A Toxicity Evaluation and Predictive System Based on Neural Networks and Wavelets

P. L. Piotrowski,[†] B. G. Sumpter,[‡] H. V. Malling,[§] J. S. Wassom,[§] P. Y. Lu,*[,†] R. A. Brothers,[§] G. A. Sega,[§] S. A. Martin,[†] and M. Parang[§]

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, Computer Science and Mathematics Division and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, and YAHSGS, LLC, 3100 George Washington Way, Suite 103, Richland, Washington 99354

Received October 31, 2006

A computational approach has been developed for performing efficient and reasonably accurate toxicity evaluation and prediction. The approach is based on computational neural networks linked to modern computational chemistry and wavelet methods. In this paper, we present details of this approach and results demonstrating its accuracy and flexibility for predicting diverse biological endpoints including metabolic processes, mode of action, and hepato- and neurotoxicity. The approach also can be used for automatic processing of microarray data to predict modes of action.

I. INTRODUCTION

Comprehensive laboratory testing has been one of the most direct approaches for meeting the daunting requirements for evaluating toxicity and assessing risks of diverse chemicals and materials. Accurately assessing risk from novel chemicals against a broad spectrum of end points requires a depth of chemical, physicochemical, and toxicological data and interpretative expertise that are prohibitive to obtain through experimental approaches. These approaches can cost millions of dollars (involving several thousand test animals) and take five or more years to complete. As a result, very few chemicals or materials have undergone the degree of testing needed to support accurate health risk assessments and informed decision making.

In a report released over two decades ago, the National Research Council noted that toxicity data suitable for conducting health-hazard assessments were unavailable for almost 80% of the chemicals in general commerce, and adequate test data existed for only 10% of the substances.[1] In 1994, the Government Accounting Office reported that the Environmental Protection Agency (EPA) had fully reviewed only about 2% of the existing chemicals in commerce.[2] There are now well over 14 million known compounds, with thousands of new ones being developed each year. Given the number of uncharacterized compounds, the production rate of new ones, and the cost of testing, it is clear that laboratory-based approaches alone will not provide the assessments needed. Meeting this demanding challenge will require the development of alternate methods to use in combination with conventional testing. Federal agencies such as the National Institutes of Health (including the National Institute of Environmental Health Sciences) and the EPA, as well as the armed forces and private industry, would benefit greatly from a robust computational synthesis and toxicity assessment/evaluation system. Such a tool would enable predictions of potential health risks resulting from exposure to new chemicals, materials, and mixtures, or chemical byproducts due to reactions or metabolic processes, interactions with other chemicals, molecular aging, or biodegradation.

An integrated toxicity evaluation and prediction system must be based on the successful implementation of a computational scheme capable of predicting chemical toxicity on diverse data, including toxicogenomic and proteomic data. Such a system requires a number of highly innovative components including (1) automatic structural characterization of chemicals, (2) accurate chemical similarity recognition, (3) integration of multimedia data from internal and external sources, (4) accurate estimation of toxicity (all categories), and (5) estimation of confidence levels for predictions. Other useful characteristics would include capabilities that support an estimation of exposures for specific operational scenarios and an integration of exposure and toxicity data to predict scenario-specific outcomes.

We have developed an integrated modular approach that incorporates many of these capabilities. This system uses computational modules based on quantitative structure−activity relationships (QSARs) that perform specific tasks and can communicate "seamlessly" with other computational modules. This paper describes the recent implementation and testing of QSAR modules and microarray-analysis modules implemented using neural networks coupled with wavelet analysis.

II. METHODS

The basic idea behind QSAR was inspired by the pioneering work of Hansch and Leo who demonstrated that the biological activity of chemical compounds is a mathematical function of their physicochemical characteristics (e.g., hydrophobicity, size, and electronic properties).[3,4] This general approach involves the calculation of a number of physico-

* Corresponding author phone: (865) 574-7587; e-mail: lupy@ornl.gov.
[†] Environmental Sciences Division, Oak Ridge National Laboratory.
[‡] Computer Science and Mathematics Division and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory.
[§] YAHSGS, LLC.

10.1021/ci6004788 CCC: $37.00
Published on Web 02/13/2007

chemical characteristics for each molecule and the application of statistical regression analyses to find the best equation(s) that correlate with a biological activity (e.g., anticarcinogenic effectiveness, ecotoxicological behavior, neuron receptor affinity, or toxicity).[3,4] Two goals of this approach are to determine structural features influencing activity and to find a model equation that can be used to predict new candidate pharmaceuticals and commercial chemicals that will be useful for their intended purposes, while causing few or no adverse health effects to humans. Computational neural networks are ideally suited for application to these problems, and our past research has demonstrated good success in this area.[5,6]

Several computational methods can be used to make the correlation(s) between the structure-based input and the biological endpoint; these include neural networks, fuzzy logic, evolutionary algorithms, and machine learning, as well as rule-based and statistics-based methods. Regardless of the method chosen, however, a description of the most important variables for the molecular species theoretically will allow a better generalization to a broader class of systems and target properties. The literature testifies to the great variety of possible molecular descriptors, ranging from topological indices to properties determined from quantum chemistry computations. The fundamental question raised by using QSAR analysis is how to determine the relevant variables that satisfactorily represent dependencies between the activity and structure. Our goal is to optimize the representation of geometrical and electronic molecular structures for use as descriptors in a neural-network-based QSAR analysis.

Three-Dimensional Characterizations. The representation of molecular structures profoundly influences the level of insight that can be gained from chemical information. In the past, chemical structures were largely represented by fragment codes and line notations. For example, the Wiswesser line notation[7] allowed a highly concise coding of chemical structures but was insufficient for representing chemical reactions in which individual bonds were broken and made. This deficiency led to the development of connection tables, a method that has gained universal acceptance.[8] This method uses a unique and unambiguous coding of a chemical structure by canonical numbering of the atoms in the molecule (see Figure 1 of the Supporting Information). Connection tables show that molecules consist of atoms and bonds, basically providing a valence-bond representation of the molecule. Although this representation has the advantage of requiring a minimal set of numbers, it provides only two-dimensional information. Additionally, connection tables are inadequate for representing structures that are better described by molecular orbital representations. Because molecules are three-dimensional (3-D) objects, any in-depth analysis of chemical information, particularly an analysis of relationships between structure and physical, chemical, or biological properties, must take this dimensionality into account. Biological activity is intimately tied to the 3-D molecular structure and to electronic properties of specific molecular sites. Thus, it is crucial to develop molecular descriptors that encode the full 3-D structural information. This structural information has been determined using X-ray analysis for more than 100 000 organic compounds and is stored in the Cambridge Crystallographic Database.[9] As noted previously, however, the number of known chemical compounds exceeds 14 million. This situation highlights a severe deficiency of data on 3-D structures.

The increased computational power offered by integrated chip technology now makes it possible to perform real-time computations of 3-D structures based on information contained in connection tables. These algorithms belong to a class of computational tools called molecular mechanics. Our molecular mechanics codes use highly efficient quasi-Newton Raphson techniques combined with a unique geometric statement function technique that optimizes the first and second derivatives to perform the minimization of a universal potential energy function. The codes are further supplemented with simulated annealing and other Newton methods. Other available software described in the literature, such as CONCORD,[10] ALCOGEN,[11] CHEM-X[12], MOLGEO,[13] COBRA,[14] CORINA,[15] and CONVERTER,[16] are based on similar ideas but often give mixed results in terms of general conversion capability, time to perform the conversion, and accuracy as compared to X-ray structures.

Conversion of 3-D Information to a Vector or Matrix. Once the 3-D structure has been obtained, the next step is to represent the information in a vector or matrix. The most obvious approach is to use the Cartesian or internal coordinates of the individual atoms of the molecule. Since each atom requires three coordinates, the size of the descriptor will reflect the number of atoms contained in the molecule ($3N$ numbers). This presents a problem since most methods used to make correlations between structure and activity require each molecule of a data set to be represented by the same number of variables. One possible solution is to use a molecular transform such as the one used in experimental electron-diffraction studies. An efficient equation can be derived with this approach, in which the $3N$ data is converted into S, where S is a resolution control and can be taken as any number; studies have shown that $S = 32$ is sufficient to distinguish most compounds.[17] We have found that this approach provides reasonably accurate descriptions of the molecule.

Another option is to use data compression techniques to convert the full $3N$ values for the structure into N values. In our investigations, the best techniques for this purpose come from signal processing. One such method uses wavelets, which offer a number of advantages, including simultaneous localization in time and frequency domains. Wavelets are mathematical functions that divide data into different frequency components and then study each component with a resolution matched to its scale (analysis according to scale).[18] A wavelet prototype function, called the *mother wavelet* is used for the analysis. Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, and frequency analysis is performed with a dilated, low-frequency version of the same wavelet. Because the original signal or function can be represented in terms of a wavelet expansion, data operations can be performed using just the corresponding wavelet coefficients. Truncation of the coefficients below a threshold gives a sparse representation of the original signal and makes wavelets an excellent tool for data compression. In the current implementation of wavelets with neural networks, we utilize this capability to reduce the length of the input vectors that are needed to describe the three-dimensional structure of the various molecular compounds. A discrete wavelet transform is used with a four-

coefficient Daubechies mother wavelet. The wavelet transform of the input vector that defines the molecular structure (consisting of 32 components) is truncated to produce a new input vector consisting of eight components. Wavelet transforms efficiently compress and de-noise the data by eliminating the small (thresholding) coefficients. This results in a 1-D vector that can be inverse-wavelet-transformed back to the original data set with good accuracy. Using the wavelet transformed data as input to the neural network will yield a prediction accuracy similar to the nontransformed data. For the blood−brain barrier penetration data (see section III), correlation coefficients of 0.88 and 0.86 were achieved with the wavelet transformed (eight components) and nontransformed (32 components) molecular structure data, respectively.

Topology and Geometry. The spatial arrangement of atoms constituting a material is specified completely by its topology and geometry. Topology reflects the pattern of interconnections between atoms and often is expressed in the form of connectivity tables. Geometry encompasses the values of the coordinates of the atoms (as discussed above). Both the topology and geometry of a system provide important, complementary types of information. The mathematical discipline of topology examines the interconnections of components but does not consider the detailed coordinates of compounds. Graph theory, a subdiscipline of topology, is used to study the chemical physics of molecular systems. This work has provided many valuable physical insights but no simple predictive methods or correlations that could be used routinely. Applications of graph theory generate connectivity indices, which are appealing because each index can be calculated exactly from valence-bond diagrams.

From our extensive studies over the past 10 years, we have found that at least four fundamental topological indices are needed to adequately specify the bond connectivity and important electronic structure characteristics.[19] Two atomic indices are needed to compute topological indices. The first atomic index is called the connectivity index and is equal to the number of non-hydrogen atoms to which a given non-hydrogen atom is bonded. The second atomic index is the valence-connectivity index that incorporates fundamental details of the electronic configuration of each non-hydrogen atom. These two atomic indices enable the computation of the zero-, first-, second-, and higher-order (a finite number) connectivity indices.

Aromaticity. Aromatic characteristics of a molecule can also affect its activity. While there is no unique definition of aromaticity, several structure-based methods can be used to estimate it. One measure is the harmonic-oscillator model of aromaticity (HOMA). This model compares the bond lengths in the molecule to an optimal bond length, R_{opt}. The equation for calculating the HOMA index is

$$\mathrm{HOMA} = 1 - \frac{\alpha}{N}\sum_{i=1}^{N}(R_{opt} - R_i)^2 \quad (1)$$

where α is a constant chosen so that HOMA = 0 for the Kekulé structures of the aromatic systems and HOMA = 1 for systems with all bond lengths equal to R_{opt}, N is the number of bonds, and R_i is the individual bond length.[20] The HOMA model can be broken down further into a bond elongation term, EN, and a bond length alternation term, GEO,

$$\mathrm{HOMA} = 1 - \mathrm{EN} - \mathrm{GEO} \quad (2)$$

$$\mathrm{EN} = f\bullet\alpha(R_{opt} - R_{av})^2 \quad (3)$$

$$f = \begin{cases} 1{:}R_{av} > R_{opt} \\ -1{:}R_{av} < R_{opt} \end{cases} \quad (3)$$

$$\mathrm{GEO} = \frac{\alpha}{N}\sum_{i=1}^{N}(R_{av} - R_i)^2 \quad (4)$$

where R_{av} is the average bond length.[20] In this case, an adjustment has to be made for heterocyclic systems to obtain a useful R_{av} term. This adjustment incorporates the Pauling concept of bond order[21] (n), which relates to the bond length:

$$R(n) - R(1) = -c\ln(n) \quad (5)$$

The constant, c, can be estimated using the typical values for single, $R(1)$, and double, $R(2)$, bonds:

$$c = \exp\{[R(1) - R(2)]/\ln(2)\} \quad (6)$$

The bond order for the individual bonds can then be calculated by

$$n = \exp\{[R(1) - R(n)]/c\} \quad (7)$$

Each bond is converted into a "virtual" C−C bond using

$$R(n) = 1.467 - 0.1702\ln(n) \quad (8)$$

These equations can be used to calculate an estimate of the degree of aromaticity for any molecule.

General Approach. The procedure we used to accurately describe molecules for input into neural networks is the following: (1) Obtain the 3-D structure from the database or from our molecular mechanics codes. (2) Transform the $3N$ data vector into a 1-D constant-length vector using a combination of a molecular transform and a wavelet transform. (3) Compute the topology and electronic structure information via topological indices (up to second order). (4) Compute the molecular weight, the number of backbone atoms (hydrogen-suppressed map), and the principal moments of inertia. (5) Compute the radial distribution function of the system (if it is a single molecule, this information will not be needed); this function is useful for liquids or polymeric materials. (6) Compute the density of states spectra (this is achieved by performing a normal-mode analysis). (7) Calculate the HOMA index.

This general approach attempts to adequately capture the full molecular Hamiltonian. Such descriptors will carry the entire information about structural and electronic properties of the molecules, which in turn will determine molecular behavior in a biochemical environment. The full input vector using all seven steps encompasses some 30 variables. We have found, however, that 18 input variables are sufficient for predicting both acute and chronic chemical toxicity. These variables, associated with steps 1−4 above, are the simplest ones to compute and include eight vector components to describe the 3-D structure (wavelet threshold of 3-D Carte-

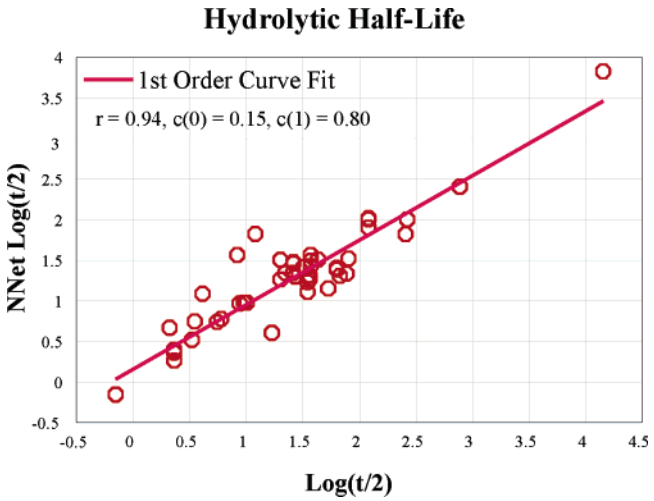

Figure 1. Neural-network-predicted metabolic half-life (in minutes), log(t/2), versus experimental values for chemicals in human blood. The neural network performed well, with a correlation coefficient of $r = 0.94$, slope of $c(1) = 0.80$, and intercept of $c(0) = 0.15$.

Table 1. Ecotoxicologically Relevant Modes of Action[a]

mode of action	number of compounds	site of action
nonpolar nonspecific toxicity	14	membranes
polar narcotics	171	membranes
uncoupling	25	mitochondria ATP-synthesis
inhibition of photosynthesis	15	photosynthesis e-transport
inhibition of AchE	14	neuromuscular end-plate receptor complex
respiratory inhibition	21	respiratory chain e-transport
SH-alkylation	9	(non)-protein SH
reactive toxicity	8	membranes, nucleophiles
estrogenic activity	9	estrogen receptor
precursors to soft electrophiles	27	
soft electrophiles	23	

[a] The neural network was able to predict the mode of action for each chemical with an overall accuracy of 95%.

sian coordinates), one component for the molecular weight, three components for the inertia tensor, and six components for topology and connectivity (including electronic structure). These 18 input variables representing the molecular Hamiltonian have consistently provided very accurate structure−property predictions for the *n*-octanol/water partition coefficient (log P),[22] carcinogenicity, mutagenicity,[22] metabolic half-life in human blood (see Figure 1), mode of action (see Table 1), glucagon receptor affinity, hepatotoxicity, partitioning of chemicals across the blood−brain barrier (see Figure 2), gene expression profiles for hepatotoxins (see Tables 2 and 3), and the NCI-60 human cancer cell lines (see section III). Including the HOMA index improves predictions for blood−brain barrier penetration (see section III).

Estimation of Error. The fundamental approach implemented was based on computational neural networks and wavelets linked to modern computational chemistry methods (primarily molecular mechanics methods) for evaluating and predicting toxicological properties of chemical compounds. Central to the success of the proposed system is the use of computational neural networks, a machine-learning method

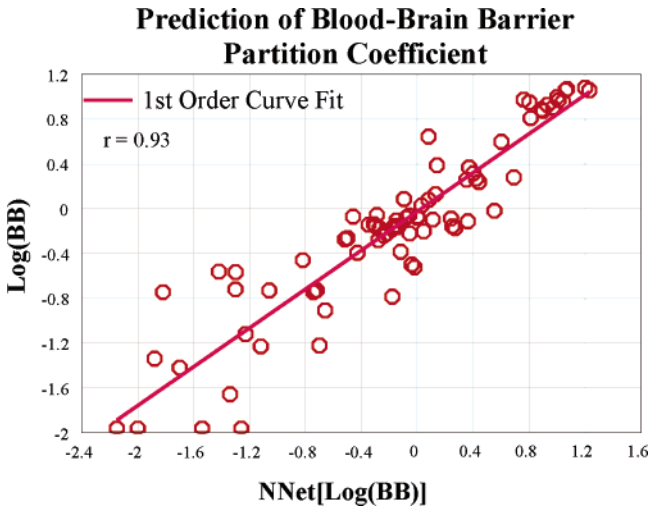

Figure 2. Blood−brain barrier partition coefficients of 106 compounds as predicted by the neural network compared to the experimentally measured values.

Table 2. Example of Some of the Predicted and Measured Gene Expression Levels for Bromobenzene[a]

gene	predicted	measured
aldehyde dehydrogenase 1A1	induced	induced
progesterone receptor membrane component 1	induced	induced
glutamate−cysteine ligase, modifier subunit	induced	induced
UDP-glucose dehydrogenase	induced	induced
aflatoxin B1 aldehyde reductase	induced	induced
glutathione-S-transferase, α type 2	induced	induced
epoxide hydrolase 1	induced	induced
NAD(P)H dehydrogenase, quinone 1	induced	induced
transketolase	induced	induced
glutathion-S-transferase, μ type 2	induced	induced
growth and transformation-dependent protein	induced	induced
v-raf-1 murine leukemia viral oncogene homolog 1	induced	induced
aminopeptidase A	induced	induced
solute carrier family 17 member 1 (Sic17a1)	induced	induced
neurotrophic tyrosine kinase receptor, type 2	induced	induced
EST	induced	induced
integrin-associated protein (Cd47)	repressed	repressed
MHC class II α chain RT1,D α	repressed	repressed
granulin	repressed	repressed
glutamine synthetase	repressed	repressed

[a] There were 66 genes exposed to 28 chemicals for a total of 1848 examples.

capable of on-line learning, prediction, and decision making. This learning-system-based approach can use a variety of techniques for exploiting computational power to systematically search through large numbers of possibilities. The goal is to deal with complex, real-world decision-making problems and predictions and to reach the correct conclusion. Research over the past 15 years has demonstrated that computational neural networks can be used to make highly accurate correlations between structure and property or activity and to efficiently identify related chemicals (i.e., molecular similarity analysis, which is useful for both QSAR and database queries).

An estimation of errors for the predicted values obtained using a neural network can be accomplished by assigning confidence limits to neural network models based on training-data distribution. Correlating the network output with an independent validation set (i.e., a set of unknown data) enables measurement of the degree of accuracy of the model.

Table 3. List of 28 Hepatotoxins and Their Modes of Action: Microarray Expression Measured for 66 Genes

chemical	mode of action
bromobenzene	oxidative stress
disulfiram	oxidative stress
clofibrate	peroxisome proliferation
zymosan A	macrophage activation
acetamidofluorene	oxidative stress
diethyl-hexylphthalate	peroxisome proliferation
methylthiazole	oxidative stress
pulegone	oxidative stress
aniline	oxidative stress
perfluoro-n-heptanoic acid	peroxisome proliferation
lipopolysaccharide	macrophage activation
perfluoro-n-octanoic acid	peroxisome proliferation
trans-anethole	oxidative stress
hexacholorcyclohexane (40 mg/kg)	oxidative stress
diiso-nonylphthlate	peroxisome proliferation
dieldrin (65 mg/kg)	oxidative stress
ethinyl estradiol	oxidative stress
benzafibrate	peroxisome proliferation
nimesulide	oxidative stress
benzbromarone	peroxisome proliferation
perfluoro-decanoic acid	peroxisome proliferation
tannic acid	oxidative stress
piperonyl butoxide	oxidative stress
hexacholorcyclohexane (65 mg/kg)	oxidative stress
dieldrin (40 mg/kg)	oxidative stress
WY14643	peroxisome proliferation
butylated hydroxytoluene	oxidative stress
precocene	oxidative stress

Standard statistical techniques (t statistic with percent confidence, variance, correlation coefficient, and mean values) can then be applied to determine whether the fit is significant. Noise in the data will generate a correlation of less than 1. Assuming that the correlation is significant, plotting the network output against the actual results should reveal a cluster of points around a straight line. This works by assuming that the regression line through the validation data pivots about the average of the network output over the entire test set so that the farther away you are from that point, the more scope for movement there is. This scope is also limited by the variance of the network output; the farther the values go toward the extremes of the distribution of network outputs and expected values, the less confident we can be. These calculations work for cases when the network output is unimodal (i.e., only one cluster of network outputs). More complicated calculations can be performed to estimate confidence limits for multimodal output.

Validation checks for the neural network can also be obtained by comparing attributes or outcomes such as physical properties with values obtained from the literature or that can be calculated from computational chemistry (e.g., fundamental thermodynamic properties). If the neural network module obtains a reasonable accuracy for these predictions, then it gives a greater confidence for its prediction of the toxicological endpoint. Methods of molecular mechanics that provide an interface to generating 3-D structures from atom and connectivity tables can be used in this manner.

III. RESULTS AND DISCUSSION

A toxic effect can be defined as any adverse effect of a chemical on a target organism or cell. A large battery of studies is needed to assess potential toxicity, including tests of absorption, distribution, metabolism, and excretion. There are many experimental variables to consider including the nature of the adverse health effects, animals used for the study, dose, and route of exposure. The picture is equally as complex biochemically, because adverse effects are mediated by different mechanisms and metabolic pathways. A toxic substance may directly affect the target site, undergo transformation into an active metabolite, or trigger the activation of some other biological receptor. Biochemical and molecular toxicology deal with events that occur at the molecular level when toxic compounds interact with processes occurring in living organisms. Defining these interactions is fundamental to our understanding of toxicity, both acute (i.e., LD_{50}) and chronic (e.g., carcinomas, cataracts, peptic ulcers, and reproductive effects). This knowledge is essential for identifying toxic hazards and for developing new therapies. It is important, therefore, to define a QSAR module at multiple levels of toxicity on the basis of responses at the cellular, metabolic, target-organ, and systemic levels. Initial responses to chemical toxicity occur at the receptor and cellular levels, and methods that allow an accurate prediction of this response are crucial in the development of an integrated toxicity evaluation and predictive system.

The rapid development of ecotoxicology in the 1970s was spurred by the realization that a certain amount of the toxicity caused by a substance could be predicted by observing its behavior when placed in *n*-octanol [i.e., the coefficient of partition between *n*-octanol and water (K_{ow} or log P)]. The rationale for this relationship stems from the fact that a compound must enter the cell in order to show biological activity, and compounds that tend to remain suspended in *n*-octanol have a greater affinity for the cell membrane and thus enter the cell more easily. An important example is the ability of some chemicals to pass the blood−brain barrier and exhibit neurotoxicity, causing particularly profound forms of chronic and acute problems (e.g., paralysis, blurred vision, and disorientation). The ability of a chemical to reach the brain depends on its solubility in lipids (a component of membranes) as well as on its molecular weight (chemicals with a molecular weight less than 500 Da can penetrate the blood−brain barrier more easily than those with a greater molecular weight). Studies have demonstrated the capability of our neural-network-based approach to accurately predict log P, as well as other indicators of acute inhalation toxicity, carcinogenicity, mutagenicity, and single-dose lethality.

Prediction of Metabolic Rates. The prediction of metabolic processes such as the enzymatic hydrolysis of noncongener carboxylic esters has often challenged most standard QSAR methods. Carboxylic ester hydrolases efficiently catalyze the hydrolysis of a variety of ester-containing chemicals to their respective free acids. These enzymes exhibit broad and overlapping substrate specificity toward esters and amides, and the same substrate is often hydrolyzed by more than one enzyme. Consequently, their classification is difficult and somewhat confusing. Studies have shown that humans express carboxylesterase in the liver, plasma, small intestine, brain, stomach, colon, macrophages, and monocytes. In vitro hydrolytic half-lives measured in rat blood have been reported to be orders of magnitude lower than those measured in human blood for esmolol or remifentanil, but the opposite was found for flestolol.[23] Thus, the extrapolation of animal results to humans is not always a

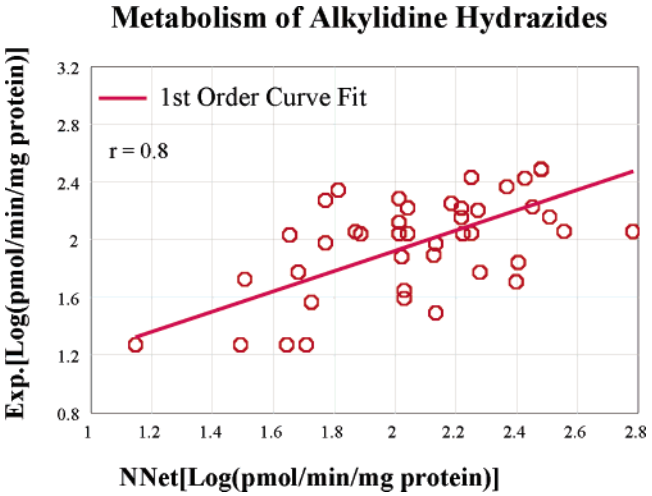

Figure 3. Neural-network-predicted metabolic rates measured in rat microsomes vs experiment.

good approach, and accurate structure−metabolism relationships are needed to predict the rate of enzymatic hydrolysis.

A total of 80 compounds belonging to seven different chemical classes were used in our study. These include two short-acting beta-blocker series, short-acting angiotensin-converting enzyme inhibitors, opioid analgesics, soft corticosteroids, antiarrhythmic agents, and buprenorphine prodrugs.[24−26] Input to a neural-network module consisted of the 18 variables described in section II, and the output was the hydrolytic half-life (log $t_{1/2}$). The neural-network module was able to accurately predict the in vitro hydrolysis rates of these chemicals in human blood ($r = 0.94$) as shown in Figure 1.

This structure−metabolism module was also tested on novel alkylidene hydrazides. In this case, the metabolic pathways were not necessarily known, but previous studies have shown a fast metabolic turnover that resulted in poor in vivo pharmacokinetic (PK) profiles; results, therefore, were based on an in vitro analysis of rat liver microsome incubations. The metabolism data was taken from Madsen et al.[27] and ranged from 14 to 606 pmol/min/mg of protein. Overall, the neural-network module was capable of predicting the metabolism rate in picamoles per minute milligram of protein to a reasonable accuracy ($r = 0.8$), as shown in Figure 3, but not as well as that discussed for enzymatic hydrolysis in Figure 1. We still view these results as a success since the hydrazide data are on newly developed compounds for which metabolic turnover was very fast. The PK profiles had to be estimated from a liquid chromatography mass spectrometry analysis of incubations with rat liver microsomes. The combined results (Figures 1 and 3) demonstrate the capability of the neural network module for making relevant predictions to metabolic processes of chemical compounds.

These preliminary results are encouraging for the development of more extensive structure−metabolism evaluation and prediction tools, including those for the cytochrome P450 system, which is important in both prokaryotic and eukaryotic cells. This system appears to interact with almost every kind of chemical bond, most often via oxidative mechanisms, but also by reduction, as is the case for prostaglandin synthetase cooxidation, flavin-containing monooxygenase, and alcohol and aldehyde dehydrogenase.

Predicting Mode of Action. Assessing the likely mode of action for a toxic compound is critical for correctly predicting toxicity. Compounds having different modes of action are toxic in different ways due to different interactions at the biomolecular level, and their eco- and biotoxic effects in any given test system generally must be predicted with different QSARs. Modes of toxic action were predicted for 336 test compounds (taken from refs 28−31) for up to 11 modes of action with 95% accuracy (see Table 1). Results for predicting the mode of action for chemical compounds on the basis of structural information (95% classification accuracy) exceed those previously reported in the literature, which generally range from 85 to 89% accuracy. This approach can be used to develop mode-of-action classification capabilities with similar accuracies for a larger set of possible modes of action. We demonstrate how this is possible using microarray data (discussed later in this section). In short, it appears that, from the fundamental structural input of chemical compounds representing their molecular Hamiltonian, sufficient information is available to formulate accurate correlations between structure and mode of action.

Predictive Hepatotoxicity. An intriguing question in toxicology is why some substances manifest significant toxicity only in certain tissues while nontarget tissues remain relatively unaffected. For example, the chemical paracetamol exhibits toxicity in the liver by necrosis, acrylamide causes toxicity in the nervous system by axonopathy, bleomycin causes damage to the respiratory system by pulmonary fibrosis, and chloroquine causes damage to the eye by retinopathy. Clearly, these are complex phenomena involving several interacting processes that include the pharmacokinetics and distribution of the toxicant, the presence of specific uptake mechanisms in susceptible tissues, the specific biochemistry of the target tissue, including the presence of the activation or deactivation enzymes, and the ability of the tissues in question to repair a particular damage or lesion elicited by the toxicant.

The liver is often the main target for chemically induced toxicities, and several factors contribute to its particular susceptibility. The liver is the organ with the highest complement of P450 in terms of quantity as well as numbers of isozymes and is the organ in which P450 enzymes are most readily induced. It is also the site of metabolism for xenobiotics absorbed from the gastrointestinal tract, the major route of absorption for most xenobiotics. Additionally, the liver may activate chemicals that can then be transported to distant tissues to affect toxicity in those organs. The liver maintains normal sugar concentration in the blood by storing glycogen and releasing glucose and synthesizes many proteins and other vital components of blood plasma. Damage to the liver or interference with its vital functions can thus be extremely harmful or even lethal.

Liver damage can be classified by the types of lesion, such as (1) fatty liver (lipid content greater than 5% by weight), (2) necrosis (chemical damage leading to cell death that can affect different areas of the organ), (3) cholestasis (obstruction to bile flow), (4) hepatitis (inflammation of liver tissue caused by diseases), (5) fibrosis (large amounts of collagen and other extracellular matrix proteins), (6) cirrhosis (chronic liver damage characterized by deposition of massive amounts of collagen), and (7) carcinogenesis (cancer of the liver).[32]

Our approach was used to examine two data sets, for chemicals known to cause acute hepatic injury, typically hepatitis, or secondary toxic effects, such as elevated liver enzyme levels. These chemicals are listed in Table 1 of the Supporting Information. Neural-network classification of these chemicals into acute or secondary liver toxicity was 100% accurate. The inclusion of chemicals that were not hepatotoxins did not reduce the accuracy.

Additional studies were carried out to investigate a broader range of general predictive capabilities. In this case, results were predicted for the lesion types: fat, necrosis, cirrhosis, carcinoma, and/or cholestasis, known to be caused by ~100 various organic hepatotoxicants that have been experimentally studied.[33] Classification to the type of lesion was 97% accurate for necrosis, 96% accurate for fatty liver, 98% accurate for cirrhosis, 94% accurate for carcinoma, and 98% accurate for cholestasis.

Predicting Neurotoxicity. The nervous system is extremely complex, and toxins can act at many different points. It has three basic functions: to detect and relay sensory information inside and outside the body, to direct motor functions of the body, and to integrate the thought processes of learning and memory. The nervous system consists of two fundamental anatomical divisions, the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS includes the brain and the spinal cord and serves as the control center. It processes and analyzes information received from sensory receptors and, in response, issues motor commands to control body functions. The PNS consists of all nervous tissue outside the CNS and contains two forms of nerves: afferent nerves, which relay sensory information to the CNS, and efferent nerves, which relay motor commands from the CNS to various muscles and glands.

The human brain is one of the most complicated systems known. It contains an estimated 100 billion neurons, each forming as many as 100 000 synapses leading to a system with up to 10^{16} connections having an astronomically large number of possible different connectivities—estimated to be larger than the number of atoms in the universe. The brain also exhibits enormous diversity with perhaps as many as 1000 different cell types, each with a highly intricate and specific communication pathway. To understand such a special organ and the myriad events that occur within its vast network of cellular interconnections requires the application of a very broad range of scientific disciplines. The union of disciplines that emerged in the past three decades to understand higher brain functions such as perception, learning, and memory is now known as the neurosciences. The 1990s, called the Decade of the Brain, witnessed an explosion of experimental discovery in this nascent field.

The nervous system is quite vulnerable to toxins since xenobiotic chemicals interacting with neurons can alter critical voltages and concentrations of ions required for signal transduction and transmission continuance. Most of the CNS, however, is protected by an anatomic barrier between the neurons and blood vessels called the blood−brain barrier (BBB). The BBB is composed of unique tight junctions of the brain capillary endothelial cells resulting from tissue-specific gene expression. These tight junctions, which exhibit electrical resistance as high as 8000 $\Omega \cdot cm^2$, eliminate a paracellular pathway of solute movement through the BBB. The virtual absence of pinocytosis across brain capillary endothelium eliminates transcellular bulk flow of a circulating solute through the BBB.

The BBB is actually composed of two adjacent membranes in brain capillary endothelial cells. The lumenal and the ablumenal membranes are separated by approximately 300 nm of endothelial cytoplasm enriched with mitochondria (some 4 times more than other capillary cells), resulting in the increased metabolic capacity that accounts for the rapid degradative processes of endogenous and exogenous agents. Additionally, there are numerous active efflux systems and enzymes for the transport and inactivation of agents. Passage through the BBB therefore requires penetration of two membrane barriers plus diffusion across 300 nm of cytoplasm full of enzymes and active efflux systems. In addition to these barriers, the BBB has a basement membrane that appears to be expressed from another type of BBB cell called the pericyte, which also expresses a number of ectoenzymes. Another type of BBB cell structure, the astrocyte foot processes, compose 99% of the BBB on the ablumenal side (separated by 20 nm that is filled by the basement membrane) and express p-glycoproteins. These p-glycoproteins cause an ATP-dependent active efflux of drugs from the cellular compartment to the extracellular space. Therefore, pericyte ectoenzymes, astrocyte foot process p-glycoproteins, and endothelial active efflux systems at the endothelial ablumenal membrane, together with the three different membrane barriers (i.e., lumenal, ablumenal, and basement), work together to prevent brain uptake of xenobiotics. Solutes can gain access to brain interstitium via one of two pathways: lipid-mediated (generally limited to small molecular weight and lipid-soluble compounds, although local hydrophobicity, flexibility, and number of accepted or donated hydrogen bonds also play a role) or catalyzed transport (carrier-mediated or receptor-mediated processes). This elaborate series of obstacles protects the brain from a barrage of bacterial, viral, and other chemical assaults that would lead to severe or even lethal consequences. The same barrier that protects us, however, also inhibits the development of therapeutics that can treat CNS disorders. Virtually all small-molecule drugs generated by receptor-based, high-throughput drug-screening programs cannot cross the BBB.

Toxic damage to the nervous system can occur at peripheral sensory receptors and sensory neurons, which affect blood and intraocular pressure, temperature, vision, hearing, taste, smell, touch, and pain. Examples of compounds that can cause damage in these areas are heavy metals (in particular, lead and mercury), several inorganic salts, and organophosphorus compounds. Another site where potential damage can occur is at motor neurons; damage to these by compounds such as isonicotinic hydrazide can cause muscular weakness and paralysis. Low levels of inorganic mercury and carbon monoxide can cause interneuronal damage and lead to significant learning deficiencies, loss of memory, loss of coordination, and emotional disorders. In general, toxic damage to the nervous system occurs by one of the following basic mechanisms: direct damage and death of neurons and glial cells, interference with neural-electrical transmission, or interference with neural-chemical transmission.

In an effort to design a predictive model for neurotoxicity, we have examined the capabilities of computational neural networks for determining structure-transport properties as it

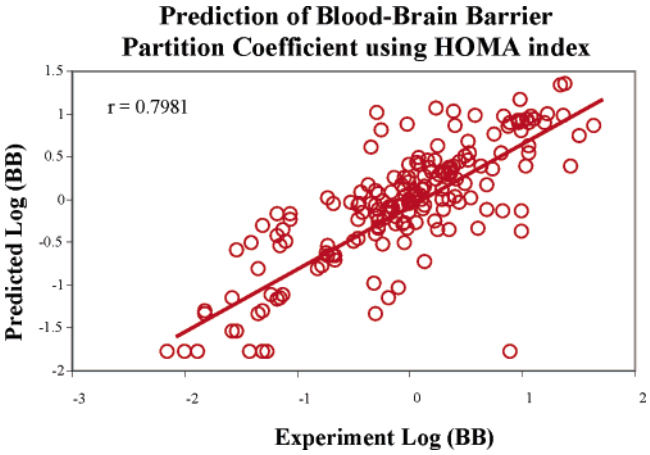

Figure 4. Neural-network-predicted log(BB) vs experimental values with input to include the HOMA index.

pertains to BBB partitioning of chemical compounds. Since poor penetration of a compound means low bioavailability, much effort has been devoted to developing accurate methods for predicting BBB partitioning of chemicals. A data set consisting of 106 compounds that have been studied in laboratory tests[34,35] was used to test our neural network module. The compounds ranged from small organics to large drugs such as indinavir and verapamil, having in vivo measured BBB partition coefficients (log BB) in rats (ratio of the concentration of the compound in the brain to that in the blood). The BBB partitioning data span a range of more than 3 orders of magnitude. While partitioning of chemicals across the BBB is thought to be strongly dependent on local hydrophobicity, molecular size, lipophilicity, and molecular flexibility, nearly all previous attempts to delineate an accurate relationship have had little success. Correlations with $r = 0.90$ appear to be among the best results reported in the literature.[36] Our results, shown in Figure 2, demonstrate the generality and accuracy of our approach to toxicological predictions; a linear correlation coefficient for the data set of $r = 0.93$ was obtained. We expect this accuracy to apply over a broad range of compounds that span the trained neural network domain: log(BB) from −2.15 to 1.44.

The HOMA index was added into the above QSAR model in an attempt to improve the overall correlation. Various combinations of the structural components discussed in section II were generated from the QSAR model and input into the neural network. The neural network was used to predict the blood−brain partition coefficient for a larger and more diverse data set which consisted of 193 molecules. The data set was compiled from three different studies[37−39] and included molecules ranging from small alcohols and ethers to larger molecules such as ranitidine and nevirapine. The results of using different input variables are shown in Figures 4 and 5. Including the HOMA index in the input variables increases the linear correlation coefficient from $r = 0.7856$ to $r = 0.7981$. The best correlation, $r = 0.8590$, was found when the molecular weight, wavelet transform, and HOMA index were used as the input variables, shown in Figure 5. While this combination works well for this particular data set, it may not be sufficient for predicting toxicity or for different sets of molecules. A sensitivity analysis performed independently of the data set will allow rapid and optimal selection of the input variables.

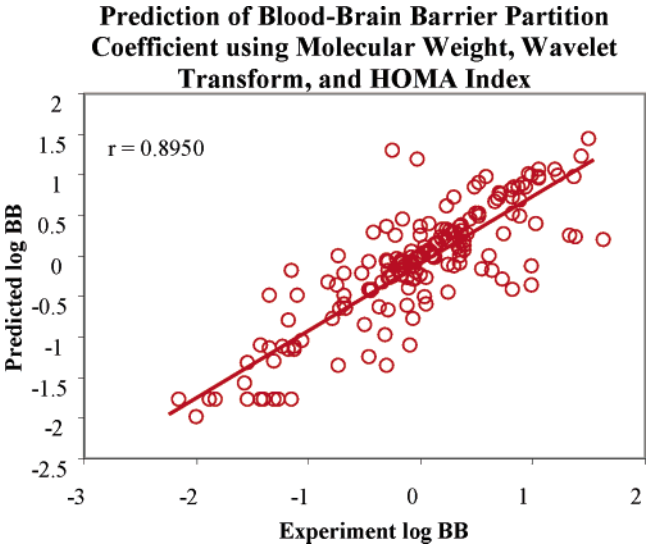

Figure 5. Neural network, with molecular weight, wavelet transform, and HOMA index as input variables, prediction of the BBB partition coefficient for 193 chemicals.

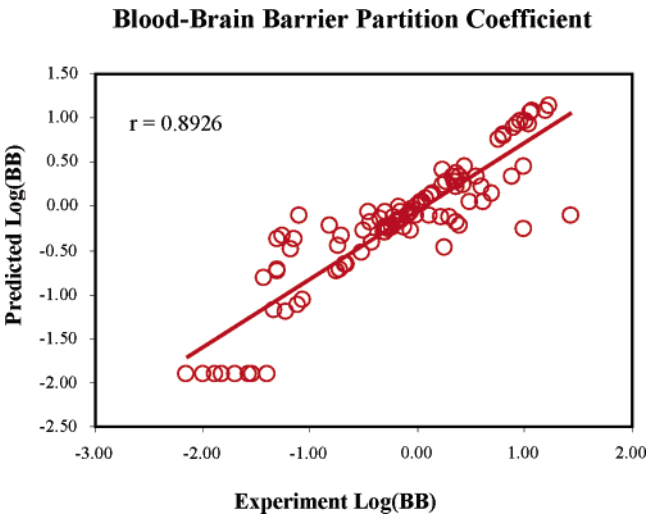

Figure 6. Neural-network-predicted log BB vs experimental values[34] with input to include the HOMA index.

Another data set of BBB partition coefficients[34] was also used to test the predictability of the neural network. The results, using all 19 input variables, are shown in Figure 6. Omitting the HOMA index decreases the correlation coefficient from $r = 0.8926$ to $r = 0.8771$. Using the three components that gave the best correlation in the previous data set did not improve the results for this set.

Gene Expression Microarray Analysis. Our approach can be applied directly to microarray data by using a combination of wavelets for spatio-temporal multiresolution analysis, noise and trend reduction, and compression with neural networks to yield reasonably accurate predictions of gene expression activities. For example, Table 2 shows a sample (only those of bromobenzene are shown) of the 1848 total examples of the predicted induced, repressed, or unchanged levels of expression for 66 genes resulting from exposure to the 28 different hepatotoxins shown in Table 3.[40]

The neural-network system correctly predicted gene expression levels for all but two cases (clofibrate and hexachlorocyclohexane at a 65 mg/kg dose), giving 93%

accuracy. These particular results were obtained from direct prediction based on chemical structure instead of DNA-microarray data. The inverse problem, using the microarray data as input into a neural network and predicting the mode of toxic action, gave 100% accuracy. Thus, predictions of the effects of chemical compounds can be made with considerable confidence by using two simultaneous predictive modules, one based on structure−toxicity modules and the other based on the analysis of microarray data.

We have also examined data corresponding to the NCI-60 cancer cell lines with drug treatments being used. Gene expression levels were expressed as log(red/green), and ratios of fluorescence measurement were corrected by computational balancing of the two channels. The data set includes expression values for 1376 genes plus 40 assessed molecular targets for the drugs (1416 clones), in the 60 different cell lines corresponding to nine different types of cancer. The gene expression data is used as input to a neural-network system trained to predict the mode of action for the set of drugs.[41] The results for the drugs that were originally classified (see Table 2 of the Supporting Information) were not as good as those reported above for hepatotoxicty but are reasonable for proof of principle (80% correct prediction for mode of action).

IV. SUMMARY AND CONCLUSIONS

There are well over 14 million known chemical compounds, and thousands of new ones are being developed each year. Accurately assessing risk from known and novel chemicals against the spectrum of endpoints that are needed is a daunting challenge. Experimental approaches using animal systems to collect and integrate the necessary chemical, physicochemical, and toxicological data and interpretative expertise are prohibitively costly and time-consuming.

Given the number of compounds needing health risk assessments, laboratory-based analysis alone will not suffice; alternative methods need to be synergistically coupled with conventional testing to meet the demand. A robust computational synthesis and toxicity assessment and evaluation can provide the capacity needed to predict potential health risks resulting from exposure to new chemicals, materials, and mixtures; biological or chemical degradative processes (i.e., molecular aging and chemical reactivity processes); or metabolic byproducts.

The computational approach described here shows promise for efficiently making reasonably accurate predictions on the toxicological properties of chemical compounds. Central to the success of this system is the use of the machine-learning method, computational neural networks, coupled with spatio-temporal analysis capabilities of wavelets. In this paper, we have demonstrated reasonable capabilities for using this approach to predict metabolic processes, mode of action, and hepato- and neurotoxicity, and for automatic processing of microarray data for predicting modes of action.

The authors wish to thank Dr. Katherine Yuracko, CEO of YAHSGS LLC, and Bill Hewitt, President of YAHSGS LLC, for their support and encouragement in the preparation of this paper. Special thanks are also given to Dr. Jerry Heindel, our project's Science Manager at the National Institute of Environmental Health Sciences for the opportunity to conduct this work and his enthusiasm, support, and encouragement. We thank Denise Casey for her technical editing skills and critical reading of the manuscript through its many drafts. Last but not least, we thank Karen Martin, Jennifer Goldberg, and Angela Edmonds for their excellent work in creating the database used as the foundation for this work project. Work supported by National Institute of Health SBIR/STTR Grant No. 5R42ES013321-04 with YAHSGS LLC, Richland, WA 99354. ORNL is managed by UT-Battelle for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Supporting Information Available: A more detailed description about computational neural networks and resampling techniques is provided. Also included is an example of a connection table, a table of chemicals associated with liver injury, and a table showing the mechanism of action of the drugs used in the gene expression study. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) National Research Council, Commission on Life Sciences. *Toxicology Testing: Strategies to Determine Needs and Priorities*; National Academy Press: Washington, DC, 1984.
(2) *Science in Focus. Evaluating Chemical Safety*; The American Chemical Society: Washington, DC, 1999. http://www.acs.org/portal/resources/ACS/ACSContent/government/publications/EvalChem.pdf (accessed Dec 13, 2006).
(3) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232.
(4) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(5) Sumpter, B. G.; Noid, D. W. On the Design, Analysis, and Characterization of Materials Using Computational Neural Networks. *Annu. Rev. Mater. Sci.* **1996**, *26*, 223−277.
(6) Sumpter, G. B.; Getino, C.; Noid, D. W. Theory and Applications of Neural Computing in Chemical Science. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439−481.
(7) Vollmer, J. J. Wiswesser Line Notation: An Introduction. *J. Chem. Educ.* **1983**, *60*, 192−196.
(8) Gasteiger, J. A Representation of π Systems for Efficient Computer Manipulation. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111−115.
(9) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Smith, J. M.; Watson, D. G. The Development of Version 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187.
(10) *CONCORD*; TRIPOS Associates: St. Louis, MO. http://www.tripos.com (accessed Dec 13, 2006).
(11) *ALCOGEN*; Chemical Concepts: Weinheim, Germany.
(12) CHEM-X; *Chemical Design Ltd.*: Oxford, England.
(13) Gordeeva, E. V.; Katritzky, A. R.; Shcherbukhin, V. V.; Zefirov, N. S. Rapid Conversion of Molecular Graphs to Three-Dimensional Representation Using the MOLGEO Program. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 102−111.
(14) *COBRA*; Oxford Molecular Ltd.: Oxford, England.
(15) *CORINA*; TORVS Research Team: Erlangen, Germany. http://www2.chemie.uni-erlangen.de/software/ (accessed Dec 13, 2006).
(16) *CONVERTER*; Biosym Technologies, Inc.: San Diego, CA.
(17) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure−Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.
(18) Newland, D. E. *Random Vibrations, Spectral & Wavelet Analysis*, 3rd ed.; Longman Scientific and Technical: Essex, England, 1993.
(19) Sumpter, B. G.; Noid, D. W. Neural Networks and Graph Theory as Computational Tools for Predicting Polymer Properties. *Macromol. Theory Simul.* **1994**, *3*, 363−378.
(20) Krygowski, T. M.; Cyranski, M. K. Structural Aspects of Aromaticity. *Chem. Rev.* **2001**, *101*, 1385−1419.
(21) Pauling, L. Atomic Radii and Interatomic Distances in Metals. *J. Am. Chem. Soc.* **1947**, *69*, 542.
(22) Piotrowski, P. L.; Sumpter, B. G.; Wassom, J.; Lu, P.-Y.; Brothers, R. Neural Networks and Wavelet Tools for Computational Molecular Toxicology. To be submitted.

(23) Bodor, N.; Buchwald, P. Soft Drug Design: General Principles and Recent Applications. *Med. Res. Rev.* **2000**, *20*, 58−101.
(24) Buchwald, P.; Bodor, N. Quantitative Structure−Metabolism Relationships: Steric and Nonsteric Effects in the Enzymatic Hydrolysis of Noncongener Carboxylic Esters. *J. Med. Chem.* **1999**, *42*, 5160−5168.
(25) Vashishtha, S. C.; Hawes, E. M.; McCann, D. J.; Ghosheh, O.; Hogg, L. Quaternary Ammonium-Linked Glucuronidation of 1-Substituted Imidazoles by Liver Microsomes: Interspecies Differences and Structure−Metabolism Relationships. *Drug Metab. Dispos.* **2002**, *30*, 1070−1076.
(26) Cupid, B. C.; Holmes, E.; Wilson, I. D.; Lindon, J. C.; Nicholson, J. K. Quantitative Structure−Metabolism Relationships (QSMR) Using Computational Chemistry: Pattern Recognition Analysis and Statistical Prediction of Phase II Conjugation Reactions of Substituted Benzoic Acids in the Rat. *Xenobiotica* **1999**, *29*, 27−42.
(27) Madsen, P.; Ling, A.; Plewe, M.; Sams, C. K.; Knudsen, L. B.; Sidelmann, U. G.; Ynddal, L.; Brand, C. L.; Andersen, B.; Murphy, D.; Teng, M.; Truesdale, L.; Kiel, D.; May, J.; Kuki, A.; Shi, S.; Johnson, M. D.; Teston, K. A.; Feng, J.; Lakis, J.; Anderes, K.; Gregor, V.; Lau, J. Optimization of Alkylidene Hydrazide Based Human Glucagon Receptor Antagonists. Discovery of the Highly Potent and Orally Available 3-Cyano-4-hydroxybenzoic Acid [1-(2,3,5,6-Tetramethylbenzyl)-1H-indol-4-ylmethylene]hydrazide. *J. Med. Chem.* **2002**, *45*, 5755−5775.
(28) Aptula, A. O.; Netzeva, T. I.; Valkova, I. V.; Cronin, M. T. D.; Schultz, T. W.; Kuhne, R.; Schuurmann, G. Multivariate Discrimination between Modes of Toxic Action of Phenols. *Quant. Struct.-Act. Relat.* **2002**, *21*, 12−22.
(29) Nendza, M.; Muller, M. Discriminating Toxicant Classes by Mode of Action: 2. Physico-Chemical Descriptors. *Quan. Struct.-Act. Relat.* **2000**, *19*, 581−597.
(30) Ren, S. Use of Molecular Descriptors in Separating Phenols by Three Mechanisms of Toxic Action *Quan. Struct.-Act. Relat*. **2002**, *21*, 486−492.
(31) Ren, S. Classifying Class I and Class II Compounds by Hydrophobicity and Hydrogen Bonding Descriptors. *Environ. Toxicol.* **2002**, *17*, 415−423.
(32) In *The Merck Manual of Diagnosis and Therapy*, 18th ed.; Beers, M. H., Ed.; Merck Research Laboratories: Whitehouse Station, NJ, 2006; Section 4, Chapter 43.
(33) Zimmerman, H. J. *Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver*; Appleton-Century-Crofts: New York, 1978; Tables 8.6 and 8.7.
(34) Rose, K.; Hall, L. H. Modeling Blood−Brain Barrier Partitioning Using the Electrotopological State. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651−666 and references therein.
(35) Engkvist, O.; Wrede, P.; Rester, U. Prediction of CNS Activity of Compound Libraries Using Substructure Analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 155−160.
(36) Garg, P.; Verma, J. In Silico Prediction of Blood Brain Barrier Permeability: An Artificial Neural Network Model. *J. Chem. Inf. Model*. **2006**, *46*, 289−297.
(37) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. J. Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D-Molecular Similarity Measures and Cluster Analysis. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2083−2098.
(38) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 3. Modeling Blood−Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137−2152.
(39) Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. Prediction of Blood−Brain Partitioning Using Monte Carlo Simulations of Molecules in Water. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 697−708.
(40) Data was obtained from a poster by Parker, J. B.; Leone, A. M.; Nie, A.; McMillian, M.; Bryant, S.; Herlich, J.; Kemmerer, M.; Yieh, L.; Bittner, A.; Liu, X.; Wan, J.; Johnson, M. D.
(41) Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold, W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O.; Weinstein, J. N. A Gene Expression Database for the Molecular Pharmacology of Cancer. *Nat. Genet.* **2000**, *24*, 236−244.

CI6004788