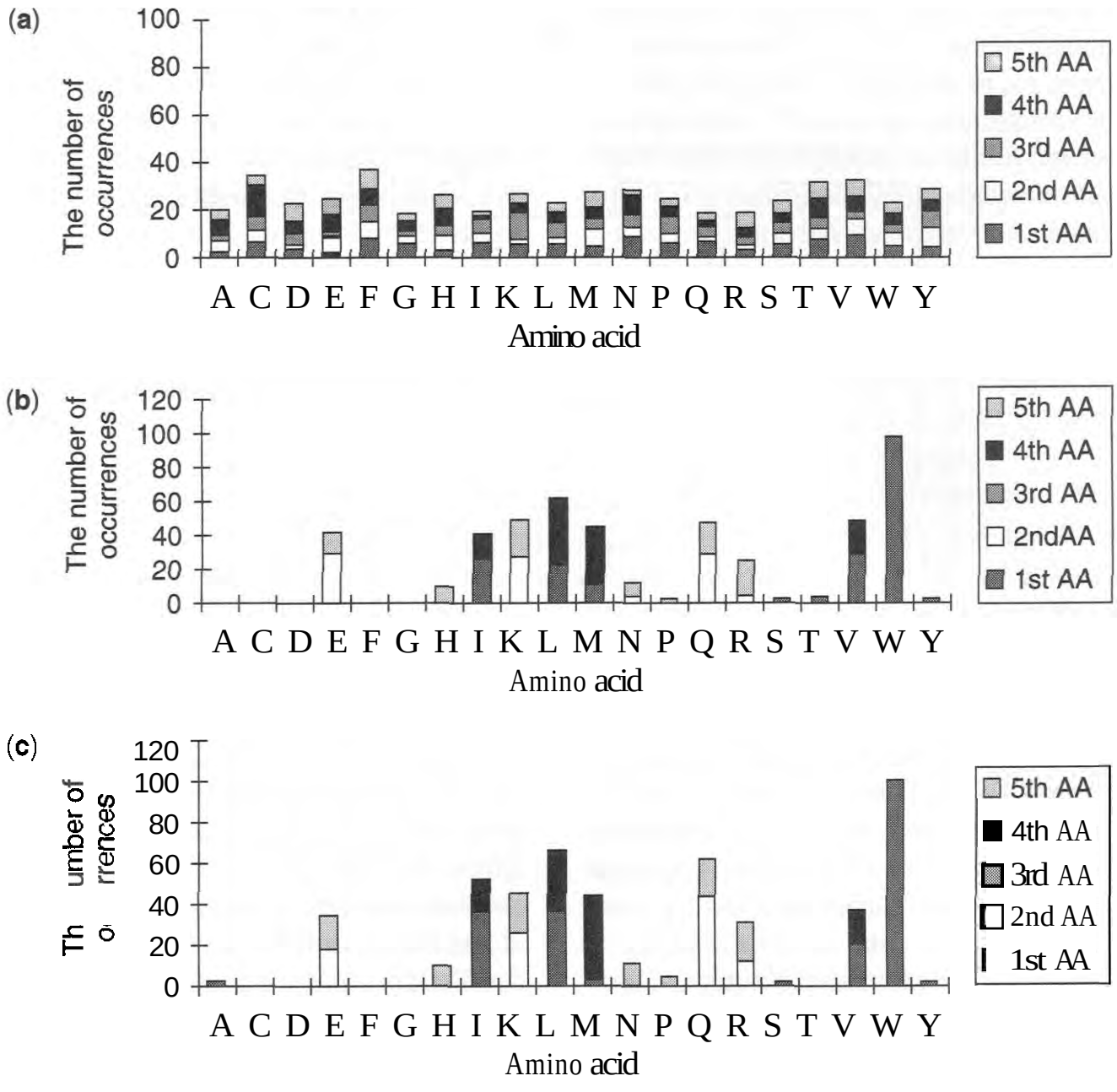


Figure 2.10. Ratonal selection of building blocks for library design by use of FOCUS-2D and a QSAR model for activity prediction: (a) initial population; (b) final population after FOCUS-2D; and (c) final population after the exhaustive search.

tion techniques. We have particularly emphasized that the true power of any QSAR model comes from its statistical significance and the model's ability to predict accurately biological properties of chemical compounds both in the training and, most important, in the test sets. One of the important research challenges in the QSPR modeling remains finding descriptor types, correlation approaches, and adequate statistical characteristics of the training set only, which may ensure high predictive power of the models.

In conclusion, we strongly advocate rigorous validation of QSAR models before their practical application or interpretation. The practical guidelines for the development of statistically robust and predictive QSAR models can be summarized as follows:

1. Establish an SAR database through the use of reliable quantitative measurements of the target property and a preferred set of molecular descriptors.
2. Divide the underlying data set into training and test sets through the use of diversity sampling algorithms.
3. Develop training set models through the use of available QSAR methods or commercial software. Characterize these models with internal validation parameters, as discussed in this chapter, and define the applicability domain for each model.
4. Validate training set models through the use of an external test set and calculate the external validation parameters, as discussed in this chapter. Ideally, repeat the

procedure of training and test selection and external validation several times to identify the QSAR model for the smallest training set that affords adequate prediction power for the biggest test set.

5. Finally, explore and exploit validated QSAR models for possible mechanistic interpretation and prediction.

In the modern age of medicinal chemistry, QSAR modeling remains one of the most important instruments of computer-aided drug design. Skillful application of various methodologies discussed in this chapter will afford validated QSAR models, which should continue to enrich and facilitate the experimental process of drug discovery and development.

REFERENCES

1. C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, E. Geiger, and M. Streich, *J. Am. Chem. Soc.*, **85, 2817(1963).**
2. T. Fujita, J. Iwasa, and C. Hansch. *J. Am. Chem. Soc.*, **86, 5175(1964).**
3. L. P. Hammett, *Chem. Rev.*, **17, 125(1935).**
4. C. Hansch and A. Leo in S. R. Heller, Ed., ***Exploring QSAR: Fundamentals and Applications in Chemistry and Biology,*** American Chemical Society, Washington, DC, **1995.**
5. C. Hansch, A. Leo, and D. Hoekman in S. R. Heller, Ed., ***Exploring QSAR: Hydrophobic, Electronic, and Steric Constants.*** American Chemical Society, Washington, DC, **1995.**
6. A. Verloop, W. Hoogenstraaten, and J. Tipker in E. J. Ariens, Ed., Drug Design, Vol. VII, Academic Press, New York, **1976, 165** pp.
7. C. Selassie, this volume, Chapter **1.**
8. H. Kubinyi in R. Mannhold, P. Krogsgaard-Larsen, and H. Timmerman, Eds., ***Methods and Principles in Medicinal Chemistry,*** Vol. **1,** VCH, New York, **1993.**
9. D. Livingstone, ***Data Analysis for Chemists: Applications to QSAR and Chemical Product Design,*** Oxford University Press, Oxford, UK, **1995.**
10. H. Kubinyi, G. Folkers, and Y. Martyn, Eds., **3D** QSAR in Drug Design, Vols. **2** and **3,** Kluwer/ESCOM, Dordrecht, The Netherlands, **1998.**
11. M. Karelson, ***Molecular Descriptors in QSAR/QSPR,*** Wiley-Interscience, New York, **2000.**
12. D. J. Livingstone, *J. Chem. Inf. Comput. Sci.*, **40, 195-209(2000).**
13. Chemical Abstracts Service (CAS), Columbus, OH. May be accessed at http://www.cas.org
14. D. S. Tan, M. A. Foley, M. D. Shair, and S. L. Schreiber, *J. Am. Chem. Soc.*, **120, 8565-8566** (1998).
15. J. Drews, ***Science,*** **287, 1960-1964(2000).**
16. J. G. Topliss and R. P. Edwards, *J. Med. Chem.*, **22, 1238** (1979).
17. U. Burkert and N. L. Allinger, ***Molecular Mechanics,*** American Chemical Society, Washington, DC, **1982.**
18. C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, **86, 1616-1626(1964).**
19. M. Randíc, *J. Am. Chem. Soc.*, **97, 6609-6615 (1975).**
20. L. B. Kier and L. H. Hall, ***Molecular Connectivity in Chemistry and Drug Research,*** Academic Press, New York, **1976.**
21. L. B. Kier and L. H. Hall, ***Molecular Connectivity in Structure-Activity Analysis,*** Research Studies Press, Chichester, UK, **1986.**
22. L. B. Kier and L. H. Hall in K. B. Lipkowitz and D. B. Boyd, Eds., ***Reviews in Computational Chemistry*** *ZZ,* VCH, Weinheim/New York, **1991,** pp. **367-422.**
23. L. B. Kier, ***Quant. Struct.-Act. Relat.,*** **4,** 109–**116 (1985).**
24. L. B. Kier, ***Quant. Struct-Act. Relat.,*** **6, 8-12** (1987).
25. L. H. Hall and L. B. Kier, ***Quant. Strut.-Act. Relat.,*** **9,** 115–131 **(1990).**
26. L. H. Hall, B. K. Mohney, and L. B. Kier, *Quant. Struct.-Act. Relat.,* **10,** 43–51 **(1991).**
27. L. H. Hall, B. K. Mohney, and L. B. Kier, *J. Chem. Inf. Comput. Sci.,* **31,** 76–82 (1991).
28. L. B. Kier and L. H. Hall, ***Molecular Structure Description: The Electrotopological State,*** Academic Press, Orlando, FL, **1999.**
29. G. E. Kellogg, L. B. Kier, P. Gaillard, and L. H. Hall, *J. Cornput.-Aided Mol. Des.,* **10, 513-520** (1996).
30. R. P. Sheridan, R. B. Nachbar, and B. L. Bush, *J. Cornput.-Aided Mol. Des.,* **8, 323-340(1994).**
31. H. Matter, *J. Med. Chem.,* **40, 1219-1229** (1997).
32. A. J. Hopfinger, *J. Am. Chem. Soc.,* **102, 7196 (1980).**
33. G. M. Crippen, *J. Med. Chem.,* **22, 988-997 (1979).**

34. G. M. Crippen, *J. Med. Chem.*, 23, 599–606 (1980).
35. L. G. Boulu and G. M. Crippen, *J. Comput. Chem.*, 10,673 (1989).
36. U. Holzbrabe and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 36, 1018 (1996).
37. A. J. Hopfinger, B. J. Burke, and W. J. Dunn, *J. Med. Chem.*, 37,3768 (1994).
38. S. Srivastava and G. M. Crippen, *J. Med. Chem.*, 36,3572 (1993).
39. M. P. Bradley and G. M. Crippen, *J. Med. Chem.*, 36,3171 (1993).
40. R. D. Cramer III, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.*, 110, 5959–5967 (1988).
41. S. Wold, A. Ruhe, H. Wold, and W. J. Dunn III, *SZAM J. Sci. Stat. Comput.*, 5, 735–743 (1984).
42. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 185, 1–17 (1986).
43. G. R. Marshall and R. D. Cramer III, *Trends Pharmacol. Sci.*, 9, 285–289 (1988).
44. C. Pérez, M. Pastor, A. R. Ortiz, and F. Gago, *J. Med. Chem.*, 41,836–852 (1998).
45. G. Klebe in H. Kubinyi, G. Folkers, and Y. C. Martin, Eds., 3D QSAR in Drug Design, Vol. 3, Kluwer/ESCOM, Dordrecht, The Netherlands, 1998, pp. 87–104.
46. H. Kubinyi, F. A. Hamprecht, T. Mietzner, *J. Med. Chem.*, 41,2553–2564 (1998).
47. S. Wold in H. van de Waterbeemd, Ed., *Chemometrics Methods in Molecule Design*, VCH, Weinheim/New York, 1995, pp. 195–218.
48. B. Hoffman, S. J. Cho, W. Zheng, S. Wyrick, D. E. Nichols, R. B. Mailman, and A. Tropsha, *J. Med. Chem.*, 42,32173226 (1999).
49. W. Zheng and A. Tropsha, *J. Chem. Inf. Comput. Sci.*, 40, 185–194 (2000).
50. Ajay, *J. Med. Chem.*, 36, 3565–3571 (1993).
51. L. S. Anker and P. C. Jurs, *Anal. Chem.*, 62, 2676 (1990).
52. P. C. Jurs, J. W. Ball, and L. S. Anker, *J. Chem. Inf. Comput. Sci.*, 32,272 (1992).
53. T. M. Nelson and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 34, 601 (1994).
54. D. T. Stanton and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 32, 109 (1992).
55. S. Hellberg, M. Sjostrom, B. Skagerberg, and S. Wold, *J. Med. Chem.*, 30, 1126–1135 (1987).
56. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, 96,916–918 (1974).
57. K. C. Chu, R. J. Feldmann, N. B. Shapiro, G. F. Harard, and R. I. Geran, *J. Med. Chem.*, 18, 539–545 (1975).
58. G. Klopman, *J. Am. Chem. Soc.*, 106, 7315–7321 (1984).
59. G. Klopman, *Quant. Struct.-Act. Relat.*, 11, 176–184 (1992).
60. T. Aoyama, Y. Suzuki, and H. Ichikawa, *J. Med. Chem.*, 33,2583–2590 (1990).
61. S.-S. So and W. G. Richards, *J. Med. Chem.*, 35, 3201–3207 (1992).
62. F. R. Burden, B. S. Rosewarne, and D. A. Winkler, *Chemom. Intel. Lab. Syst.*, 38, 127–137 (1997).
63. G. Bolis, L. Di Pace, and F. Fabrocini, *J. Comput.-Aided Mol. Des.*, 5, 617–628 (1991).
64. R. D. King, S. H. Mugglfton, R. A. Lewis, and M. J. E. Sternberg, *Proc. Natl. Acad. Sci. USA*, 89,11322–11326 (1992).
65. R. D. King, S. H. Muggleton, A. Srinivasan, and M. J. E. Sternberg, *Proc. Natl. Acad. Sci. USA*, 93,438–442 (1996).
66. S. Clementi and S. Wold in H. van de Waterbeemd, Ed., *Chemometrics Methods in Molecular Design*, VCH, Weinheim/New York, 1995, pp. 319–338.
67. J. M. Sutter, S. L. Dixon, and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 35, 77 (1995).
68. D. Rogers and A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, 34,854–866 (1994).
69. H. Kubinyi, *Quant. Struct.-Act. Relat.*, 13, 285–294 (1994).
70. H. Kubinyi, *Quant. Struct.-Act. Relat.*, J3, 393–401 (1994).
71. B. T. Luke, *J. Chem. Inf. Comput. Sci.*, 34, 1279–1287 (1994).
72. S.-S. So and M. Karplus, *J. Med. Chem.*, 39, 1521–1530 (1996).
73. K. Hasegawa, T. Kimura, and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 39,112–120 (1999).
74. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi, *Quant. Struct.-Act. Relat.*, 12, 9–20 (1993).
75. S. J. Cho and A. Tropsha, *J. Med. Chem.*, 38, 1060–1066 (1995).
76. L. B. Kier, L. H. Hall, and J. W. Frazer, *J. Chem. Inf. Comput. Sci.*, 33,143 (1993).
77. L. H. Hall, L. B. Kier, and J. W. Frazer, *J. Chem. Inf. Comput. Sci.*, 33,148 (1993).
78. L. H. Hall, R. S. Dailey, and L. B. Kier, *J. Chem. Inf. Comput. Sci.*, 33, 598 (1993).
79. L. H. Hall and L. B. Kier in K. B. Lipkowitz and D. B. Boyd, Eds., *Reviews in Computational Chemistry II*, VCH, Weinheiflew York, 1991, pp. 367–422.

80. A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch, *J.* Med. Chem., **34,786–797 (1991).**

81. A. N. Jain, K. Koile, and D. Chapman, **J.** Med. Chem., **37**, 2315–2327 **(1994).**

82. F. **R.** Burden, Quant. *Struct.-Act.* Relat., **15, 7–11 (1996).**

83. P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, and J. Mockus, *J.* Chem. *Inf.* Comput. Sci., **23**, 93–102 **(1983).**

84. R. E. Carhart, D. H. Smith, and R. Venkataraghavan, **J.** Chem. *Inf.* Comput. Sci., **25, 64–73 (1985).**

85. R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, *J.* Chem. *Inf.* Comput. Sci., **27**, 82–85 **(1987).**

86. C. A. Pepperrell and P. Willett, **J.** *Comput.-*Aided Mol. Des., **5,455–474** (1991).

87. R. Nilakantan, N. Bauman, and R. Venkataraghavan, *J.* Chem. *Inf.* Comput. Sci., **33, 79–85 (1993).**

88. R. P. Sheridan, M. D. Miller, D. J. Underwood, and S. K. Kearsley, **J.** Chem. *Inf.* Comput. Sci., **36,128–136** (1996).

89. F. R. Burden, Quant. *Struct.-Act.* Relat., **16, 309–314 (1997).**

90. B. D. Silverman and D. E. Platt, **J.** Med. Chem., **39,2129–2140 (1996).**

91. D. A. Winkler, F. R. Burden, and A. Watkins, Quant. *Struct.-Act.* Relat., **17, 14–19 (1998).**

92. R. D. Brown and Y. C. Martin, *J.* Chem. *Inf.* Comput. Sci., **37, 1–9** (1997).

93. D. A. Winkler and F. R. Burden, Quant. *Struct.-Act.* Relat., **17, 224–231** (1998).

94. G. M. Downs and P. Willett in K. B. Lipkowitz and D. B. Boyd, Eds., Reviews in Computational Chemistry, Vol. **7,** VCH, Weinheim/New York, **1996,** pp. **1–65.**

95. Molconn-Z version **3.5,** Hall Associates Consulting, Quincy, MA.

96. M. Petitjean, **J.** Chem. *Inf.* Comput. Sci., **32, 331–337 (1992).**

97. H. Wiener, *J.* Am. Chem. *Soc.*, **69**, 17 **(1947).**

98. J. R. Platt, *J.* Phys. Chem., **56,328 (1952).**

99. C. Shannon and W. Weaver, Mathematical Theory of Communication, University of Illinois, Urbana, **1949.**

100. D. Bonchev, O. Mekenyan, and N. Trinajstic, **J.** Comput. Chem., **2**, 127–148 **(1981).**

101. The program Sybyl is available from Tripos Associates, St. Louis, MO.

102. M. Shen, A. LeTiran, Y. Xiao, H. Kohn, and A. Tropsha, **J.** Med. Chem., **45, 2811–2823** (2002).

103. F. H. Allen, J. E. Davies, J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith, and D. G. Watson, **J.** Chem. *Inf.* Comput. Sci., **31,187–204(1991).**

104. F. H. Allen, S. Bellard, M. D. Brice, B. A. Cartwright, A. Doubleday, H. Higgs, T. Hummelink, B. G. Hummelink-Peters, O. Kennard, W. D. S. Motherwell, J. R. Rodgers, and D. G. Watson, *Acta* Crystallogr. Sect. B, **B35,** 2331–**2339 (1979).**

105. A. Rusinko III, J. M. Skell, R. Balducci, C. M. McGarity, and R. S. Pearlman, Concord, *A Pro*gram for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures, The University of Texas at Austin and Tripos Associates, St. Louis, MO, **1988.**

106. R. S. Pearlman, Chem. Des. Aut. News, **2**, 1–6 (1987).

107. J. Gasteiger, **C.** Rudolph, and J. Sadowski, Tetrahedron Comput. Methodol., **3, 537–547 (1990).**

108. G. R. Marshall, C. D. Barry, H. E. Bosshard, R. **A.** Dammkoehler, and D. A. Dunn in E. C. Olson and R. E. Christoffersen, Eds., Computer-Assisted Drug Design, Vol. **112,** American Chemical Society, Washington DC, **1979,** pp. **205–226.**

109. G. E. Kellogg, S. F. Semus, and D. J. Abraham, **J.** *Comput.-Aided Mol.* Des., **5,545–552** (1991).

110. P. J. Goodford, **J.** Med. Chem., 28, **849–857 (1985).**

111. A. Agarwal, P. P. Pearson, E. W. Taylor, H. B. Li, T. Dahlgren, M. Herslof, Y. Yang, G. Lambert, D. L. Nelson, J. W. Regan, and A. R. Martin, **J.** Med. Chem., **36,4006–4014 (1993).**

112. C. L. Waller, T. I. Oprea, A. Giolitti, and G. R. Marshall, **J.** Med. Chem., **36, 4152–4160 (1993).**

113. A. K. Debnath, C. Hansch, K. H. Kim, and Y. C. Martin, **J.** Med. Chem., **36, 1007–1016** (1993).

114. M. Y. Brusniak, R. S. Pearlman, K. A. Neve, and R. E. Wilcox, **J.** Med. Chem., **39,850–859 (1996).**

115. Y. C. Martin, Methods Enzymol., **203,587–613 (1991).**

116. Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico, and P. **A.** Pavlik, **J.** *Comput.-*Aided Mol. Des., **7, 83–102** (1993).

117. S. J. Cho, M. G. Serrano, J. Bier, and A. Tropsha, **J.** Med. Chem., **39,5064–5071** (1996).

118. J. L. Sussman, M. Harel, F. Frolow, C. Oefner, A. Goldman, L. Toker, and I. Silman, Science, **253,8872–8879 (1991).**

119. M. Harel, I. Schalk, L. Ehret-Sabatier, F. Bouet, M. Goeldner, C. Hirth, P. H. Axelsen, I. Silman, and J. L. Sussman, Proc. Natl. *Acad.* Sci. USA, **90,9031–9035 (1993).**

120. R. D. Cramer III, S. A. DePriest, D. E. Patterson, and P. Hecht in H. Kubinyi, Ed., *3D* QSAR in Drug Design: Theory, Methods, and Applications, ESCOM Scientific, Leiden, The Netherlands, **1993,** pp. **443–485.**

121. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi, Quant. Strut.-Act. Relat., **12, 9–20 (1993).**

122. M. Pastor, G. Cruciani, I. McLay, S. Pickett, and S. Clementi, J. Med. Chem., **43,3233–3243 (2000).**

123. P. Ehrlich, Dtsch. Chem. Ges., **42, 17 (1909).**

124. C. Humblet and G. R. Marshall, *Annu.* Rep. Med. Chem., **15,267–276 (1980).**

125. S. A. DePriest, D. Mayer, C. B. Naylor, and G. R. Marshall, **J. Am.** Chem. *Soc.*, **115**, 5372–**5384 (1993).**

126. S. Wang, D. W. Zaharevitz, R. Sharma, V. E. Marquez, N. E. Lewin, L. Du, P.M. Blumberg, and G. W. A. Milne, *J.* Med. Chem., **37**, 4479–**4489 (1994).**

127. I. Motoc, R. A. Dammkoehler, and G. R. Marshall, Mathematics and Computational Concepts in Chemistry, Ellis Horwood, Chichester, UK, **1985,** pp. **222–251.**

128. D. Mayer, C. B. Naylor, I. Motoc, and G. R. Marshall, J. Cornput.-Aided Mol. Des., **1, 3–16 (1987).**

129. R. P. Sheridan, R. Nilakantan, J. S. Dixon, and R. Venkataraghavan, J. Med. Chem., **29**, 899–**906 (1986).**

130. G. W. A. Milne, M. C. Nicklaus, J. S. Driscoll, S. Wang, and D. Zaharevitz, J. Chem. *Inf. Com*put. Sci., **34, 1219–1224 (1994).**

131. Catalyst/Hypo Tutorial, version **2.0,** BioCAD Corp., Mountain View, CA, **1993.**

132. P. W. Sprague, Perspect. Drug Discov. Des., **3, 1–20 (1995).**

133. D. Barnum, J. Greene, A. Smellie, and P. Sprague, J. Chem. *Inf.* Comput. Sci., **36**, 563–**571 (1996).**

134. HipHop Tutorial, version **2.3,** Molecular Simulation Inc., Sunnyvale, CA, **1995.**

135. V. Golender and B. Vesterman, Network Science (http://www.netsci.org/Science/Compchem/feature09. html).

136. (a) Available from the author's WWW home page at http://mmlin1.pha.unc.edu/~jin/QSAR/ (b) A. Tropsha, S. J. Cho, and W. Zheng in A. L. Parrill and M. R. Reddy, Eds., Rational Drug Design: Novel Methodology and Practical Applications, ACS Symposium Series **719, 1999, pp. 198–211.**

137. T. A. Andrea and H. Kalayeh, *J.* Med. Chem., **34,2824–2836 (1991).**

138. J. D. Hirst, R. D. King, and M. J. Sternberg, J. Cornput.-Aided Mol. Des., **8, 405–420** (1994).

139. J. D. Hirst, R. D. King, and M. J. Sternberg, J. Cornput.-Aided Mol. Des., **8, 421–432 (1994).**

140. I. V. Tetko, V. Yu. Tanchuk, N. P. Chentsova, S. V. Antonenko, G. I. Poda, V. P. Kukhar, and A. I. Luik, *J.* Med. Chem., **37, 2520–2526 (1994).**

141. D. T. Manallack, D. D. Ellis, and D. J. Livingstone, J. Med. Chem., **37**, 3758–3767 **(1994).**

142. D. J. Maddalena and G. A. Johnston, *J.* Med. Chem., **38,715–724 (1995).**

143. G. Bolis, L. Pace, and F. A. Fabrocini, J. *Com*put.-Aided Mol. Des., **5**, 617–628 **(1991).**

144. R. D. King, S. Muggleton, R. A. Lewis, and M. J. Sternberg, Proc. Natl. Acad. Sci. USA, **89,11322–11326 (1992).**

145. A. N. Jain, T. G. Dietterich, R. H. Lathrop, D. Chapman, R. E. Critchlow Jr., B. E. Bauer, T. A. Webster, and T. Lozano-Perez, J. *Com*put.-Aided Mol. Des., **8,635–652 (1994).**

146. J. D. Hirst, J. Med. Chem., **39, 3526–3532 (1996).**

147. V. S. Rose, J. Wood, and H. J. H. MacFie in H. van de Waterbeemd, Ed., Advanced *Computer-*Assisted Techniques in Drug Discovery, VCH, Weinheim/New York, **1995,** pp. **228–242.**

148. Y. Hamamoto, S. Uchimura, and S. Tomita, ZEEE Trans. Pattern Anal. Machine Zntell., **19, 73–79 (1997).**

149. A. Djouadi and E. Bouktache, ZEEE Trans. Pattern Anal. Machine Zntell., **19, 277–282 (1997).**

150. O. Strouf, Chemical Pattern Recognition, Research Studies Press, Chichester, UK, **1986.**

151. M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn, J. Mol. Biol., **265**, 445–464 **(1997).**

152. S. C. Basak and G. D. Grunwald, *SAR* QSAR Environ. Res., **3, 265–277 (1995).**

153. S. C. Basak, S. Bertelsen, and G. D. Grunwald, Toxicol. Lett., **79,239–250 (1995).**

154. S. C. Basak and G. D. Grunwald, Chemosphere, 31,2529–2546 (1995).

155. S. C. Basak and G. D. Grunwald, New J. Chem., 19,231 (1995).

156. X. Gironés, A. Gallegos, and C.-D. Ramon, J. Chem Inf. Comput. Sci., 46, 1400–1407 (2000).

157. B. Bordás, T. Kömíves, Z. Szántó, and A. Lopata, J. Agric. Food Chem., 48, 926–931 (2000).

158. Y. Fan, L. M. Shi, K. W. Kohn, Y. Pommier, and J. N. Weinstein, J. Med. Chem., 44, 3254–3263 (2001).

159. M. Randíc and S. C. Basak, J. Chem. Inf. Comput. Sci., 40,899–905 (2000).

160. T. Suzuki, K. Ide, M. Ishida, and S. Shapiro, J. Chem. Inf. Comput. Sci., 41, 718–726 (2001).

161. M. Recanatini, A. Cavalli, F. Belluti, L. Piazzi, A. Rampa, A. Bisi, S. Gobbi, P. Valenti, V. Andrisano, M. Bartolini, and V. Cavrini, J. Med. Chem., 43,2007–2018 (2000).

162. J. A. Morón, M. Campillo, V. Perez, M. Unzeta, and L. Pardo, J. Med. Chem., 43, 1684–1691 (2000).

163. A. Golbraikh and A. Tropsha, J. Mol. Graphics Model., 20,269–276 (2002).

164. J. Huuskonen, J. Chem. Inf. Comput. Sci., 41, 425–429 (2001).

165. I. V. Tetko, V. V. Kovalishyn, and D. J. Livingstone, J. Med. Chem., 44, 2411–2420 (2001).

166. W. Wu, B. Walczak, D. L. Massart, S. Heuerding, F. Erni, I. R. Last, and K. A. Prebble, Chemom. Intell. Lab. Syst., 33, 35–46 (1996).

167. A. Yasri and D. Hartsough, J. Chem. Inf. Comput. Sci., 41, 1218–1227 (2001).

168. P. Bernard, D. B. Kireev, J. R. Chretien, P. L. Fortier, and L. Coppet, J. Cornput.-Aided Mol. Des., 13,355–371 (1999).

169. Y. Takeuchi, E. F. B. Shands, D. D. Beusen, and G. R. Marshall, J. Med. Chem., 41, 3609–3623 (1998).

170. G. V. Kauffmanand P. C. Jurs, J. Chem. Inf. Comput. Sci., 41, 1553–1560 (2001).

171. B. E. Mattioni and P. C. Jurs, J. Chem. Inf. Comput. Sci., 42, 94–102 (2002).

172. J. Gasteiger and J. Zupan, Angew. Chem., 32, 503 (1993).

173. Y. L. Loukas, J. Med. Chem., 44, 2772–2783 (2001).

174. P. Bernard, M. Pintore, J. Y. Berthon, and J. R. Chretien, Eur. J. Med. Chem., 36, 1–19 (2001).

175. F. R. Burden and D. A. Winkler, J. Med. Chem., 42,3183–3187 (1999).

176. F. R. Burden, M. G. Ford, D. C. Whitley, and D. A. Winkler, J. Chem. Inf. Comput. Sci., 40, 1423–1430 (2000).

177. M. J. Adams, Chemometrics in Analytical Spectroscopy, The Royal Society of Chemistry, London, 1995.

178. T. Potter and H. Matter, J. Med. Chem., 41, 478–488 (1998).

179. M. Lajiness, M. A. Johnson, and G. M. Maggiora in J. L. Fauchere, Ed., Quantitative Structure-Activity Relationships in Drug Design, Alan R. Liss, New York, 1989, pp. 173–176.

180. R. Taylor, J. Chem. Inf. Comput. Sci., 35, 59–67 (1995).

181. M. Snarey, N. K. Terrett, P. Willett, and D. J. Wilton, J. Mol. Graphics Model., 15, 372385 (1997).

182. R. W. Kennard and L. A. Stone, Technometrics, 11,137–148 (1969).

183. B. Bourguignon, P. F. Deaguiar, K. Thorre, and D. L. Massart, J. Chromatogr. Sci., 32, 144–152 (1994).

184. B. Bourguignon, P. F. Deaguiar, M. S. Khots, and D. L. Massart, Anal. Chem., 66, 893–904 (1994).

185. S. Hellberg, L. Eriksson, J. Jonsson, F. Lindgren, M. Sjostrom, B. Skagerberg, S. Wold, and P. Andrews, Int. J. Pept. Protein Res., 37,414–424 (1991).

186. L. Eriksson and E. Johansson, Chemom. Intell. Lab. Syst., 34, 1–19 (1996).

187. R. Carlson, Design and Optimization in Organic Synthesis, Elsevier, Amsterdam/New York, 1992.

188. E. J. Martin and R. E. Critchlow, J. Comb. Chem., 1, 32–45 (1999).

189. A. Miller and N.-K. Nguyen, Appl. Stat., 43, 669–678 (1994).

190. T. J. Mitchell, Technometrics, 42, 48–54 (2000).

191. S. Wold and L. Eriksson in H. van de Waterbeemd, Ed., Chemometrics Methods in Molecular Design, VCH, Weinheim/New York, 1995, pp. 309–318.

192. E. Novellino, C. Fattorusso, and G. Greco, Pharm. Acta Helv., 70, 149–154 (1995).

193. U. Norinder, J. Chemom., 10, 95–105 (1996).

194. N. S. Zefirov and V. A. Palyulin, J. Chem. Inf. Comput. Sci., 41, 1022–1027 (2001).

195. L. Sachs, Applied Statistics: A Handbook of Techniques, Springer-Verlag, Berlin/New York, 1984.

196. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Adavnces in Knowledge Discovery and Data Mining, AAAI* Press/The MIT Press, Cambridge, MA, **1996.**

197. G. H. John, ***Enhancements to the Data Mining Process,*** Ph.D thesis, Stanford University, **1997.**

198. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, ***From Data Mining to Knowledge Discovery,*** AAAI Press/The MIT Press, Cambridge, MA, **1995.**

199. E. Simoudis, ***ZEEE Expert.*, 11, 26-33 (1996).**

200. M. A. Razzak and R. C. Glen, ***J. Cornput.-Aided Mol. Des.*, 6,349–383 (1992).**

201. R. D. King, J. D. Hirst, and M. J. E. Sternberg, *Appl. Artif. Intell.*, *9*, 213–234 (1994).

202. S. S. Young and D. M. Hawkins, ***J. Med. Chem.*, 38,2784–2788** (1995).

203. S. S. Young and D. M. Hawkins, ***SAR QSAR Environ. Res.*, 8, 183–193 (1998).**

204. D. M. Hawkins, S. S. Young, and A. Rusinko, ***Quant. Struct.-Act. Relat.*, 16, 1–7** (1997).

205. R. King, R. Henery, C. Feng, and A. Sutherland in D. Michie, S. Muggleton, and F. Furukawa, Eds., Machine Intelligence and Inductive Learning, Vol. **13,** Oxford University Press, Oxford, UK, **1994.**

206. S. S. Young, M. Farmen, and A. Rusinko, ***Network Science*** (http://www.netsci.org/Science/Screening/feature09.html)

207. J. M. Barnard and G. M. Downs, ***Perspect. Drug Discov. Des.*, 7/8, 13–30 (1997).**

208. W. P. Walters, M. T. Stahl, and M. A. Murcko, ***Drug Discov. Today*, 3,160–178** (1998).

209. J. M. Barnard, ***J. Chem. Inf. Comput. Sci.*, 33, 532–538 (1993).**

210. S. A. Cook, Proceedings of the Third Annual ACM Symposium on the Theory of Computing, ACM, New York, **1971,** pp. **151–158.**

211. L. C. Ray and R. A. Kirsch, ***Science, 126,*** 814–**819 (1957).**

212. X. Jun and Z. Maosen, ***Tetrahedron Comput. Methodol.*, 2, 75–83** (1989).

213. A. Dengler and I. Ugi, ***Comput. Chem.*, 15, 103–107 (1991).**

214. E. H. Sussenguth, ***J. Chem. Doc.*, 5, 36–43 (1965).**

215. J. Figueras, ***J. Chem. Doc.*, 12,237–244 (1972).**

216. J. R. Ullmann, ***J. Assoc. Comput. Mach.*, 23, 31–42 (1976).**

217. A. Von Scholley, ***J. Chem. Inf. Comput. Sci.*, 24, 235–241 (1984).**

218. P. H. A. Sneath and R. R. Sokal, ***Numerical Taxonomy,*** Freeman, San Francisco, **1973.**

219. P. Willett and V. A. Winterman, ***Quant. Strut.-Act. Relat.*, 5, 18–25 (1986).**

220. M. A. Gallop, R. W. Barret, W. J. Dower, S. P. A. Fodor, and E. M. Gordon, ***J. Med. Chem.*, 37,1233–1251 (1994).**

221. E. M. Gordon, R. W. Barret, W. J. Dower, S. P. A. Fodor, and M. A. Gallop, ***J. Med. Chem.*, 37,** 1385–1401 **(1994).**

222. W. A. Warr, ***J. Chem. Inf. Comput. Sci.*, 37, 134–140 (1997).**

223. R. P. Sheridan and S. K. Kearsley, ***J. Chem. Inf. Comput. Sci.*, 35,310–320 (1995).**

224. A. K. Ghose and V. N. Viswanadhan, Eds., ***Combinatorial Library Design and Evaluation for Drug Discovery: Principles, Methods, Software Tools and Applications,*** Marcel Dekker, New York, **2001.**

225. W. Zheng, S. J. Cho, and A. Tropsha, ***J. Chem. Inf. Comp. Sci.*, 38,251–258 (1998).**

226. S. J. Cho, W. Zheng, and A. Tropsha, ***J. Chem. Inf. Comp. Sci.*, 38,** 259–268 **(1998).**

227. A. Tropsha, S. J. Cho, and W. Zheng in A. L. Parrill and M. R. Reddy, Eds., ***Rational Drug Design: Novel Methodology and Practical Applications,*** ACS Symposium Series **719,** American Chemical Society, Washington, DC, **1999,** pp. **198–211.**

228. I. O. Bohachevsky, M. E. Johnson, and M. L. Stein, ***Technometrics, 28,209–217 (1986).***

229. J. H. Kalivas, J. M. Sutter, and N. Roberts, ***Anal. Chem.*, 61,2024–2030** (1989).

230. D. E. Goldberg, ***Genetic Algorithm in Search, Optimization, and Machine Learning,*** Addison-Wesley, Reading, MA, **1989.**

231. J. H. Holland, ***Sci. Am.*, 267, 66–72 (1992).**

232. S. Forrest, ***Science, 261,872–878 (1993).***