Environmental Toxicology and Chemistry, Vol. 22, No. 8, pp. 000–000, 2003
© 2003 SETAC
Printed in the USA
0730-7268/03 $12.00 + .00

Annual Review

STRUCTURE–ACTIVITY RELATIONSHIP APPROACHES AND APPLICATIONS

WEIDA TONG,† WILLIAM J. WELSH,‡ LEMING SHI,§ HONG FANG,‖ and ROGER PERKINS*‖
†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, 675 Hoes Lane, Piscataway, New Jersey 08854, USA
§BASF Corporation, P.O. Box 400, Princeton, New Jersey 08543–0400, USA
‖Logicon ROW Sciences, 3900 NCTR Road, MC 910, Jefferson, Arkansas 72079, USA

(*Received* 23 *April* 2001; *Accepted* 2 *January* 2003)

?1

Abstract—New techniques and software have enabled ubiquitous use of structure–activity relationships (SARs) in the pharmaceutical industry and toxicological sciences. We review the status of SAR technology by using examples to underscore the advances as well as the unique technical challenges. Applying SAR involves two steps: Characterization of the chemicals under, and application of chemometric approaches to explore data patterns or to establish the relationships between structure and activity. We describe generally but not exhaustively the SAR methodologies popular in use toxicology, including representation of chemical structure, and chemometric techniques where models are both unsupervised and supervised. The utility of SAR technology is most evident when supervised methods are used to predict toxicity of untested chemicals based only on chemical structure. Such models can predict on both an ordinal scale (e.g., active vs inactive) or a continuous scale (e.g., median lethal dose [LD50] dose). The reader is also referred to a companion paper in this issue that discusses quantitative structure–activity relationship (QSAR) methods that have advanced markedly over the past decade.

Keywords—Structure–activity relationship Predictive toxicology Computational toxicology Chemometrics

INTRODUCTION

In recent years, the pharmaceutical industry has invested a significant amount of intellectual and monetary capital in combinatorial chemistry, high throughput screening (HTS), and both microarray and protein array systems. These highly automated technologies have quite literally revolutionized the drug discovery paradigm in both dimension and scale. By enabling the rapid synthesis and biological evaluation of new chemical entities, combinatorial chemistry and HTS have led to the creation of vast libraries of ‘‘drug like’’ chemicals and associated biological data [1,2]. Microarray and protein array systems are just as rapidly discovering genes and proteins, many of which will serve as the key targets for drugs that treat unmet medical needs [3–10].

Given the explosion of data culminating from these aforementioned technologies, structure–activity relationship (SAR) methods have become increasingly essential as tools for organizing, mining, and interpreting these data to guide further experimentation and discovery. At the same time, these new experimental paradigms have made a profound impact in recent years on the practical utility of SAR techniques. Driven by the mandate to process the continually expanding body of chemical and biological data, speed is replacing accuracy as the criterion of paramount importance for SAR techniques. This trend will surely persist for the foreseeable future [11].

Structure–activity relationship techniques are currently employed in a wide range of applications, including: In silico design of virtual chemical libraries that explore molecular diversity for subsequent synthesis and screening [1,12–14]; screening proprietary, commercially available, and public databases for lead discovery [15–24]; and, mining gene expression data from microarray experiments for target identification [25]. It is obvious from these examples that SAR technology now fulfills expanding roles in handling large and expanding sources of data. The success of drug discovery efforts within the pharmaceutical industry depends heavily on utilization of SAR techniques for these and related purposes.

In contrast to drug discovery, the fields of toxicity screening and environmental risk assessment have been impacted to a smaller degree by SAR techniques up to now [26]. In this review, we explore the status of SAR technology by using examples to underscore the advances as well as the unique technical challenges. Our primary objective is to bring a general awareness of currently popular SAR technology to the wider toxicology community. No attempt is made to survey all available methods or to exhaustively cover the extensive published literature on this topic. Rather, emphasis is placed on methods that are representative or that the authors have learned through direct experience.

Structure–activity relationship technology is based conceptually on the ‘‘similar property’’ principle [27], which states that chemicals with similar structures are likely to exhibit similar biological activities. The general procedure in applying SAR involves two steps: Characterization of the molecules under investigation using computational, chemical, and biological methods, and application of chemometric approaches to explore data patterns or to establish the relationships between structure and activity (or property).

CHARACTERIZATION OF CHEMICALS

Chemicals can be characterized at three different levels, as shown in Figure 1: Molecular structure (S); physicochemical properties (P); and biological activity (A). Molecular structure

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

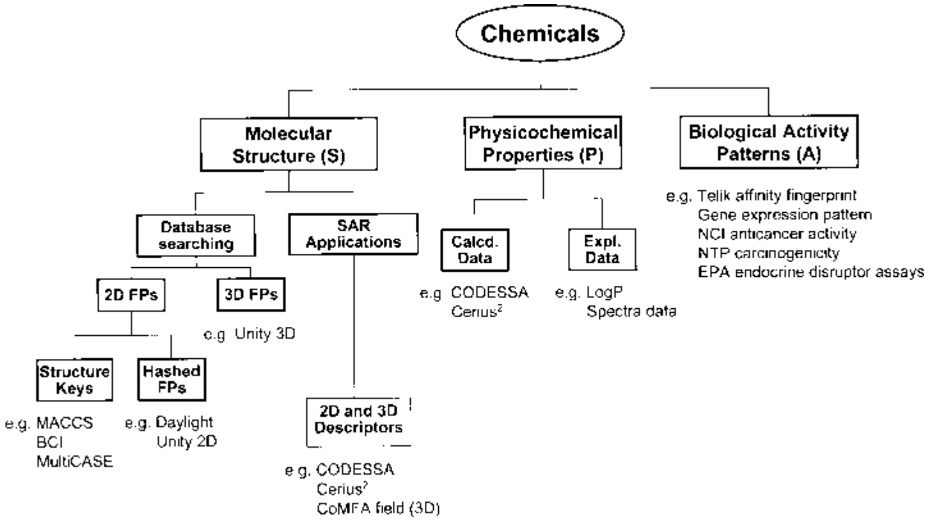

Fig. 1. Three levels of approach to characterize chemicals.

can be represented by a variety of structural descriptors, while a chemical's physicochemical properties can be either measured by experimental methods or calculated using computational approaches. The biological response induced by a chemical can be tested by a variety of assay techniques. One can argue from the principles of chemistry that the molecular structure of a chemical is key to understanding its physicochemical properties and ultimately its biological activity and influence on organisms. Because both S and P are associated with the chemical itself, the relationship between S and P should be apparent and, therefore, more accessible using SAR techniques. In contrast, the A of a chemical is an induced response that is influenced by numerous factors dictated by the level of biological complexity of the system under investigation. The relationship between S (and P or both S and P) and A is thus more implicit and thereby poses a more challenging problem in SAR applications.

It is important to note that these three levels of chemical information can be used separately or together to predict the biological effects of chemicals on animals, including humans. Needless to say, this is an extremely critical yet daunting quest in toxicological research that deserves, and is now attracting, considerably more attention. Our experience suggests that no single level of chemical information is likely sufficient for assessing or predicting a chemical's diverse effects on a complex biological system. Rather, this elusive goal will be reached only by pooling all three levels of chemical information.

Molecular structure (S)

Structural descriptors can be considered the most fundamental information about a chemical's composition. Considerable effort has focused on exploring different paradigms and techniques to rapidly generate structural descriptors that, hopefully, also encode for biological activity. With respect to their intended application, the computer-generated descriptors can be divided into two separate, yet related categories: database searching and SAR applications.

Database searching. Most of the techniques in this category were developed for managing or mining 2- or 3-dimensional (2D or 3D) databases of chemical structures. A chemical structure can be encoded as a binary string called a structural fingerprint (FP). Each element of the FP denotes the absence or presence of a specific structural feature, which can be either a 2D substructure or a 3D feature. By representing a chemical database in terms of structural FPs, we can accelerate structural searches against queries through similarity comparison. In most cases, the majority of database entries can be eliminated from further consideration by simply comparing the structural FPs of a query with those of each database entry.

Two-dimensional FPs were originally designed for searching 2D structural databases which encode each molecule merely in terms of atom types and their connectivities. Two-dimensional FPs are independent of the molecule's conformation, an attribute that greatly facilitates speed in their computation and versatility in their use. Two-dimensional FPs can be further divided into two subclasses: 2D structural keys and 2D hashed FPs.

With regard to 2D structural keys, the first step is to define a set of structural keys (i.e., fragments) after which each chemical can be indexed using a binary code (1,0) in terms of the presence (1) or absence (0) of these keys. The database representing the presence or absence of this set of keys is usually very sparse because most of the data entries are zeros. The criteria employed in selecting an appropriate set of structural keys for efficient database searching is a topic of considerable discussion and debate [28–30]. The MACCS 2D-database system from MDL (San Leandro, CA, USA; http://www.mdli.com/), in which 167 keys have been carefully selected, is perhaps the widely employed set of 2D structural keys in database searching. Another example is Barnard Chemical Information's Program MAKEBITS (Sheffield, UK), which automatically generates all possible fragments of user-defined length and type and then selects a subset through statistical analysis of the distribution of the keys [31,32]. A drawback common to all 2D structure-key representations is that practical limitations in disk storage space impose restrictions on the number of keys that can be considered out of the almost countless possibilities.

Two-dimensional hashed FPs [33] were conceived in part to solve the storage problem. A hashing algorithm maps each fragment of a structure to a specific location in the FP. To apply this method, each fragment of user-defined length (i.e., atom count) and atom-type is assigned an integer ID number that is generated by a cyclic redundancy check (CRC) algorithm [34]. The ID is used to identify the bin in the FP. Each fragment can turn on one bit, and the resulting FP consists of a string of 1's and 0's. The length of the FP is usually adjustable and much shorter than the number of possible fragments from a typical database of nontrivial chemicals. Therefore, coding a database of chemicals using hashed FP representation is much more compact than that using 2D structure-key representation. Two of the most widely used hashed FPs are Daylight's FP (Mission Viejo, CA, USA; http://www.daylight.com/) and the Unity 2D FP from Tripos (St Louis, MO, USA; http://www.tripos.com/). Because each chemical can be decomposed into many fragments, in most cases each chemical will set many bits in the hashed FP to the on status. When the length of the hashed FP is sufficiently long, it is unlikely that two different chemicals will generate FPs with identical patterns. However, it is not uncommon that different fragments from the same chemical will set the same bit on; this is the so-called collision problem. The occurrence of collisions makes it impossible to establish a one-to-one mapping between fragments and the settings of the bits. Even if two chemicals set the same bit on, it is not guaranteed that they share the same fragments. Because of the collision problem, employing the 2D hashed FPs for (dis)similarity searching is problematic [16].

As mentioned above, 2D FPs encode molecular structural information in only two dimensions. The relative position of atoms in 3D space is not encoded, even though the 3D configuration of atoms in a molecule is paramount in determining its biological activity (e.g., through ligand-receptor binding). Three-dimentional FPs were originally developed mainly for the purpose of 3D pharmacophore searching [35,36]. The 3D FP, which is similar in some ways to the 2D FP, primarily encodes distance information between atoms that are deemed relevant to biological activity. Information on atom types can also be encoded. The FP is segmented into bins, with each bin holding distance information for a specific pair of atom types. Each bin consists of many bits that determine the distance range in which the pair belongs. An important step in constructing a 3D FP is to define the atom types to be encoded. It is a time-consuming operation to fingerprint a database of chemicals and then search the resulting 3D database using a 3D FP. An example of 3D FPs is the Unity 3D FPs from Tripos.

Brown and Martin evaluated the information content of 2D and 3D FPs relevant to ligand-receptor binding [16]. For their 2D FPs, these workers selected two 2D structural keys (MACCS keys with 153 bits, SSKEYS with 960 bits) and two 2D hashed FPs (Daylight and Unity). The Daylight 2D hashed FP hashes fragments with two to seven atom counts to 1040 bits while the Unity 2D hashed FP (http://www.tripos.com) hashes fragments with two to six atom counts to 992 bits. The 3D FPs examined in the study were: (1) the rigid Unity 3D FP that encodes distances between pairs of heteroatoms, ring centroids and normals, and carbonyl extension points; (2) the flexible Unity 3D FP which is similar to the rigid Unity 3D FP but allows for rotation of all rotatable bonds; (3) the so-called ppp pairs that encode distances between pairs of predefined features (e.g., hydrogen bond acceptors and donors, positive and negative charges, and hydrophobic centroids); and (4) the ppp triangles that encode all distances between triplets of features in the ppp pairs. The relevance of these FPs to ligand-receptor binding were assessed indirectly through examination of the relationship between the FPs and the binding force (e.g., hydrophobic, dispersion, electrostatic, steric, and hydrogen bonding interactions). One of the conclusions from their study was that 3D FPs did not perform well in separating active chemicals from inactive chemicals. Their findings by no means indicate that 3D information is less valuable, but rather that the current methods for encoding 3D information leave room for improvement.

The CASE/MultiCASE Program [37,38], a popular tool employed in computational toxicology, is most closely related to the concept of structural keys. This program automatically generates a list of structural fragments embedded in a molecule, in which the keys (biophores or biophobes) have been preselected based on their relevance to the bioactivity under consideration. The statistical significance of the distribution of the substructural keys present in active versus inactive training chemicals is employed as the selection criterion. Therefore, (qualitative) knowledge of the biological activity of the training set is required to derive the set of keys. To obtain a set of high-quality keys generally requires a large training set of chemicals.

SAR applications. Although database searching employs FPs, SAR applications are more interested in abstract descriptors that directly encode 2D and 3D molecular structure information. The validity of SAR applications will depend heavily on the quality and information content of these descriptors. A large number of molecular descriptors have been reported in the literature, and many commercial software products are available for calculating scores of descriptors. For example, the COmprehensive DEscriptors for Structural and Statistical Analysis (CODESSA) package (http://www.semichem.com/) calculates more than 400 molecular descriptors, categorized as: (1) constitutional, (2) topological, (3) geometrical, (4) electrostatic, (5) quantum–chemical, and (6) thermodynamic. Cerius^2 from MSI (http://www.msi.com) calculates more than 250 descriptors categorized as: (1) conformational, (2) electronic, (3) information content, (4) quantum–mechanical, (5) receptor related, (6) shape related, (7) spatial, (8) thermodynamic, and (9) topological. The interested reader can find a more detailed discussion of these descriptors in our companion paper in this volume, ''Quantitative Structure–Activity Relationship (QSAR) Methods: Perspectives on Drug Discovery and Toxicology'' (Perkins et al., this issue). These descriptors have been used for a wide range of applications, including recently for diversity analysis of combinatorial libraries and for QSAR/QSPR studies [39–43].

Recognizing the critical role of molecular shape in determining most ligand–receptor interactions, descriptors based on the 3D structural information of molecules have proven highly successful in a variety of applications. Comparative Molecular Field Analysis (CoMFA), a 3D-QSAR technique, is widely recognized as a versatile and powerful tool for a broad range of applications including rational drug design [40,44–47]. The descriptors generated by CoMFA correspond to the steric and electrostatic interaction energies between a probe atom (e.g., Csp3+1) and the molecule at every intersection of a 3D grid within a box that encompasses the molecule.

Physicochemical properties (P)

The physicochemical properties of a molecule are related to its molecular structure. Unlike the familiar Hammett substituent constants [48] that describe properties of only part of a molecule, physicochemical property descriptors are global in that they describe the nature of the whole molecule. These properties are generally suitable for characterizing structurally diverse sets of chemicals, making it possible to conduct QSAR analysis on noncongeneric datasets. Although some of these property data can be measured directly in the laboratory, it is more common and generally more convenient to compute them using theoretical calculations. Most molecular structural descriptors used in SAR applications can be broadly defined under the heading of physicochemical descriptors.

Foremost among physicochemical properties, Log*P* is recognized as a key parameter for virtually all biological systems. A strong correlation between biological activity and Log*P* is frequently observed; however, the variation of biological activity with Log*P* for a set of drug-like molecules against a common target is more often parabolic than linear [49]. This observation is not unexpected because molecules which are either too hydrophobic (high Log*P* value) or too hydrophilic (low Log*P* value) are generally not good drug candidates [50]. In a SAR study of a large and diverse set of natural, synthetic and environmental estrogens, Fang et al. [51] found that Log*P* is only important in the estrogen receptor (ER) binding activity when the key pharmacophores are already present. Where a direct comparison can be made, strong estrogens tend to be more hydrophobic. Many studies have been reported which further our understanding of Log*P* [52–54].

Direct measurement of Log*P* requires a tedious and laborious process that includes chemical synthesis followed by wet-laboratory experiments, a process that is not practical for routine processing of large numbers of chemicals. Hence, many methods for estimating Log*P* computationally based on a chemical's structure have been proposed [52,55–69], some of which are available through commercial software. For example, the fragment-based Clog*P* method developed by Hansch and Leo [52,55–57] and revised by Rekker [61] is available in Sybyl (http://www.tripos.com), while the atom-based Alog*P* method described by Ghose and Crippen [58–60] is available in Cerius^2 (http://www.msi.com). By virtue of its accuracy and low price, the atom/fragment contribution method known as logK_{ow} (http://esc.syrres.com/) developed by Meylan and Howard [70] has also attracted many users within the toxicology community. A new method for estimating Log*P* was recently developed by Tetko et al. [71] that is accessible online (http://www.lnh.unil.ch/~itetko/logp). The particular method selected for calculating the Log*P* when handling large chemical databases often depends on the desired level of accuracy and applicability of the method to the chemicals under investigation. Unlike logK_{ow}, for example, Clog*P* encounters difficulties in decomposing structures into appropriate fragments whose constants are not available. This is often described as the ‘‘missing fragment’’ problem [70].

Spectral data, for example, ultraviolet, infrared (IR), nuclear magnetic resonance (NMR), and mass spectroscopy (MS), can be reasonably well predicted from molecular structures and readily measured experimentally. This type of information, if accessible, can be used for the characterization of chemicals in QSAR studies of diverse datasets [72,73]. Using experimentally determined ^{1}H NMR, MS, IR spectra, and simulated IR and ^{13}C NMR spectra, Bursi et al. [74] performed QSAR studies on 45 diverse progestagens by means of partial least squares. Beger et al. [75] developed a classification model for 108 estrogens using experimental ^{13}C NMR and mass spectrometric data.

Biological activity patterns (A)

Chemicals can be tested across many endpoints, either pharmacologically or toxicologically. The pattern of activity across these endpoints encodes information about the biological nature of the tested chemical. The advent of HTS techniques, together with the large and growing diversity of chemicals obtained from combinatorial libraries or from other chemical collections (e.g., natural products), has assured that the vast arrays of activity data will only grow in importance and familiarity within both the pharmaceutical industry and the regulatory agencies. Examples of endpoints include ligand binding affinities, gene expression profiles, in vitro cell culture responses, and in vivo animal testing results, each of which represents a distinct level of biological complexity.

For example, Terrapin Technologies (http://www.terrapintech.com/), now renamed as Telik (http://informagen.com/Resource_Informagen/Full/3349.html), has developed a method for predicting ligand binding to proteins by affinity fingerprinting from a small panel of reference proteins [76,77]. Following preliminary testing of over 300 proteins from a variety of sources, the authors selected panels of 8 to 18 proteins that displayed the broadest binding affinities for a set of over 5,000 chemicals. Computational surrogates, modeled by multiple linear regression, were built based on the affinity fingerprints of a training set of molecules across the reference set of proteins. These models can be used to predict the binding potencies of additional chemicals. The affinity fingerprint database, which provides a rich source of data defining operational similarities among proteins, is useful for efficient prescreening of a large number of chemicals against target proteins in order to select promising candidates for further study.

Using DNA chip technology [5,6,78–80], changes in gene expression in a tissue or cell caused by a chemical can now be measured using cDNA or oligonucleotide microarrays. Such gene expression profiles provide a significantly large body of data to understand chemical-dependent toxicity, drug efficacy, or both. For example, Scherf et al. [25] used cDNA microarrays to assess gene expression profiles in 60 human cancer cell lines and to relate these profiles with the activity patterns responding to drugs. A cluster analysis of this expression data using K-mean is also illustrated in Figure 2.

One of the most notable examples in this category is the U.S. National Cancer Institute anticancer drug screening program (http://dtp.nci.nih.gov/), in which each chemical is tested in vitro for anticancer activity against a panel of 60 human cancer cell lines [81–83]. More than 63,000 chemicals have been tested in this screening program. The anticancer activity of each chemical against each of the 60 cell lines is expressed as $-\log(\text{GI50})$, where GI50 is the concentration in mol/L required to inhibit cancer cell growth by 50% compared with untreated controls. The activity pattern for each chemical is composed of 60 such values. Although cell growth inhibition for a single cell line is not informative, activity patterns across the 60 cell lines have been shown to encode critical information about the molecular and biological properties of tested chemicals [42,84–88]. Various statistical and artificial intelligence methods have been used to mine this large database for the discovery of anticancer drugs and for studies of the molecular pharmacology of cancer [42,86–91].

The U.S. National Toxicology Program (http://ntp-server.niehs.nih.gov/), successor to the National Cancer Institute Bioassay Program (1971–1978), initiated long-term animal assays for carcinogenicity of selected chemicals in 1978. Chemicals are chosen primarily on the basis of human exposure, level of production, and chemical structure; selection per se is not an indicator of a chemical's carcinogenic potential. Generally, each chemical is tested in two species (rats and mice) in both

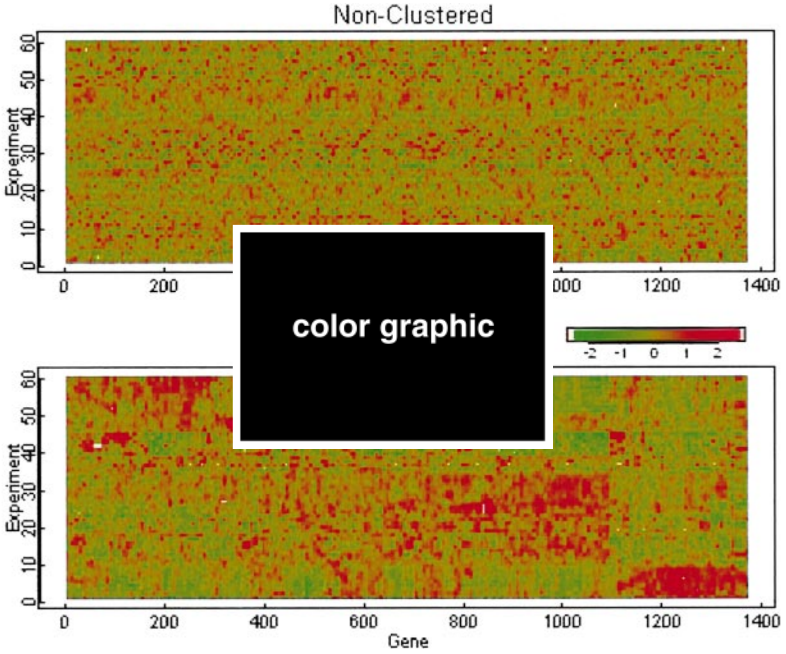

Fig. 2. Gene expression pattern in the 60 human cancer cell lines used in a drug discovery screen by the National Cancer Institute based on clustering analysis. The *Y*-axis contains 60 microarray experiments associated with 60 different cancer cell lines. The *X*-axis contains the genes being evaluated in the experiment. Red points indicate the upregulated genes while green points the down-regulated genes. The expression data is publicly available from http://genome-www.stanford.edu/nci60.

sexes over a period of two years. Earlier studies recorded the level of carcinogenicity using three levels: Positive, equivocal, and negative. Since 1983, five categories of evidence of carcinogenic activity have been used in the U.S. National Toxicology Program Technical Report series to summarize the strength of the evidence observed in each experiment: Two categories for positive results (clear evidence and some evidence); one category for uncertain findings (equivocal evidence); one category for no observable effects (no evidence); and one category for experiments that because of major flaws cannot be evaluated (inadequate study). Combined with the results from mutagenic assays, including the Ames *Salmonella* assay, a pattern of carcinogenicity/mutagenicity for each tested chemical is documented. This pattern characterizes the biological features of a tested chemical. This rich source of information should prove important for human health risk assessment [92].

Another example is the ongoing project for the identification of endocrine disrupting chemicals (EDCs) coordinated by the U.S. Environmental Protection Agency (U.S. EPA) (http://www.epa.gov). Growing concern exists among the scientific community, government regulators, and the public that EDCs in the environment are adversely affecting the health of humans and wildlife by disrupting endocrine function [93,94]. The scientific debate surrounding EDCs has grown contentious, in part due to the fact that some suspected EDCs are economically important chemicals produced in high volume. The public and regulatory concerns led to government regulatory actions and expanded research across Europe, Japan, and North America [95,96]. In response to Congressional mandate, the U.S. EPA established the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC); EDSTAC recommended a plan to screen and test for estrogenic, androgenic, and thyroid endpoints for a large number of chemicals. To accomplish this, chemicals will be screened (Tier I) using a multiple-endpoint strategy that includes more than 20 different in vitro and in vivo assays recommended by EDSTAC [97]. Although more than approximately 87,000 chemicals were initially selected for evaluation, many were polymers or otherwise unlikely to bind to steroid receptors, leaving about 58,000 chemicals for evaluation in Tier 1. The number that will progress to the testing step (Tier 2) [98] is not known. Processing chemicals through both tiers will require many years and extensive resources (www.epa.gov/scipoly/oscpendo/). It is expected that a massive amount of assay data across different levels of biological complexity will be generated through this project. Researchers at the FDA's National Center for Toxicological Research (NCTR) have been actively engaged in the application of various techniques from computational chemistry, chemometrics, and chemoinformatics to mine relevant data currently available from the literature for the purpose of human risk assessment [47,51,99–105].

CHEMOMETRICS APPROACHES

Chemometrics is the discipline concerned with the application of statistical and mathematical methods and methods based on mathematical logic, to chemistry [106]. This is a mature discipline and its fundamental theory and methodology has been fully established in the 1980s. Even though the field of chemometrics continues to enjoy steady growth, only a few new methods have been introduced into SAR community in recent years. Growing demand exists for chemometrics tools for SAR applications to interpret and mine a larger volume of data generated by combinatorial synthesis, HTS, and microarray technology. This can be demonstrated by examples in which recursive partitioning [107] or cluster analysis [108] are used as part of an integrated process to rapidly model data and promptly give feedback for HTS in drug discovery.

Most commercial molecular modeling software programs now include chemometrics tools of some type. A number of chemometrics methods are also included in bioinformatics software for analysis and interpretation of microarray exper-

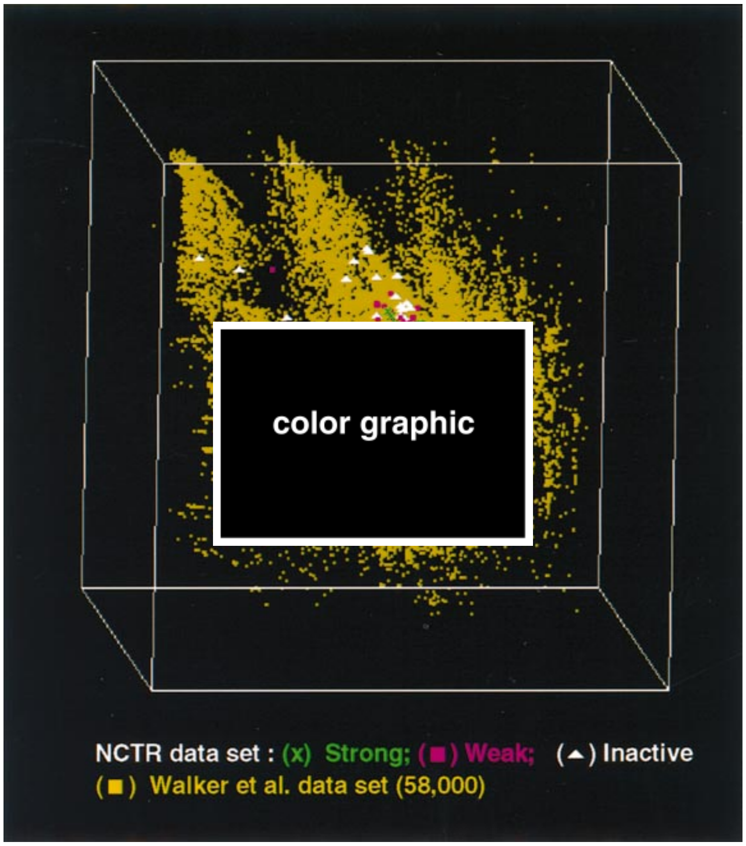

Fig. 3. Chemistry space of the National Center for Toxicological Research [101] and Walker et al. [111] data sets based on BCUT descriptors.

iments, such as GeneSpring from Silicon Genetics (http://www.sigenetics.com/) and GeneSight from BioDiscovery (http://www.biodiscovery.com). In addition, a variety of specialized chemometrics software products are commercially available. Even some statistical software programs (e.g., SAS, S-Plus), more commonly used by statisticians, now feature a user-friendly interface in addition to their standard command line operation to attract general users. In light of the growing popularity and broader utility of SAR techniques, most users are more interested in applying chemometrics without struggling with details about the theory and process. However, these techniques should not be viewed as black boxes. In contrast, sufficient knowledge on underlining principle of chemometrics tools is crucial for successful use of the SAR techniques to interpret science.

The focus of most SAR applications in chemometrics is the development of linear or nonlinear multiple regression models that relate molecular representations (descriptors) as independent variables to an activity (or property) as the dependent variable. These techniques can be categorized into three general classes in terms of their application: Clustering, classification, and QSAR models. Clustering applies unsupervised learning techniques to explore data patterns that may or may not be associated with the problem in question. In contrast, classification methods are supervised learning techniques, which group chemicals into categorical endpoints on the basis of descriptors. Most applications of QSAR attempt to establish a quantitative relationship between structure and activity using multivariate linear regression methods. Quantitative structure–activity relationship approaches and applications are covered in this issue by Perkins et al., hence, this review will focus on data exploratory and classification techniques.

Classification and clustering are two related mathematical methods for extracting information and knowledge by analyzing data [109,110]. The key distinction between the two methods is that classification requires a dependent variable (biological data) for calculation, while clustering does not.

Clustering

Clustering methods, also called exploratory techniques or unsupervised learning, use the independent variables (*X*) to conduct its analysis. Clustering methods are employed primarily to identify patterns or regularities in the dataset. While not providing categorical information on samples (chemicals), they are useful as visualization tools in the early stage of data analysis especially for uncovering key information (e.g., variance) in the original data. In practice, the reasons for using clustering procedures for grouping or classifying a database of chemicals are threefold: Selecting a subset of representative chemicals from a large set prior to conducting costly and time-consuming biological testing for subsequent QSAR model development; obtaining an optimized classification model so that active chemicals can be effectively separated from inactive ones; therefore, reasonable prediction of activity can be made on a larger number of untested chemicals; selecting subsets of structurally similar chemicals from a diverse database for traditional QSAR studies which generally give improved results for congeneric datasets.

Clustering methods are nonsupervised classification techniques in that no preassigned categorical information is required for the objects to be clustered. Clustering of chemicals based on molecular structure (dis)similarity has been discussed in the literature. The most comprehensive discussion can be found in Willett's well-known monograph [22]. Several informative reviews on this topic are also available [15,17,18,31]. Mirkin has given a comprehensive overview of clustering algorithms [110]. Within this category, principal component analysis (PCA) and clustering analysis are the two most commonly used approaches.

Principal component analysis. For investigations involving a large number of variables, it is often useful to simplify the analysis by considering a smaller number of linear combinations of the original variables. Principal component analysis is a mathematical transformation that finds these linear combinations, called principal components (PCs), which are orthogonal and taken together explain all the variance of the original data. The PCA decomposition algorithm ensures that the first PC explains the maximal amount of variance of the original data, the second PC explains the maximal remaining variance in the data subject to being orthogonal to the first PC, and so on. Principal component analysis performs two important tasks. First, it provides a way to reduce the dimensionality of the data. Second, it is a powerful visualization tool to enable graphically representation of intersample and intervariable relationships for data exploratory analysis.

Even though PCA is not a new technique, it is playing an increasing role in combinatorial chemistry and HTS. One application is to construct chemistry space for chemical selection using the first three PCs. The BCUT descriptors in the Concord software (Tripos, St Louis, MO, USA) have been widely used in the drug discovery community for defining chemistry space. Once chemistry space is defined for a given database, a small set of representative chemicals can be selected in this chemistry space using several methods. For example, in the cell-based approach, chemistry space is divided into a number of cells, then the chemical located closest to the center of a cell is selected as the representative for the cell. The number of cells generated depends on the number of chemicals to be selected.

An important use of PCA is as a visualization tool to compare structural diversity between two or more datasets. Figure 3 illustrates the chemistry space of two datasets: The NCTR dataset [101] that contains 230 chemicals together with available ER binding activity data, and the Walker et al. database [111] containing approximately 58,000 chemicals that require testing for their estrogenic activity. It appears that the NCTR dataset of chemicals covers a significant portion of the chemistry space circumscribed by the Walker et al. dataset. Active chemicals tend to cluster together, whereas inactive chemicals are scattered over the defined space. This PCA analysis thus provides evidence that SAR and QSAR models constructed from the NCTR dataset would be statistically valid for estimating the estrogen binding activity of chemicals in the Walker et al. dataset. In other words, the NCTR dataset constitutes a representative subset of the larger Walker et al. dataset.

Clustering analysis. Clustering of chemicals based on molecular structural (dis)similarity has been discussed in the literature. The most comprehensive discussion can be found in Willett's well-known monograph [22]. Several good reviews on this topic are also available [15,17,18,31]. Mirkin has given a comprehensive overview of clustering algorithms [110]. A schematic view of the various clustering approaches mentioned in the chemical literature is shown in Figure 4 [15]. Clustering analysis methods can be categorized as hierarchical clustering analysis (HCA) and nonhierarchical clustering analysis (nonHCA) based on the structure of output (Fig. 4).

The primary purpose of HCA is to present data in a manner that emphasizes natural groupings in the original data. In contrast to techniques that group samples into pre-existing categories, HCA seeks to define those categories in the first place. It groups a dataset of objects into subsets such that a member in a subset is somewhat more similar to other members in the same subset than to members in other subsets [109,110,112].

Hierarchical clustering analysis generates a hierarchy about the relationships among chemicals. The result is generally presented in the form of a dendrogram known commonly as a cluster tree (Fig. 5). At the base of the tree (i.e., the root), all chemicals are considered to belong in a single class; at the other extreme (leaf), each chemical is considered as a separate

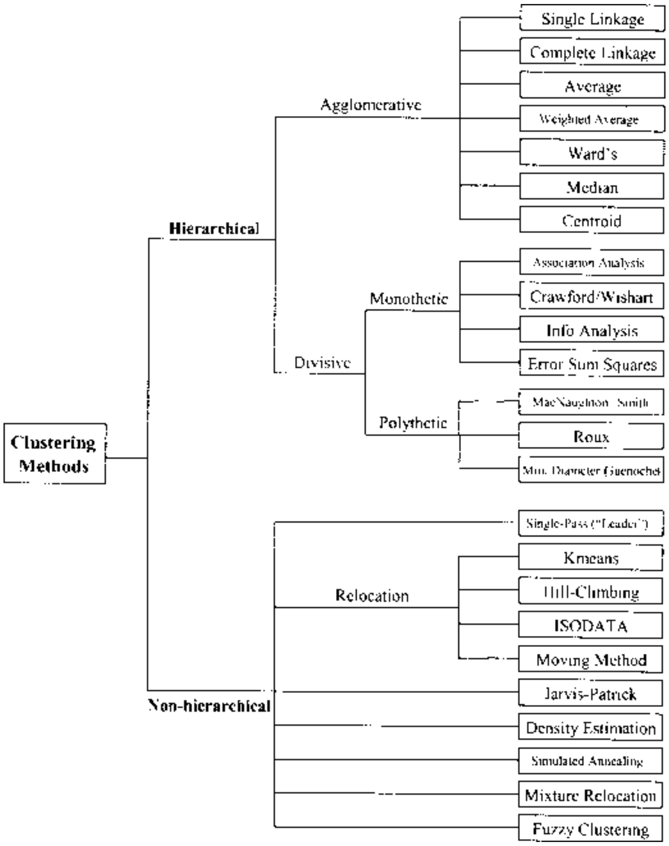

Fig. 4. Clustering methods. Modified from Barnard and Downs [110].

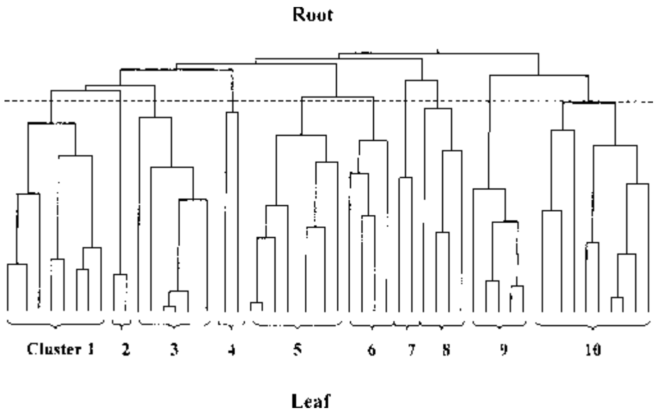

Fig. 5. A schematic view of a dendrogram from hierarchical clustering analysis. At the root, all chemicals (samples) are considered to belong in a single class; at the leaf, each chemical (sample) is considered as a separate class. The number of clusters can be set anywhere between these two extremes by adjusting the dash lines.

class. The number of clusters can be set anywhere between these two extremes without having to rerun the clustering procedure. Depending on the definition of the clustering process, HCA methods can be divided into agglomerative and divisive subcategories. Agglomerative HCA methods have been used much more widely in chemistry than divisive HCA methods.

Agglomerative HCA proceeds from the leaves to the root of the tree. Initially, each cluster contains only a single chemical. At each stage the two nearest clusters are merged to form a single larger cluster. This process continues to aggregate clusters together until all chemicals have merged into a single cluster at the root of the tree. Agglomerative HCA methods differ from each other in the manner used to determine the distance for merging two groups. For single linkage methods, it is the minimum distance between any pair in the two groups; in complete linkage methods, it is the maximum distance. Other methods employ different forms of averaging. Ward's method, for example, minimizes the within-cluster sum of squares over all variables.

In contrast to agglomerative HCA, divisive HCA proceeds from the root of the tree to individual leaves by binary splitting of the dataset into smaller subsets. A set of *N* chemicals has
?2 $2^{N\text{-}1}$-1 ways to split the dataset. The challenge is to determine an appropriate way of splitting the dataset. Monothetic divisive HCA methods use only one descriptor to perform the splitting. Monothetic methods are much faster than polythetic divisive HCA methods that use all descriptors to split the dataset and are very computationally intensive. The fastest algorithm for polythetic divisive HCA is Guenoche's minimum diameter method [113].

An example of HCA is described by Li Xing et al. [114] in a study of the relative binding affinities for a structurally diverse series of ligands to the ER α and β isoforms. Hierarchical clustering analysis was performed on the CoMFA-generated steric and electrostatic fields for 31 training-set chemicals. The high-level clustering of these fields is indicative of the rich structural diversity of these training-set chemicals. The dendrogram corresponding to the ER-α ligands reveals a strong grouping of five ligands with long side chains (clomifene, tamoxifene, 4-OH-tamoxifen, nafoxidine, ICI-164384). The largest single cluster is composed of steroidal and tamoxifen-like structures that are devoid of side chains. The remaining long spikes, which converge at high levels of the hierarchical tree, correspond to bisphenol-A, zearanol, coumestrol, genistein, and methoxychlor. This single example illustrates the exceptional ability of clustering methods, HCA in particular, to clarify the similarities and differences among a series of chemicals both conceptually and visually.

Unlike HCA, nonHCA methods classify a dataset into a predefined number of clusters through partitioning [15,110]. Because a systematic evaluation of all possible partitions is unfeasible, many heuristics have been used to allow the identification of reasonable, but possibly sub-optimal, partitions. NonHCA methods are generally much less demanding of computational resources. Within this class, the K-means and Jarvis-Patrick methods are the most widely used in chemical applications. Discussion of other methods can be found in Barnard and Downs [15]. Based on comparison of several clustering procedures, including Ward's, group average, Guénoche, Jarvis-Patrick and enhanced Jarvis-Patrick, on several datasets using MACCS keys, Brown and Martin [115] concluded that Ward's clustering procedure produces the most accurate prediction, followed by group average and then Guénoche.

Self-Organizing Maps. A discussion on clustering would be incomplete without at least mentioning the growing role of artificial neural networks (ANNs), particularly the self-organizing maps originated by Kohonen [116], as a computational tool for clustering in many scientific endeavors including recently in chemistry [117], food science [118], drug discovery [119,120], and microarray technology [121]. The self-organizing map is particularly well suited to the task of identifying a small number of prominent classes in a dataset [122]. It finds an optimal set of centroids around which the data points appear to aggregate. It then partitions the dataset, with each centroid defining a cluster consisting of the data points nearest it. Although applications of self-organizing maps and related ANN techniques are still sparse in the areas of toxicology and environmental science, they are sure to grow in popularity as awareness of the power and versatility of these methods becomes more widespread.

Classification

Classification methods, a type of supervised learning, use dependent (activity) and independent (descriptors) variables in their calculation. A classification model based on a training set can be used for the prediction of external datasets. Of the several notable classification methods that fall into this category, K-nearest neighbor (KNN), soft independent modeling of class analogy (SIMCA), decision tree methods, and ANNs are discussed in this article.

K-nearest neighbor. K-nearest neighbor is a widely used pattern-recognition technique that can categorize an unknown chemical based on its proximity to samples already placed in categories [123]. Specifically, the predicted class (i.e., active or inactive) of an unknown chemical depends on the distribution of class assignment of its KNNs in the training set, which accounts for the name of the technique. The nearness is generally measured by a Euclidean distance metric in an *N*-dimensional space of molecular descriptors, although other distance metrics can be applied. In a fashion analogous to polling, each of the KNN training set samples votes once for its class; the unknown chemical is then assigned to the class with the most votes. With a chosen distance metric, the most important part of the KNN process is to determine an optimal K-value for the final model development. In one particular KNN procedure [104], leave-one-out cross validation is employed in the following manner to select the optimal K-value: (1) remove a chemical from the dataset; (2) calculate the distance metric between the omitted chemical and all remaining chemicals in the dataset; (3) select K chemicals nearest (or similar) to the omitted chemical using the calculated distances; (4) assign the omitted chemical to the activity class corresponding to the majority of the *K* chemicals; (5) count the number of misclassifications by comparing the predefined class with the predicted class of the omitted chemical; and (6) repeat steps 1 through 5 until each chemical in the dataset is omitted once for prediction and the rate of misclassification is summarized for the predefined K value. Theoretically, the K value can vary from 1 to *N*-1, where *N* is the size of the dataset. Normally, repeating steps 1 through 6 for each K value from 3 to 10 should be sufficient to determine the optimal K value corresponding to the smallest rate of misclassification.

Although KNN is very fast, it is generally less insightful compared with other classification methods. Shi et al. [104] evaluated four classification methods, including KNN, SIMCA, a decision tree, and ANN (vide infra) to classify 230 estrogen ligands into active and inactive categories. While these approaches are different in a number of ways, they generally produced similar classification results. The nature of the descriptors used, and more particularly the effectiveness with which they encode the structural features of the molecules related to the activity, was far more critical than the specific method employed.

Soft independent modeling of class analogy. The SIMCA method was first introduced by Wold [124]. In contrast to KNN, which is based on distances between pairs of samples, SIMCA splits the dataset into subsets according to predefined activity categories. Principal component analysis models are then constructed for each subset. The categorical activity of an untested sample (chemical) is determined based on which PCA model it best fits.

A very attractive feature of SIMCA is its more realistic prediction options when compared to the rather limited choices in KNN. K-nearest neighbor assigns every sample to exactly one class in the training set, regardless of whether this nearest class is close or far in distance in an absolute sense. Soft independent modeling of class analogy, however, provides three possible prediction outcomes: A sample can be assigned to none, one, or several of the predefined classes. In addition, because these decisions are made on the basis of statistical tests, confidence limits can be placed on outcomes.

Reliable classification of untested samples is the ultimate goal of SIMCA; however, the technique also provides a rich set of diagnostics that addresses other interesting aspects of classification. An example is the modeling power parameter, which measures to what degree a particular variable (i.e., descriptor) contributes to the PCs in terms of class separation. Another is the Mahalanobis distance, which provides a probabilistic means of identifying and ranking outliers. A third example is the variance structure of each class, which yields clues about category complexity and can sometimes reveal the underlying factors that cause one category to differ from another.

Chemometrics approaches have found utility as pattern recognition tools for purposes of regulatory monitoring of pharmaceuticals. In collaboration with the U.S. FDA, Welsh et al. compared KNN, SIMCA, and ANN techniques in terms of their ability to determine the sameness and difference between same-product manufacturers of the over-the-counter supplement L-tryptophan based solely on analysis of their trace organic impurity patterns obtained from high-performance liquid chromatography [125–128]. Described as pharmaceutical fingerprinting, the strategies developed were successful in distinguishing not only among different manufacturers but also day-to-day, lot-to-lot, and column-to-column variations in high-performance liquid chromatography data for the same manufacturer. Various preprocessing schemes were implemented to extract inputs from the high-performance liquid chromatography data, including wavelet packets [126], an automatic windowing technique [125], and windowing followed by pruning to extract the most information-rich data entries as input for classification [127]. Among the three classifiers tested, ANNs performed better than SIMCA and KNN. Pruning of the data inputs improved ANN speed and performance. A highly innovative and more recent approach known as Phase Space yielded the best performance for this particular application [129,130]. The polynomial neural network [131], which integrates many of the attributes of partial least-squares regression and standard ANNs into a single entity, is only mentioned here but is reviewed in detail by Perkins et al. in this issue.

Decision tree methods. Combinatorial chemistry, HTS, and computational chemistry are making a profound impact on the drug discovery process. These technologies enable parallel synthesis and biological evaluation of hundreds of thousands of chemicals then simultaneously generate thousands of descriptors for each chemical. Most traditional SAR techniques are not suitable for such large datasets. For this reason, there is growing interest in applying decision tree methods for this purpose [107,132,133]. This technology enjoys several notable attributes not seen in other SAR techniques: (1) it rapidly finds a solution; (2) it can identify the most relevant and discriminating descriptors for a particular application; (3) it can handle noisy data; (4) it is not strongly influenced by missing data; (4) it is capable of modeling nonlinear problems; and (5) the rules generated by the decision tree methods are straightforward and easy to understand by chemists and toxicologists even those who are not experts in chemometrics and related areas.

The automatic construction of decision trees dates back to early 1960s [134]. Breiman et al. [135,136] had a seminal influence both in bringing the scheme to the attention of statisticians and in proposing new algorithms for constructing trees. The Classification and Regression Tree (CART) developed by Breiman et al. and marketed by Salford Systems (http://www.salford-systems.com) is widely used in various disciplines. Other decision tree methods include automatic interaction detection (AID) [134], chi-squared automatic interaction detection (CHAID) [137], and C4.5 [138]. Decision trees, under the name of recursive partitioning, are typically applied in drug discovery for mining large datasets [107,132,133].

?3

A decision (classification or regression) tree is composed of a collection of many production rules, expressed as premise and conclusion (in the form: IF. . . THEN. . .) and displayed in the form of a tree containing only binary branching. For examples, a simple rule could be IF molecular weight $>$ 300, then the chemical is active. A decision tree model provides an alternative to linear and additive models for regression problems, and to linear and additive logistic models for classification problems. Depending on the nature of the activity data (endpoint), the tree can be constructed for either regression or classification. Each end node (leaf of the tree) of a regression tree gives a quantitative prediction, while the classification tree gives categorical predictions. The classification tree is used most commonly in data analysis, where the endpoint is usually binomial (yes/no). The development of a decision model consists of two steps, tree construction and tree pruning. In the tree construction process, a parent population is split into two child nodes that become parent populations for further splits. The splits are selected to maximally distinguish the response variable in the left and right nodes. Splitting continues until chemicals in each node are either in one activity category or can not be split further to improve the model. To avoid overfitting the training data, the tree needs to be cut down to a desired size using tree cost-complexity pruning.

An example of a decision tree model constructed using a method described by Clark and Pregibon [139] and implemented in the S-Plus software (http://www.insightful.com/) is illustrated in Figure 6. In this application, a tree-based model classified 230 chemicals into active and inactive ER binding categories based on five descriptors. The model identified the Phenolic Ring Index, which indicates the presence and absence

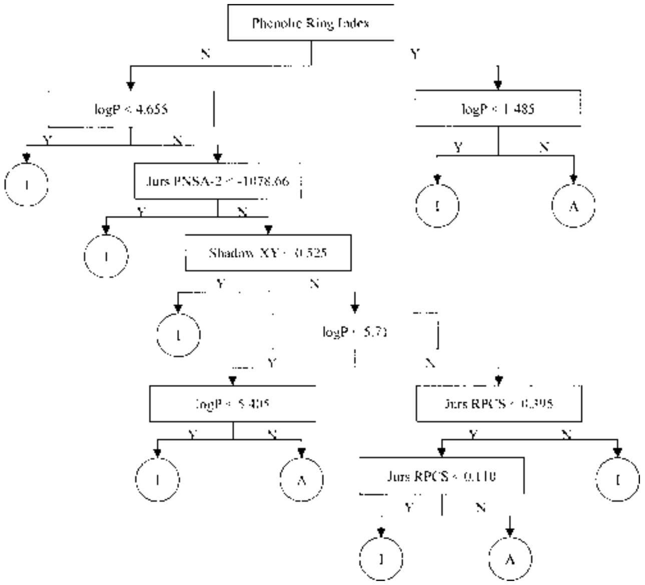

Fig. 6. Tree-based model. The model displays a series of YES/NO (Y/N) rules to classify chemicals into active (A) and inactive (I) categories based on 5 descriptors: Phenolic Ring Index, Log*P*, Jurs PNSA-2, Shadow XY, and Jurs RPCS. The squares represent the rules while the circles represent the categorical results.

of a phenolic substructure in structure, as the most important descriptor for ER binding. This finding is consistent with the understanding that the phenolic ring is often associated with estrogenic activity [140], and the contribution of the phenolic ring in binding is much more significant than any other structural feature [51]. For example, by overlaying the crystal structures of four ligand-estrogen receptor (ER) complexes (estradiol-ER, 4-hydroxytamoxifen-ER, raloxifene-ER, and diethylstilbesterol-ER complexes) based on their common protein residues at the binding site, it was found that the phenolic rings of all four ligands are similarly positioned within the binding pocket to allow hydrogen bond interactions with Glu 353 and Arg 394 of the receptor and a highly conserved water molecule [105]. Chemicals containing a phenolic moiety and a Log*P* value larger than 1.49 were more likely to be ER binders. In contrast, chemicals without a phenolic moiety were less likely to be ER binders unless they had relatively larger hydrophobicity (Log*P*), charge surface area (represented by the Jurs-PNSA-2 and Jurs-RPCS descriptors [141]), and breadth of the structure (represented by the Shadow-XY descriptors [142]).

Artificial neural networks. Over the past 10 years an increasing number of scientists have explored the feasibility of more sophisticated nonlinear regression methods for development of SARs. Artificial neural networks are now being employed for a wide range of applications in such diverse areas of chemistry, engineering, biology, food science, and viticulture, and even toxicology and the environmental sciences for purposes of property prediction, pattern recognition, and data analysis [143]. As their name implies, ANNs attempt to mimic the structure and function of the human brain. Like the human brain, the ANN is composed of an interconnected network of neurons. The most common ANN configuration, known as feed-forward back-propagation, consists of three layers: Input layer, hidden layer, and output layer. The input layer contains the same number of neurons as input variables (i.e., descriptors). The output layer usually contains one neuron for each predicted output (i.e., biology activity, toxicity). The hidden layer may consist of several sublayers, each containing a variable number of neurons.

Training of ANNs is achieved through repeated presentations of input and output pairs. By back propagating the error between the ANN-predicted output and the correct output, the ANN learns to establish and optimize associations between input and output. Learning is achieved by adjusting the transfer-function weights that modulate the signal between connected neurons in different layers until the error between the ANN-predicted and correct output values falls below a preassigned convergence criterion. Once trained in this fashion, the ANN is capable of predicting the output target property for a set of inputs not yet encountered (e.g., the toxicity of an untested chemical). Artificial neural networks have found increasing utility in drug design, particularly as a nonlinear approach to developing QSAR models. Several comprehensive reviews and commentaries on the subject can be found in the recent literature [144,145].

MINING GENE EXPRESSION DATA

In recent years, microarray experiments have proved invaluable for the study of gene expression patterns in cell lines induced by chemicals (drugs or toxicants) [25] and in tissues associated with disease [146]. An extensive bibliography and comprehensive survey on microarray technology and applications is maintained online by Shi [147] at http://www.gene-chips.com. Unlike the traditional biological technologies, such as Northern Blot and Southern Blot, microarray technology can monitor expression of tens and thousands of genes simultaneously in a single experiment. A major challenge facing microarray technology concerns the huge amount of data generated from these experiments, in particular how to use SAR techniques and related data mining tools to extract knowledge [148]. It is safe to assume that no single tool is ideal for all problems, even for a single category of information such as gene expression data. Thus, two gene expression datasets (http://waldo.wi.mit.edu/MPR/data_sets.html and http://cellcycle-www.stanford.edu) were recently made available to encourage the multidisciplinary scientific community to evaluate various statistical approaches for treating these data [149,150]. This effort culminated in a workshop (December 18–19, 2000) on ‘‘Critical Assessment of Techniques for MicroArray Data Analysis (CAMDA’00)’’ organized by Duke University (http://www.bioinformatics.duke.edu/CAMDA/CAMDA.html).

The primary objectives of microarray data analysis are [151]: To identify statistically significant up- or down-regulated genes; to discern functional groupings of genes by discovering (dis)similarity among genes expression profiles; and to predict the biochemical and physiological pathways of previously uncharacterized genes. Today, clustering methods have been used more often than any other data mining approaches. Clustering is normally performed on either samples (cell lines or tissues) or variables (genes). Clustering of samples reveals similarities among the samples whereas clustering of variables pinpoints intervariable relationships.

An example of clustering methods taken from an analysis of gene expression data obtained from microarray experiments is illustrated in Figure 2. These data were reported by Scherf et al. as part of a drug discovery screen by the U.S. National Cancer Institute to identify a pattern between anticancer activity and gene expression in 60 human cancer cell lines. Prior to conducting clustering analysis on the data, no expression

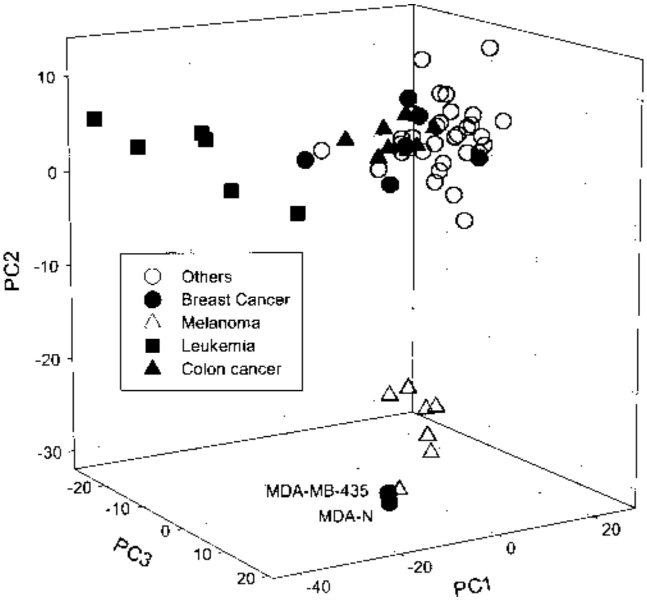

Fig. 7. Principal component analysis on the data set presented in Figure 2. Two breast cancer cell lines, MDA-N and MDA-MB-435, are closely grouped with melanoma cell lines.

pattern can be observed. After clustering analysis using the nonHCA, however, the relationship between genes becomes apparent from which co-expressed genes could be identified or new functions for uncharacterized genes discovered. Moreover, the similarity between 60 cell lines can also be examined by conducting PCA on their gene expression patterns. As shown in Figure 7, two breast cancer cell lines, MDA-N and MDA-MB-435, are closely grouped with melanoma cell lines, indicating that they should behave like melanoma cells rather than breast cancer cells in biological testing. The findings are consistent with the recent similarity analysis of the 60 cell lines in terms of their responses to the 25,023 chemicals in an in vitro anticancer screening using clustering analysis [89].

TRANSFORMING DATA INTO KNOWLEDGE: THE NEW PARADIGM

A recursive process combining screening assays and SAR modeling has become prevalent throughout much of the pharmaceutical industry. In lead discovery, the process has been called sequential screening [108] as depicted in Figure 8. The process starts with assay data for an initial set of chemicals from an existing chemical library. The resulting data for active chemicals, and sometimes for inactive chemicals, are then used for initial SAR modeling. The number of models developed can vary from one to several. Next, the preliminary SAR models are used prospectively to define a set of active chemicals. The new chemicals are assayed, and these data are then used to refine the set of SAR models. The models can be used to identify new chemicals in many ways, such as searching a library of existing chemicals or assisting in the design of a virtual combinatorial library. The chemical synthesis stage might employ combinatorial chemistry, and HTS can be employed for the new bioassays. Whether engaged in lead discovery, lead optimization, or drug development, the process is essentially the same. As a logical extension of this process, Neurogen Corporation employs a proprietary scheme by which high-speed synthesis and screening, as well as SAR development, are carried out continuously and in parallel. Results are updated regularly and available online (www.aidd.com).

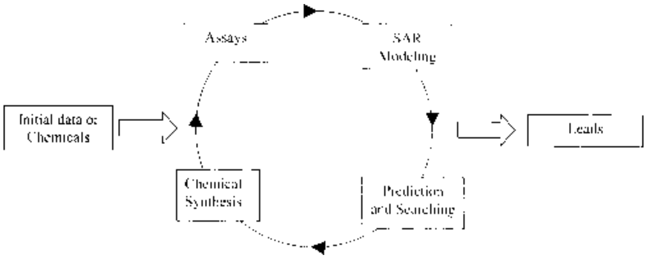

Fig. 8. Depiction of the sequential screening process now prevalent in the pharmaceutical industry where structure–activity relationship (SAR) development and prediction are integral to the process. The output of the process is a drug lead that may be further developed.

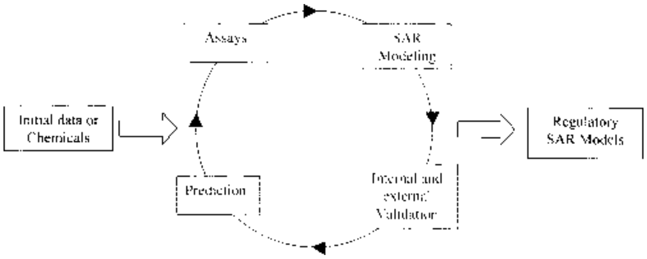

Fig. 9. Depiction of the sequential process used by the National Center for Toxicological Research to develop models for predicting estrogen receptor binding. The process output is a model for use in toxicological regulation. Validation of the model is critical, particularly with respect to confirming minimal false-negative predictions.

The cyclical process now prevalent in the drug industry is an equally suitable paradigm for toxicology. In fact, the U.S. FDA at the National Center for Toxicological Research (NCTR), in a cooperative program with the U.S. EPA, has adopted this process to develop SAR models that may support priority setting of tens of thousands of chemicals that will undergo assessment for potential as EDCs. The NCTR's Endocrine Disruptor Knowledge Base Program applies essentially the same cyclical process except, as depicted in Figure 9, a validation step is specifically emphasized [103]. The requirements for accuracy in toxicity predictions, and particularly for minimizing false negative predictions, is stringent for regulatory purposes compared to requirements in drug development [152]. Concomitantly, more rigorous model validation is required.

In order to make the prediction of potential for ER binding more tractable for some 80,000 chemicals, the Endocrine Disruptor Knowledge Base Program has developed a hierarchical suite of models that progress from the computationally fast, but less accurate, to the computationally intensive, but more accurate [153]. In the first level, several rejection filters are used to exclude chemicals most unlikely to be active. The next level consists of a set of pharmacophore models, structural alerts, classification models (each with varied strengths for certain chemical classes), which are used in combination to exclude additional inactives and to rank order potential actives in accordance with likelihood of activity. The models composing the first two levels are constructed in a manner to minimize false negatives, which is an issue of particular concern in regulatory matters [104]. At the highest level of the model hierarchy is a robust CoMFA model that provides statistically robust quantitative prediction of activity [105]. Applying this procedure, a cyclical process was followed to design a training set of data that covers the broad range ligand structures that bind ER [103]. All data were obtained from a validated competitive binding assay that used replication of data points [101,154]. Ultimately, the models will serve to blind-predict hundreds of chemicals in a validation dataset, and these data will be used to refine the models where inaccuracies are encountered.

FUTURE DIRECTIONS AND CHALLENGES

Ongoing efforts at the U.S. FDA and U.S. EPA, such as those cited above, are demonstrating how SAR technologies can foster and advance toxicological sciences. At the present time, serial testing of chemicals in a suite of in vitro and in vivo assays still remains in large part the prevailing paradigm in the realm of toxicology. In many cases the resultant data are disparate across species, endpoints, protocols, and laboratories and, therefore, are often unsuitable for development of predictive SAR models. However, the extraordinary technological advances in genomics, in high-throughput synthesis and screening, and in information technology are now making a profound and lasting impact on the way toxicology is perceived and practiced.

Toxicogenomics, the study of alterations in gene expression during toxicity, is emerging as a well-defined subdiscipline at the interface between the fields of toxicology and genomics [9]. Experimental genomics methods, in combination with the rapidly growing body of sequence information, promise to revolutionize the way cells, cellular processes, and diseases are studied. These efforts are providing enormous volumes of data about the state of the cell under both normal and disease conditions. Computational methods are essential tools for organizing these data into meaningful information and for guiding the transformation of this information into useable knowledge. In this context, it is hoped that this review has succeeded in drawing attention to the tremendous utility of SAR and associated QSAR methods particularly with respect to the field of computational toxicology.

Within the pharmaceutical industry today, drug discovery and toxicology are becoming permanently intertwined [155]. The pharmaceutical industry is facing a multitude of significant challenges, including the need to compress timescales for converting new chemical entities to marketable drugs, the management of the surfeit of data flowing both from genomics and from combinatorial chemistry and HTS, and the economic imperative to fill their pipelines with a large number of safe yet effective drug candidates. To meet these challenges, more and more emphasis on the early assessment of absorption, distribution, metabolism, excretion, and toxicity properties of potential drug candidates has become routine in many drug discovery programs. The ultimate goal is to design clinical candidates packaged with the desired potency, selectivity, bioavailability, absorption, distribution, metabolism, excretion, and toxicity properties together with innumerable other requirements, and to achieve this goal in the shortest amount of time. As a consequence, a growing trend exists in the pharmaceutical industry toward integrated workflow that emphasizes collecting the preclinical data associated with these properties in time to influence research plans. A salient example is the confluence of pharmacological and toxicological information to guide the drug design process during early lead optimization. When enough data are collected and correlated to known mechanisms of action, it should be possible to enhance empirical experimentation with computational toxicology. In fact, computational approaches make it possible to apply the acquired knowledge to both real and virtual chemical collections.

Although environmental toxicology and drug discovery differ in their ultimate objectives, the computational tools, strategies, and schemes employed are much the same in both disciplines. This fact attests to the incredible versatility of these computational tools that span a wide range of disciplines including drug discovery, toxicology, and environmental science. Still the choice of computational tools, and the precise way they are implemented, will vary from one discipline to another.

The utility of SAR-based computational models for a wide range of tasks, including drug discovery, risk assessment, and environmental toxicology, will depend on a number of factors, including knowledge of the SAR methods, understanding of experimental designs, and the quality of data. Ultimately, it will highly relay on the quality data available to build the predictive models. As advocated in recent communications by Johnson [156] and by Richard [157], this outcome will be reached more swiftly if the pharmaceutical industry were to explore and adopt creative policies for sharing in-house generated toxicological data with the outside scientific community even if only for the sake of enlightened self-interest.

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 Success Program.

REFERENCES

1. Warr WA. 1997. Combinatorial chemistry and molecular diversity: An overview. *J Chem Inf Comput Sci* 37:134–140.
2. Williard X, Pop I, Bourel L, Horvath D, Baudelle R, Meynyk P, Deprez B, Tartar A. 1996. Combinatorial chemistry: A rational approach to chemical diversity. *Eur J Med Chem* 31:87–98.
3. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274: 610–614.
4. Covacci A, Kennedy GC, Cormack B, Rappuoli R, Falkow S. 1997. From microbial genomics to meta-genomics [review]. *Drug Dev Res* 41:180–192.
5. Ramsay G. 1998. DNA chips—state-of-the-art [review]. *Nat Biotechnol* 16:40–44.
6. Wallace RW. 1997. DNA on a chip—serving up the genome for diagnostics and research. *Mol Med Today* 3:384–389.
7. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglut T, Hubbell E, Robinsin E, Mittman M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280(5366): 1077–1082.
8. 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.
9. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. 1999. Microarrays and toxicology: The advent of toxicogenomics. *Mol Carcinog* 24:153–159.
10. MacBeath G, Schreiber SL. 2000. Printing proteins as microarrays for high-throughput function determination. *Science* 289: 1760–1763.
11. Martin EJ, Spellmeyer DC, Critchlow RE Jr, Blaney JM. 1997. Does combinatorial chemistry obviate computer-aided drug design? In Lipkowitz KB, Boyd DB, eds, *Reviews in Computational Chemistry*, Vol 10. VCH, New York, NY, USA pp 75–100.
12. Blaney JM, Martin EJ. 1997. Computational approaches for

combinatorial library design and molecular diversity analysis. *Curr Opin Chem Biol* 1:54–59.

13. Brown RD, Martin YC. 1997. Designing combinatorial library mixtures using a genetic algorithm. *J Med Chem* 40:2304–2313.
14. Martin EJ, Blaney JM, Siani MA, Spellmeyer DC, Wong AK, Moos WH. 1995. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J Med Chem* 38: 1431–1436.
15. Barnard JM, Downs GM. 1992. Clustering of chemical structures on the bases of two-dimensional similarity measures. *J Chem Inf Comput Sci* 32:644–649.
16. Brown RD, Martin YC. 1996. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* 36:572–584.
17. Downs GM, Willett P, Fiasanick W. 1994. Similarity searching and clustering of chemical-structure databases using molecular property data. *J Chem Inf Comput Sci* 34:1094–1102.
18. Downs GM, Willett P. 1995. Clustering of chemical structure databases for compound selection. In van de Waterbeemd H, ed, *Advanced Computer-Assisted Techniques in Drug Discovery,* Vol 3–Methods and principles in medicinal chemistry. VCH, Weinheim, Germany, pp 111–130.
19. Hodes L. 1981. Computer-aided selection of compounds for antitumor screening: Validation of a statistical-heuristic method. *J Chem Inf Comput Sci* 21:128–132.
20. Hodes L. 1989. Clustering a large number of compounds. 1. Establishing the method on an initial sample. *J Chem Inf Comput Sci* 29:66–71.
21. Shemetulskis NE, Dunbar JB Jr, Dunbar BW, Moreland DW, Humblet C. 1995. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J Comput-Aided Mol Des* 9:407–416.
22. Willett P. 1987. *Similarity and Clustering in Chemical Information Systems,* Bawden D, ed. John Wiley, New York, NY, USA, pp 1–254.
23. Lewis RA, Mason JS, McLay IM. 1997. Similarity measures for rational set selection and analysis of combinatorial libraries: The diverse property-derived (DPD) approach. *J Chem Inf Comput Sci* 37:599–614.
24. Matter H. 1997. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J Med Chem* 40:1219–1229.
25. 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.
26. Lawson RG, Jurs PC. 1990. Cluster analysis of acrylates to guide sampling for toxicity testing. *J Chem Inf Comput Sci* 30:137–144.
27. Johnson M, Maggiora GM. 1990. *Concepts and Applications of Molecular Similarity.* John Wiley, New York, NY, USA.
28. Feldman A, Hodes L. 1975. An efficient design for chemical structure searching. 1. The screens. *J Chem Inf Comput Sci* 15: 147–152.
29. Hodes L. 1976. Selection of descriptors according to discrimination and redundancy. Application to chemical substructure searching. *J Chem Inf Comput Sci* 16:88–93.
30. Hodes L. 1981. Selection of molecular fragment feactures for structure–activity studies in antitumor screening. *J Chem Inf Comput Sci* 21:132–136.
31. Barnard JM, Downs GM. 1997. Chemical fragment generation and clustering software. *J Chem Inf Comput Sci* 37:141–142.
32. Downs GM, Barnard JM. 1997. Techniques for generating descriptive fingerprints in combinatorial libraries. *J Chem Inf Comput Sci* 37:59–61.
33. Wipke WT, Krishnan S, Ouchi GI. 1978. Hash functions for rapid storage and retrieval of chemical structures. *J Chem Inf Comput Sci* 18:32–37.
34. Knuth D. . *Sorting and Searching.* Addison-Wesley, Reading, MA, USA.
35. Martin YC, Bures MG, Willett P. 1990. Searching databases of three-dimensional structures. In Lipkowitz KB, Boyd DB, eds, *Reviews in Computational Chemistry,*Vol 1. VCH, New York, NY, USA, pp 213–263.
36. Good AC, Mason JS. 1996. Three-dimensional structure datase searches. In Lipkowitz KB, Boyd DB, eds, Reviews in Computational Chemistry, Vol 7. VCH, New York, NY, USA, pp 67–117.
37. Klopman G. 1992. Multicase: A hierarchical computer automated structure evaluation program. *Quant Struct-Act Relat* 11: 176–184.
38. Klopman G, Shi LM, Ramu A. 1997. Quantitative structure–activity relationship of multidrug resistance reversal agents. *Mol Pharmacol* 52:323–334.
39. 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.
40. 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.
41. Hong H, Tong W, Fang H, Shi L, Qian X, Wu J, Perkins R, Walker JD, Braham W, Sheehan D. 2002. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated computational approach. *Environ Health Perspect* 110:29–36.
42. Shi LM, Fan Y, Myers TG, O'Conner PM, Paull KD, Friend SH, Weinstein JN. 1998. Mining the nci anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J Chem Inf Comput Sci* 38:189–199.
43. Karelson M, Lobanov VS, Katritzky AR. 1996. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem Rev* 96:1027–1043.
44. 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.
45. 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.
46. 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. *Japanese Chemistry Today* 2:50–57.
47. 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. *Endocrine* 138:4022–5.
48. Hammett LP. 1940. *Physical Organic Chemistry.* McGraw-Hill, New York, NY, USA.
49. Dietrich SW, Dreyer ND, Hansch C, Bentley DL. 1980. Confidence interval estimators for parameters associated with quantitative structure–activity relationships. *J Med Chem* 23:1201–1205.
50. 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.
51. Fang H, Tong W, Shi L, Blair R, Perkins R, Branham WS, Dial SL, Moland CL, Sheehan DM. 2001. Structure activity relationship for a large diverse set of natural, synthetic and environmental chemicals. *Chem Res Toxicol* 14:280–294.
52. Leo A, Hansch C, Elkins D. 1971. Partition coefficients and their uses. *Chem Rev* 71:525–616.
53. Hansch C, Kim D, Leo AJ, Novellino E, Silipo C, Vittoria A. 1989. Toward a quantitative comparative toxicology of organic compounds. *Crit Rev Toxicol* 19:185–226.
54. Kellogg GE, Abraham DJ. 2000. Hydrophobicity: Is log*P*(o/w) more than the sum of its parts? *Eur J Med Chem* 35:651–661.
55. Hansch C, Leo A. 1979. *Substituent Constraints for Correlation Analysis in Chemistry and Biology.* Wiley Interscience, New York, NY, USA.
56. Leo A. 1993. Calculating logpoct from structures. *Chem Rev* 93:1281–1306.
57. Hansch C, Leo A. 1995. *Exploring QSAR—Fundamentals and Applications in Chemistry and Biology.* American Chemical Society, Washington, DC, USA.
58. 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.

59. Ghose AK, Prichett A, Crippen GM. 1988. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. III. Modeling hydrophic interactions. *J Comput Chem* 9:80–90.
60. 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.
61. Rekker RF. 1977. *The Hydrophobic Fragment Constant.* Elsevier, New York, NY, USA.
62. Wang R, Fu Y, Lai L. 1997. A new atom-additive method for calculating partition coefficients. *J Chem Inf Comput Sci* 37: 615–621.
63. Chou JT, Jurs PC. 1979. Computer-assisted computation of partition coefficient from molecular structures using fragment constrants. *J Chem Inf Comput Sci* 19:172–178.
64. Suzuki T, Kudo Y. 1990. Automatic logP estimation based on combined additive modeling methods. *J Comput-Aided Mol Des* 4:155–198.
65. Moriguchi I, Hirono S, Liu Q, Nakagome I, Matsushita Y. 1992. Simple method of calculating octanol/water partition coefficient. *Chem Pharm Bull* 40:127–130.
66. Klopman G, Wang S. 1992. A computer automated structure evaluation (CASE) approach to calculation of partition coefficient. *J Comput Chem* 12:1025–1032.
67. Klopman G, Li J-Y, Wang S, Dimayuga M. 1994. Computer automated logp calculation based on an extended group contribution approach. *J Chem Inf Comput Sci* 34:752–781.
68. Kamlet MJ, Doherty RM, Abboud JL, Abraham MH, Taft RW. 1986. Linear solvation energy relationships: 36. Molecular properties governing solubilities of organic nonelectrolytes in water. *J Pharm Sci* 75:338–349.
69. Kamlet MJ, Doherty R, Abraham M, Marcus Y, Taft R. 1988. Linear solvation energy relationships. 46. An improved equation for correlation and prediction of octanol/water partition coefficients of organic nonelectrolytes (including strong hydrogen bond donor solutes). *J Phys Chem* 92:5244–5255.
70. Meylan W, Howard P. 1995. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J Pharm Sci* 8:83–92.
71. Tetko IV, Tanchuk VY, Kasheva TN, Villa AE. 2001. Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds. *J Chem Inf Comput Sci* 41: 246–252.
72. Calleja MC, Geladi P, Persoone G. 1994. QSAR models for predicting the acute toxicity of selected organic chemicals with diverse structures to aquatic non-vertebrates and humans. *SAR/QSAR Environ Res* 2:193–234.
73. Mazerska Z, Augustin E, Dziegielewski J, Cholody MW, Konopa J. 1996. QSAR of acridines, III. Structure–activity relationship for antitumour imidazoacridinones and intercorrelations between in vivo and in vitro tests. *Anti-Cancer Drug Des* 11:73–88.
74. Bursi R, Dao T, van Wijk T, de Gooyer M, Kellenbach E, Verwer P. 1999. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J Chem Inf Comput Sci* 39:861–867.
75. Beger RD, Freeman JP, Lay JO Jr, Wilkes JG, Miller DW. 2001. Use of 13c nmr spectrometric data to produce a predictive model of estrogen receptor binding activity. *J Chem Inf Comput Sci* 41:219–224.
76. Kauvar LM, Higgins DL, Villar HO, Sportsman JR, Engqvist-Goldstein AE, Bukar R, Bauer KE, Dilley H, Rocke DM. 1995. Predicting ligand binding to proteins by affinity fingerprinting. *Chem Biol* 2:107–118.
77. Kauvar LM, Higgins DL, Villar HO, Sportsman JR, Engqvist-Goldstein A, Buckar R, Bauer KE, Dilley H, Rocke DM. 1995. Predicting ligand binding to proteins by affinity fingerprinting. *Chem Biol* 2:107–118.
78. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. 1999. Microarrays and toxicology: The advent of toxicogenomics. *Mol Carcinog* 24:153–159.
79. Marshall A, Hodgson J. 1998. DNA chips—An array of possibilities. *Nat Biotechnol* 16:27–31.
80. Livache T, Fouque B, Roget A, Marchand J, Bidan G, Teoule R, Mathis G. 1998. Polypyrrole DNA chip on a silicon device—example of hepatitis c virus genotyping. *Anal Biochem* 255: 188–194.
81. Boyd MR. 1989. Status of the nci preclinical antitumor drug discovery screen. In DeVita VT, Hellman S, Rosenberg SA, eds, *Cancer: Principles and Practice of Oncology Update,* Vol 3. Lippincott Williams & Wilkins, Philadelphia, PA, USA, pp 1–12.
82. Boyd MR, Paull KD. 1995. Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen. *Drug Dev Res* 34:91–109.
83. Boyd MR. 1997. The NCI in vitro anticancer drug discovery screen: Concept, implementation, and operation, 1985–1995. In Teicher BA, ed, *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval.* Humana Press, Totowa, NJ, USA, pp 23–42.
84. Paull KD, Shoemaker RH, Hodes L, Monks A, Scudiero DA, Rubinstein L, Plowman J, Boyd MR. 1989. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: Development of mean graph and compare algorithm. *J Natl Cancer Inst* 81:1088–1092.
85. Paull KD, Hamel E, Malspeis L. 1995. Prediction of biochemical mechanism of action from the in vitro antitumor screen of the national cancer institute. In Foye WO, ed, *Cancer Chemotherapeutic Agents.* American Chemical Society, Washington, DC, USA, pp 9–45.
86. Shi LM, Myers TG, Fan Y, O'Connor PM, Paull KD, Friend SH, Weinstein JN. 1998. Mining the national cancer institute anticancer drug discovery database: Cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity. *Mol Pharmacol* 53:241–251.
87. Shi LM, Fan Y, Myers TG, Waltham M, Paull KD, Weinstein JN. 1998. Mining the anticancer activity database generated by the nci anticancer drug discovery program using statistical and artificial intelligence techniques. *Mathematical Modelling and Scientific Computing* Vol 9.
88. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD. 1997. An information-intensive approach to the molecular pharmacology of cancer. *Science* 275:343–349.
89. Shi LM, Fan Y, Lee JK, Waltham M, Andrews DT, Scherf U, Paull KD, Weinstein JN. 2000. Mining and visualizing large anticancer drug discovery databases. *J Chem Inf Comput Sci* 40:367–79.
90. Weinstein JN, Kohn KW, Grever MR, Viswanadhan VN, Rubinstein LV, Monks AP, Scudiero DA, Welch L, Koutsoukos AD, Chiausa AJ, and others. 1992. Neural computing in cancer drug development: Predicting mechanism of action. *Science* 258: 447–451.
91. Weinstein JN, Myers T, Buolamwini J, Raghavan K, van Osdol W, Licht J, Viswanadhan VN, Kohn KW, Rubinstein LV, Koutsoukos AD, Monks A, Scudiero DA, Anderson NL, Zaharevitz DW, Chabner BA, Grever MR, Paull KD. 1994. Predictive statistics and artificial intelligence in the U.S. National cancer institute's drug discovery program for cancer and aids. *Stem Cells* 12:13–22.
92. Jackson MA, Stack HF, Waters MD. 1997. Activity profiles of carcinogenicity data—Application in hazard identification and risk assessment. *Mutat Res* 394:113–124.
93. Hileman B. 1994. Environmental estrogens linked to reproductive abnormalities, cancer. *Chem Eng News* 72:19–23.
94. Hileman B. 1997. Hormone disrupter research expands. *Chem Eng News* 75:24–25.
95. U.S. Code. 1996. Food Quality Protection Act: PL 104–170. et seq.
96. U.S. Code. 1996. Safe Drinking Water Act: Public law 104–182. Title 21 U.S. Congress §346a.97.
97. Gray LE Jr. 1998. Tiered screening and testing strategy for xenoestrogens and antiandrogens. *Toxicol Lett* 102–103:677–80.
98. Patlak M. 1996. A testing deadline for endocrine disrupters. *Environ Sci Technol* 30:540A–544A.
99. 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.

100. Fang H, Tong W, Perkins R, Soto A, Prechtl N, Sheehan DM. 2000. Quantitative comparison of in vitro assays for estrogenic activity. *Environ Health Perspect* 108:723–729.

101. Blair R, Fang H, Branham WS, Hass B, Dial SL, Moland CL, Tong W, Shi L, Perkins R, Sheehan DM. 2000. Estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands. *Toxicol Sci* 54:138–153.

102. Tong W, Perkins R, Fang H, Hong H, Xie Q, Branham W, Sheehan D, Anson J. 2002. Development of quantitative structure–activity relationships (QSARs) and their use for priority setting in testing strategy of endocrine disruptors. *Regulatory Research Perspectives* 1:1–16.

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). ?4

104. Shi L, Tong W, Fang H, Xie Q, Hong H, Perkins R, Wu J, Tu M, Blair RM, Branham WS, Waller C, Walker J, Sheehan DM. 2002. An integrated ''4-phase'' approach for setting endocrine disruption screening priorities—phase I and II predictions of estrogen receptor binding affinity. *SAR/QSAR Environ Res* 13: 69–88.

105. 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* 41:186–195.

106. Brown SD, Blank TB, Sum ST, Weyer LG. 1994. Chemometrics. *Anal Chem* 66:315R–359R.

107. Hawkins DM, Young SS, Rusinko III A. 1997. Analysis of large structure–activity data set using recursive partitioning. *Quant Struct-Act Relat* 16:296–302.

108. Engels MF, Thielemans T, Verbinnen D, Tollenaere JP, Verbeeck R. 2000. Cerberus: A system supporting the sequential screening process. *J Chem Inf Comput Sci* 40:241–245.

109. Arabie P, Hubert LJ, De Soete G. 1996. *Clustering and Classification.* World Scientific Singapore, Singapore.

110. Mirkin BG. 1996. *Mathematical Classification and Clustering.* Kluwer Academic, Boston, MA, USA.

111. Timm GE, Darr JF, Flessner CJ, Kennedy PW, Maciorowski AF, O'Bryan TR, Walker D. 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. Murtagh E. 1985. *Multidimensional Clustering Algorithms.* Physica-Verlag, Vienna, Austria.

113. Guenoche A, Hassen P, Jaumard B. 1991. Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of Classification* 8:5–30.

114. Xing L, Welsh WJ, Tong W, Perkins R, Sheehan DM. 1999. Comparison of estrogen receptor alpha and beta subtypes based on comparative molecular field analysis (CoMFA). *SAR/QSAR Environ Res* 10:215–237.

115. Brown RD, Martin YC. 1997. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J Chem Inf Comput Sci* 37:1–9.

116. Kohonen T. 1995. *Self-Organization Maps.* Springer-Verlag, Berlin, Germany.

117. Simon V, Gasteiger J, Zupan JA. 1993. Combined application of two different neural network types for the prediction of chemical reactivity. *J Am Chem Soc* 115:9148–9159.

118. Polanski J, Jarzembek K, Gasteiger J. 2000. Self-organizing neural networks for screening and development of novel artificial sweetener candidates. *Comb Chem High Throughput Screening* 3:481–495.

119. Tetko IV, Kovalishyn VV, Livingstone DJ. 2001. Volume learning algorithm artificial neural networks for 3D QSAR studies. *J Med Chem* 44:2411–2420.

120. Sadowski J. 2000. Optimization of chemical libraries by neural networks. *Curr Opin Chem Biol* 4:280–282.

121. Sherlock G. 2000. Analysis of large-scale gene expression data. *Curr Opin Immunol* 12:201–205.

122. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96:2907–2912.

123. Livingstone D. 1995. *Data Analysis for Chemists—Applications to QSAR and Chemical Product Design.* Oxford University Press, New York, NY, USA.

124. Wold S. 1976. Pattern recognition by means of disjoint principal components models. *Pattern Recogn* 8:127–139.

125. Welsh WJ, Lin W, Tersigni SH, Collantes E, Duta R, Carey MS, Zielinski WL, Brower J, Spencer JA, Layloff TP. 1996. Pharmaceutical fingerprinting: Evaluation of neural networks and chemometric techniques for distinguishing among same-product manufacturers. *Anal Chem* 68:3473–3482.

126. Collantes ER, Duta R, Welsh WJ, Zielinski WL, Brower J. 1997. Preprocessing of HPLC trace impurity patterns by wavelet packets for pharmaceutical fingerprinting using artificial neural networks. *Anal Chem* 69:1392–1397.

127. Tetko IV, Villa AE, Aksenova TI, Zielinski WL, Brower J, Collantes ER, Welsh WJ. 1998. Application of a pruning algorithm to optimize artificial neural networks for pharmaceutical fingerprinting. *J Chem Inf Comput Sci* 38:660–668.

128. Zielinski WL, Brower JF, Welsh WJ, Collantes ER, Layloff TP. 1998. A strategy for developing consistent hplc data for assessing sameness and difference in consistency of pharmaceutical products. *American Pharmaceutical Reviews* 1:44–54.

129. Aksenova TI, Tetko IV, Ivakhnenko AG, Villa AE, Welsh WJ, Zielinski WL. 1999. Pharmaceutical fingerprinting in phase space. 1. Construction of phase fingerprints. *Anal Chem* 71: 2423–2430.

130. Aksenova TI, Tetko IV, Ivakhnenko AG, Villa AE, Welsh WJ, Zielinski WL. 1999. Pharmaceutical fingerprinting in phase space. 2. Construction of phase fingerprints. *Anal Chem* 71: 2431–2439.

131. Tetko IV, Aksenova TI, Volkovich VV, Kasheva TN, Filipov DV, Welsh WJ, Livingstone DJ, Villa AEP. 2000. Polynomial neural network for linear and non-linear model selection in quantitative-structure activity relationship studies on the internet. *SAR/QSAR Environ Res* 11:263–280.

132. Rusinko A, III, Farmen MW, Lambert CG, Brown PL, Young SS. 1999. Analysis of a large structure/biological activity data set using recursive partitioning. *J Chem Inf Comput Sci* 39: 1017–1026.

133. Izrailev S, Agrafiotis D. 2001. A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J Chem Inf Comput Sci* 41:176–180.

134. Morgan JN, Sonquist JA. 1963. Problems in the analysis of survey data, and a proposal. *J Am Stat Assoc* 58:415–434.

135. Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees.* Wadsworth, Pacific Grove, CA, USA.

136. Breiman L, Friedman J, Olshen R, Stone C, Steinberg D, Colla P. 1995. Cart: Classification and regression trees, 4.0. Salford System, Stanford, CA, USA.

137. Kass GV. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29:119–127.

138. Quinlan J. 1993. *C4.5: Programs for Machine Learning.* Morgan Kauffman, San Francisco, CA, USA.

139. Clark LA, Pregibon D. 1992. *Tree-Based Models.* Thomson/Wadsworth, Stamford, CT, USA.

140. Anstead GM, Carlson KE, Katzenellenbogen JA. 1997. The estradiol pharmacophore: Ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* 62:268–303.

141. 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.

142. Rohrbaugh RH, Jurs PC. 1987. Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Analytica Chimica Acta* 11:220. ?5

143. Zupan JA, Gasteiger J. 1993. *Neural Networks for Chemists.* VCH, Weinheim, Germany.

144. Kovesdi I, Dominguez-Rodriguez MF, Orfi L, Naray-Szabo G, Varro A, Papp JG, Matyus P. 1999. Application of neural networks in structure–activity relationships. *Medical Research Reviews* 19:249–269.

145. Manallack DT, Livingstone DJ. 1999. Neural networks in drug discovery: Have they lived up to their promise? *Eur J Med Chem* 34:195–208.

146. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D,

Levine AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 96:6745–6750.

147. NetWatch. 1999. Site visit: Dipping into DNA chips. *Science* 285:799.

148. Bassett DE Jr, Eisen MB, Boguski MS. 1999. Gene expression informatics–it's all in your mine. *Nat Genet* 21(Suppl 1):51–55.

149. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9:3273–3297.

150. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531–537.

151. Kalocsai P, Shams S. 1999. Visualization and anaylsis of gene expression data. *Journal of the Association of Laboratory Automation* 4:58–61.

152. Hong H, Tong W, Fang H, Shi LM, Xie Q, Wu J, Perkins R, Walker J, Branham W, Sheehan D. 2002. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated computational approach. *Environ Health Perspect* 110:29–36.

153. Walker JD, Fang H, Perkins R, Tong W. 2003. QSARs for EDPSD 2: The Integrated 4-Phase Model. *Quant-Struct Act Relat* (in press). ?6

154. Branham WS, Dial SL, Moland CL, Hass B, Blair R, Fang H, Shi L, Tong W, Perkins R, Sheehan DM. 2000. Phytoestrogen and mycoestrogen binding to rat uterine estrogen receptor. *Am J Nutr* 132:658–664.

155. Matthews EJ, Benz RD, Contrera JF. 2000. Use of toxicological information in drug design. *J Mol Graph Model* 18:605–615.

156. Johnson DE. 2001. The optimal fragmentation principle. *Drug Discov Today* 6:175.

157. Richard AM. 2001. The optimal fragmentation principle-reply. *Drug Discovery Today* 6:235–237.