

*Annual Review*QUANTITATIVE STRUCTURE–ACTIVITY RELATIONSHIP METHODS:
PERSPECTIVES ON DRUG DISCOVERY AND TOXICOLOGY

ROGER PERKINS,*† HONG FANG,† WEIDA TONG,‡ and WILLIAM J. WELSH§

†Logicon ROW Sciences, 3900 NCTR Road, MC 910, Jefferson, Arkansas 72079, USA

‡Center for Toxicoinformatics, National Center for Toxicological Research, Food and Drug Administration,
3900 NCTR Road, Jefferson, Arkansas, 72079, USA§Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey,
675 Hoes Lane, Piscataway, New Jersey 08854, USA

(Received 11 April 2001; Accepted 2 January 2003)

Abstract—Quantitative structure–activity relationships (QSARs) attempt to correlate chemical structure with activity using statistical approaches. The QSAR models are useful for various purposes including the prediction of activities of untested chemicals. Quantitative structure–activity relationships and other related approaches have attracted broad scientific interest, particularly in the pharmaceutical industry for drug discovery and in toxicology and environmental science for risk assessment. An assortment of new QSAR methods have been developed during the past decade, most of them focused on drug discovery. Besides advancing our fundamental knowledge of QSARs, these scientific efforts have stimulated their application in a wider range of disciplines, such as toxicology, where QSARs have not yet gained full appreciation. In this review, we attempt to summarize the status of QSAR with emphasis on illuminating the utility and limitations of QSAR technology. We will first review two-dimensional (2D) QSAR with a discussion of the availability and appropriate selection of molecular descriptors. We will then proceed to describe three-dimensional (3D) QSAR and key issues associated with this technology, then compare the relative suitability of 2D and 3D QSAR for different applications. Given the recent technological advances in biological research for rapid identification of drug targets, we mention several examples in which QSAR approaches are employed in conjunction with improved knowledge of the structure and function of the target receptor. The review will conclude by discussing statistical validation of QSAR models, a topic that has received sparse attention in recent years despite its critical importance.

Keywords—Quantitative structure–activity relationship Toxicology Drug design Chemoinformatics Chemometric

INTRODUCTION

Quantitative structure–activity relationship (QSAR) models are quantitative regression methods that attempt to relate chemical structure to biological activity. Quantitative structure–activity relationship and related methods have been applied extensively in a wide range of scientific disciplines, including chemistry, biology, and toxicology [1,2]. In both drug discovery and environmental toxicology [3], QSAR models are now regarded as a scientifically credible tool for predicting and classifying the biological activities of untested chemicals. As we enter the new millennium, QSAR has become inexorably embedded as an essential tool in the pharmaceutical industry, from lead discovery and optimization to lead development [4,5]. For example, a growing trend is to use QSAR early in the drug discovery process as a screening and enrichment tool to eliminate from further development those chemicals lacking druglike properties [6] or those chemicals predicted to elicit a toxic response. This developing scenario portends the spread of QSAR beyond the pharmaceutical industry to human and environmental regulatory authorities for use in toxicology [7–13].

Computer hardware and software improvements have been enabling technologies in QSAR development during the past decade. Within the pharmaceutical industry alone, the enormous financial incentives to accelerate the drug discovery process and to improve the odds of success by enriching the drug

pipeline with more effective and less toxic candidates are powerful driving forces that have led to improved QSAR approaches and associated software. The integration of QSAR modeling with recent advances in hardware and software for data storage and management has further stimulated its wider implementation [14]. The algorithms used in QSAR software also improved markedly, particularly with respect to the large and growing pool of descriptors used to characterize molecular structure and properties.

The fundamental assumption of QSAR is that variations in the biological activity of a series of chemicals that target a common mechanism of action are correlated with variations in their structural, physical, and chemical properties [15]. Since presumably these structurally related properties of a chemical can be determined by experimental or computational means much more efficiently than its biological activity using *in vitro* or *in vivo* approaches, a statistically validated QSAR model is capable of predicting the biological activity of a new chemical within the same series in lieu of the time-consuming and labor-intensive processes of chemical synthesis and biological evaluation. Applied judiciously, QSAR can save substantial amounts of time, money, and human resources.

QSAR MODELING

This generally involves three steps: (1) collect or, if possible, design a training set of chemicals; (2) choose descriptors that can properly relate chemical structure to biological activity; and (3) apply statistical methods that correlate changes in structure with changes in biological activity. Obtaining a good-

* To whom correspondence may be addressed
(rperkins@nctr.fda.gov).

quality QSAR model with the ability to predict activity of a chemical outside the training set depends on many factors in the approach and execution of each of the three steps.

Quality of data

Data should come from the same assay protocol, and care should be taken to avoid interlaboratory variability. Any bad data points will tend to corrupt the proper correlation of structure and activity. Rules of thumb for a good QSAR data set are that the dose–response relationship should be smooth, the potency (or affinity) should be reproducible, the activity range should span two or more orders of magnitude from the least active to the most active chemical in the series, the number of chemicals used to build the QSAR model should be sufficiently large to ensure statistical stability, the activities of the chemicals should be evenly distributed across the range of activity, and the chemicals selected for the training set should possess enough structural diversity to span the range of chemistry space associated with the biological activity under study.

Descriptor selection

Many types of chemical structure descriptors are available from commercial software. Obtaining a statistically robust model is very much dependent on how well the selected descriptors can encode the variation of activity with structure. The more that is known at the molecular level about the biological mechanism of action of the chemicals, the better the chemist is able to select among the wide variety and types of specific molecular descriptors. Commercially available molecular modeling programs often include statistical tools to help in evaluating which descriptors best encode structure–activity variation. Some of these tools include the genetic algorithm (GA) in its various incarnations, which employs the evolutionary rules of natural selection to select the optimal (fittest) subset of descriptors amongst its wide set for a particular problem.

Statistical methods

It is also critical that the QSAR method selected to develop the structure–activity correlation be suitable. Although the relationship between a molecular descriptor and biological activity may be linear or nonlinear, it is still common practice today to deploy linear approaches such as multiple (or multivariate) linear regression (MLR) or partial least squares (PLS) regression to construct the QSAR model. For nonlinear modeling, the Polynomial Neural Network (PNN) offers an alternative that combines the best features of Artificial Neural Networks (ANNs) and MLR/PLS by providing the inherent non-linearity of the ANN with the desired analytical regression equation furnished by MLR and PLS [16]. Several statistical approaches that are commonly used in QSAR modeling are listed in the Appendix along with a brief description of their theory and procedure. The most common scenario encountered in practice is for the number of possible descriptors to exceed the number of chemicals, a situation that can lead to chance correlations. Fortunately, soft modeling methods such as PLS reduce the risk of encountering chance correlations by transforming the dimensionality of the regression problem from chemical-descriptor space to so-called principal components (PCs) space.

QSAR models are useful in research for purposes beyond prediction [17]. Additional benefits that may accrue include leveraging existing structure–activity data, providing insights

into mechanism or identifying an alternative mechanism (e.g., metabolism) of action, identifying important structural characteristics, suggesting new design strategies and synthetic targets, narrowing the dose range for a planned assay, assisting in generation of new hypotheses to guide further research, and revealing chemicals that deviate from the QSAR model.

TRANSFORMATION OF DRUG DISCOVERY

The drug discovery and development process used by pharmaceutical companies has undergone a radical metamorphosis over the past decade. What used to be likened to the search for a needle in a haystack, one chemical at a time and one target at a time, is now more akin to a search for many needles in many haystacks. The transformation was spawned by the advent of combinatorial chemistry and high throughput screening (HTS) that generate more needles and of genomics that generate more haystacks. Quantitative structure–activity relationship figures prominently in this evolved paradigm [18,19]. Combinatorial library design often uses diversity analysis and QSAR to select chemicals for synthesis and testing. The rapid turnaround to produce data in this new paradigm requires concomitantly fast QSARs to process information derived from the chemical library design–combinatorial synthesis–HTS cycle. Data culminating from the HTS assays are then available to build more robust QSAR models that can be used to guide more refined lead discovery, optimization, and development.

Within the new paradigm is also a growing trend to build models to predict not only potency and selectivity but also absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties, that is, to predict with QSAR the pharmacokinetic and pharmacodynamic properties that make a chemical druglike or properties that may induce toxic side effects [20]. Unfavorable ADMET properties are a major cause of a chemical's removal from development, and compelling financial incentives exist to eliminate those leads that are not viable as commercial products as early as possible in the drug development process. For example, QSAR can be used to estimate such key properties as permeability [21,22], solubility [23,24], and cytochrome P450 metabolism [25,26].

The term QSAR is sometimes interpreted broadly to include methodologies that predict activity on an ordinal or categorical scale. For the purpose of this review, we adopt a more conservative definition of QSARs, that is, methodologies that predict activities strictly on an interval or continuous scale. A companion paper by Tong et al. in this volume reviews qualitative predictive methods that encompass the broader definition.

Several commercial software products have been successfully used in the areas of acute toxicity prediction [3], carcinogenicity prediction [27], skin sensitization prediction, and so on [28]. These include CASE/MultiCASE (MultiCASE, Beachwood, OH, USA) [29,30], TOPKAT (Accelrys, San Diego, CA, USA) [31,32], COREPA [33,34], and rule-based expert systems, such as DEREK (LHASA Group, Chemistry Department of Harvard University, Cambridge, MA, USA) [35,36] and ONCOLOGIC (Logichem, Boyertown, PA, USA) [37]. Most of these software products employ algorithms, based on either 2D or sometimes 3D structure fragments, that produce a qualitative rather than a quantitative prediction. A number of good review papers can be found in the literature [28,38], and these methods are not further covered in this review.

Quantitative structure–activity relationship methods have undergone rapid change during the past decade. The primary aim of this review is to summarize the status of QSAR with emphasis on conveying the utility and limitations of QSAR technology together with examples of representative applications. Since it is impossible to cover such a broad topic in depth in a single communication, a robust bibliography is provided for the interested reader. We will first review 2D QSAR together with a discussion of the availability and appropriate selection of molecular descriptors. We will then proceed to describe 3D QSAR and key issues associated with this technology, then compare the relative suitability of 2D and 3D QSAR for different applications. Given the recent technological advances in biological research for rapid identification of drug targets, we mention several examples in which QSAR approaches are employed in conjunction with improved knowledge of the structure and function of the targeted receptor. The review will conclude by discussing statistical validation of QSAR models, a topic that has received sparse attention in recent years despite its critical importance. In this regard, it is essential to note that the utility of a QSAR model, regardless of the inherent sophistication of the methods employed and the time expended in its development, is only as good as the quality of the data modeled.

2D QSAR

Investigation of the effect of physicochemical properties on activity and toxicity dates back to the 19th century [39,40]. In 1935, the Hammett constant (σ) was successfully used to correlate the equilibrium constants and reactivity of organic acids and bases [41,42]. However, difficulties were encountered when investigators attempted to apply Hammett-type relationships to biological systems, indicating that other structural determinants needed consideration. In 1969, the seminal work by Hansch inaugurated a new era of QSARs in which the hydrophobicity expressed as the octanol–water partition coefficient ($\log P$) was found highly valuable for predicting various biological observations [43]. $\log P$ or other measures of hydrophobicity are now used extensively in drug discovery and predictive toxicology. The Hansch-type approach that correlates physicochemical properties with activity using multivariable regression method has been widely applied to problem areas such as toxicity, enzyme inhibition, ligand–receptor binding, carcinogenicity, mutagenesis, and metabolism [1].

Fundamental to QSAR is development of a model relating the molecular structures of a set of chemicals with their biological activities. The nature of the descriptors used, and the extent to which they encode the structural features of the molecules that are related to the biological activity, is a crucial component of any QSAR study. Early QSAR studies concentrated on establishing a correlation between biological activity and experimentally derived physicochemical properties, such as σ , $\log P$, pK_a , and molar refractivity (MR). The predominant method of correlation was multiple linear regression (see Appendix) [1,2,44]. Although this approach is still widely applied, the experimental physicochemical parameters have been replaced largely by scores of computer-generated descriptors each of which encodes a particular molecular feature. For instance, the CODESSA program (Semichem, Shawnee, KS, USA) and the Cerius² program (Accelrys, San Diego, CA, USA) each can generate hundreds of calculated descriptors to represent the structure of a molecule.

Types of descriptors

In general, molecular descriptors used in 2D QSAR can be grouped into three categories: (1) 2D (e.g., molecular connectivity), (2) 3D (e.g., molecular surface area), and (3) physicochemical properties (e.g., $\log P$). The 2D descriptors are calculated solely on the basis of 2D structure, in which graph theory is widely employed for this purpose. In contrast, 3D descriptors require the 3D geometry of structures and are sensitive to the particular conformation adopted by a flexible molecule. Physicochemical descriptors characterize the properties of the entire molecule in a single value.

The descriptors for 2D QSAR can also be categorized according to their nature as well as calculation method, such as constitutional, topological, geometrical, electrostatic, quantum-chemical, and thermodynamic descriptors. The simplest descriptor type is constitutional, which reflects the molecular composition without regard to geometric or electronic structure (e.g., atom counts, molecular weight, the number of rotatable bonds, and the number of hydrogen-bond donors and acceptors). Topological descriptors, which include the Kier and Hall, Randic, and Wiener indices, encode molecular connectivity (bond information between a pair of atoms in a molecule). In particular, the E-State descriptors have recently gained considerable popularity in QSAR studies [45–48]. Geometrical descriptors (sometimes called spatial descriptors), such as moment of inertia, molecular surface area, shadow indices, and molecular density and volume, require the 3D coordinates of a structure. Quantities such as sum of atomic partial charges and sum of atomic polarization are examples of non-3D electrostatic descriptors. Dipole moment and the Jurs charge partial surface area [49,50] are 3D electrostatic descriptors that reflect particular aspects of charge distribution of a molecule. Quantum-chemical descriptors such as the highest-occupied molecular orbital and the lowest-unoccupied molecular orbital enhance the conventional descriptors by providing information about the internal electronic properties of molecules. The thermodynamic descriptors, found in many early Hansch-type QSARs, are all non-3D, although some contain 3D information. These include molar refractivity (MR) as a combined measure of molecule size and polarizability, $\log P$ to characterize the hydrophobicity of the molecule, heats of formation, and the (de)solvation free energies for water and for octanol.

For the calculation of molecular descriptors used in 2D QSAR, it is usually sufficient to generate the necessary structural information either from experimental methods (e.g., X-ray crystallography or nuclear magnetic resonance spectroscopy) or from calculations using molecular mechanics (MM). However, some descriptors (e.g., highest-occupied molecular orbital and the lowest-unoccupied molecular orbital energies) require calculations by quantum mechanics (QM). To the best of our knowledge, few studies have attempted to compare the information content of the MM-based descriptors with that of the QM-based descriptors. Recently, Tong et al. [51] compared two QSAR models that employed different geometries (one MM calculated and another QM calculated) to calculate values for the same set of molecular descriptors. Variable selection in both cases was achieved using the Genetic Function Approximation [52] ([GFA]; see the description of the method in the next section). A slightly better QSAR model was obtained for the QM-based descriptors than for the MM-based descriptors.

Variable selection

One key for obtaining a useful QSAR model is to select descriptors (variables or factors) that are information rich. Commercial molecular modeling programs make it possible to choose from hundreds or even thousands of molecular descriptors. The challenge is to select that subset of descriptors that is capable of representing the most critical structural and physicochemical features associated with activity. Often, a small ensemble of descriptors will be sufficient to capture most of the variation between structure and activity [53]. Effective descriptor selection, otherwise known as variable selection, is thus an integral and inseparable part of the QSAR modeling process. In fact, most improvements in 2D QSAR have been in the development and use of statistical approaches to make the selection of descriptors more effective.

Among descriptor selection methods, the GA approach has been particularly effective and efficient [54,55]. Genetic algorithm, as the name implies, is based on the principle of Darwinian evolution. A GFA approach developed by Rogers et al. [52] and implemented in Cerius² (Accelrys) is a popular GA-based statistical approach that is now widely used in QSAR model development. The GFA starts with random selection of sets of descriptors from an original descriptor pool. A multivariate regression technique is then used to develop a number of QSAR equations for each set of descriptors to form a QSAR equation pool (100 equations is default). Next, the quality of each individual equation is estimated using a lack-of-fit fitness function [54,55], and the equations are rank-ordered in accordance with fitness. Following Darwinian precepts, the two best QSAR equations (parents) are mated to produce offspring QSAR equations (children); a crossover process takes descriptors (alleles) from each parent to form the offspring descriptors. Finally, the genetically improved offspring QSAR equations replace the worst equations in the original equation pool. The overall process is repeated many times (typically 20,000 times) until no significant improvement is observed in the model and good combinations of descriptors are discovered and are dispersed throughout the population of QSAR models. We have found that the descriptors used by the majority of equations in the final equation pool are most relevant to the biological mechanism of action [51].

Although the general process is the same, GA-based QSAR approaches vary in a number of ways. Differences are found in the mating process, statistical method used, and the fitness function. For example, in a GA-PLS method reported by Cho et al. [56] (<http://mmlin1.pha.unc.edu/~jin/QSAR/GA-PLS/gapls.html>), PLS is used to construct QSAR models. The fitness function, $1 - (N - 1)(1 - q^2)/(N - PC)$ where N is the data size and PC is the optimal number of PCs, is directly weighted by q^2 , the cross-validated r^2 . A descriptor crossover operation is performed for two parents that are randomly selected from the original QSAR population to produce two offspring QSAR equations. Next, each offspring is subjected to a random single-point mutation where one descriptor is either added or removed. The fitness function is applied to each offspring to determine if the mutation increased or decreased fitness compared to the parents. If the offspring are better than the parents, they replace the parents; otherwise, the parents are retained. A statistically robust 2D QSAR model ($q^2 = 0.85$) was developed for a set of dopamine transporter ligands using the GA-PLS approach [57].

Zheng and Tropsha [58] reported an automated variable

selection QSAR method that is based on the k -Nearest Neighbor (kNN) principle. In this so-called k NN-QSAR method, a chemical's activity is estimated as the mean activity value of its k nearest neighbor based on Euclidean distance in a multidimensional descriptor coordinate system. An optimum subset of descriptors is determined using a simulated annealing method where a QSAR equation with randomly selected descriptors is compared with the same equation that is perturbed by randomly removing and replacing a small number of descriptors. If the perturbed equation is better, it is accepted. If the perturbed equation is worse, its acceptance or rejection is based on the Metropolis criterion. The method was tested on a number of data sets demonstrating its effectiveness and generality.

Fragment-based QSAR

Predicting chemical properties and activities based solely on the fragments (or substructures) of molecules has attracted considerable attention by virtue of its simplicity and speed of application. The simplest form of structural representation, the simplified molecular input line entry specification (SMILES) notation, is the only structural information needed for this type of modeling. No need exists for time-consuming determination of 3D structure, putative binding conformations, and molecular alignment, as is the case for some 2D QSAR and all 3D QSAR methods. Fragment-based QSAR approaches are particularly useful for rapidly screening chemical libraries against a given target and a given assay system—needs that are often encountered in both drug discovery and toxicology.

The earliest attempt to utilize substructural fragments to predict activity was the Free-Wilson Method in 1956 [59]. The method assumes that activity (or property) is linearly correlated with the additive and constant contribution of the substituents on a basic molecular structure. A similar approach was used [60] for CLOGP (BioByte, Claremont, CA, USA) calculation where a molecule's octanol-water partition coefficient P (given as $\log P$) is estimated by adding the $\log P$ contribution from each individual fragment comprising the molecule. The atom-based Alog P method developed by Ghose and Crippen [61–63] is calculated in a similar fashion using the equation $\log P = \sum n_i a_i$, where n_i is the number of atoms of type i and a_i is the atomic $\log P$ contribution. The CASE program developed by Klopman and Wang [64] analyzes the statistical significance of the distribution of the fragments presented in active versus inactive chemicals to determine key fragments (biophores and biophobes) that are associated with the activity under study.

Recently, a novel fragment-based QSAR approach, Hologram QSAR (HQSAR), was introduced by Tripos (St. Louis, MO, USA). In HQSAR, each molecule in the data set is divided into structural fragments that are then counted in bins of a fixed-length array to form a molecular hologram. This process is similar to the generation of molecular fingerprints often used for database searching and molecular diversity calculations [65]. The bin occupancies of the molecular hologram are structural descriptors (independent variables) encoding compositional and topological molecular information. A linear regression equation that correlates variation in structural information (as encoded in the hologram for each molecule) with variation in activity data is derived through PLS regression analysis to produce a QSAR model. Unlike other fragment-based methods [66], HQSAR encodes all possible molecular fragments (linear, branched, and overlapping). Optionally, additional 3D infor-

mation such as hybridization and chirality may also be encoded in the molecular holograms. Molecular holograms are generated in the same manner as hashed fingerprints where different unique fragments may populate the same holographic bin allowing the use of a fixed-length hologram fingerprint. This hashing procedure emphasizes the importance of patterns of fragment distribution within the hologram bins, which represents the nature of chemical structures more appropriately. Hologram QSAR has several attributes, including speed, reproducibility, and ease of use, that suggest its potential utility for prioritizing large numbers of chemicals for subsequent testing [67].

3D QSAR

The 3D QSAR, which correlates spatially localized features across a chemical series with biological activity, has attracted considerable attention in QSAR over the past decade. Because the descriptors used to represent chemical structure usually encode location-dependent structural characteristics that account for activity, the molecular structures need to be aligned across the series. Descriptor types distinguish the two primary types of 3D-QSAR methods: lattice-based descriptors and surface-based descriptors. Among the lattice-based methods, Comparative Molecular Field Analysis (CoMFA) is by far the most studied and applied 3D-QSAR method. In CoMFA, alignment of chemical structure is of paramount importance yet somewhat subjective. Alignment is easier and more accurate given knowledge of the mechanism of action at the molecular level to guide the definition of meaningful molecular alignment rules. For example, the availability of crystal structure data for a bound ligand-receptor complex will assist in identification of similar structural features across a series, such as the location at which critical hydrogen bonds or steric interactions occur. Thus, alignment is more or less a knowledge-guided process. Notwithstanding the drawbacks associated with the subjective nature of alignment, efforts to develop alignment-free 3D-QSAR methods (e.g., CoMMA [68,69], WHIM [70–72], and EVA [73,74]) have been only partially successful. For example, CoMMA utilizes the zeroth-, first-, and second-order spatial moments of the charge (e.g., quadrupolar moments) as well as the mass distribution (e.g., moments of inertia) to capture 3D structural information. When applied to the same steroid data set used in the first published CoMFA model [75], CoMMA produced comparable statistical results. One of the key components in CoMMA is the principal quadrupolar axes calculated with respect to the molecular center of dipole. A unique feature of this method consists of the descriptors known as dx, dy, and dz, which measure the displacement between the center of mass and center of dipole with respect to the principal inertial axes. Unfortunately, the value of these descriptors equals infinity for symmetric molecules whose dipole moment is zero, a drawback that may account for the limited number of published CoMMA applications.

CoMFA

CoMFA is accessible through the Comparative Molecular Field Analysis software. To construct a CoMFA model, a collection of chemicals with known activities (the training set) are first aligned together, usually employing structure similarity as the basis for alignment. The aligned molecules are then embedded in a 3D grid, after which the steric and electrostatic fields are computed for each chemical at every grid

point surrounding the molecules. The variations in these steric/electrostatic fields are then correlated with variations in the observed biological activity using PLS.

Since CoMFA was first introduced by Cramer et al. [75] in 1988, a large number of applications have been reported in the literature. The Dialog Science Database (Dialog, Cary, NC, USA) that reports scientific citations shows an ever-increasing number of CoMFA applications: 66 between 1988 and 1992, 392 during the period of 1993 to 1996, and 652 from 1997 to the present. With such rapid and widespread usage, it is not surprising that CoMFA has been the subject of numerous critiques and review articles [5,53,76,77]. It has been found that CoMFA is relatively insensitive to variations in geometry as calculated using various methods. For example, no appreciable differences in CoMFA results were observed using molecular geometry generated with MM when compared to geometry generated with semiempirical QM [78]. Similarly, no significant advantage has been identified for calculation of the atomic partial charges using more rigorous QM methods compared with simple methods such as Gasteiger-Marsili [79]. With respect to calculating geometry and charge distribution at least, it appears more important that the methods employed be consistent rather than precise or correct in the absolute sense.

Because locally steep spatial gradients can exist in the calculated steric and electrostatic fields, several studies have focused on the effects of grid spacing and lattice position [78]. CoMFA superimposes a 3D lattice on the molecules, and the steric and electrostatic fields are computed at each grid point located in the intersection of the lines comprising the lattice. The grid spacing is the constant distance between any two adjacent, parallel lines of the lattice, while lattice position determines the spatial position and orientation of the rigidly aligned molecules with respect to the lattice lines within a box of fixed dimensions. For a particular series of aligned molecules, Cho and Tropsha [80] reported a dependency on lattice position that could be as large as 0.5 in q^2 . Such sensitivity may result from a fixed grid spacing close to a molecule surface where steric and electrostatic fields vary dramatically. The larger the grid spacing is, the larger the field change between grid points, and a significant loss of information can occur if the grid is too coarse [80]. Many studies have shown that more stable and statistically robust models are obtained with a 1Å versus a 2Å grid spacing [81]; these studies used training sets that were either small, comprised a congeneric series, or both. More recently, we [12] reported a CoMFA model with a training set comprising 130 highly diverse estrogens that had been aligned with the guidance of crystal structure information for several chemicals. Using all-orientation search and all-placement search methods [82] to explore every possible orientation and placement of the aligned molecules in the CoMFA region, no significant variation of q^2 with lattice position was found. Moreover, no significant improvement was found in the model by switching the grid spacing from 2Å to 1Å. One explanation for this insensitivity to grid spacing may be that a large and properly aligned training set of structurally diverse chemicals provides a much higher signal-to-noise ratio in a 3D grid for PLS modeling.

Another active area in CoMFA development has been methods to improve model performance through variable selection. A number of methods have been proposed, including generating optimal linear PLS estimation-guided region selection [83], q^2 -guided region selection (q^2 -GRS) [80,84], a region-focusing approach available in SYBYL (Tripos), and so forth.

77

78

These separate but related methods did yield some improvement when implemented on small training data sets. However, no improvement in CoMFA performance was realized when region focusing was applied to a large and diverse data set [12].

CoMFA uses Lennard-Jones and Coulomb potentials, respectively, to calculate the steric and electrostatic field energies. A number of other fields have been used separately or in combination with the standard CoMFA fields in several studies [85–87]. In particular, several attempts have been made to augment the standard CoMFA fields with field descriptors representing hydrophobicity [85], other electrostatic characteristics (molecular orbital fields [88], E-state fields [86]), and hydrogen bonding [87]. For example, Kim [87] has reported the use of the direction-dependent 6 to 4 function of the GRID program [89] to generate hydrogen bonding fields. The hydrophobic interaction (HINT) technique [86] calculates the hydrophobicity field for a given molecule at each grid intersection point using an empirical equation:

$$A_i = S_i \times a_i \times R_{it} \quad (1)$$

where S_i and a_i are solvent-accessible surface area and hydrophobic atom constant for atom i , respectively, while $R_{it} = e^{-r}$ (r is the distance between atom i and field point t). Furthermore, Kellogg et al. suggest that E-state fields based on a 3D field can be used alone or in combination with HINT in CoMFA [86].

After reviewing 364 CoMFA models reported during the 1993–1996 period, Kim et al. [77] concluded that selection of the bioactive conformations and their alignments are the two most crucial steps in CoMFA. If bound ligand-receptor crystal structures are available for every chemical under study, these can readily be used as the bioactive conformation. Moreover, by superimposing corresponding amino acids in the receptor-active site, the alignment of the ligands is apparent. However, the availability of such a wealth of crystal data is a rare event, such that determination of the bioactive conformation(s) and alignment rules is critical to obtaining a high-quality CoMFA model.

Selection of the bioactive conformation

While in most cases the actual bioactive conformations for the training set chemicals are unknown, a single conformation for each chemical is normally chosen a priori. Lacking experimental data, alignment typically proceeds by aligning common-core structural frameworks. For a training set with a core framework, where the primary differences are in position and lengths of the side chains, the selection of the bioactive conformation will likely have minimal impact on the statistical performance of the resulting model, provided the side chains are consistently aligned.

If receptor-bound ligand structures are known for one or more chemicals, the bound conformations can be used as templates to determine the conformations of chemicals with corresponding structural frameworks [12]. Furthermore, identifying pharmacophores that are likely to be associated with activity can aid in selection of conformations for chemicals that are not similar to the templates [90]. Initial conformations can be determined by adjusting the rotatable bonds in such a way that key features (e.g., hydrogen bonds, hydrophobic centers) are proximal in alignment to that of the templates [91]. Next, a routine energy minimization is applied to derive the putative bioactive conformations, or the chemist's scientific

intuition is used to conjecture a higher-energy active conformation. In some cases, multiple ways exist to overlay key features on the templates, in which case conformation selection is integral to determination of the alignment rules [92]. For example, in the study of two classes of acetylcholinesterase inhibitors, *N*-benzylpiperidine benzisoxazoles and 1-benzyl-4-[2-(*N*-benzoyl-amino)ethyl]-piperidines (NBEPs), two conformations for the NBEPs were examined. With two distinct classes of chemicals involved, the first conformation maximized similarity on the basis of steric field and the other on the basis of electrostatic field. It was found that both conformations were plausible, given that the active site of the enzyme is relatively large and thus may allow inhibitors to bind in multiple conformations [92].

If no experimental data are available for a ligand-macromolecule complex, conformationally restricted chemicals and/or crystal structures of small molecules from the Cambridge Structural Database (CSD) can reasonably be taken as the starting points for ascertaining or inferring the bioactive conformations. It is also common practice to use the putative global minimum-energy conformation, which might be found by applying, for example, the following procedure [93–95]: Each molecule is optimized to its local minimum-energy conformation, the energy-minimized structure is subjected to systematic search over all rotatable bonds, and the molecule is re-minimized after each rotatable torsion angle is set to its low-energy conformation. In cases where multiple conformers are identified with similar energies, trial CoMFA models might be constructed for some or all possible combinations of these unique conformers. The combination yielding the statistically best model is usually retained. However, it is not uncommon to find examples where widely different choices of putative bioactive conformations have yielded CoMFA models with virtually identical statistical validity [96].

Alignment rules

In a CoMFA study, proper alignment of the molecules is critical yet often problematic. An optimal alignment can be defined as that achieving the maximum superposition of steric and electrostatic fields of a set of molecules. In reality, what is usually derived is a set of rules to be applied to a set of molecules, thereby ensuring consistency in the procedure. The alignment proceeds one molecule at a time, and alignment varies from molecule to molecule based on consideration of structural similarity or diversity. Alignment determines to what extent the steric and electrostatic fields differ from one molecule to the next. Hence, alignment substantially influences the results of the model, and significant and relevant results should be expected only for valid alignments.

For a set of congeneric chemicals, it is reasonable to assume that their desolvation energies and entropy effects will be approximately the same such that they can be aligned on the basis of their structural commonality. For a structurally diverse set of chemicals, no straightforward alignment rule based on such commonality exists. The identification of common structural features based on knowledge or assisted by computational software, such as CATALYST (Accelrys) or DISCO [97], may be helpful for selection of key pharmacophore elements for superposition [90]. For example, in a study of 130 diverse chemicals [12] that can be divided into more than six classes for estrogen receptor binding, six pharmacophore elements derived from the template molecule, 17 β -estradiol, were identified as important for estrogen receptor binding. When each

chemical class was aligned to the template based on the corresponding pharmacophore elements using a least-squares fitting method, the model exhibited good statistics and yielded accurate predictions for two external validation data sets.

In another recent study, Shim et al. [98] employed the DISCO module, accessed through the SYBYL program, to help identify the corresponding pharmacophore elements in a diverse series of CB₁ cannabinoid receptor ligands that included both classical cannabinoids and aminoalkylindoles (AAIs), which belong to two distinctly different chemical classes. This analysis enabled these workers to build a unified pharmacophoric map for the CB₁ cannabinoid receptor that encompassed both chemical classes. To extend the utility of this concept, Shim et al. [99] applied the concept of this unified pharmacophoric map to construct a unified CoMFA model for a mixed training set of cannabinoids and aminoalkylindoles that exhibited excellent self-consistency ($r^2 > 0.98$) and predictive ability ($q^2 \approx 0.5$).

Chemistry offers many examples of chemicals that pose unique challenges in terms of molecular alignment by virtue of their distinctive molecular structure, and the field of environmental toxicology is no exception. A case in point would be the polycyclic aromatic hydrocarbons (PAHs), which, unlike chemicals encountered in drug discovery programs, are notable for their flat geometry and absence of any notable pharmacophoric features. In such cases, ingenuity must be called on to derive a basis for molecular alignment. In this regard, Welsh et al. [100] employed the moments of inertia as a basis for aligning a large and structural diverse series of PAHs to construct separate CoMFA models for predicting their sublimation enthalpy and formation enthalpy. In a related study, Collantes et al. [101] demonstrated that alignment using moments of inertia could be extended to construct CoMFA models for predicting the chromatographic retention of these PAHs.

Because of its sensitivity to alignment, a CoMFA model is strongly dependent on the experience of the investigators. Largely for this reason, different investigators normally cannot reproduce CoMFA models. Realizing the difficulties associated with the manual alignment procedure, several efforts have aimed at development of an automatic alignment algorithm that systematically evaluates different alignment rules. The Steric and Electrostatic Alignment (SEAL) program offers an alternative approach reducing human judgment in alignment by using the rigid body alignment procedure. According to this algorithm, all molecules in the training set are aligned to a designated template molecule, with only steric and electrostatic interactions considered, although additional factors (e.g., hydrogen bonding, hydrophobicity) can be included. A fast Monte Carlo search procedure is used to identify all alignments that maximize overlap of both steric and electrostatic features. Satisfactory results are obtained using the following Lorentzian functional form to compute the similarity score A_F for a given alignment:

$$A_F = -\sum \sum w_{ij}/(1 + \alpha r_{ij}^2) \quad (2)$$

where the subscript i runs over the atoms in the first structure and j runs over the atoms in the second structure and r_{ij} is the distance between atom i of the first structure and atom j of the second structure. Two parameters need to be manually adjusted: the attenuation factor α , which controls the coarseness or fineness of molecular feature recognition, and the weighting factor, w_{ij} , which determines the percentage of the

contribution between steric and electrostatic interaction to the alignment. Obviously, SEAL is not an entirely automatic procedure but does offer a number of advantages compared to the full manual scheme [102]. For example, preassignment of atom–atom alignments are not required; thus, counterintuitive alignments are not eliminated as possibilities. Also, since all alignments are rank-ordered in accordance to a similarity score, modelers have the flexibility to test what they believe based on their understanding of the mechanism to be a more biological meaningful alignment, even though it might not rank the highest in score. Tong et al. [102] have compared results of manual and SEAL alignment for representative molecules from various chemical classes of estrogens in their minimum-energy conformation with respect to the template 17 β -estradiol. The superposition solutions for the two methods are very close for both steroids and phytoestrogens that have a backbone structure and overall shape similar to the template. However, salient differences exist in superposition solutions for those chemicals that are less similar to the template [95,102,103].

FlexS is a rapid and automatic approach for superimposing a target molecule to a template structure. It was developed by Lemmen et al. [104,105] and has been implemented in SYBYL software. FlexS first decomposes the target molecule into many small and relatively rigid fragments from which a base (or anchor) fragment is selected to place on the template molecule that is considered to be a rigid structure. Then the remaining fragments of the target molecule are incrementally added to the base fragment in a stepwise fashion to reconstruct the target molecule. At each construction step, flexibility is considered by allowing all possible conformations. Finally, the superposition quality for each solution is ranked using a scoring function. This is a rapid process with a mean computing time per target molecule of about 90 s on a standard present-day work station, making the method particularly suitable for virtual screening. The quality of superposition using FlexS is dependent on a number of factors, including the selection of the base fragment and the size of target molecules. A good base fragment should have a rigid structure with multiple pharmacophore sites (H-bonding, hydrophobic and salt bridges). Too flexible or too large of a target structure may result in poor superposition.

Other 3D QSAR approaches

The ubiquity and popularity of CoMFA gives testimony to the effectiveness of 3D QSAR methods for both drug discovery and toxicity prediction. However, a number of drawbacks are associated with the technology, most obviously the neglect of solvation/desolvation, receptor flexibility, ligand flexibility, and entropic effects.

Continued development of new 3D QSAR approaches and refinements of the existing 3D technologies are needed, and several notable efforts have already been made. Hopfinger et al. [106] developed a 4D-QSAR method with the aim of overcoming the weakness of lack of conformational flexibility. The method integrates conformational and alignment variability into the development of 3D QSAR models. A similar approach has also been reported by Vedani et al. [107,108].

Considered in some ways an extension of CoMFA, Comparative Molecular Similarity Indices Analysis (CoMSIA) replaces the field descriptors in CoMFA with descriptors that recognize the spatial (dis)similarity of aligned molecules [109]. In the CoMSIA approach, the similarity of molecules

is evaluated via the similarity of each molecule in the data set with a common probe atom that represents a certain type of property (e.g., steric, electrostatic). When CoMSIA and CoMFA were compared using several data sets, similar statistical results were observed for both approaches. However, unlike the Lennard-Jones potential used in CoMFA, the Gaussian function used in CoMSIA has no cutoff value for the interaction close to the molecular surface. Consequently, the color contour plots generated by CoMSIA may be biologically more meaningful [109].

QSARS BASED ON LIGAND-RECEPTOR INTERACTION

The recent growth of lab-on-chip (microarray and protein array) technology [110–112] and advanced recombinant DNA technology (cDNA cloning, Southern blotting, PCR, and so on) has in the past decade enabled rapid identification of biological targets as well as their expression at sufficient purity and quantities adequate for structure determination. Today, some 14,265 structures are available in the Brookhaven Protein Data Bank (on January 23, 2001), or three times as many as there were three years ago [113]. This dramatic increase in the availability of 3D structures for many macromolecules has greatly expanded the list of potential drug targets. This increase has also provided a rich source of information for QSAR and related computational techniques [114,115]. Scoring functions to estimate docking potential are often built into molecular modeling software programs, and such algorithms produce descriptors for ligand-receptor interaction that may be suitable for QSAR. Calculations of ligand-receptor interaction energies will likely play an important role in the future of QSARs [116].

Quantitative structure–activity relationship methods that incorporate information on ligand-receptor interactions have been investigated by a number of groups. The receptor coordinates are required either from crystal structure data or from homology modeling analysis, as is energy minimization of the bound ligand-receptor complex. VALIDATE [117] uses physicochemical properties of both ligands and ligand-receptor complexes as descriptors for QSAR. The descriptors representing the ligand-receptor complex encode information on the following: steric and energetic intermolecular interactions, ligand transfer from solution to the binding site, conformational entropy and enthalpy, and ligand-receptor contact surface areas.

The Comparative Binding Energy analysis (COMBINE) approach developed by Ortiz and Wade [118] also uses descriptors that encode ligand-receptor interactions for QSAR with PLS regression. The ligand-receptor interactions are decomposed according to physical type (van der Waals, electrostatic, and so on) for each interaction between defined fragments of the ligand and defined regions of the receptor. Even though the efficiency of using the decomposed intermolecular interaction energies as descriptors compared to other traditional descriptors for QSAR remains debatable, the recognition of the contribution of specific regions and/or fragments to the activity may be advantageous. Recently, DNA binding specificity was analyzed with COMBINE for 16 different DNA response elements and 20 mutant glucocorticoid receptors [119]. The COMBINE analysis indicated that the most important properties for determining binding specificity are the changes of the solvation free energies of the mutated base on binding, together with electrostatic interactions of the mutated nucleotides with certain charged amino acids.

Direct prediction of ligand–receptor binding affinity can

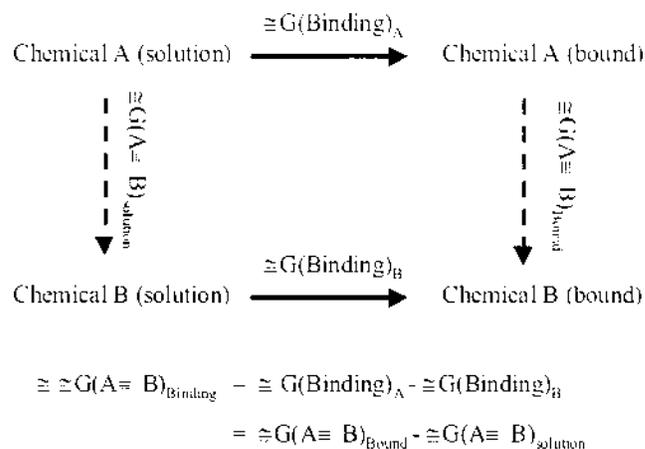


Fig. 1. Thermodynamic cycle used in free energy perturbation calculations.

also be accomplished simply by calculating the difference between the potential energy of the ligand–receptor complex and the potential energy of the ligand and receptor separately. Of course, the free energy of binding can be calculated more precisely using the Free Energy Perturbation (FEP) method [120]. In this method, the relative binding energies for pairs of chemicals are determined using a thermodynamic cycle in which the structure of one chemical (A) is perturbed into the structure of another (B), both in the receptor site and in solvent, as shown in Figure 1.

The nonphysical path between two states (or structures) A and B is simulated using thermodynamic integration of the relative free energy perturbation method, in which a coupling parameter λ is used to adequately describe a continuous conversion between two states. Thus, the relative binding free energies are given as the difference of two relative free energies: relative free energies of two ligands in solution versus the relative free energies of those two ligands bound to the receptor. This approach has been reported to yield relative free energies that are accurate to ± 1 kcal/mol when compared with experiment results [120]. However, the substantial amount of computer time required makes the method impractical for routine screening and assessment of many chemicals. Recently, Oostenbrink et al. [121] reported a single-step perturbation method allowing the calculation in a single simulation of relative free energy for a large number of polyaromatic hydrocarbons binding to the estrogen receptor α -subtype. The agreement between the calculated and experimental results had a maximum deviation of only 3.3 kJ/mol. Moreover, this method is between four and six times less computationally intensive as the thermodynamic integration method.

COMPARISON OF 2D AND 3D QSAR METHOD

Choosing between 2D and 3D QSAR for a particular problem depends on a number of factors. Since 2D QSAR requires no time-consuming structural alignment of the molecules such as required for CoMFA, it appears more suitable to support the high-speed technological processes that use combinatorial synthesis and HTS for lead discovery. The same would hold true in an environment that required rapid screening for toxicity of a large number of chemicals. Since some ADMET properties are directly associated with global properties of a molecule, 2D QSAR has often proved to be more efficient than 3D QSAR for this application [19]. On the other hand, 3D

QSAR is likely more accurate for a biological endpoint that is dependent on the geometry of key structural features. Generally, 3D QSAR is most successful for predicting small molecule–receptor interactions.

Another consideration for choosing between 2D and 3D QSAR is the value of information that is gained other than the value of prediction per se. For example, since 3D QSAR descriptors generally are lattice or surface grid descriptors, they might provide considerable insight into the SAR, such as the mechanism of action. This insight, in turn, with the aid of modern scientific visualization, can foster communication and understanding among multidisciplinary teams leading to the design of more efficacious and/or less toxic drug candidates.

Hoffman et al. [122] studied several 2D and 3D QSAR methods that were applied to a relatively small data set of 29 chemically diverse D1 dopamine antagonists. Two 2D QSAR methods, kNN-QSAR [58] and GA-PLS, were compared with the traditional and q^2 -GRS CoMFA [80]. All four approaches were found to yield reasonable predictive models with q^2 values of 0.57 for CoMFA, 0.54 for q^2 -GRS CoMFA, 0.73 for GA-PLS, and 0.79 for kNN-QSAR. However, relatively poor self-consistency was observed for GA-PLS, kNN-QSAR, and q^2 -GRS CoMFA with r^2 values of 0.72, 0.56, and 0.49, respectively. In contrast, traditional CoMFA yielded much higher r^2 value (0.94). One possible explanation for the poorer r^2 values typically obtained from the GA-based techniques is that their fitness function is mathematically dependent on q^2 but not r^2 . Hence, the fittest QSAR models generated by these GA-based techniques are aimed at maximizing q^2 even at the cost of lowering r^2 .

Tong et al. [67] compared the results from two 2D QSAR methods with those from CoMFA for three estrogenic activity data sets. The two 2D QSAR methods were the classic QSAR using descriptors generated from CODESSA and HQSAR. Data sets 1 and 2 contained the same set of structurally diverse molecules but differed with respect to biological endpoints. Data set 3 was composed of a set of congeners exhibiting several degrees of conformational flexibility. All three QSAR methods used PLS to derive the regression model; consequently, the only substantive difference among the three methods involved the nature of their chemical descriptors. Specifically, CoMFA employs steric and electrostatic field descriptors that encode detailed information concerning intermolecular interactions in a 3D grid surrounding each molecule. CODESSA program calculates molecular descriptors on the basis of 2D and 3D geometrical input optionally supplemented by quantum-chemical properties. HQSAR calculates exclusively fragment-based molecular descriptors. The statistical quality of the CoMFA and HQSAR models were comparable and generally better than the classical QSAR model. More recently, HQSAR and CoMFA were further compared by using a relatively large data set that contained 130 structurally diverse estrogens [12]. The models' performance was investigated using both internal validation and external validation. For the internal validation, the models' consistency and robustness were tested using a leave- N -out cross-validation procedure where the percentage of chemicals left out for prediction was up to 50%. In the external validation, two data sets containing over 40 chemicals not included in the training set were used to test the models' predictive ability. In both validation processes, CoMFA showed superior results over HQSAR.

MODEL VALIDATION

The current challenge in QSARs is no longer in constructing a statistically robust model but in developing a model with the capability to accurately predict the activity. The issue is how we quantify the quality of the model and validate that quality.

Generally, we can distinguish two types of model validation: internal validation and external validation. For internal validation, the quality of a QSAR model can be assessed in terms of several statistical measures. The values of r^2 and q^2 are normally accepted as measurements of the goodness of fit and predictive ability of the model, respectively. A model is generally deemed statistically significant if $r^2 \geq 0.9$. The value of q^2 is derived from a cross-validation procedure in which a fraction of chemicals in the training set are excluded and then predicted by the model generated from the remaining chemicals. When each chemical is left out one at a time and the process repeated for each chemical, this is known as leave-one-out (LOO) cross validation. If the training set is divided into N groups with approximately equal numbers of chemicals, the process is called leave- N -out (LNO) cross validation. Both LOO and LNO methods test the stability of the model through perturbation of the regression coefficients by consecutively omitting chemicals during the model generation procedure.

In recent years, q_{LOO}^2 derived from the LOO process has become the de facto measure of a model's predictive ability, that is, its ability to extrapolate beyond the training set [123]. A QSAR model with a value of $q_{LOO}^2 > 0.5$ is normally considered to possess significant predictive ability. A preferred approach for internal validation proposed on theoretical grounds [124] is to use the LNO method. While the LOO process is fast and reproducible, it tends to overestimate predictive ability compared with LNO [125]. Since the LNO procedure allows more chemicals to be omitted for prediction to test the model's stability, q_{LNO}^2 tests the ability of the model to extrapolate more so than does q_{LOO}^2 . Given the fact that chemicals in each left-out group are randomly selected from the training set in the LNO process, each LNO run will yield a different q_{LNO}^2 value. It has been proposed that it is necessary to run LNO 100 times for each random N groups ($N = 10$ is recommended) for a valid statistical analysis. A recent comparison of CoMFA with HQSAR [12] using LNO with N ranging from 2 to 65 (groups) of a training set composed of 130 chemicals (1.5% – 50% of the chemicals in the training set), each 100 times, showed markedly better extrapolation quality for CoMFA. While the mean q_{LNO}^2 values were consistently higher, indicating better extrapolation for CoMFA than for HQSAR, the decreases in q_{LNO}^2 with increasing N were very comparable for both methods. In addition, the standard deviation of q_{LNO}^2 was consistently smaller for CoMFA than for HQSAR, indicating that CoMFA provided much more robust QSAR models.

It is important to point out that although r^2 and q^2 are useful for validating the quality of a QSAR model, these parameters alone fail to account for other factors, particularly when PLS is used to construct a model. One factor is the number of PCs (or the number of descriptors in Hansch-type QSARs) used to construct a model that corresponds to the degrees of freedom. This factor holds particular importance when comparing different QSAR methods applied to the same data set. Since r^2 generally increases as more PCs are included in the model, it seems reasonable to scale r^2 by the number of PCs. Another

factor is the range of biological activity within the data set, which also should be considered during the comparison of the quality of QSAR models across different data sets. Given that two QSAR models have the same r^2 (or q^2) value, the model derived from the data set with the larger biological activity range is more valid than that with the smaller activity range.

Alternatively, the standard error and cross-validated standard error can be used as measures of goodness of fit and predictive ability [67]. While several ways exist to calculate the standard error for a regression equation, the number of degrees of freedom should be factored in when comparing different models. A more effective measure of model goodness of fit is the ratio of the standard error to the activity range. One advantage of explicitly including the range of biological activity is that the performance of separate QSAR models can be compared across different data sets.

Even if the model is validated as high quality by internal cross validation, uncertainty will remain regarding its ability to predict chemicals not in the training set. To address this question, external validation data sets are required. Most experts in the QSAR field, as well as the present authors, assert that a model's predictive capability can be fully tested and validated when robust prediction has been demonstrated with an external validation data set. Of course, one rarely enjoys the luxury of setting aside a sufficient number of test-set chemicals for use in external validation (10–20% of the data set is recommended) since in most cases data sets contain barely enough chemicals to create a statistically robust model in the first place. Furthermore, many data sets, such as those taken from *in vivo* studies or agricultural field tests [126], must first be pruned using creative measures to eliminate outliers that would otherwise obscure the underlying QSAR model.

Various studies have attempted to compare QSAR methods [67,68,122]. However, most comparisons have been made on small data sets. For example, the steroid data set reported in the original CoMFA paper [75] is often used as the benchmark data set for comparison with CoMFA [53]. Since this congeneric data set would probably not present a great challenge to most QSAR approaches, caution is warranted in judging the quality of a new method based solely on this comparison. Rather, a far more robust comparison would be demonstrated if a data set that contains a large number of structurally diverse chemicals with a wide range of activity was to be used.

FUTURE EXPECTATIONS

The current time has been called the golden age of biomedical research. Certainly, major breakthroughs in the understanding of the mechanisms of disease and toxicity in the postgenomics era are anticipated with much excitement. We see the prodigious data from microarrays and HTS being combined in large, integrated databases for exploitation with QSAR methods. New or enhanced QSAR methods should continue to evolve, enabled by ever-faster microprocessors and driven by the financial incentives to cost-effectively design safer, more efficacious drugs. Use of QSAR in toxicology, particularly in the regulatory arena, may lag the private sector, but ultimately computer-based prediction of toxicity, based on chemical structure alone, will become increasingly prevalent. The scenario is likely to unfold, mechanism by mechanism, organ by organ.

APPENDIX:

Common Statistical Approach Used In QSAR Study

Several statistical methods are available for QSAR study. If the number of chemicals and descriptors is small, then simple linear or multiple linear regression is the good choice. With a larger number of chemicals and larger pool of descriptors, partial least squares (PLS) or principal components regression (PCR) is either preferred or required. Additionally, several of the methods described here can be used in combination with various variable selection methods to develop more robust QSAR equation. For example, the combination of genetic algorithm (GA) with multiple linear regression (MLR) [52,127–129], PLS [56], and Artificial Neural Networks (ANNs) [130,131] is reported as effective approaches for QSAR and other applications.

Simple linear regression

The simple linear regression method correlates each individual descriptor with the activity using a standard linear regression calculation to generate a set of QSAR equations. This method is good for exploring simple relationships between structure and activity. Comparing goodness of fit measured commonly by the r^2 value, the approach can also be used as a simple tool for descriptor selection.

MLR

The MLR method calculates a QSAR equation by performing a standard multivariable regression calculation using a number of descriptors in a single equation. To avoid chance correlation, the number of descriptors should not be more than one-fifth the number of chemicals in the training set. The fewer descriptors that are used, the more robust and biological relevant the QSAR equation is. A trial-and-error approach might be used to determine a set of descriptor for a final QSAR equation. More common practice to select an optimal set of descriptors for a model is by implementing a variable selection technique in the model development process, such as GA.

PLS

PLS regression is particularly effective when the number of descriptors is large compared to the number of chemicals modeled. Many of the descriptor sets available in commercial software are both large and covariant. The PLS method reduces such large volume of descriptors to several principal components (PCs) or latent variables that are most correlative with the activity. The number of PCs defines a descriptor space of reduced dimension. Selecting the number of PCs for producing a most predictive model can be done in several ways. Normally, it is determined by the leave-one-out cross-validation procedure. The number of PCs can also be selected by the user, with three to seven seen as typical.

PCR

Principal components regression is a technique like PLS to handle the situation where the number of descriptors is larger than the number of chemicals modeled. Principal component analysis on the original descriptors is first done to derive several PCs that incorporate a significant portion of the variance in the descriptors. The PCs become the new descriptors that are correlated with activity using MLR. It is worthwhile to note that each PC is a linear combination of the original descriptors and describes only variance not contained in other

PCs; thus, the PCs are not covariant like the original descriptors. Additionally, the first few PCs are likely to explain most of the variance in the data set such that a few PCs (two or three) are normally used to construct a QSAR equation. An additional benefit of PCR is the ability to graphically observe which of the original descriptors contain a large amount of variance, which in turn can provide insights into the mechanism of action.

Stepwise MLR

The stepwise MLR method calculates quantitative structure-activity relationship (QSAR) equations by adding one descriptor at a time and testing each addition for significance in improving the result. This regression method is especially useful when the number of descriptors is large but the key descriptors are not known. As is always the case when using MLR, the number of descriptors should be much less (typically one-fifth) than the number of chemicals in order to minimize the possibility of a chance correlation. This technique could also be considered one type of variable selection method.

ANN

The ANN is a supervised learning method. It is best thought of as a nonparametric method to model complex response surfaces. Each neuron in a network applies a simple transformation to weighted sum of its inputs. By taking the outputs from several such neurons and using them as inputs to other neurons and so on, a very complex response can be modeled. One impressive feature of ANNs is that several methods to determine the weights on the sums exist, based solely on the data. This means that the form of the response model does not need to be specified beforehand. However, care must be taken to avoid overfitting the training set.

Acknowledgement—William J. Welsh wishes to acknowledge the financial support provided by a grant from the U.S. Environmental Protection Agency's Science to Achieve Results program.

REFERENCES

1. Hansch C, Leo A. 1995. *Exploring QSAR—Fundamentals and Applications in Chemistry and Biology*. American Chemical Society, Washington, DC.
2. Hansch C, Telzer BR, Zhang L. 1995. Comparative QSAR in toxicology: Examples from teratology and cancer chemotherapy of aniline mustards. *Crit Rev Toxicol* 25:67–89.
3. Bradbury S. 1995. Quantitative structure-activity relationship and ecological risk assessment: An overview of predictive aquatic toxicology research. *Toxicol Lett* 79:229–237.
4. Hopfinger AJ. 1997. Practical applications of computer-aided drug design. In Charifson SP, ed. *Practical Applications of Computer-Aided Design*. Marcel-Dekker, New York, NY, USA, pp 105–164.
5. Kubinyi H, Folkers G, Martin YC. 1998. 3D QSAR in drug design—Recent advances. *Perspect Drug Discov Des* 12:R5–R7.
6. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Delivery Rev* 23:3–25.
7. Bradbury SP. 1994. Predicting modes of toxic action from chemical structure: An overview. *SAR QSAR Environ Res* 2:89–104.
8. Russom CL, Bradbury SP, Carlson AR. 1995. Use of knowledge bases and QSARs to estimate the relative ecological risk of agrichemicals: A problem formulation exercise. *SAR QSAR Environ Res* 4:83–95.
9. Benigni R, Richard AM. 1998. Quantitative structure-based modeling applied to characterization and prediction of chemical toxicity. *Methods* 14:264–276.
10. Schultz TW, Seward JR. 2000. Health-effects related structure-

- toxicity relationships: A paradigm for the first decade of the new millennium. *Sci Total Environ* 249:73–84.
11. Hansch C, Hoekman D, Leo A, Zhang L, Li P. 1995. The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicol Lett* 79:45–53.
12. Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair R, Branham W, Sheehan D. 2001. QSAR models using a large diverse set of estrogens. *J Chem Inf Comput Sci* 4:186–195.
13. Tong W, Perkins R, Wu J, Shi L, Tu M, Fang H, Blair R, Branham W, Sheehan DM. 1999. An integrated computational approach for prioritizing potential estrogens. *Proceedings, International Symposium on Environmental Endocrine Disruptors*, December 9–11, Environment Agency, Government of Japan, Kobe, Japan, pp XX–XX.
14. Bone RGA, Firth MA, Sykes RA. 1999. SMILES extensions for pattern matching and molecular transformations: Applications in chemoinformatics. *J Chem Inf Comput Sci* 39:846–860.
15. Johnson M, Maggiora GM. 1990. *Concepts and Applications of Molecular Similarity*. John Wiley, New York, NY, USA.
16. Tetko IV, Aksenova TI, Volkovich VV, Kasheva TN, Filipov DV, Welsh WJ, Livingstone DJ, Villa AEP. 2000. Polynomial neural network for linear and nonlinear model selection in quantitative-structure activity relationship studies on the internet. *SAR QSAR Environ Res* 11:263–280.
17. McKinney JD, Richard A, Waller C, Newman MC, Gerberick F. 2000. The practice of structure activity relationships (SAR) in toxicology. *Toxicol Sci* 56:8–17.
18. Blaney JM, Martin EJ. 1997. Computational approaches for combinatorial library design and molecular diversity analysis. *Curr Opin Chem Biol* 1:54–59.
19. Hopfinger AJ, Duca JS. 2000. Extraction of pharmacophore information from high-throughput screens. *Curr Opin Biotechnol* 11:97–103.
20. Lipnick RL. 1999. Correlative and mechanistic QSAR models in toxicology. *SAR QSAR Environ Res* 10:239–248.
21. Wessel MD, Jurs PC, Tolan JW, Muskal SM. 1998. Prediction of human intestinal absorption of drug compounds from molecular structure. *J Chem Inf Comput Sci* 38:726–735.
22. Yoshida F, Topliss JG. 2000. QSAR model for drug human oral bioavailability. *Journal of Medical Chemistry* 43:2575–2585.
23. Dai J, Xu M, Wang L. 2000. Prediction of octanol/water partitioning coefficient and sediment sorption coefficient for benzaldehydes by various molecular descriptors. *Bull Environ Contam Toxicol* 65:190–199.
24. Gombar VK. 1999. Reliable assessment of log P of compounds of pharmaceutical relevance. *SAR QSAR Environ Res* 10:371–380.
25. Lewis DF, Ioannides C, Parke DV, Schulte-Hermann R. 2000. Quantitative structure-activity relationships in a series of endogenous and synthetic steroids exhibiting induction of CYP3A activity and hepatomegaly associated with increased DNA synthesis. *J Steroid Biochem Mol Biol* 74:179–185.
26. Lozano JJ, Pastor M, Cruciani G, Gaedt K, Centeno NB, Gago F, Sanz F. 2000. 3D-QSAR methods on the basis of ligand-receptor complexes: Application of combine and grid/golpe methodologies to a series of CYP1a2 ligands. *J Comput-Aided Mol Des* 14:341–353.
27. Contrera JF, Jacobs AC, DeGeorge JJ. 1997. Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals. *Reg Toxicol Pharmacol* 25:130–145.
28. Richard AM. 1998. Commercial toxicology prediction systems: A regulatory perspective. *Toxicol Lett* 102–103:611–616.
29. Klopman G. 1992. 1. A hierarchical computer automated structure evaluation program. *QSAR* 11:176–184.
30. Rosenkranz HS, Klopman G. 1990. Structural basis of carcinogenicity in rodents of genotoxicants and non-genotoxicants. *Mutat Res* 228:105–124.
31. Enslein K, Blake BW, Borgstedt HH. 1990. Prediction of probability of carcinogenicity for a set of ongoing ntp bioassays. *Mutagenesis* 5:305–306.
32. Enslein K, Gombar VK, Blake BW. 1994. International commission for protection against environmental mutagens and carcinogens: Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutat Res* 305:47–61.
33. Mekenyan OG, Kamenska V, Schmieder PK, Ankley GT, Bradbury SP. 2000. A computationally-based identification algorithm

210

22

- for estrogen receptor-ligands. Part II. Evaluation of a heRa binding affinity model. *Toxicol Sci* 58:270–281.
34. Bradbury SP, Kamenska V, Schmieder PK, Ankley GT, Mekenyang OG. 2000. A computationally-based identification algorithm for estrogen receptor-ligands. Part I. Predicting heRa binding affinity. *Toxicol Sci* 58:253–269.
 35. Sanderson DM, Earnshaw CG. 1991. Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum Exp Toxicol* 10:261–273.
 36. Ridings JE, Barratt MD, Cary R, Earnshaw CG, Eggington CE, Ellis MK, Judson PN, Langowski JJ, Marchant CA, Payne MP, et al. 1996. Computer prediction of possible toxic action from chemical structure: An update on the DEREK system. *Toxicology* 106:267–279.
 37. Woo YT, Lai DY, Argus MF, Arcos JC. 1995. Development of structure-activity relationship rules for predicting carcinogenic potential of chemicals. *Toxicol Lett* 79:219–228.
 38. Benfenati E, Gini G. 1997. Computational predictive programs (expert systems) in toxicology. *Toxicology* 119:213–225.
 39. Borman S. 1990. New QSAR techniques eyed for environmental assessments. *Chem Eng News* 68:20–23.
 40. Lipnick RL. 1986. Charles Ernest Overton: Narcosis studies and a contribution to general pharmacology. *Trends Pharmacol Sci* 7:161–164.
 41. Hammett LP. 1940. *Physical Organic Chemistry*. McGraw-Hill, New York, NY, USA.
 42. Hansch C, Leo A, Taft RW. 1991. A survey of Hammett substituent constraints and resonance and field parameters. *Chem Rev* 91:165–195.
 43. Hansch C. 1969. A quantitative approach to biological structure-activity relationships. *Accounts of Chemical Research* 2:232–239.
 44. Fujita T, Iwasa J, Hansch C. 1964. A new substituent constant, π , derived from partition coefficient. *J Am Chem Soc* 86:5175–5180.
 45. Hall LH. 1995. Molecular similarity based on novel atom-type electrotopological state indices. *J Chem Inf Comput Sci* 35:1074–1080.
 46. Hall LH. 1995. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci* 35:1039–1045.
 47. De Gregorio C, Kier LB, Hall LH. 1998. QSAR modeling with the electrotopological state indices: Corticosteroids. *J Comput-Aided Mol Des* 12:557–561.
 48. Hall LH, Kier LB. 2000. The E-state as the basis for molecular structure space definition and structure similarity. *J Chem Inf Comput Sci* 40:784–791.
 49. Rohrbaugh RH, Jurs PC. 1987. Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal Chim Acta* 11:220.
 50. Stanton DT, Jurs PC. 1990. Development and use of charge partial surface area structural descriptors in computer-aided quantitative structure-property relationship studies. *Anal Chem* 62:2323–2329.
 51. Shi LM, Tong W, Fang H, Perkins R, Wu J, Tu M, Blair R, Branham W, Waller C, Walker J, Sheehan D. 2002. An integrated “Four-Phase” approach for priority setting of endocrine disruptors—Phase I and II for prediction of potential estrogenic endocrine disruptor. *SAR/QSAR Environ Res* 13:69–88.
 52. Rogers D, Hopfinger AJ. 1994. Application of genetic function approximation to quantitative structure-activity relationship models. *J Chem Inf Comput Sci* 34:854–866.
 53. Coats EA. 1998. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect Drug Discov Des* 12:199–214.
 54. Clark DE, Westhead DR. 1996. Evolutionary algorithms in computer-aided molecular design. *J Comput-Aided Mol Des* 10:337–358.
 55. Forrest S. 1993. Genetic algorithms—principles of natural selection applied to computation. *Science* 261:872–878.
 56. Cho SJ, Cummins D, Bentley J, Andrews CW, Tropsha A. 2000. An alternative to 3D QSAR: Application of genetic algorithms and partial least squares to variable selection of topological indices. *J Comput Aided Mol Des* (in press).
 57. Hoffman BT, Kopajtic T, Katz JL, Newman AH. 2000. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of molconn z descriptors. *J Med Chem* 43:4151–4159.
 58. Zheng W, Tropsha A. 2000. A novel variable selection QSAR approach based on the *k*-Nearest Neighbor principle. *J Chem Inf Comput Sci* 40:185–194.
 59. Bruice TC, Kharasch N, Winzler RJ. 1956. A correlation of thyroxine-like activity and chemical structure. *Arch Biochem Biophys* 62:305–317.
 60. Leo A, Hansch C, Elkins D. 1971. Partition coefficients and their uses. *Chem Rev* 71:525–616.
 61. Ghose A, Crippen G. 1986. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity. *J Med Chem* 7:565–578.
 62. Ghose AK, Prichett A, Crippen GM. 1988. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. III. Modeling hydrophobic interactions. *J Comput Chem* 9:80–90.
 63. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. 1989. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application. *J Chem Inf Comput Sci* 29:163–172.
 64. Klopman G, Wang S. 1992. A computer automated structure evaluation (CASE) approach to calculation of partition coefficient. *J Comput Chem* 12:1025–1032.
 65. Turner D, Tyrell S, Willett P. 1997. Rapid quantification of molecular diversity for selective database acquisition. *J Chem Inf Comput Sci* 37:18–22.
 66. Rosenkranz HS, Cunningham A, Klopman G. 1996. Identification of a 2-D geometric descriptor associated with non-genotoxic carcinogens and some estrogens and antiestrogens. *Mutagenesis* 11:95–100.
 67. Tong W, Lowis DR, Perkins R, Chen Y, Welsh WJ, Goddette DW, Heritage TW, Sheehan DM. 1998. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J Chem Inf Comput Sci* 38:669–677.
 68. Silverman BD, Platt DE. 1996. Comparative molecular moment analysis (COMMA): 3D-QSAR without molecular superposition. *J Med Chem* 39:2129–2140.
 69. Platt DE, Silverman BD. 1996. Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching. *Journal of Computational Chemistry* 17:358–366.
 70. Todeschini R, Lasagni M, Marengo E. 1994. New molecular descriptors for 2D and 3D structures theory: Part 1. *J Chemom* 8:263–272.
 71. Todeschini R, Gramatica P, Provenzani R, Marengo E. 1995. Weighted holistic invariant molecular descriptors: Part 2. Theory development and application on modeling physico-chemical properties of polyaromatic hydrocarbons. *Chemom Intell Lab Syst* 27:221–229.
 72. Todeschini R, Gramatica P. 1997. 3D-modelling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies. *QSAR* 16:120–125.
 73. Ferguson AM, Heritage T, Jonathon P, Pack SE, Phillips L, Rogan J, Snaith PJ. 1997. A new theoretically based molecular description for use in QSAR/QSPR analysis. *J Comput-Aided Mol Des* 11:143–152.
 74. Ginn CMR, Turner DB, Willett P, Ferguson AM, Heritage T. 1997. Similarity searching in files of three-dimensional chemical structures: Evaluation of the EVA descriptor and combination of rankings using data fusion. *J Chem Inf Comput Sci* 37:23–37.
 75. Cramer RD, Patterson DE, Bunce JD. 1988. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110:5959–5967.
 76. Martin YC. 1998. 3D QSAR-current state, scope, and limitations. *Perspect Drug Disc Des* 12:3–23.
 77. Kim KH, Greco G, Novellino E. 1998. A critical review of recent CoMFA applications. *Perspect Drug Disc Des* 12:257–315.
 78. Horwitz JP, Massova I, Wiese TE, Besler BH, Corbett TH. 1994. Comparative molecular field analysis of the antitumor activity of 9h-thioxanthene-9-one derivatives against pancreatic ductal carcinoma 03. *J Med Chem* 37:781–786.
 79. Recanatini M. 1996. Comparative molecular field analysis of

- nonsteroidal aromatase inhibitors related to fadrozole. *J Comput-Aided Mol Des* 10:74–82.
80. Cho SJ, Tropsha A. 1995. Cross-validated r^2 -guided region selection for comparative molecular field analysis: A simple method to achieve consistent results. *J Med Chem* 38:1060–1066.
 81. Wiese TE, Brooks SC. 1994. Molecular modeling of steroidal estrogens: Novel conformations and their role in biological activity. *J Steroid Biochem Mol Biol* 50:61–73.
 82. Wang R, Gao Y, Lin L, Lai L. 1998. All orientation search and all-placement search in comparative molecular field analysis. *J Molecular Modeling* 4:276–283.
 83. Cruciani G, Watson KA. 1994. Comparative molecular field analysis using grid force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J Med Chem* 37:2589–2601.
 84. Cho SJ, Tropsha A, Suffness M, Cheng YC, Lee KH. 1996. Antitumor agents. 163. Three-dimensional quantitative structure-activity relationship study of 4'-*o*-demethylepipodophyllotoxin analogs using the modified CoMFA/q2-GRS approach. *J Med Chem* 39:1383–1395.
 85. Kellogg GE, Semus SF, Abraham DJ. 1991. HINT: A new method of empirical hydrophobic field calculation for CoMFA. *J Comput-Aided Mol Des* 5:545–552.
 86. Kellogg GE, Kier LB, Gaillard P, Hall LH. 1996. E-state fields: Applications to 3D QSAR. *J Comput-Aided Mol Des* 10:513–520.
 87. Kim KH. 1993. 3D-quantitative structure-activity relationships: Describing hydrophobic interactions directly from 3D structures using a comparative molecular-field analysis (CoMFA) approach. *Quant Struct-Act Relat* 12:232–238.
 88. Waller CL, Marshall GR. 1993. Three-dimensional quantitative structure-activity relationship of angiotensin-converting enzyme and thermolysin inhibitors. II. A comparison of CoMFA models incorporating molecular orbital fields and desolvation free energies based on active-analog and complementary-receptor-field alignment rules. *J Med Chem* 36:2390–2403.
 89. Goodford PJ. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28:849–857.
 90. Norinder U. 1995. The alignment problem in 3D-QSAR: A combined approach using catalyst and a 3D-QSAR technique. In Sanz F, Giraldo J, Manaut F, eds, *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*. Prous Science, Barcelona, Spain, pp 433–438.
 91. Tong W, Collantes ER, Welsh WJ, Berglund BA, Howlett AC. 1998. Derivation of a pharmacophore model for anandamide using constrained conformational searching and comparative molecular field analysis. *J Med Chem* 41:4207–4215.
 92. Tong W, Collantes ER, Chen Y, Welsh WJ. 1996. A comparative molecular field analysis study of *n*-benzylpiperidines as acetylcholinesterase inhibitors. *J Med Chem* 39:380–387.
 93. Tong W, Perkins R, Xing L, Welsh WJ, Sheehan DM. 1997. QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. *Endocrinology* 138:4022–4025.
 94. Tong W, Perkins R, Strelitz R, Collantes ER, Keenan S, Welsh WJ, Branham WS, Sheehan DM. 1997. Quantitative structure-activity relationships (QSARs) for estrogen binding to the estrogen receptor: Predictions across species. *Environ Health Perspect* 105:1116–1124.
 95. Waller CL, Oprea TI, Chae K, Park HK, Korach KS, Laws SC, Wiese TE, Kelce WR, Gray LE Jr. 1996. Ligand-based identification of environmental estrogens. *Chem Res Toxicol* 9:1240–1248.
 96. Howlett AC, Mukhopadhyay S, Shim JY, Welsh WJ. 1999. Signal transduction of eicosanoid cb1 receptor ligands. *Life Sci* 65: 617–625.
 97. Martin YC, Bures MG, Danaher EA, DeLazzer J, Lico I, Pavlik PA. 1993. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J Comput-Aided Mol Des* 7:83–102.
 98. Shim JY, Collantes ER, Welsh WJ, Howlett AC. 1999. Unified pharmacophoric model for cannabinoids and aminoalkylindoles derived from molecular superimposition of CB1 cannabinoid receptor agonists CP55244 and WIN55212-2. In Parrill AL, Reddi MR, eds, ACS symposium series on rational drug design: *Novel Methodology and Practical Applications*, ACS Symposium Series Volume American Chemical Society, Washington, DC, pp 165–184.
 99. Shim JY, Collantes ER, Welsh WJ, Howlett AC. 2000. Unified pharmacophoric model for cannabinoids and aminoalkylindoles. In Gundertofte K, Jorgensen FS, eds, *Molecular Modeling and Prediction of Bioactivity*. Kluwer Academic/Plenum, New York, NY, USA, pp 201–206.
 100. Welsh WJ, Tong WD, Collantes ER, Chickos JS, Gagarin SG. 1997. Enthalpies of sublimation and formation of polycyclic aromatic hydrocarbons (PAHs) derived from comparative molecular field analysis (CoMFA)—Application of moment of inertia for molecular alignment. *Thermochem Acta* 290:55–64.
 101. Collantes ER, Tong W, Welsh WJ, Zielinski WL. 1996. Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. *Anal Chem* 68:2038–2043.
 102. Tong W, Perkins R, Sheehan DM. 1999. Perspectives on three-dimensional quantitative structure-activity relationship (3D-QSAR)/comparative molecular field analysis (CoMFA) in determining estrogenic effects. *Japan Chemistry Today* 2:50–57.
 103. Walker JD, Fang H, Perkins R, Tong W. 2003. QSARs for EDPSD 2: The integrated 4-phase model. *QSAR Comb Sci* 22 (in press).
 104. Lemmen C, Lengauer T. 1997. Time-efficient flexible superposition of medium-sized molecules. *J Comput-Aided Mol Des* 11:357–368.
 105. Lemmen C, Lengauer T, Klebe G. 1998. FLEXs: A method for fast flexible ligand superposition. *J Med Chem* 41:4502–4520.
 106. Hopfinger AJ, Wang S, Tokarski JS, Jin G, Albuquerque M, Madhav PJ, Duraiswami C. 1997. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J Am Chem Soc* 119:10509–10524.
 107. Vedani A, McMasters DR, Dobler M. 1999. Genetic algorithms in 3D-QSAR: The use of multiple ligand orientations for improved predictions of toxicity. *Altex Arch* 16:142–145.
 108. Vedani A, Briem H, Dobler M, Dollinger H, McMasters DR. 2000. Multiple-conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. *J Med Chem* 43:4416–4427.
 109. Klebe G. 1998. Comparative molecular similarity indices analysis-CoMSIA. In Kubinyi H, Folkers G, Martin YC, eds, *3D QSAR in Drug Design*, Vol 3. Kluwer, Dordrecht, The Netherlands, pp 87–104.
 110. Sosnowski RG, Tu E, Butler WF, O'Connell JP, Heller MJ. 1997. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proc Natl Acad Sci USA* 94:1119–1123.
 111. Timm GE, Darr JF, Flessner CJ, Kennedy PW, Maciorowski AF, O'Bryan TR, Walker JD. 2002. Priority setting and framework for endocrine disruptor screening. In: Katz SA, Salem H, eds, *Alternative Toxicological Methods for the New Millennium*. CRC, Boca Raton, FL, USA (in press).
 112. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. 2000. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 24:236–244.
 113. Rognan D. 1998. Molecular dynamics simulations: A tool for drug design. In Kubinyi H, ed, *3D QSAR in Drug Design*, Vol 2. Kluwer, Dordrecht, The Netherlands, pp 181–209.
 114. Kuntz ID. 1992. Structure-based strategies for drug design and discovery. *Science* 257:1078–1082.
 115. Blundell TL. 1996. Structure-based drug design. *Nature* 384(Suppl. 6604):23–26.
 116. Jayatilke PR, Nair AC, Zauhar R, Welsh WJ. 2000. Computational studies on HIV-1 protease inhibitors: Influence of calculated inhibitor-enzyme binding affinities on the statistical quality of 3D-QSAR CoMFA models. *J Med Chem* 43:4446–4451.
 117. Head RD, Smythe ML, Oprea TI, Waller CL, Green SM, Marshall GR. 1996. Validate: A new method for the receptor-based prediction of binding affinities of novel ligands. *J Am Chem Soc* 118:3959–3969.
 118. Ortiz AR, Pisabarro MT, Gago F, Wade RC. 1995. Prediction of drug binding affinities by comparative binding energy analysis. *J Med Chem* 38:2681–2691.
 119. Tomic S, Nilsson L, Wade RC. 2000. Nuclear receptor-DNA

24

25

- binding specificity: A COMBINE and Free-Wilson QSAR analysis. *J Med Chem* 43:1780–1792.
120. Kollman P. 1993. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem Rev* 93:2395–2417.
 121. Oostenbrink BC, Pitera JW, van Lipzig MM, Meerman JH, van Gunsteren WF. 2000. Simulations of the estrogen receptor ligand-binding domain: Affinity of natural ligands and xenoestrogens. *J Med Chem* 43:4594–4605.
 122. Hoffman BT, Cho SJ, Zheng W, Nicols DE, Tropsha A. 1999. Quantitative structure-activity relationship modeling of dopamine D1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and *k* Nearest Neighbor methods. *J Med Chem* 42:3217–3226.
 123. Wold S. 1991. Validation of QSARs. *Quant Struct-Act Relat* 10:191–193.
 124. Shao J. 1993. Linear model selection by cross-validation. *J Am Stat Assoc* 88:486–494.
 125. Baroni M, Costantino G, Cruciani G, Riganelli D, Valigi R, Clementi S. 1992. Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant Struct-Act Relat* 12:9–20.
 126. Collantes ER, Xing L, Miller PC, Welsh WJ, Profeta S Jr. 1999. Comparative molecular field analysis as a tool to evaluate mode of action of chemical hybridization agents. *J Agric Food Chem* 47:5245–5251.
 127. Rogers D. 1994. Genetic function approximation: A genetic approach to building quantitative structure-activity relationship models. In Sanz F, Giraldo J, Manaut F, eds, *QSAR and Molecular Modelling: Computational Tools and Biological Applications*. Prous Science, Barcelona, Spain, pp 420–426.
 128. Rogers D. 1995. Development of the genetic function approximation algorithm. *Proceedings*, 6th International Conference on Genetic Algorithms. Kaufmann, San Mateo, CA, USA, pp 589–596.
 129. Rogers D. 1996. Some theory and examples of genetic approximation with comparison to evolutionary techniques. In Devillers J, ed, *Genetic Algorithms in Molecular Modeling*. Academic, London, UK, pp 87–107.
 130. Parbhane RV, Unniraman S, Tambe SS, Nagaraja V, Kulkarni BD. 2000. Optimum DNA curvature using a hybrid approach involving an artificial neural network and genetic algorithm. *J Biomol Struct Dyn* 17:665–672.
 131. Dybowski R, Weller P, Chang R, Gant V. 1996. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet* 347:1146–1150.