Interpretation of Quantitative Structure−Property and −Activity Relationships

Alan R. Katritzky,* Ruslan Petrukhin, and Douglas Tatham

Department of Chemistry, University of Florida, Gainesville, Florida 32611

Subhash Basak

Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunkhwy, Duluth, Minnesota 55811

Emilio Benfenati

Instituto Mario Negri, Via Eritrea 62, 20157 Milan, Italy

Mati Karelson and Uko Maran

Department of Chemistry, Tartu University, 2 Jakobi Street, Tartu EE51014, Estonia

Received September 27, 2000

The potential utility of data reduction methods (e.g. principal component analysis) for the analysis of matrices assembled from the related properties of large sets of compounds is discussed by reference to results obtained from solvent polarity scales, ongoing work on solubilities and sweetness properties, and proposed general treatments of toxicities and gas chromatographic retention indices.

INTRODUCTION

Quantitative structure−property relationships (QSPRs) now correlate chemical structure to a wide variety of physical, chemical, biological (including biomedical, toxicological, ecotoxicological), and technological (glass transition temperatures of polymers, critical micelle concentrations of surfactants, rubber vulcanization rates) properties. Suitable correlations, once established, can be used to predict properties for compounds as yet unmeasured or even unknown. Some of the present authors have recently reviewed the literature on QSPRs for diverse data sets of technologically important properties;[1] their utility is clear, and such applications will undoubtedly expand rapidly. The present article suggests that extensions to this methodology can significantly increase our understanding of how structure determines the property−behavior phenomena of chemical compounds: it attempts to bring together in one place various techniques already utilized by us and others and demonstrate their wide potential.

It is widely recognized that QSPR equations, whether they be derived in a purely empirical fashion from an arbitrary set of molecular descriptors or from a preselected set of descriptors selected on theoretical grounds for a connection with a particular property, can give considerable insight into the manner by which chemical structure controls physical and biological properties of compounds.

The present article summarizes the way in which data reduction methods, like principal component analysis (PCA) of a matrix formed by the assembly of related properties for a large set of structures, provide insight into how these related properties depend on each other in a quantitative manner. We first review the already published results obtained by using this technique on solvent polarity scales and then indicate how a similar treatment could clarify the phenomenon of solubility in general. Further potential applications of this methodology are illustrated by reference to chromatographic retention times, taste related properties, and general toxicities.

APPLICABILITY OF PCA TO QSPR TREATMENTS

It is well-established that PCA can be a tool in extracting uncorrelated and useful information from predictors (independent variables) used in quantitative structure−activity relationship (QSAR) development. For example, Basak et al.[2a,b] reported a PCA of 90 topological indices calculated for a subset of chemicals in the toxic substances control act (TSCA) inventory consisting of 3692 molecules. The first 4 principle components (PCs) explained 78.3% and the first 10 92.6% of the variance. Such PCs can be used as independent variables in principle component regressions (PCR) or as axes for defining *n*-dimensional spaces to select analogues or to predict properties of structurally similar chemicals.[2a,3a−h] Basak et al.[3b,4] also used PCs to select analogues of chemicals based on their Euclidean distance in the *n*-dimensional PC space and applied *K*-nearest neighbor (KNN) approach in estimating properties of chemicals from properties of their *K* (=1, 2, ..., 50) selected neighbors as well as in the discrimination of structurally related isospectral graphs.[5]

PCs can classify a diverse set of toxic chemicals with different modes of action (MOAs) into subsets consisting of distinct MOA classes.[3a] Applied to larger databases of toxicants, this leads to a two-tier QSAR approach for toxicity by first assigning the proper MOA class and then developing class-specific QSARs. The United States Environmental

* To whom correspondence should be addressed. Phone: (352) 392-0554. E-mail: katritzky@chem.ufl.edu.

10.1021/ci000134w CCC: $20.00
Published on Web 02/23/2001

Protection Agency (USEPA) uses two major methods in their premanufacture hazard assessment notification (PMN) of chemicals: (a) class specific QSARs if available and (b) chemical analogues.[6] PCA and PCs derived from computed molecular descriptors can be used for both purposes.

The data reduction capability of PCA has also provided a "synthetic and holistic view" of different solvent polarity scales, insight into the action of structural classes of sweeteners,[7a,b] and the solubility of chemicals in diverse solvents.[8] Analogy indicates that such approaches will yield useful results for the various toxicity scales that have been developed for the assessment of hazard posed by natural and anthropogenic chemicals to human and environmental health.

A fundamental goal of QSAR/QSPR studies is to predict complex physical, chemical, biological, and technological properties of chemicals from simpler "descriptors", preferably those calculated solely from molecular structure, excluding experimental data.[9] To this end, numerous experimental and computed descriptors have been developed for QSAR/QSPR studies.[10] Any descriptor, whether experimental or calculated, associates a real number with a chemical and then orders the set of chemicals according to the numerical value of the specific property. Each descriptor or property provides a scale for a particular set of chemicals. Thus an experimentally determined solvent polarity scale orders a set of solvents according to the magnitude of the solvent polarity as defined by the scale. Similarly, the magnitude of the molecular complexity descriptor (e.g., the first-order information content, IC_1) maps a set of chemicals into a corresponding set of real numbers, and orders them into a scale.[11a,b] If such a scale (independent variable), experimental or calculated, is linearly or nonlinearly related to the magnitude (scale) of a particular physical, chemical, biological, or technological property (dependent variable) of interest, this provides a successful QSAR/QSPR model. Multiple linear regressions have been very popular in the formulation of QSARs/QSPRs.

The partial least squares (PLS) method is particularly suited for the extraction of a few highly significant formal correlational factors from large homogeneous sets of descriptors such as molecular field grid data (cf. comparative molecular field analysis, CoMFA).[12a] However, this approach is often less appropriate in cases of large diverse descriptor sets, as its use can result in the selection of too many formal factors.[12b] Therefore, in this paper we consider an alternative approach that combines the PCA and multilinear regression analysis. The new approach is outlined on the basis of the previous work by our groups.

FURTHER MODEL BUILDING METHODS

For more complicated situations, several statistical methods can be used for flexible nonlinear modeling, including the following: polynomial regression; tree-based models; Bayesian methods.

For instance, Trinajstic and co-workers used nonlinear multivariate regression to predict biological and pharmacological properties.[13a] Methods of machine learning have also been used in the development of QSAR/QSPR models. In the 1990s, regression methods based on neural networks (NNs) offered new possibilities to QSARs, accounting for nonlinear structure−activity relationships and dealing with nonlinear dependencies.[13b] Repeatedly, NNs proved to be equal or superior to multivariate regression.[13a,b,14a,b] Artificial intelligence offers advantages in dealing with numerical continuous values and also with categories and rules.[15] Machine learning research seeks to develop algorithms that learn predictive relationships from data by data mining and knowledge discovery techniques. Fuzzy logic can be used to keep into account the uncertainty of the property of interest, e.g., the magnitude of a toxicological value.

Numerous QSAR/QSPR models apply both statistical and neural net methods, with a single or a small set of independent variables, to small and structurally related compound sets. Complex properties such as solvent polarity, sweetness, and the toxicity of structurally diverse chemicals (e.g. those in the TSCA inventory) call for broader integrated approaches, rather than such piecemeal methods. Techniques such as PCA provide holistic approaches for combining many independent variable (descriptor) scales for deriving QSAR/QSPR models for complex properties and larger sets of compounds.

The data of p descriptors for n chemicals form an $n \times p$ matrix X. Each chemical is now a point in the p-dimensional space, R^P. Since many descriptors, whether experimental or calculated, are significantly intercorrelated, the points in R^P will in fact define a subspace of lower dimension than p, and, as discussed above, PCA can provide PCs which represent reduced data and efficiently combine diverse predictor variables. The PCs can then be used in the prediction of properties, quantification of structural similarity/dissimilarity of chemicals, and the clustering of large and diverse combinatorial libraries of chemicals.[16]

PCs may find applications in the clustering and classification techniques for complex properties such as toxicity. While classification methods may appear crude, compared to multilinear analysis and NN, given the huge variability of the toxic effects, they will be suitable for the preliminary treatment of large sets of data.[17]

SOLVENT POLARITY SCALES[8,18]

Solvent polarity is widely recognized to be of great importance in many fundamental and applied areas of research. However, the precise definition of solvent polarity has proved difficult. More than a hundred quantitative solvent polarity scales have been proposed on the basis of diverse properties, including reaction rates, solvatochromic effects, and entropies. In recent joint work of two of our laboratories, a matrix was formed from 40 of the most important scales and 40 of the most important solvents. However, there were many gaps in this database. A QSPR was established for each of the 40 scales,[18] and this was then used to fill in all the gaps in the matrix. The principal component analysis of this matrix[8] showed that the first three principal components accounted for about 75% of the total variance. These components described 22 of the scales very well (greater or equal to 90% of variance), another 14 were well or fairly well described (70−89% of the variance), and 4 were rather poorly described, the 54−65% of variance.

A three-dimensional plot using the loadings of the first three PCs as axes gave very useful information on the scales; see Figure 1. In particular, most scales fell into five groups as follows: (i) expression of dielectric constant; (ii) charge-transfer effects on electronic spectra; (iii) other UV spectral

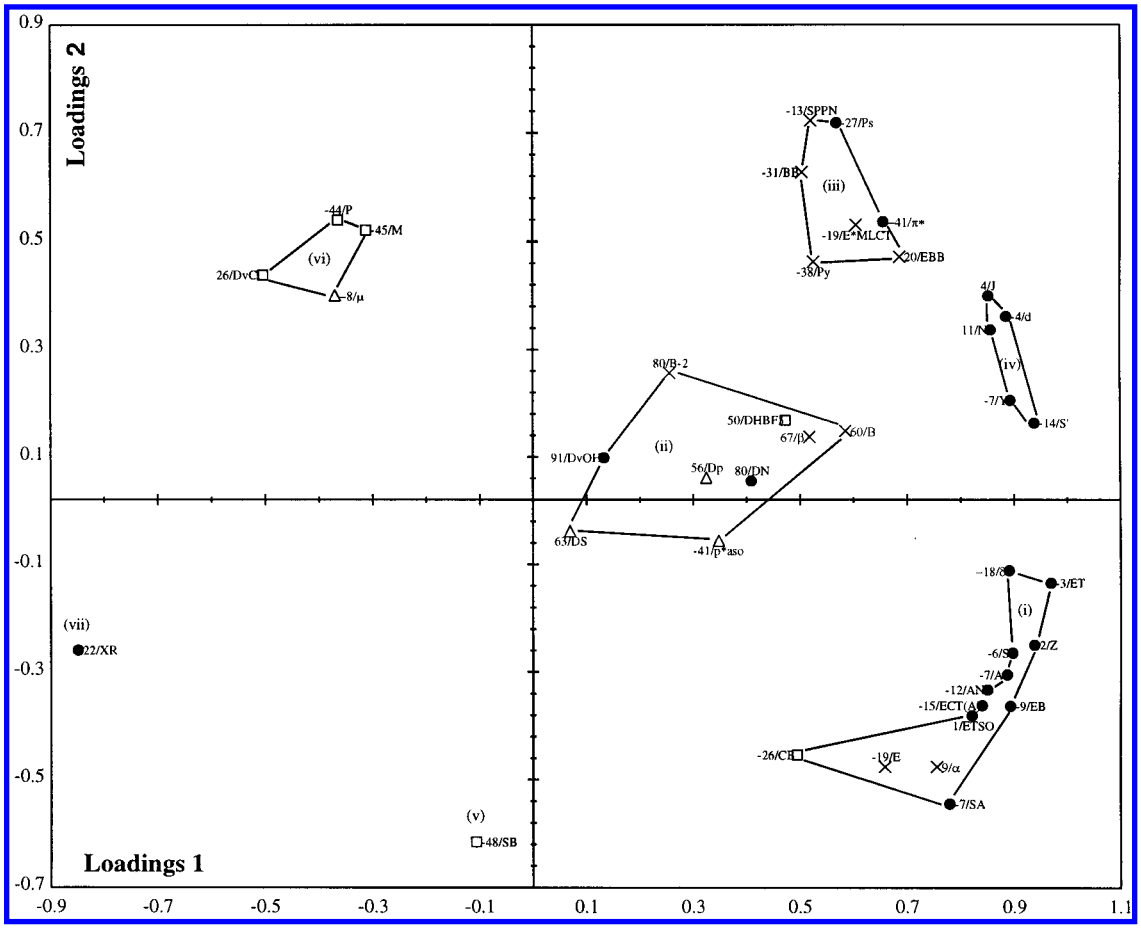

Figure 1. Loadings of the second PCA component plotted versus the loadings of the first component with the third component loading and scale classification given as labels to the data points. Reprinted with permission from ref 8. Copyright 1999 American Chemical Society.

effects; (iv) expression of solvent basicity; (v) expression of solvent refractive index.

Similarly, a three-dimensional plot of the scores of the first three PCs (Figure 2) gave information on the solvents. In particular, the hydroxylic solvents appeared in one group, the dipolar−aprotic solvents in another, and polar solvents in yet another group (well-separated from the nonpolar solvents). Only formamide remained as a single group.

In this way considerable information and rationalization was obtained for both solvents and solvent polarity scales.

General Treatment of Solubility. Consideration of the solubilities of solids or liquids in a liquid solvent is complicated by the need to consider intermolecular interactions in the bulk solute in addition to those in the bulk solvent and between the solute and the solvent. Thus, it is easier to treat the solubilities of vapors and gases, i.e., gas−liquid partition coefficients. Moreover, such gas−liquid partition coefficients are extremely important from an environmental point of view, especially when the liquid is water. Therefore it is of great utility to have the structure-based chemical information on gas solubilities generalized.

Water-gas phase partition coefficients of diverse organic compounds can be adequately described using descriptors based solely on the chemical structures of the organic molecules: a web site is available (http://clogp.pomona.edu/medchem/chem/qsar-db). The partitioning of two sets of organic gases and vapors between water and air (L_w) has been studied using the CODESSA program.[19] For a set of 95 alkanes, cycloalkanes, alkylarenes, and alkynes, excellent predictions were obtained with a two-parameter correlation equation ($R^2 = 0.977$, $R^2_{cv} = 0.975$) that adequately represented the effective dispersion and cavity formation effects for the solvation of nonpolar solutes in water. A set of 406 structurally diverse organic compounds (including structures containing N, O, S, and halogen atoms) was successfully correlated by a five-parameter equation ($R^2 = 0.941$, $R^2_{cv} = 0.939$),[19] which accounts for the dispersion energy of polar solutes in solution, the electrostatic part of the solute−solvent interaction, and hydrogen-bonding interactions in liquids.

We recently obtained similar equations for the solubilities of organic molecules in methanol and ethanol.[20] The solubilities of 87 gases and vapors in methanol resulted in a four-parameter equation ($R^2 = 0.945$, $R^2_{cv} = 0.938$) that adequately represents the solute−solvent interactions described by the polarizability, dipole moment, hydrogen bonding, and lipophilicity. The solubilities in ethanol of 61 gases and vapors also yielded a four-parameter equation ($R^2 = 0.969$, $R^2_{cv} = 0.964$), where the solute−solvent intercorrelations, similar to those of methanol, include electrostatic and hydrogen-bonding interactions.

We plan to extend this work to a variety of other solvents, including polar aprotic solvents such as dimethylformamide, dimethyl sulfoxide, nitrobenzene; polar solvents such as chloroform and ethyl acetate; and nonpolar solvents such as hexane and benzene. This will provide a matrix between the solvents and solutes; vacancies in the matrix will be calculated using the correlations already obtained. A principal

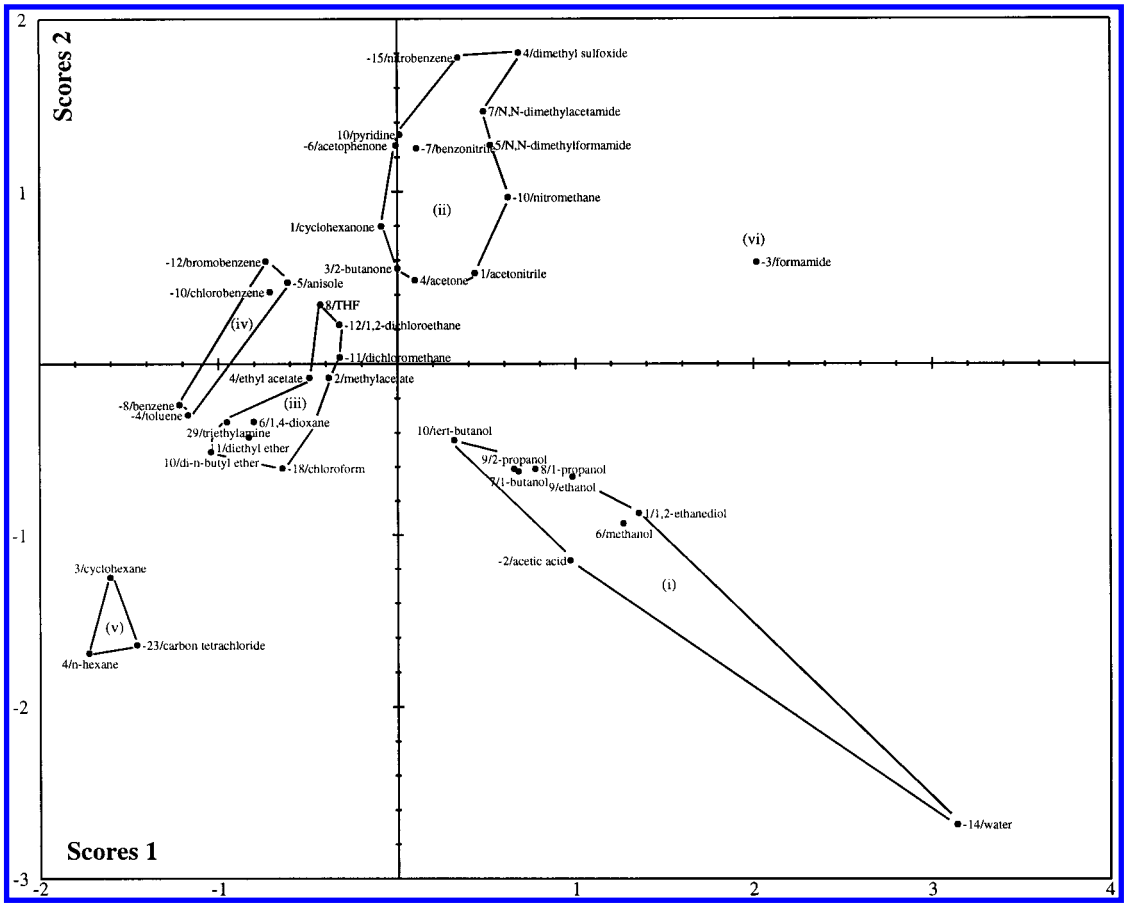

Figure 2. Plot of the scores of the second component versus the scores of the first component with the third component loading and scale classification given as labels to the data points. Reprinted with permission from ref 8. Copyright 1999 American Chemical Society.

component analysis on this matrix, similar to that described above for polarity scales, should provide a set of loadings which will characterize the solvents and a set of scores that will characterize the solutes. We believe that examination of the patterns for the loadings and the scores will present useful information and insight into the general phenomenon of solubility.

Gas Chromatographic Retention Times. It would be advantageous to systematize gas chromatographic (GC) retention times based on the chemical structure. We recently reviewed the enormous amount of data on the QSPR and related analyses of GC retention times.[1] A systematic treatment should illuminate the structural dependencies between the eluted compound and various stationary phases in GC.

Several authors have estimated retention indices using topological descriptors.[21a−e] Charged partial surface area (CPSA) descriptors[22] have also been successfully combined with topological and geometrical descriptors to predict retention indices of substituted pyrazines,[23] polycyclic aromatic compounds,[24] stimulants and narcotics,[25] and anabolic steroids.[26] The CPSA descriptors encode information about charge distribution and surface areas, which relates to interactions between the eluted compounds and molecules in the stationary phase.

As the polarity of the stationary phase changes, the influence of the charge distribution of the eluted molecules changes, and different descriptors become important. Therefore, each phase has to be modeled separately. Diverse classes of descriptors extend the pool of information and consequently should result in a better description of the property. Indeed, topological descriptors can be successfully combined with quantum-chemical descriptors to predict GC retention indices.[27a−c] Quantum chemical descriptors also encode information about the charge distribution and polarity of molecules and were capable of handling specific effects of the stationary phase. Even alone, quantum chemical descriptors can be useful for this type of study, as indicated by the theoretical linear solvation energy relationship (TLSER) established for GC retention indices.[28]

Our QSPR analysis of GC retention times utilized a mixed set of topological and quantum-chemical descriptors to model 152 structures, including a wide cross-section of classes of organic compounds.[27b] A forward procedure for the selection of molecular descriptors in the multilinear regression analysis in the CODESSA program gave a six-parameter model ($R^2 = 0.959$, $R^2_{cv} = 0.955$), with polarizability being the most important descriptor in the model. These results were recently reevaluated using improved procedures in CODESSA and new methods for the efficient selection of variables in the multilinear regression analysis.[27c] In more recent work,[29] we analyzed a set of 178 methyl-branched hydrocarbons to give a four-parameter model ($R^2 = 0.9585$, $R^2_{cv} = 0.9543$) combining topological and quantum chemical descriptors.

Considering the amount of information encoded into descriptors, insight into the general phenomenon of gas−solid absorption could be obtained by combining QSPRs and subsequent PCAs of a matrix of retention times of a diverse set of compounds using a range of solid phases in GC. It would of course be necessary to make all the GC measure-

ments under the same experimental conditions such as the length of the column, the temperature of the column, the nature of the carrier gas, and the speed of the carrier gas.

Sensory Properties. In unpublished work,[7b] we have provided QSPRs for the sweetness property, defined as the dimensionless ratio of the concentration of the alternative sweetener to the concentration of sucrose, which has an equally sweet taste. For a comprehensively referenced set of 348 natural and artificial sweeteners, the treatment of data using the linear and nonlinear regression methods of the CODESSA software package resulted in a global three-parameter correlation with $R^2 = 0.71$. Significantly more reliable models were developed for various subclasses of compounds (peptides, aldoximes, acesulfamates and sulfamates, guanidines, ureas and thioureas, and various natural sweeteners).

Following the general idea now expounded, it would be of substantial interest to extend this investigation by applying the QSPR treatment to other sensory properties of compounds. It is known that taste reception is localized in four regions of the tongue, corresponding to the sensations of sweetness, saltiness, sourness, and bitterness, each related to different receptors.[30] Nevertheless, according to the approach described above for other properties, all these gustatory properties should be treatable simultaneously using a combination of QSPR with PCA. Furthermore, it has been observed that the gustatory properties of certain compounds can be interrelated with the corresponding olfactory properties.[7a,31a,b] Consequently, a combined QSPR/PCA treatment may also be feasible for the sets of data on both sensory properties. Extensive data on olfactory properties have been collected and systematized using the QSAR approach.[32]

General Treatment of Toxicities. A general treatment could determine underlying relationships between different measures of toxicity. Although toxicity is far more complex than the topics previously discussed in this paper, we believe that the method could make a significant contribution to the analysis, classification, and understanding of toxicity.

A comparison with the treatment of solvent polarity scales mentioned above is illuminating. Between 100 and 200 solvent polarity scales have been formulated, and perhaps 400 or 500 solvents were examined. The numbers for toxicity are far larger: many different measures of toxicity have been used depending on species, concentration, mode of application, and duration. The number of compounds, on which at least one measure of toxicity has been obtained, ranges up to six figures. Despite this complexity, the method could investigate (i) general interrelationships between various types of toxicity and (ii) interrelationships between structures in determining toxicity.

The enormous amount of experimental data available makes this attempt challenging. Moreover, considering the data as a matrix of compounds against toxicities, the matrix is very fragmentary: there are far more missing than available data points. Work on multidimensionality problems so far has centered on much simpler topics. It is very difficult to compare the performances of the multitude of different models reported for the prediction of toxicity because they refer to a multitude of situations: different toxicological endpoints, chemical descriptors, mathematical algorithms, and data sets. Some toxicological endpoints can be explained more easily than others. For instance, narcosis in fish is related to nonspecific mechanisms, while carcinogenesis is the result of several complex phenomena involving many biological and chemical steps. Furthermore, it is easier to model the toxicity of a congeneric set of compounds, for instance, a homologous series, while it is more difficult to extrapolate the behavior of chemicals of vastly diverse chemical classes.

In addition, there is the problem of the variability of the biological data arising from the chemical purity of the compound under study, the variability of the protocol, and biological variability. We may be able to avoid much of the variability resulting from the chemical purity and protocol, but the reproducibility of biological tests is much lower than that for other properties, such as gas chromatographic retention times. Such variability is particularly relevant when dealing with reduced sets of data.

QSAR models for various toxicities have been collected.[33] Hermens co-coordinated a project in which QSAR models for aquatic toxicity were reviewed:[34] log *P* was the parameter most frequently related to toxicity, but it is insufficient to explain all the toxicological properties.[35a,b] Many other descriptors can be used in order to predict toxicity better; for example, Basak et al. compared topological, geometrical, and quantum-chemical parameters in predicting mutagenicity,[36] aquatic toxicity,[36] and dermal penetration[37] of chemicals.[38] Of course, chemical descriptors can be combined and selected, to take advantage of the most useful parameters; this has been done, for instance, in a study of genotoxicity using multilinear regression.[39]

A huge number of parameters describing a compound can be measured or computed, but how to deal with this high-dimensional information is a problem. In many cases no *a priori* knowledge on the role of parameters in determining a property is available. In this situation, a selection of the variables is needed to reduce the complexity of the description, using for instance PCAs (which imply linearity of the model) or genetic algorithms (which may also keep into account nonlinearity).

The complexity of toxicity stems from the following: the toxicological aspect, the chemical information, the mathematical approach, the dimension, and the diversity of the set of chemicals. To investigate all of these points, in an ongoing project a data set of compounds presenting six different toxicological endpoints has been compiled. About 200 chemical descriptors were calculated for these compounds, and different computational models were evaluated. Preliminary results[40a,b] indicated the feasibility of the approach; however, a wider data set is required, both for the number of compounds and for the number of toxicological endpoints.

Quo Vadis? We are suggesting a transition from the familiar one-dimensional QSAR/QSPR treatments, where the variation of a single property with structure is studied, to a general multidimensional treatment. This implies the simultaneous study of many descriptors or the study of the utilization of orthogonal variables extracted from many descriptors in the development of QSAR/QSPR models. Such models should be based solely on parameters that can be calculated directly from the molecular structure using computer algorithms without any input of experimental data. This is essential because even the simplest experimental properties are not available for many known environmental

pollutants and most chemicals of real or virtual combinatorial libraries. This more general approach is also advantageous for real applications: for example, it is much more useful to have a model that takes into account the numerous toxicological endpoints related to an aquatic ecosystem than a model which accounts only for a single endpoint such as lethality in *Daphnia*. Thus, the approach should provide additional insight into QSAR/QSPR by the application of data-reduction methods such as PCA to property/structure matrices.

ACKNOWLEDGMENT

We thank Dr. Dennis Hall for comments on the draft manuscript.

REFERENCES AND NOTES

(1) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(2) (a) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Topological Indexes−Their Nature, Mutual Relatedness, and Applications. *Math Model.* **1987**, *8*, 300−305. (b) Basak S. C.; Niemi G. J.; Regal, R. R.; Veith, G. D. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indexes. *Discrete Appl. Math.* **1988**, *19* (*1*−*3*) 17−44.

(3) (a) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical, and Neural Methods for Predicting Toxic Modes of Action. *Environ. Toxicol. Chem.* **1998**, *17*, 1056−1064. (b) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Estimation of Molecular-Properties. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 366−372. (c) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243−272. (d) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modell. Sci. Comput.*, in press. (e) Basak, S. C.; Grunwald, G. D. Development and Application of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modell. Sci. Comput.*, in press. (f) Basak, S. C.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Math. Modell. Sci. Comput.*, in press. (g) Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Sci.* **2000**, *40*, 801−809. (h) Basak, S. C.; Grunwald, G. D. Tolerance Space and Molecular Similarity. *SAR QSAR Environ. Res.* **1995**, *3*, 265−277.

(4) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the Molecular Similarity of Chemicals Using Topological Invariants. In *Advances in Molecular Similarity*; JAI Press: Greenwich, CT, 1996; Vol. 2, pp 171−185.

(5) Balasubramanian, K.; Basak, S. C. Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367−373.

(6) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure−Activity Relationships (SAR) Under TSCA, Section 5. *Environ. Health Perspect.* **1990**, *87*, 183−197.

(7) (a) Jurs, P. C.; Bakken, G. A.; McClelland, H. E. Computational Methods for the Analysis of Chemical Sensor Array Data from Volatile Analytes. *Chem. Rev.* **2000**, *100*, 2649−2678. (b) Katritzky, A. R.; Petrukhin, R.; Karelson, M.; Prakash, I.; Desai, N. Sweetness Correlations Using CODESSA. Part I. Manuscript in preparation.

(8) Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. A Unified Treatment of Solvent Properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 692−698.

(9) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(10) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.

(11) (a) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609. (b) Johnson, M.; Basak, S. C.; Maggiora, G. A Characterization of Molecular Similarity Methods for a Property Prediction. *Math. Comput. Modell.* **1988**, *11*, 630−634.

(12) (a) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am Chem. Soc.* **1988**, *110*, 5959−5967. (b) Höskuldsson, A. PLS Regression Methods. *J. Chemom.* **1988**, *2*, 211−228.

(13) (a) Lucic, B.; Trinajstic, N. Multivariate Regression Versus Artificial Neural Networks in QSAR. Second Indo-US Workshop on Mathematical Chemistry. Duluth, MN, May 30−June 3, 2000. (b) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. Correlation between Structure and Normal Boiling Points of Haloalkanes C_1–C_4 Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118−1121.

(14) (a) Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; Bruschi, M. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1076−1080. (b) Basak, S. C.; Gute, B. D.; Grunwald, G. D.; Optiz, D. W.; Balasubramanian, K. Use of Statistical and Neural Net Methods in Predicting Toxicity of Chemicals: A Hierarchical QSAR Approach. In *Predictive Toxicology of Chemicals*: *Experiences and Impact of AI Tools*; Gini, G. C., Katritzky, A. R., Eds.; AAAI 1999 Spring Symposium Series; AAAI Press: Menlo Park, CA, 1999; pp 108−111.

(15) Benfenati, E.; Gini, G. Computational Predictive Programs (Expert Systems) in Toxicology. *Toxicology* **1997**, *119*, 213−225.

(16) Basak, S. C.; Mills, D.; Gute, B. D.; Balaban, A. T.; Basak K.; Grunwald, G. D. Use of Mathematical Structural Invariants in Analyzing Combinatorial Libraries: A Case Study with Psoralen Derivatives. In *Aspects of Mathematical Chemistry*; Sinha, D. K., Basak, S. C., Mohanty, R. K., Basumallick, I. N., Eds.; Visva Bharati University Press: in press.

(17) Benfenati, E.; Lorenzini, P.; Grasso, P.; Gini, G. Classification Experiments for the Prediction of Pesticide Ecotoxicity. Second Indo-US Workshop on Mathematical Chemistry. Duluth, MN, May 30−June 3, 2000.

(18) Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. QSPR Treatment of Solvent Scales. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 684−691.

(19) Katritzky, A. R.; Mu, L.; Karelson, M. A QSPR Study of the Solubility of Gases and Vapors in Water. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162−1168.

(20) Katritzky, A. R.; Tatham, D. B.; Maran, U. The Correlation of the Solubility of Gases and Vapors in Methanol and Ethanol with Their Molecular Structures. *J. Chem. Inf. Comput. Sci.* Accepted for publication.

(21) (a) Michotte, Y.; Massart, D. L., Molecular Connectivity and Retention Indexes. *J. Pharm. Sci.* **1977**, *66*, 1630−1632. (b) Bonchev, D.; Mekenjan, O.; Protic, G.; Trinajstic, N. Application of Topological Indices to Gas Chromatographic Data: Calculation of the Retention Indices of Isomeric Alkylbenzenes. *J. Chromatogr.* **1979**, *176*, 149−156. (c) Kier, L. B.; Hall, L. H. Molecular Connectivity Analysis of Structure Influencing Chromatographic Retention Indices. *J. Pharm. Sci.* **1979**, *68*, 120−122. (d) Duvenbeck, Ch.; Zinn, P. List Operations on Chemical Graphs. 3. Development of Vertex and Edge Models for Fitting Retention Index Data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 211−219. (e) Duvenbeck, Ch.; Zinn, P. List Operations on Chemical Graphs. 4. Using Edge Models for Prediction of Retention Index Data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 220−230.

(22) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(23) Stanton, D. T.; Jurs, P. C. Computer-Assisted Prediction of Gas Chromatographic Retention Indices of Pyrazines. *Anal. Chem.* **1989**, *61*, 1328−1332.

(24) Whalen-Pedersen, E. K.; Jurs, P. C. Calculation of Linear Temperature Programmed Capillary Gas Chromatographic Retention Indices of Polycyclic Aromatic Compounds. *Anal. Chem.* **1981**, *53*, 2184−2187.

(25) Georgakopoulos, C. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Stimulants and Narcotics. *Anal. Chem.* **1991**, *63*, 2021−2024.

(26) Georgakopoulos, C. G.; Tsika, O. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas-Chromatographic Relative Retention Times of Anabolic Steroids. *Anal. Chem.* **1991**, *63*, 2025−2028.

(27) (a) Buydens, L.; Massart, D. L.; Geerlings, P. Prediction of Gas Chromatographic Retention Indexes with Topological, Physicochemical, and Quantum Chemical Parameters. *Anal. Chem.* **1983**, *55*, 738−744. (b) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relashionship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807. (c) Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multire-

gression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 610−621.
(28) Donovan W. H.; Famini, G. R. Using Theoretical Descriptors in Structure Activity Relationships: Retention Indices of Sulfur Vesicans and Related Compounds. *J. Chem. Soc., Perkin Trans. 2* **1996**, 83−89.
(29) Katritzky, A. R.; Chen, K.; Maran, U.; Carlson, D. A. QSPR Correlation and Predictions of GC Retention Indexed for Methyl-Branched Hydrocarbons Produced by Insects. *Anal. Chem*. **2000**, *72*, 101−109.
(30) Shallenberger, R. S. Taste Recognition Chemistry. *Pure Appl. Chem*. **1997**, *69*, 659−666.
(31) (a) Nahon, D. F.; Roozen, J. P.; De Graaf, C. Sensory Evaluation of Mixtures of Maltitol or Aspartame, Sucrose and an Orange Aroma. *Chem. Senses* **1998**, *23*, 59−66. (b) Nahon, D. F.; Roozen, J. P.; De Graaf, C. Sensory Evaluation of Mixtures of Sodium Cyclamate, Sucrose and an Orange Aroma. *J. Agric. Food Chem.* **1998**, *46*, 3426−3430.
(32) Rossiter, K. J. Structure−Odor Relationships. *Chem. Rev*. **1996**, *96*, 3201−3240.
(33) http://clogp.pomona.edu/medchem/chem/qsar-db/search.html.
(34) Hermens, J. QSAR for Prediction of Fate and Effects of Chemicals in the Environment. Final Report, European Commission, Project EV5V-CT92-0211.
(35) (a) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Models of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales Promelas*). *Environ. Toxicol. Chem*. **1997**, *16*, 948−967. (b) Klopman, G.; Saiakhov, R.; Rosenkranz, H. S.; Hermens, J. L. M. Multiple Computer-Automated Structure Evaluation Program Study of Aquatic Toxicity 1: Guppy. *Environ. Toxicol. Chem*. **1999**, *18*, 2497−2505.
(36) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Assessment of the Mutagenicity of Aromatic Amines from Theoretical Structural Parameters: A Hierarchical Approach. *SAR QSAR Environ. Res*. **1999**, *10*, 117−129.
(37) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res*. **1999**, *10*, 1−15.
(38) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Relative Effectiveness of Topological, Geometrical, and Quantum-Chemical Parameters in Investigating Mutagenicity of Chemicals. In *Quantitative Structure−Activity Relationships in Environmental Sciences, VII*; Chen, F., Schuurmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 245−261.
(39) Maran, U.; Karelson, M.; Katritzky, A. R. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat*. **1999**, *18*, 3−10.
(40) (a) Benfenati, E.; Pelagatti, S.; Grasso, P.; Gini. G. COMET: The Approach of a Project in Evaluating Toxicity. In *Predictive Toxicology of Chemicals*: *Experiences and Impact of AI Tools*; Gini, G. C., Katritzky, A. R., Eds.; AAAI 1999 Spring Symposium Series; AAAI Press: Menlo Park, CA, 1999; pp 40−43. (b) Gini, G.; Lorenzini, M.; Vittore, A.; Benfenati, E.; Grasso, P. Some Results for the Prediction of Carcinogenicity Using Hybrid Systems. In *Predictive Toxicology of Chemicals*: *Experiences and Impact of AI Tools*; Gini, G. C., Katritzky, A. R., Eds.; AAAI 1999 Spring Symposium Series; AAAI Press: Menlo Park, CA, 1999; pp 139−143.

CI000134W