Progress Report
of the AFOSR
AASERT Project

Covering research period 8/1/98 to 7/31/99

Agency No: DOD/F49620-96-1-0330
U of M No: 9606754
NRRI Technical Report No: NRRI/TR-99/14

**QUANTITATIVE CHARACTERIZATION OF
MOLECULAR SIMILARITY SPACES:
TOOLS FOR COMPUTATIONAL TOXICOLOGY**

Submitted by:

Subhash C. Basak, Ph.D.
Principal Investigator
Natural Resources Research Institute
University of Minnesota, Duluth
5013 Miller Trunk Highway
Duluth, MN 55811
Tel: (218)720-4230
Fax: (218)720-4328
Email: sbasak@wyle.nrri.umn.edu

ARY

TABLE OF CONTENTS

OBJECTIVES

STATUS OF EFFORT

ACCOMPLISHMENTS/NEW FINDINGS

TASK 1: Characterization of molecular similarity spaces

TASK 2: Selection of Analogs

TASK 3: Similarity-based estimation of properties

PERSONNEL SUPPORTED

PUBLICATIONS

APPENDIX 1

..... 3

..... 3

..... 3

..... 3

..... 4

..... 5

..... 6

..... 6

..... 8

OBJECTIVES

(Same as the original proposal)

STATUS OF EFFORT

During the third year of the project, our work on the first three tasks of the project; viz., a) characterization of molecular similarity spaces, b) selection of analogs, and c) similarity-based estimation of properties; has continued. However, the focus of our work has shifted to the fourth and final task of the project, viz., the application of neural networks in property estimation.

In the area of Task 1, the effectiveness of theoretical molecular descriptors *vis-a-vis* experimental physicochemical properties in quantifying intermolecular similarity has been explored for several sets of compounds with varying physicochemical and biological properties. In Task 2, the various structure spaces developed in Task 1 have been used in the selection of analogs for specific probe compounds. In Task 3, we have used the *k*-nearest neighbor (KNN) method to estimate properties of chemicals from various databases. For these experiments, *k* has been varied from 1-40. The results showed that, for different physicochemical, toxicological and biochemical properties, optimal property estimation is generally obtained in the range of $k = 5-10$. Finally, in Task 4, we have used neural networks for the prediction of toxicological endpoints. In addition, we examined several methods for feature (independent variable) selection using a machine learning techniques, GEFS (genetic ensemble feature selection), based on genetic algorithms. The results show that neural networks, in general, show slight improvement in modeling power over statistical methods, but the use of GEFS to select relevant features for modeling greatly improves the performance of the neural networks.

ACCOMPLISHMENTS/NEW FINDINGS

Described below are the accomplishments of the four project tasks that have been pursued during this reporting period.

TASK 1: Characterization of molecular similarity spaces

Molecular similarity spaces were constructed using computed molecular descriptors. These descriptors included atom pairs, topological indices, geometrical indices, semi-empirical quantum chemical parameters, and, when available, physicochemical property data and *ab initio* quantum chemical parameters. Atom pairs and topological indices were calculated using in-house software packages, *APProbe* and *POLLY 2.3* respectively; geometrical parameters were calculated by *Sybyl 6.4* using an SPL (Sybyl Programming Language) program developed in-house; and the quantum chemical indices were calculated by *MOPAC 6.00*. Additional physicochemical property data were taken from the literature and *ab initio* calculations were conducted using *Gaussian 98W*.

As part of this task we have also begun the development of an expanded set of molecular descriptors. For adequate characterization of molecular similarity spaces, we must be sure that we have parameters that adequately represent the pertinent molecular features. Currently, we have added nearly one-hundred additional novel indices to our predictor set (Pub. #1) and we plan to continue to expand this set in the future.

Two statistical methods were used to derive non-redundant information from the calculated parameters, viz., principal components analysis (PCA) and variable clustering (VC). The results of these studies were reported in two peer-reviewed manuscripts (Pub. # 2 & 3, See Publications below) and in chapters in two books: a volume of the *Discrete Mathematics and Theoretical Computer Science* series (Pub. # 4) and volume 2 of the *Advances in Molecular Similarity* series (Pub. #5).

These similarity spaces, constructed from theoretical descriptors and physicochemical property data, are distinct in the sense that they select different sets of analogs for a given probe chemical. The similarity spaces constructed in our studies were used in the selection of analogs and estimation of toxicologically-relevant properties for diverse sets of chemicals (See Task 2 and 3 below).

The optimal characterization of molecular structure is prerequisite to the creation of useful similarity spaces and the prediction of the toxicity of chemicals for which very little experimental data is available. A novel, hierarchical approach was used in selecting orthogonal structural information from calculated topostructural, topochemical, geometrical, and semi-empirical quantum chemical descriptors. The resultant orthogonal structural information was used to develop hierarchical quantitative structure-activity relationship (QSAR) models for predicting properties such as inhibition of the complement system by benzamidines and the dermal penetration of polycyclic aromatic hydrocarbons. Results of this research have been reported in two recent publications (Pub. # 6-8). This hierarchical approach has also been employed in the development of similarity spaces (Pub. #2 & 5).

TASK 2: Selection of analogs

The similarity spaces created using the atom pair (AP) and PCA methods were used in the selection of analogs for probe compounds. In one study (Pub. #2), five distinct similarity spaces were created from: a) calculated topostructural indices (TSI) only, b) calculated topochemical indices (TCI) only, c) a combination of both TSI and TCI, d) calculated atom pairs, and e) physicochemical property data taken from the literature. In another study (Pub. #5), three distinct similarity spaces were created from: a) calculated TSI only, b) calculated TCI only, and c) a combination of both TSI and TCI.

One study (Pub #2) attempted to quantify the degree of overlap between similarity measures. To this end, the analogs selected for a set of 76 compounds were compared and the methods were scored in a pair-wise fashion to demonstrate the degree of overlap. This study resulted in the discovery that, for this particular set of compounds, even though the degree of overlap between the groups of analogs selected by theoretical descriptor spaces is relatively high, the similarity space constructed from physicochemical property data provided relatively unique groups of analogs.

This demonstrates that if one is attempting to determine the optimal characterization for a similarity space it is best to employ two or three distinct methods, e.g., one theoretical space and one property space, rather than two theoretical spaces that may have a high degree of overlap.

Further investigation is needed to determine which of these similarity techniques is most capable of estimating toxicological properties of chemicals from the toxicity data of their selected neighbors. It should also be noted that while similarity spaces derived from physicochemical property data seem to be unique as compared to theoretically-derived similarity spaces, relevant physicochemical data is not always readily available for all the compounds in a given set. In technology transfer, this finding will have important implications. Many drug companies are using molecular similarity methods in their drug discovery process. This research, one aim of which is to derive molecular similarity methods which are non-redundant, and further pursuit of this issue could be beneficial to these companies and others involved in the design, synthesis, and testing of new chemicals.

TASK 3: Similarity-based estimation of properties

Similarity spaces created in Task 1 have been used in the estimation of properties using the *k*-nearest neighbor (KNN) method. One such study used the KNN approach for the classification of a set of 113 compounds as mutagens and non-mutagens (Pub. #4). Both the AP and PCA methods were employed to predict mutagenic activity with comparable results for both methods.

This research will have important implications both in computational toxicology and pharmaceutical drug discovery. In toxicology, most of the chemicals in commerce and new chemical entities do not have the data necessary for proper risk assessment. Similarity methods can be used in the quick estimation of properties in such cases. Combinatorial chemistry, which produces thousands of chemicals per week, is fast growing as THE method for drug discovery and lead optimization. Only certain bioassays that can be run in a 96-well plate at micromolar concentrations are carried out for these new compounds. Few if any of these chemicals have simple property data such as boiling point or vapor pressure, let alone the more complex pharmacokinetic or pharmacodynamic data. However, all of these chemicals have a known molecular structure. Our molecular similarity methods, based on the AP or PCA methods, or utilizing the hierarchical approach and the newly developed hierarchical QSAR approach, can be enormously beneficial in such situations for the rapid and reasonable estimation of necessary properties.

TASK 4: Application of neural networks in property estimation

Neural networks have been constructed for the estimation of acute aquatic toxicity (LC₅₀) in fathead minnow (*Pimephales promelas*) (Pubs. #9-11). In the first two studies, two standard backpropagation neural networks were constructed for the estimation of toxicity: a) a network using 95 topological, geometrical, and quantum chemical parameters, and b) a network using a subset of 23 of the 95 parameters based on a statistical method for variable selection (VC) (Pub. #9 & 10). The performance of these models was on par with the performance of linear statistical methods from an earlier

study. However, the neural network using only 23 parameters showed a slight improvement in model performance over the model using all 95 parameters.

The third study focused on the use of a machine learning technique, rather than traditional statistical approaches, for the selection of a reduced set of model parameters. Seeing the improvement made by using a reduced feature set (set of molecular descriptors) in our first two studies, we decided to try other techniques for limiting the feature set. This study compared the estimation of aquatic toxicity between three models: a) a neural network using all 95 parameters, b) a statistical analysis using 23 parameters selected through the variable clustering procedure, and c) a neural network utilizing a genetic ensemble feature selection (GEFS) algorithm. The neural network using the GEFS algorithm developed by David Opitz showed significant improvement over both the linear statistical model and the "standard" neural network model.

PERSONNEL SUPPORTED

Brian Gute – Graduate Research Assistant
Ph.D. student in computational toxicology

Jennifer Maki – Undergraduate Research Assistant

Jason Dagit – Undergraduate Research Assistant

Denise Mills – Undergraduate Research Assistant

PUBLICATIONS

The following peer-reviewed papers, which are currently either published, in press, or submitted, report results of research carried out between August 1, 1998 and July 31, 1999.

1. Topological indices: Their nature and mutual relatedness, S.C. Basak, A.T. Balaban, G. D. Grunwald, and B. D. Gute, *J. Chem. Inf. Comput. Sci.*, submitted, 1999.
2. Quantitative comparison of five molecular structure spaces in selecting analogs of chemicals, S.C. Basak, B.D. Gute, and G.D. Grunwald, *Mathl. Model. Comput. Sci.*, **8**, in press, 1999.
3. Assessment of the mutagenicity of chemicals from theoretical structural parameters: A hierarchical approach, S.C. Basak, B.D. Gute, and G.D. Grunwald, *SAR QSAR Environ. Res.*, in press, 1999.

4. Use of graph invariants in QMSA and predictive toxicology, S.C. Basak and B.D. Gute, in: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, in press, 1999.
5. Characterization of the molecular similarity of chemicals using topological invariants, S. C. Basak, B. D. Gute, and G. D. Grunwald, in: *Advances in Molecular Similarity, Volume 2*, eds. R. Carbo-Dorca, P.G. Mezey, JAI Press, Stamford, CT, 1998, p 171-185.
6. A QSAR study of complement-inhibitory activity of benzamidines, S.C. Basak, B. D. Gute, B. Lucic, S. Nikolic, and N. Trinajstić, *Computers and Chemistry*, submitted, 1999.
7. Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters, S.C. Basak, B.D. Gute, and S. Ghatak, *J. Chem. Inf. Comput. Sci.*, **39**, 255-260, 1999.
8. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): a hierarchical QSAR approach, B. D. Gute, G. D. Grunwald, and S. C. Basak, *SAR QSAR Environ. Res.*, **10**, 1-15, 1998.
9. Use of statistical and neural net methods in predicting toxicity of chemicals: A hierarchical QSAR approach, S.C. Basak, B.D. Gute, G.D. Grunwald, D.W. Opitz and K. Balasubramanian, in: *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools - Papers from the 1999 AAAI Symposium*, AAAI Press, Menlo Park, CA, 1999, p 108-111.
10. Use of statistical and neural net approaches in predicting toxicity of chemicals, S.C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, and D. Opitz, *J. Chem. Inf. Comput. Sci.*, submitted, 1999.
11. Hazard assessment modeling: An evolutionary ensemble approach, D.W. Opitz, S.C. Basak, and B.D. Gute, in: *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference*, eds. W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, & R.E. Smith, Morgan Kaufmann: San Francisco, accepted, 1999.

Copies of reprints are attached. Copies of the manuscripts in the various stages of review and pre-publication have not been attached for the sake of brevity.

APPENDIX 1

- Appendix 1.1 Characterization of the molecular similarity of chemicals using topological invariants, S. C. Basak, B. D. Gute, and G. D. Grunwald, in: *Advances in Molecular Similarity, Volume 2*, eds. R. Carbo-Dorca, P.G. Mezey, JAI Press, Stamford, CT, 1998, p 171-185.
- Appendix 1.2 Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters, S.C. Basak, B.D. Gute, and S. Ghatak, *J. Chem. Inf. Comput. Sci.*, **39**, 255-260, 1999.
- Appendix 1.3 Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): a hierarchical QSAR approach, B. D. Gute, G. D. Grunwald, and S. C. Basak, *SAR QSAR Environ. Res.*, **10**, 1-15, 1998.
- Appendix 1.4 Use of statistical and neural net methods in predicting toxicity of chemicals: A hierarchical QSAR approach, S.C. Basak, B.D. Gute, G.D. Grunwald, D.W. Opitz and K. Balasubramanian, in: *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools - Papers from the 1999 AAAI Symposium*, AAAI Press, Menlo Park, CA, 1999, p 108-111.

CHARACTERIZATION OF THE MOLECULAR SIMILARITY OF CHEMICALS USING TOPOLOGICAL INVARIANTS

Subhash C. Basak, Brian D. Gute, and
Gregory D. Grunwald

Abstract	172
I. Introduction	172
II. Methods	173
A. Database	173
B. Calculation of Indices	173
C. Classification of the Indices	178
D. Statistical Methods and Computation of Similarity	178
III. Results	179
A. Principal Component Analysis	179
B. Analogue Selection	183
C. <i>K</i> -Nearest-Neighbor Property Estimation	183
IV. Discussion	183
Acknowledgments	184
References	184

Advances in Molecular Similarity,
Volume 2, pages 171–185.
Copyright © 1998 by JAI Press Inc.
All rights of reproduction in any form reserved.
ISBN: 0-7623-0258-5

ABSTRACT

Three similarity spaces were used in the selection of analogues and *K*-nearest-neighbor (KNN)-based estimation of normal boiling points for a diverse set of 2926 chemicals. The similarity spaces consisted of principal components derived from (1) 40 topostructural indices, (2) 61 topochemical parameters, and (3) the full set of 101 topostructural and topochemical indices. The three methods selected sets of analogues with a substantial number of structurally analogous molecules. For the KNN method of property estimation, the similarity space that used the full set of indices was superior to either of the subsets (topostructural or topochemical). For all three methods, *K* = 6–10 gave the best estimated values for boiling point.

I. INTRODUCTION

Interest in quantifying the similarity of molecules using computational methods has increased.^{1–8} In particular, a recent trend in the characterization of similarity/dissimilarity of chemicals makes use of graph invariants. Molecular structures can be represented by planar graphs, $G = [V, E]$, where the nonempty set *V* represents the set of atoms and the set *E* generally represents covalent bonds.⁹ These graphs can be used to adequately represent the pattern of connectedness of atoms within a molecule. Graph invariants, values derived from planar graphs, are graph theoretic properties which are identical for isomorphic graphs. A numerical graph invariant or topological index maps a chemical structure into the set of real numbers.

Various graph invariants have been used in ordering and partial ordering of sets of molecules.^{1,4–8} Various topological indices (TIs) and principal components (PCs) derived from TIs have been used in quantifying the similarity/dissimilarity of molecules and in the similarity-based estimation of physical and toxicological properties.^{4,5,10–17} Such TIs include those derived from simple planar graphs which contain adjacency and distance information for vertices. These TIs could be considered topostructural indices. Other TIs, which are derived from weighted chemical graphs, could be regarded as topochemical indices because they contain explicit information regarding the chemical nature of the atoms (vertices) and bonds (edges) in the molecular structure, in addition to quantifying the adjacency and distance relationships within the graph.

Our earlier studies made use of a combination of topostructural and topochemical indices to select analogues of chemicals and estimate properties of molecules in large and diverse databases using the *K*-nearest-neighbor (KNN) method. In this paper we have carried out a comparative analysis of similarity-based analogue selection and KNN-based estimation of normal boiling point using: (1) a set of 40 topostructural indices, (2) a group of 61 topochemical indices, and (3) the combined set of 101 indices.

II. METHODS

A. Database

The normal boiling point database consisted of 2926 compounds taken from the U.S. EPA ASTER¹⁸ system. The data comprised a set for which chemical structures and normal boiling values were available, and for which it was possible to compute all 101 TIs.

B. Calculation of Indices

The TIs calculated for this study are listed in Table 1 and include Wiener number,¹⁹ molecular connectivity indices as calculated by Randić²⁰ and Kier and Hall,²¹ frequency of path lengths of varying size, information theoretic-indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić²² as well as those of Raychaudhury et al.,²³ parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,²⁴⁻²⁶ and Balaban's *J* indices.²⁷⁻²⁹ The majority of the TIs were calculated using POLLY 2.3.³⁰ The *J* indices were calculated using software developed by the authors.

The Wiener index (*W*), the first topological index reported in the chemical literature,¹⁹ may be calculated from the distance matrix *D*(*G*) of a hydrogen-suppressed chemical graph *G* as the sum of the entries in the upper triangular distance submatrix. The distance matrix *D*(*G*) of a nondirected graph *G* with *n* vertices is a symmetric *n* × *n* matrix (*d*_{*ij*}), where *d*_{*ij*} is equal to the distance between vertices *v*_{*i*} and *v*_{*j*} in *G*. Each diagonal element *d*_{*ii*} of *D*(*G*) is zero. We give below the distance matrix *D*(*G*₁) of the unlabeled hydrogen-suppressed graph *G*₁ of *n*-propanol (Figure 1):

$$D(G_1) = \begin{matrix} & \begin{matrix} (1) & (2) & (3) & (4) \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{bmatrix} \end{matrix}$$

W is calculated as

$$W = 1/2 \sum_{ij} d_{ij} = \sum_h h \cdot g_h \quad (1)$$

where *g*_{*h*} is the number of unordered pairs of vertices whose distance is *h*. Thus, for *D*(*G*₁), *W* has a value of ten.

Randić's connectivity index,²⁰ and higher order connectivity path, cluster, path-cluster, and chain types of simple, bond and valence connectivity parameters were

Table 1. Symbols, Definitions, and Classifications of Topological Parameters

<i>Topostructural</i>	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
hX	Path connectivity index of order $h = 0-6$
hX_c	Cluster connectivity index of order $h = 3-6$
$^hX_{PC}$	Path-cluster connectivity index of order $h = 4-6$
$^hX_{Ch}$	Chain connectivity index of order $h = 3-6$
P_h	Number of paths of length $h = 0-10$
J	Balaban's J index based on distance
<i>Topochemical</i>	
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
$^hX^b$	Bond path connectivity index of order $h = 0-6$
$^hX_c^b$	Bond cluster connectivity index of order $h = 3-6$
$^hX_{Ch}^b$	Bond chain connectivity index of order $h = 3-6$
$^hX_{PC}^b$	Bond path-cluster connectivity index of order $h = 4-6$
$^hX^v$	Valence path connectivity index of order $h = 0-6$
$^hX_c^v$	Valence cluster connectivity index of order $h = 3-6$
$^hX_{Ch}^v$	Valence chain connectivity index of order $h = 3-6$
$^hX_{PC}^v$	Valence path-cluster connectivity index of order $h = 4-6$
J^b	Balaban's J index based on bond types
J^x	Balaban's J index based on relative electronegativities
J^r	Balaban's J index based on relative covalent radii

Figure 1. The unlabeled hydrogen-suppressed graph (G_1) of n -propanol.

calculated using the method of Kier and Hall.²¹ The generalized form of the simple path connectivity index is as follows:

$${}^hX = \sum_{\text{paths}} (v_i v_j \dots v_{h+1})^{-1/2} \quad (2)$$

where $v_i, v_j, \dots, v_{h+1}$ are the degrees of the vertices in the path of length h . The path length parameters (P_h), number of paths of length h ($h = 0, 1, \dots, 10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Information-theoretic TIs are calculated by the application of information theory on chemical graphs. An appropriate set A of n elements is derived from a molecular graph G depending on certain structural characteristics. On the basis of an equivalence relation defined on A , the set A is partitioned into disjoint subsets A_i of order n_i ($i = 1, 2, \dots, h; \sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, \dots, A_h$$

$$p_1, p_2, \dots, p_h$$

where $p_i = n_i/n$ is the probability that a randomly selected element of A will occur in the i th subset.

The mean information content of an element of A is defined by Shannon's relation:³¹

$$IC = - \sum_{i=1}^h p_i \log_2 p_i \quad (3)$$

The logarithm is taken at base 2 for measuring the information content in bits. The total information content of the set A is then $n \times IC$.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar et al.³² calculated the information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by stereoelectronic characteristics of distant neighbors, i.e., neighbors of neighbors, it was deemed

essential to extend this approach to account for higher order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If r is any nonnegative real number and v is a vertex of the graph G , then the open sphere $S(v, r)$ is defined as the set consisting of all vertices v_i in G such that $d(v, v_i) < r$. Therefore, $S(v, 0) = \phi$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r)$ is the set consisting of v and all vertices v_i of G situated at unit distance from v , if $1 < r < 2$.

One can construct such open spheres for higher integral values of r . For a particular value of r , the collection of all such open spheres $S(v, r)$, where v runs over the whole vertex set V , forms a neighborhood system of the vertices of G . A suitably defined equivalence relation can then partition V into disjoint subsets consisting of vertices that are topologically equivalent for r th-order neighborhood. Such an approach has been developed and the information-theoretic indices calculated based on this idea are called indices of neighborhood symmetry.²⁶

In this method, chemicals are symbolized by weighted linear graphs. Two vertices u_0 and v_0 of a molecular graph are said to be equivalent with respect to r -th-order neighborhood if and only if corresponding to each path $u_0, u_1, \dots, u_r$ of length r , there is a distinct path $v_0, v_1, \dots, v_r$ of the same length such that the paths have similar edge weights, and both u_0 and v_0 are connected to the same number and type of atoms up to the r th-order bonded neighbors. The detailed equivalence relation has been described in earlier studies.^{26,33}

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by Eq. 2. Basak et al. defined another information-theoretic measure, structural information content (SIC_r), which is calculated as

$$SIC_r = IC_r / \log_2 n \quad (4)$$

where IC_r is calculated from Eq. 2 and n is the total number of vertices of the graph.²⁴

Another information-theoretic invariant, complementary information content (CIC_r), is defined as

$$CIC_r = \log_2 n - IC_r \quad (5)$$

CIC_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r .²⁵

In Figure 2, the calculation of IC_2 , SIC_2 , and CIC_2 is demonstrated for the labeled hydrogen-filled graph (G_2) of n -propanol.

The information-theoretic index on graph distance, I_D^W , is calculated from the distance matrix $D(G)$ of a chemical graph G as follows:²²

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h \quad (6)$$

Second order neighbors:

Subsets:

Probability:

$$\begin{aligned}
 IC_2 &= 5 \cdot 1/12 \cdot \log_2 12 + 2 \cdot 2/12 \cdot \log_2 12/2 + 3/12 \cdot \log_2 12/3 = 2.855 \text{ bits} \\
 SIC_2 &= IC_1 / \log_2 12 = 0.796 \text{ bits} \\
 CIC_2 &= \log_2 12 - IC_2 = 0.730 \text{ bits}
 \end{aligned}$$

Figure 2. Calculation of the indices IC_2 , SIC_2 , and CIC_2 for the hydrogen-filled, labeled graph (G_2) of *n*-propanol.

The mean information index, I_D^W , is found by dividing the information index I_D^W by W . The information-theoretic parameters defined on the distance matrix, H^D and H^V , were calculated by the method of Raychaudhury et al.²³

Balaban defined a series of indices based on distance sums within the distance matrix for a chemical graph which he designated as J indices.²⁷⁻²⁹ These indices are highly discriminating with low degeneracy. Unlike W , the J indices have a range of values that is independent of molecular size. The general form of the J index calculation is as follows:

$$J = q(\mu + 1)^{-1} \sum_{i,j, \text{ edges}} (s_i s_j)^{-1/2} \quad (7)$$

where the cyclomatic number μ (or number of rings in the graph) is $\mu = q - n + 1$ with q edges and n vertices, and s_i is the sum of the distances of atom i to all other atoms and s_j is the sum of the distances of atom j to all other atoms.²⁷ Variants were proposed by Balaban for incorporating information on bond type, relative electronegativities, and relative covalent radii.^{28,29}

C. Classification of the Indices

The set of 101 TIs was partitioned into two distinct subsets: topostructural indices and topochemical indices. Topostructural indices encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of atom type or factors such as hybridization states and number of core/valence electrons in individual atoms. Topochemical indices quantify information regarding specific chemical properties of the atoms comprising a molecule as well as the topology (connectivity of atoms). Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These subsets are shown in Table 1.

D. Statistical Methods and Computation of Similarity

Data Reduction

Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some TIs may be several orders of magnitude greater than other TIs.

A principal component analysis (PCA) was used on the transformed indices to minimize intercorrelation of indices. The PCA analysis was accomplished using the SAS procedure PRINCOMP.³⁴ The PCA produces linear combinations of the TIs, called principal components (PCs) which are derived from the correlation matrix. The first PC has the largest variance, or eigenvalue, of the linear combination of TIs. Each subsequent PC explains the maximal index variance orthogonal to the previous PCs, eliminating any redundancies that could occur within the set

of TIs. The maximum number of PCs generated is equal to the number of TIs available. For the purposes of this study, only PCs with eigenvalues greater than one were retained. A more detailed explanation of this approach has been provided in a previous study by Basak et al.⁴ These PCs were subsequently used in determining similarity scores as described below.

Similarity Measures

Intermolecular similarity was measured by the Euclidean distance (*ED*) within an *n*-dimensional space. This *n*-dimensional space consisted of orthogonal variables (PCs) derived from the TIs as described above. *ED* between molecules *i* and *j* is defined as

$$ED_{ij} = \left[\sum_{k=1}^n (D_{ik} - D_{jk})^2 \right]^{1/2} \quad (8)$$

where *n* is the number of dimensions or PCs retained from the PCA. *D_{ik}* and *D_{jk}* are the data values of the *k*th dimension for chemicals *i* and *j*, respectively.

K-Nearest-Neighbor Selection and Property Estimation

Following the quantification of intermolecular similarity of the 2926 chemicals, the *K*-nearest neighbors (*K* = 1–10, 15, 20, 25) were determined on the basis of *ED*. This procedure can be used to select structural analogues (neighbors) of a probe compound or the neighbors can be used in property estimation. In estimating the normal boiling point of the probe compound, the mean observed normal boiling point of the *K*-nearest neighbors was used as the estimate and the standard error (*s*) of the estimate was used to assess the efficacy of the set of indices.

III. RESULTS

A. Principal Component Analysis

From the PCA of the 40 topostructural indices, seven PCs with eigenvalues greater than one were retained. These seven PCs explained, cumulatively, 90.8% of the total variance within the TI data. Table 2 lists the eigenvalues of the seven PCs, the proportion of variance explained by each PC, the cumulative variance explained, and the three TIs most correlated with each individual PC.

The PCA of the 61 topochemical indices resulted in the selection of ten PCs, all having eigenvalues greater than one. The ten PCs explain a total of 92.1% of the variance within the TI data. Table 3 presents a summary of the information regarding these ten PCs.

Table 2. Summary of Principal Component Analysis of 40 Topostructural Indices for 2926 Chemicals

PC	Eigenvalue	Proportion of Explained Variance	Cumulative Explained Variance	Top Three Correlated Indices
1	28.2	46.2	46.2	$P_1, P_0, {}^1X$
2	11.0	18.0	64.3	${}^4X_{PC}, {}^5X_{PC}, {}^6X_{PC}$
3	5.9	9.6	73.9	${}^3X_C, {}^5X_C, {}^4X_{PC}$
4	4.1	6.7	80.6	$J, {}^6X_{Ch}, {}^4X_C$
5	2.8	4.6	85.2	${}^4X_{Ch}, {}^5X_{Ch}, {}^3X_{Ch}$
6	1.9	3.1	88.3	${}^3X_{Ch}, {}^4X_{Ch}, {}^5X_{Ch}$
7	1.5	2.4	90.8	${}^6X_C, P_{10}, P_9$

Table 3. Summary of Principal Component Analysis of 61 Topochemical Indices for 2926 Chemicals

PC	Eigenvalue	Proportion of Explained Variance	Cumulative Explained Variance	Top Three Correlated Indices
1	20.4	33.5	33.5	${}^1X^b, {}^2X^b, {}^3X^b$
2	10.8	17.8	51.2	SIC_4, SIC_3, SIC_5
3	8.1	13.3	64.6	${}^3X_C^b, {}^4X_C^b, {}^4X_{PC}^b$
4	6.1	9.9	74.5	${}^5X_{Ch}^b, {}^5X_{Ch}^b, {}^4X_{Ch}^b$
5	3.0	5.0	79.5	${}^3X_{Ch}^b, {}^3X_{Ch}^b, {}^4X_{Ch}^b$
6	2.4	3.9	83.4	IC_0, SIC_0, IC_1
7	1.7	2.8	86.2	${}^6X_C^b, {}^5X_C^b, {}^6X_C^b$
8	1.4	2.2	88.4	${}^4X_C^b, {}^2X^b, {}^6X_C^b$
9	1.2	2.0	90.4	${}^5X_C^b, {}^6X_C^b, {}^4X_C^b$
10	1.1	1.8	92.1	${}^4X_C^b, {}^4X_C^b, {}^6X_{PC}^b$

Table 4. Summary of Principal Component Analysis of 101 Topological Indices for 2926 Chemicals

PC	Eigenvalue	Proportion of Explained Variance	Cumulative Explained Variance	Top Three Correlated Indices
1	42.6	41.6	41.6	$P_1, P_0, {}^1X$
2	13.3	13.0	54.7	${}^4X_{PC}^b, {}^4X_{PC}, {}^3X_C$
3	11.4	11.1	65.8	SIC_5, SIC_6, CIC_6
4	8.9	8.7	74.5	${}^5X_{Ch}^b, {}^5X_{Ch}, {}^5X_{Ch}^b$
5	5.1	5.0	79.6	$J, {}^4X_{Ch}, {}^4X_{Ch}^b$
6	3.7	3.6	83.2	IC_0, SIC_0, SIC_1
7	2.6	2.6	85.8	${}^6X_C^b, {}^6X_C, {}^5X_C^b$
8	2.0	1.9	87.7	${}^4X_C^b, {}^5X_{Ch}, {}^6X_{Ch}^b$
9	1.7	1.7	89.4	${}^4X_C^b, IC_0, SIC_0$
10	1.4	1.4	90.8	${}^5X_C^b, {}^4X_{PC}, {}^6X_C^b$
11	1.1	1.1	91.9	IC_1, J, IC_0
12	1.0	1.0	92.8	P_9, P_{10}, P_8

Twelve PCs were retained from the PCA of the full set of 101 TIs. Each of these PCs had an eigenvalue greater than one and, cumulatively, they explained 92.8% of the variance within the full set of TIs. These PCs are summarized in Table 4.

Figure 3. The five analogues selected for the probe 3-methyl-4-chlorophenol using three molecular similarity spaces: topostructural, topochemical, and all indices. The numbers under the structures indicate the ranking of the analogues and the Euclidean distance to the probe.

Table 5. Comparison of the Three Sets of TIs and Their Derivative PCs for Prediction of Normal Boiling Point (°C) Using *K*-Nearest-Neighbors (*n* = 2926)

<i>Indices</i>	<i>K</i>	<i>r</i>	<i>s</i>
Topostructural	10	0.881	39.0
Topochemical	6	0.883	38.6
Topostructural + topochemical	8	0.896	36.6

Figure 4. Pattern of (top) correlation (*r*) and (bottom) standard error (*s*) of the estimates according to the *K*-nearest-neighbor selection for 2926 normal boiling points using three molecular similarity spaces.

B. Analogue Selection

Figure 3 shows an example of analogue selection using PCs to derive a Euclidean distance space. The first five analogues (neighbors) for the probe compound, 3-methyl-4-chlorophenol, are presented for each of the three similarity spaces. The analogues selected by the topostructural model show a repetition of the same skeletal structure, ignoring substituents, throughout the first five analogues. In the topochemical model and the full set model some variability in the skeletal structure arises (chemical analogues 2 and 5, full set analogue 4). Also of interest is the repetition of chemicals between the sets of analogues. While the ordering varies between the methods, the topostructural and topochemical models select two identical structures, the topostructural and the full set have three analogues in common, and the topochemical and full set select four of the same analogues. 2-Chloro-5-methylphenol appears in all three sets, while there are only three unique compounds (topostructural analogues 4 and 5, topochemical analogue 5).

C. *K*-Nearest-Neighbor Property Estimation

Figure 4 presents the correlation (r) and the standard error (s) of the prediction of the normal boiling points for the 2926 chemicals for the three groups of indices over the full range of K values examined ($K = 1-10, 15, 20, 25$). Table 5 shows the best normal boiling point model for each set of indices. The best boiling point estimates for all three sets were for K in the range of 6 to 10. The full set of indices gave the best result, although there was only a small difference between models.

IV. DISCUSSION

The purpose of this paper was to study the relative effectiveness of three similarity spaces derived from graph invariants in the selection of structural analogues and in the KNN-based estimation of properties. The similarity spaces were created using a PCA of calculated graph invariants. Tables 2–4 summarize the results of the PCA of the three sets of indices. The first PC is always correlated with indices that quantify molecular size. In the case of the topostructural indices, the second PC is most correlated with branching indices. In the case of PCs derived from either topochemical or the full set of topostructural and topochemical parameters, the first PC was strongly correlated with molecular size, while the second PC was highly associated with the molecular complexity indices. These results are in line with our earlier studies on different sets of chemicals.^{4,5,11,35,36}

All three spaces were used in the selection of five analogues of a particular structure (Figure 3). Perusal of the three sets of structures shows that there is a substantial degree of similarity among the three groups of five chemicals selected. It is interesting to note that all five nearest neighbors of the probe selected by the topostructural method had isomorphic skeletal graphs when hydrogen atoms are

suppressed. For the two similarity spaces created by topochemical indices alone and the combined set of topostructural and topochemical indices, four of the five selected neighbors are common (Figure 3) although the ordering of the molecules is different. This shows that these two similarity methods are not intrinsically very different. Our earlier results showed that analogues selected by similarity methods derived from experimental physical properties, atom pairs, and TIs select very similar sets of analogues.¹⁰

In the case of KNN-based estimation of boiling points of chemicals from their analogues, K was varied from 1 to 25. The best estimated value was obtained in the range of $K = 6-10$. This is in line with our earlier studies with different properties.^{11,12}

In conclusion, the three similarity spaces derived in this paper have reasonable power for selecting analogous molecules from a very diverse database of chemicals. The KNN-based estimation shows that selected analogues can be used for the estimation of boiling points of diverse chemicals if more accurate methods are not available.

ACKNOWLEDGMENTS

This is contribution number 161 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported herein was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, a grant from Exxon Corporation, and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

REFERENCES

1. Johnson, M. A.; Maggiora, G. M. Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
2. Carbó, R.; Leyda, L.; Arnau, M. *Int. J. Quantum Chem.* **1980**, *17*, 1185.
3. Bowen-Jenkins, P. E.; Cooper, D. L.; Richards, G. *J. Phys. Chem.* **1985**, *89*, 2195.
4. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discrete Appl. Math.* **1988**, *19*, 17.
5. Basak, S. C.; Bertelsen, S.; Grunwald, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270.
6. Rum, G.; Herndon, W. C. *J. Am. Chem. Soc.* **1991**, *113*, 9055.
7. Willett, P.; Winterman, V. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18.
8. Wilkins, C. L.; Randić, M. *Theor. Chim. Acta* **1980**, *58*, 45.
9. Trinajstić, N. *Chemical Graph Theory Vols. I & II*; CRC Press: Boca Raton, FL, 1983.
10. Basak, S. C.; Grunwald, G. D. *Math. Model. Sci. Comput.*, in press.
11. Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1994**, *2*, 289.
12. Basak, S. C.; Grunwald, G. D. *New J. Chem.* **1995**, *19*, 231.
13. Basak, S. C.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 366.
14. Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1995**, *3*, 265.
15. Basak, S. C.; Grunwald, G. D. *Chemosphere* **1995**, *31*, 2529.
16. Basak, S. C.; Gute, B. D.; Grunwald, G. D. *Croat. Chim. Acta* **1996**, *69*, 1159.
17. Lajiness, M. S. In: *Computational Chemical Graph Theory*; Rouvray, D. H., Ed. Nova Science Publishers: New York, 1990, p. 300.

18. Russom, C. L. *Assessment Tools for the Evaluation of Risk (Aster)* v. 1.0; U.S. Environmental Protection Agency, 1992.
19. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
20. Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
21. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Hertfordshire, U.K., 1986.
22. Bonchev, D.; Trinajstić, N. *J. Chem. Phys.* **1977**, *67*, 4517.
23. Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581.
24. Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the Second International Conference on Mathematical Modelling*; Avula, X. J. R.; Bellman, R.; Luke, Y. L.; Rigler, A. K., Eds.; University of Missouri-Rolla, 1980, p. 851.
25. Basak, S. C.; Magnuson, V. R. *Arzneim.-Forsch. Drug Res.* **1983**, *33*, 501.
26. Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R.; Kalman, R. E.; Liapis, A. I.; Rodin, E. Y., Eds.; Pergamon Press: New York, 1984, p. 745.
27. Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
28. Balaban, A. T. *Pure and Appl. Chem.* **1983**, *55*, 199.
29. Balaban, A. T. *Math. Chem. (MATCH)* **1985**, *21*, 115.
30. Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY* v. 2.3 (copyright University of Minnesota), 1988.
31. Shannon, C. E. *Bell Syst. Tech. J.* **1948**, *27*, 379.
32. Sarkar, R.; Roy, A. B.; Sarkar, R. K. *Math. Biosci.* **1978**, *39*, 299.
33. Magnuson, V. R.; Harriss, D. K.; Basak, S. C. In *Studies in Physical and Theoretical Chemistry*; King, R. B., Ed.; Elsevier: Amsterdam, 1983, p. 178.
34. SAS Institute Inc. In *SAS/STAT User's Guide, Release 6.03 Edition*; SAS Institute Inc.: Cary, NC, 1988, p. 751.
35. Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1991**, *7*, 243.
36. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Model.* **1987**, *8*, 300.

Prediction of Complement-Inhibitory Activity of Benzamidines Using Topological and Geometric Parameters

Subhash C. Basak, Brian D. Gute, and Shibnath Ghatak

Natural Resources Research Institute, The University of Minnesota,
5013 Miller Trunk Highway, Duluth, Minnesota 55811, and
Department of Biology, Dana Laboratory, Tufts University,
Medford, Massachusetts 02155

Journal of
**Chemical
Information and
Computer Sciences[®]**

Reprinted from
Volume 39, Number 2, Pages 255–260

Prediction of Complement-Inhibitory Activity of Benzamidines Using Topological and Geometric Parameters

Subhash C. Basak,^{*,†} Brian D. Gute,[†] and Shibnath Ghatak[‡]

Natural Resources Research Institute, The University of Minnesota, 5013 Miller Trunk Highway, Duluth, Minnesota 55811, and Department of Biology, Dana Laboratory, Tufts University, Medford, Massachusetts 02155

Received April 15, 1998

A hierarchical approach to quantitative structure–activity relationship (QSAR) modeling has been used to estimate the complement-inhibitory potency of 105 benzamidines. This hierarchical approach uses topostructural, topochemical, and geometric parameters in a stepwise fashion to build increasingly more complex models. The results show that topostructural indices alone, specifically P^D , predict inhibitory potency reasonably well. The addition of topochemical and geometrical parameters to the set of descriptors provides only marginal improvement in predictive power. However, when taken alone, the geometric parameter ^{3D}W provides a more stable model than the topostructural one.

1. INTRODUCTION

A recent trend in structure–activity relationships (SAR) is the use of topological and geometric parameters in predicting the physicochemical, biochemical, and toxicological properties of molecules.^{1–23} Topological indices (TIs) are numerical descriptors of molecular topology and encode information regarding the size, shape, branching, and symmetry of molecular graphs.²³ TIs and substructural parameters have been very useful in the development of quantitative structure–activity relationship (QSAR) models, in the quantification of the structural similarity of chemicals and in the similarity-based estimation of numerous physical and biological properties of diverse sets of molecules.^{24–39} On the other hand, geometric variables such as total surface area, volume, and three-dimensional Wiener index have been employed in QSARs pertaining to biomedical and toxicological action of molecules with good results.^{3,14,40–44}

One interesting area of research in biochemistry, pharmacology, and toxicology is the rationalization of the action of classes of chemicals with specialized modes of action. Specificity in enzymology, immunology, and toxicology arises out of specific structural features which lead to particular types of interactions between ligands and their biotargets. Topological and geometric parameters have been used in the development of QSARs of many groups of molecules with specific modes of action.^{3,7,9,10,13,14,16,17,31–33,42–44}

Complement is a system of factors occurring in normal serum which are characteristically activated by antibody–antigen interactions and which subsequently mediate a number of biologically significant consequences.⁴⁵ The factors of the complement system include at least 20 chemically distinct serum proteins and glycoproteins. These

Table 1. Conflicting Data for Structure and Log 1/C for Four Benzamidines

no.	X	obsd log 1/C
77	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -3 ^a -SO ₂ F	4.23
95	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -3-SO ₂ F	4.51
97	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-SO ₂ F	4.57
108	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-SO ₂ F	5.21

^a This SO₂F group should be *meta*- instead of *para*-.

Figure 1. Neutral base structure for the 107 benzamidines.

factors, which normally exist in an inactive form, are activated by “classical” and “alternative” pathways. Both pathways generate macromolecular membrane attack complexes which lyse a variety of cells, bacteria, and viruses.⁴⁶ Products of this activation result in inflammatory reactions at the site of antibody–antigen interaction. This is especially pronounced in the case of organ specific and systemic autoimmune disorders. Therefore, control of unregulated complement activation is important, at least in the case of autoimmune disease.

Hansch and Yoshimoto⁴⁷ carried out a QSAR study of a set of 108 benzamidine derivatives using linear free-energy related (LFER) parameters. This series of compounds are inhibitors of the complement system. In view of the fact that LFER parameters are not routinely available for any arbitrary chemical, real or hypothetical, it was of interest to see whether computable parameters such as TIs and geometric indices can give a reasonable QSAR for the set of benzamidines. Therefore, in this paper we have carried out a comparative study of the utility of topological indices vis-à-vis calculated geometric parameters in predicting the complement-inhibitory potencies of this set of benzamidines.

* All correspondence to be addressed to Dr. Subhash C. Basak, Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811. Phone: (218) 720-4230. E-mail: sbasak@wyle.nrrri.umn.edu.

† The University of Minnesota.

‡ Tufts University.

Table 2. Side-Chain Structures and Biological Property Data for 107 Benzamidines

no.	X	1/log C			no.	X	1/log C		
		obsd	predict. ^a	resid			obsd	predict. ^a	resid
1	3,5-(OCH ₃) ₂	-0.452	-0.367 ^b	-0.085	55	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₃ -2-Cl-6-SO ₂ F	-0.255	-0.237	-0.018
2	CH ₃	-0.444	-0.405	-0.040	56	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₅	-0.255	-0.249	-0.006
3	4-(CH ₃) ₂	-0.425	-0.389	-0.036	57	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₃ -2-Cl-5-SO ₂ F	-0.250	-0.236	-0.014
4	H	-0.418	-0.417	-0.002	58	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₃ -4-SO ₂ F	-0.250	-0.228	-0.022
5	3-OH	-0.415	-0.402	-0.012	59	3-O(CH ₃)OC ₆ H ₄ -2-NHCONH-C ₆ H ₂ -2,4-(CH ₃) ₂ -5-SO ₂ F	-0.248	-0.229	-0.019
6	3-NHCO(CH ₂) ₂ C ₆ H ₅	-0.412	-0.302 ^b	-0.110	60	3-O(CH ₃)OC ₆ H ₄ -4-COOCH ₃	-0.247	-0.271	0.025
7	3-CF ₃	-0.410	-0.369	-0.041	61	3-O(CH ₃)OC ₆ H ₃ -3-NO ₂ -4-CH ₃	-0.245	-0.273	0.028
8	3-NO ₂	-0.410	-0.378	-0.032	62	3-O(CH ₃)OC ₆ H ₄ -3-CF ₃	-0.245	-0.273	0.028
9	3-Br	-0.405	-0.401	-0.004	63	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₄ -4-CH ₃ -3-SO ₂ F	-0.245	-0.229	-0.015
10	3-CH ₃	-0.398	-0.402	0.004	64	3-O(CH ₃)OC ₆ H ₄ -4-NHCOC ₆ H ₅	-0.244	-0.246	0.002
11	3-OCH ₃	-0.397	-0.389	-0.008	65	3-O(CH ₃)OC ₆ H ₄ -2-NHCOCH ₂ OC ₆ H ₄ -4-SO ₂ F	-0.244	-0.227	-0.017
12	3-CH ₂ C ₆ H ₅	-0.373	-0.339	-0.034	66	3-O(CH ₃)OC ₆ H ₄ -4-NHCOC ₆ H ₄ -4-OCH ₃	-0.243	-0.236	-0.007
13	3,5-(CH ₃) ₂	-0.361	-0.389	0.028	67	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₄ -3-SO ₂ F	-0.243	-0.238	-0.005
14	3-OC ₃ H ₇	-0.355	-0.362	0.007	68	3-O(CH ₃)OC ₆ H ₄ -2-NHCOCH ₂ C ₆ H ₄ -4-SO ₂ F	-0.243	-0.233	-0.010
15	3- <i>i</i> -C ₃ H ₁₁	-0.355	-0.353	-0.002	69	3-O(CH ₃)OC ₆ H ₄ -3-COOCH ₃	-0.242	-0.272	0.030
16	3-OC ₄ H ₉	-0.351	-0.349	-0.001	70	3-O(CH ₃)OC ₆ H ₄ -2-NHCO(CH ₂) ₂ C ₆ H ₄ -4-SO ₂ F	-0.242	-0.227	-0.014
17	3-C ₆ H ₅	-0.338	-0.362	0.024	71	3-O(CH ₃)OC ₆ H ₄ -4-NHCOC ₆ H ₄ -4-NO ₂	-0.239	-0.232	-0.007
18	3-CH=CHC ₆ H ₅	-0.339	-0.325	-0.014	72	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₄ -4-NO ₂	-0.239	-0.241	0.002
19	3-OCH ₂ C ₆ H ₅	-0.331	-0.326	-0.005	73	3-O(CH ₃)OC ₆ H ₄ -4-NHCONHC ₆ H ₅	-0.237	-0.241	0.004
20	3-(CH ₂) ₂ C ₆ H ₅	-0.330	-0.326	-0.004	74	3-O(CH ₃)OC ₆ H ₄ -4-NHCOC ₆ H ₄ -3-NO ₂	-0.237	-0.233	-0.005
21	3-OC ₆ H ₁₃	-0.329	-0.327	-0.002	75	3-O(CH ₃)OC ₆ H ₄ -2-NHCO(CH ₂) ₂ C ₆ H ₄ -4-SO ₂ F	-0.237	-0.217	-0.020
22	3-O(CH ₂) ₄ OC ₆ H ₅	-0.325	-0.288	-0.037	76	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.237	-0.233	-0.004
23	3-O(CH ₂) ₂ OC ₆ H ₅	-0.323	-0.306	-0.017	77	3-O(CH ₃)OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.236	-0.225	-0.011
24	3-C ₆ H ₅	-0.323	-0.347	0.025	78	3-O(CH ₃)OC ₆ H ₄ -2-NHCONH(CH ₂) ₂ C ₆ H ₄ -4-SO ₂ F	-0.236	-0.223	-0.014
25	3-O(CH ₃) ₂ OC ₆ H ₄ -4-COOH	-0.321	-0.277	-0.044	79	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-SO ₂ F	-0.236	-0.223	-0.013
26	3-OC ₅ H ₁₁	-0.320	-0.338	0.017	80	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₃ -4-Cl-3-SO ₂ F	-0.235	-0.229	-0.006
27	3-O- <i>i</i> -C ₅ H ₁₁	-0.318	-0.341	0.022	81	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₃ -4-CH ₃ -3-SO ₂ F	-0.235	-0.229	-0.006
28	3-O(CH ₂) ₂ OC ₆ H ₇ -α	-0.312	-0.283	-0.030	82	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₂ -2,4-(CH ₃) ₂ -5-SO ₂ F	-0.234	-0.233	-0.001
29	3-O(CH ₂) ₄ OC ₆ H ₄ -4-NH ₂	-0.306	-0.282	-0.024	83	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₂ -2,4-Cl ₂ -5-SO ₂ F	-0.234	-0.233	-0.001
30	3-(CH ₂) ₄ C ₆ H ₅	-0.302	-0.306	0.004	84	3-(CH ₂) ₄ C ₆ H ₄ -2-NHCONHC ₆ H ₄ -3-SO ₂ F	-0.234	-0.239	0.005
31	3-O(CH ₂) ₃ OC ₆ H ₄ -4-NO ₂	-0.301	-0.277	-0.024	85	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-OCH ₃	-0.233	-0.237	0.004
32	3-O(CH ₂) ₃ OC ₆ H ₄ -4-NH ₂	-0.300	-0.290	-0.010	86	3-(CH ₂) ₄ C ₆ H ₄ -2-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.233	-0.239	0.007
33	3-(CH ₂) ₂ -4-C ₃ H ₄ N	-0.299	-0.326	0.026	87	3-O(CH ₃)OC ₆ H ₄ -4-NHCOC ₆ H ₄ -4-Cl	-0.232	-0.241	0.009
34	3-O(CH ₂) ₃ OC ₆ H ₅	-0.299	-0.297	-0.003	88	3-O(CH ₃)OC ₆ H ₄ -2-NHCOC ₆ H ₃ -2-CH ₃ -5-SO ₂ F	-0.232	-0.236	0.004
35	3-O(CH ₂) ₃ C ₆ H ₅	-0.296	-0.306	0.010	89	3-O(CH ₃)OC ₆ H ₄ -4-NHCONHC ₆ H ₃ -2-OCH ₃ -5-SO ₂ F	-0.232	-0.214	-0.018
36	3-(CH ₂) ₂ -3-C ₅ H ₄ N	-0.294	-0.326	0.032	90	3-O(CH ₃)OC ₆ H ₄ -4-C ₆ H ₅	-0.230	-0.261	0.031
37	3-(CH ₂) ₄ C ₆ H ₄ -4-NHAc	-0.294	-0.273	-0.021	91	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₄ -3-SO ₂ F	-0.230	-0.233	0.003
38	3-(CH ₂) ₂ -2-C ₅ H ₄ N	-0.291	-0.326	0.035	92	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -3-SO ₂ F	-0.230	-0.230	-0.000
39	3-O(CH ₂) ₃ OC ₆ H ₄ -2-NH ₂	-0.283	-0.291	0.009	93	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -3-SO ₂ F	-0.229	-0.236	0.007
40	3-O(CH ₂) ₃ OC ₆ H ₄ -4-NHAc	-0.278	-0.265	-0.012	94	3-O(CH ₃)OC ₆ H ₄ -4-CH ₃ -3-NHCOC ₆ H ₄ -4-SO ₂ F	-0.229	-0.226	-0.003
41	3-(CH ₂) ₄ -3-C ₅ H ₄ N	-0.276	-0.306	0.030	95	3-O(CH ₃)OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -3-SO ₂ F	-0.222	-0.226	0.004
42	3-O(CH ₂) ₄ C ₆ H ₅	-0.276	-0.297	0.020	96	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₃ -4-SO ₂ F	-0.220	-0.226	0.006
43	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NHAc	-0.270	-0.267	-0.003	97	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-SO ₂ F	-0.219	-0.229	0.010
44	3-O(CH ₂) ₃ OC ₆ H ₃ -3,4-Cl ₂	-0.265	-0.283	0.018	98	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₃ -2-Cl-5-SO ₂ F	-0.217	-0.230	0.013
45	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NH ₂	-0.265	-0.290	0.025	99	3-O(CH ₃)OC ₆ H ₄ -3-NHCOCH ₂ OC ₆ H ₄ -4-SO ₂ F	-0.217	-0.219	0.002
46	3-O(CH ₂) ₃ OC ₆ H ₄ -2-NHCOC ₆ H ₄ -4-SO ₂ F	-0.265	-0.237	-0.028	100	3-O(CH ₃)OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.216	-0.231	0.015
47	3-O(CH ₂) ₃ OC ₆ H ₄ -2-NHCOC ₆ H ₅	-0.265	-0.253	-0.012	101	3-O(CH ₃)OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.215	-0.220	0.005
48	3-O(CH ₂) ₃ OC ₆ H ₄ -4-OCH ₃	-0.262	-0.283	0.022	102	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-NO ₂	-0.214	-0.233	0.019
49	3-O(CH ₂) ₃ OC ₆ H ₄ -4-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.260	-0.219	-0.040	103	3-O(CH ₃)OC ₆ H ₄ -3-NHCOC ₆ H ₄ -4-SO ₂ F	-0.214	-0.235	0.021
50	3-O(CH ₂) ₃ OC ₆ H ₄ -2-NHCOC ₆ H ₃ -2-OCH ₃ -5-SO ₂ F	-0.260	-0.233	-0.027	104	3-O(CH ₃)OC ₆ H ₄ -2-NHCONHC ₆ H ₃ -2-Cl-5-SO ₂ F	-0.207	-0.225	0.018
51	3-O(CH ₂) ₃ OC ₆ H ₄ -4-Cl	-0.257	-0.290	0.033	105	3-O(CH ₃)OC ₆ H ₄ -3-NHCONHC ₆ H ₄ -4-NO ₂	-0.204	-0.230	0.025
52	3-O(CH ₂) ₃ OC ₆ H ₄ -2-NO ₂	-0.257	-0.281	0.024	106	3-O(CH ₃)OC ₆ H ₄ -4-CH ₃ -3-NHCONHC ₆ H ₄ -4-SO ₂ F	-0.204	-0.223	0.018
53	3-O(CH ₂) ₃ OC ₆ H ₄ -3-NO ₂	-0.257	-0.278	0.021	107	3-O(CH ₃)OC ₆ H ₄ -3-NHCONH(CH ₂) ₂ C ₆ H ₄ -4-SO ₂ F	-0.193	-0.215	0.022
54	3-O(CH ₂) ₃ OC ₆ H ₄ -3-OCH ₃	-0.256	-0.283	0.027					

^a Predicted values based on eq 2. ^b Values for compounds excluded from final modeling, provided to show lack of fit.

2. METHODS

2.1. Database. The 107 benzamidines used in this study are those presented in the work of Hansch and Yoshimoto.⁴⁷ These data were compiled from a series of five articles by B. R. Baker,^{48–52} in which Baker and his students determined experimentally the inhibition of guinea pig complement by benzamidines. Hansch and Yoshimoto provide the structures and measured log 1/C values, where *C* is the micromolar

ease of comparison. In the process of coding the data, it became evident that two of the compounds had structural duplicates with distinctly different values for log 1/C (see Table 1). Through close examination of Baker's work, it became evident that there was a typographic mistake in compound 77, while the error in compound 108 could not be accounted for. Thus, compound 108 was discarded from the set, leaving 107 benzamidines derivatives. The base

Table 3. Symbols and Definitions of Topological and Geometrical Parameters

I_D^W	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	degree complexity
H^V	graph vertex complexity
H^D	graph distance complexity
$\bar{I}_C$	information content of the distance matrix partitioned by frequency of occurrences of distance h
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O	order of neighborhood when I_C reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
I_{C_r}	mean information content or complexity of a graph based on the r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi$	path connectivity index of order $h = 0-6$
${}^h\chi_C$	cluster connectivity index of order $h = 3-6$
${}^h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
${}^h\chi_{Ch}$	chain connectivity index of order $h = 6$
${}^h\chi^b$	bond path connectivity index of order $h = 0-6$
${}^h\chi_C^b$	bond cluster connectivity index of order $h = 3-6$
${}^h\chi_{Ch}^b$	bond chain connectivity index of order $h = 6$
${}^h\chi_{PC}^b$	bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	valence path connectivity index of order $h = 0-6$
${}^h\chi_C^v$	valence cluster connectivity index of order $h = 3-6$
${}^h\chi_{Ch}^v$	valence chain connectivity index of order $h = 6$
${}^h\chi_{PC}^v$	valence path-cluster connectivity index of order $h = 4-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^r	Balaban's J index based on relative covalent radii
V_w	van der Waal's volume
${}^{3D}W$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

2.2. Calculation of Topological Indices (TIs). Topological indices used in this study have been calculated by POLLY 2.3.⁵³ These indices include Wiener index,⁵⁴ connectivity indices,^{16,55} information theoretic indices defined on distance matrices of graphs,^{56,57} and a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs^{10,58-60} as well as Balaban's J indices.⁶¹⁻⁶³ Table 3 gives brief definitions for the topological indices included in this study.

2.3. Calculation of Geometrical Indices. Volume (V_w) was calculated using the Sybyl⁶⁴ package from Tripos Associates, Inc. The 3-D Wiener numbers were calculated using Sybyl with an SPL (Sybyl Programming Language) program developed in our lab. Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using CONCORD 3.0.1.⁶⁵ Two variants of the 3-D Wiener number were calculated. For ${}^{3D}W_H$, hydrogen atoms are included in the computations, and for ${}^{3D}W$, hydrogen atoms are excluded from the computations.

2.4. Data Reduction. Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices. This scaling was also done for the geometric indices for consistency.

TSIs are topological indices which encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors such as hybridization states of atoms, number of core/valence electrons in individual atoms, etc. TCIs are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. TCIs are derived from weighted molecular graphs where each vertex (atom) is properly weighted with relevant chemical/physical properties. Table 4 shows the breakdown of the topological indices into structural and chemical indices.

The sets of TSIs and TCIs were further divided into subsets, or clusters, based on the correlation matrix by using the SAS procedure VARCLUS.⁶⁶ The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional.

From each cluster we selected the TI most correlated with the cluster as well as any TIs which were poorly correlated with the cluster ($R < 0.70$). These TIs were then used in the modeling of benzamidine-mediated inhibition of guinea pig complement. The variable clustering and selection of TIs was performed independently for both the TSI and TCI sets of indices.

2.5. Statistical Analysis. Regression modeling was

Figure 2. Scatterplot for observed $1/\log C$ versus predicted $1/\log C$ using eq 2 for the set of 107 benzamidines.

Table 4. Classification of Parameters Used in Developing Models for Complement Inhibition

topological	topochemical	geometric
I_D^W	I_{ORB}	V_w
I_D^W	$IC_0 - IC_6$	$3D W$
W	$SIC_0 - SIC_6$	$3D W_H$
I^D	$ClC_0 - ClC_6$	
H^V	$0\chi^b - 6\chi^b$	
H^D	$0\chi^b_C - 6\chi^b_C$	
$\overline{IC}$	$6\chi^b_{Ch}$	
O	$4\chi^b_{PC} - 6\chi^b_{PC}$	
M_1	$0\chi^v - 6\chi^v$	
M_2	$0\chi^v_C - 6\chi^v_C$	
$0\chi - 6\chi$	$6\chi^b_{Ch}$	
$3\chi_C - 6\chi_C$	$4\chi^b_{PC} - 6\chi^b_{PC}$	
$6\chi_{Ch}$	J^B	
$4\chi_{PC} - 6\chi_{PC}$	J^X	
$P_0 - P_{10}$	J^Y	
J		

would be necessary to perform an alternative transformation of the data. Using Hansch and Yoshimoto's Log $1/C$ transformation resulted in residual plots that showed that the variance of the errors correlated with the predictions. To deal with this problem, we back transformed the data to the initial value C and then tried several other transformations, finally settling on $1/\log C$ which resulted in an uncorrelated residual plot. All subsets linear regression was then carried out on three distinct sets of indices: set I—three TSIs; set II—the TSI used in model I and four TCIs; and set III—the TSI retained in model I and the three geometrical indices. The regression analysis resulted in the final selection of TIs for the models.

3. RESULTS

Using only the topostructural class of indices, all-subsets regression resulted in a one parameter model to estimate I_{50} :

This parameter was added to the set of topochemical parameters. Again, all-subsets regression was used to develop a model using this new set of independent variables. The best model for estimation of I_{50} once again used only I^D . This being the case, topochemical parameters were dropped from the modeling procedure.

Using all-subsets regression on the one parameter from eq 1 and the three geometrical parameters resulted in the selection of a different one parameter model:

$$1/\log C = -0.6428 + 0.0490(3D W) \quad (2)$$

$$n = 105, r = 0.943, r_c = 0.940, s = 0.0196, F = 824$$

Compounds 1 and 6 were removed from all models, as they were both strongly influential and were classified as outliers as defined by the studentized range. The predicted values from eq 2 for all 107 benzamidines, including the results predicted for the two outliers, are presented in Table 2.

A scatter plot of the experimental data for the 107 benzamidines versus the values predicted using eq 2 is presented in Figure 2. Predicted values for the two outliers have been included.

4. DISCUSSION

The objective of this paper was to study the relative effectiveness of topostructural, topochemical, and geometrical parameters in estimating the complement inhibitory potency of a set of benzamidines based solely on their chemical structures. Theoretical structural indices can be derived from distinct models of molecules. Also, various indices defined on the same representation of the molecule can quantify various aspects of molecular architecture. Recently, we have advocated the use of a "hierarchical QSAR approach" involving the TSI, TCI, geometrical, and quantum chemical indices in the successful development of predictive models.⁶⁷⁻⁷¹

examining our retransformed results with the results of their best equation.

	<i>n</i>	<i>r</i>	<i>s</i>
Basak et al.	105	0.943	0.264
Hansch and Yoshimoto	108	0.935	0.258

However, the LFER approach used by Hansch and Yoshimoto required experimental data for all compounds in the study and significant input from a human expert for the determination of the three "structural" indicator variables. One strength of our approach to this problem is the use of nonempirical theoretical descriptors which can be calculated solely from the chemical structure. With these purely theoretical descriptors we have modeled the inhibition of complement by benzamidines as successfully as Hansch and Yoshimoto using their LFER approach.

It is clear from this study of 107 benzamidines that the TSI indices are sufficient to explain most of the variance in bioactivity. The addition of TCI and geometrical parameters did not substantially increase the predictive power of the models. However, quantum chemical indices were not used for model development with this set of compounds.

TSIs encode information about generalized size and shape of a molecule. The success of TSI parameters in explaining most of the complement-inhibitory action of these benzamidines indicates that the general shape and size of these molecules largely determines their bioactivity. In some of our other studies we have found that the addition of quantum chemical indices can improve the correlation in cases of specific bioactivity. Further studies will focus on the contribution of quantum chemical indices in explaining the bioactivity of benzamidines.

ACKNOWLEDGMENT

This paper is contribution number 226 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force.

REFERENCES AND NOTES

- (1) Basak, S. C.; Frane, C. M.; Rosen, M. E.; Magnuson, V. R. *Molecular Topology and Acute Toxicity: A QSAR Study of Monoketones*. *Med. Sci. Res.* **1987**, *15*, 887-888.
- (2) Basak, S. C. Binding of Barbiturates to Cytochrome P₄₅₀: A QSAR Study Using Log P and Topological Indices. *Med. Sci. Res.* **1988**, *16*, 281-282.
- (3) Charton, M. In *Steric Effects in Drug Design*; Charton, M., Motoc, M., Eds.; Springer-Verlag: Berlin, 1983; pp 107-118.
- (4) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243-272.
- (5) Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. Prediction of Octanol-Water Partition Coefficient (*K_{ow}*) Using Algorithmically-Derived Variables. *Environ. Toxicol. Chem.* **1992**, *11*, 893-900.
- (6) Balaban, A. T.; Bertelsen, S.; Basak, S. C. New Centric Topological Indexes for Acyclic Molecules (Trees) and Substituents (Rooted Trees) and Coding of Rooted Trees. *Math. Chem.* **1994**, *30*, 55-72.
- (7) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett.* **1995**, *79*, 239-250.

- (9) Basak, S. C.; Grunwald, G. D. In *Proceeding of the XVI International Cancer Congress*, R. S. Rao, M. G. Deo, L. D. Sanghui, Eds.; Monduzzi: Bologna, Italy, 1995; pp 413–416.
- (10) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* 1987, 15, 605–609.
- (11) Basak, S. C.; Niemi, G. J.; Veith, G. D. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; NOVA: New York, 1990; pp 235–277.
- (12) Basak, S. C. In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic: Dordrecht/Boston/London, 1990; pp 83–103.
- (13) Basak, S. C. In *Proceedings of the NATO Advanced Study Institute (ASI) on Pharmacokinetics*; Pub: Sicily, in press.
- (14) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of Graph-Theoretic and Geometrical Molecular Descriptors in Structure–Activity Relationships. In *From Chemical Topology to Three-Dimensional Molecular Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73–116.
- (15) Basak, S. C.; Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal *p*-Hydroxylation of Anilines by Alcohol: A Molecular Similarity Approach. In *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health*; Johnson, B. L., Xintaras, C., Andrews, J. S., Jr., Eds.; Princeton Scientific Publishing Co., Inc.: Princeton, NJ, 1997; pp 492–504.
- (16) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, U.K., 1986.
- (17) Basak, S. C.; Rosen, M. E.; Magnuson, V. R. Molecular Topology and Mutagenicity: A QSAR Study of Nitrosamines. *IRCS Med. Sci.* 1986, 14, 848–849.
- (18) Basak, S. C.; Gieschen, D. P.; Magnuson, V. R.; Harriss, D. K. Structure–Activity Relationships and Pharmacokinetics: A Comparative Study of Hydrophobicity, van der Waals' Volume and Topological Parameters. *IRCS Med. Sci.* 1982, 10, 619–620.
- (19) Basak, S. C.; Grunwald, G. D. Use of Graph Invariants, Volume and Total Surface Area in Predicting Boiling Point of Alkanes. *Mathl. Modelling Sci. Computing* 1993, 2, 735–740.
- (20) Rouvray, D. H.; Pandey, R. B. The Fractal Nature, Graph Invariants and Physicochemical Properties of Normal Alkanes. *J. Chem. Phys.* 1986, 85, 2286–2290.
- (21) Randić, M. Resolution of Ambiguities in Structure–Property Studies by Use of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci.* 1991, 31, 311–320.
- (22) Randić, M. Nonempirical Approaches to Structure–Activity Studies. *Int. J. Quantum Chem: Quantum Biol. Symp.* 1984, 11, 137–153.
- (23) Trinajstić, N. *Chemical Graph Theory*, Klein, D. J., Randić, M., Eds.; CRC Press: Boca Raton, FL, 1992.
- (24) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure–Activity Studies. *J. Chem. Inf. Comput. Sci.* 1994, 34, 270–276.
- (25) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analogue Selection and Property Estimation Using Graph Invariants. *SAR QSAR Environ. Res.* 1994, 2, 289–307.
- (26) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Estimation of Normal Boiling Points of Haloalkanes Using Molecular Similarity. *Croat. Chim. Acta* 1996, 69, 1159–1173.
- (27) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals using Graph-Theoretic Indices. *Discrete Appl. Math.* 1988, 19, 17–44.
- (28) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogues. *Mathl. Modelling Sci. Computing* 1998, in press.
- (29) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity from Structural Similarity. *New J. Chem.* 1995, 19, 231–237.
- (30) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Estimation of Molecular Properties. *J. Chem. Inf. Comput. Sci.* 1995, 35, 366–372.
- (31) Basak, S. C.; Grunwald, G. D. Tolerance Space and Molecular Similarity. *SAR QSAR Environ. Res.* 1995, 3, 265–277.
- (32) Basak, S. C.; Grunwald, G. D. Predicting Mutagenicity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* 1995, 31, 2529–2546.
- (33) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* 1985, 25, 64–73.

- (35) Lajiness, M. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova: New York, 1990; pp 299–316.
- (36) Wilkins, C. L.; Randić, M. A Graph Theoretic Approach to Structure–Property and Structure–Activity Correlations. *Theor. Chim. Acta (Berl.)* 1980, 58, 45–68.
- (37) Basak, S. C.; Gut, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods using Nonempirical Parameters. *Mathl. Modelling Sci. Computing*, in press.
- (38) Fisanick, W.; Cross, K.; Ruzinko, III, A. Similarity Searching on CAS Registry Substances. I. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* 1992, 32, 664–674.
- (39) Randić, M. Similarity Based on Extended Basis Descriptors. *J. Chem. Inf. Comput. Sci.* 1992, 32, 686–692.
- (40) Mekenyan, O.; Peitchev, D.; Bonchev, D.; Trinajstić, N.; Bangov, I. Modelling the Interaction of Small Organic Molecules with Biomacromolecules. *Arzneim. Forsch.* 1986, 36, 176–183.
- (41) Mihlic, Z.; Trinajstić, N. The Algebraic Modelling of Chemical Structures: On the Development of Three-Dimensional Molecular Descriptors. *J. Mol. Struct. (Theochem.)* 1991, 232, 65–78.
- (42) Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. The Utility of Information Content (IC), Structural Information Content (SIC), Hydrophobicity (log P) and van der Waals' Volume (Vw) in the Design of Barbiturates and Tumor-Inhibitory Triazines: A Comparative Study. *Arzneim.-Forsch.* 1983, 33, 352–356.
- (43) Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. A Quantitative Structure–Activity Relationship (QSAR) Study of Barbiturates, Spasmolytics and Diphen-hydramines using van der Waals' Volume. *Acta Ciencia Indica* 1981, 4, 187–192.
- (44) Koyama, M.; Ohtani, N.; Kai, F.; Moriguchi, I.; Inouye, S. Synthesis and Quantitative Structure–Activity Relationship Analysis of N-triiodoallyl and N-iodopropargylazole: New Antifungal Agents. *J. Med. Chem.* 1987, 30, 552–562.
- (45) Lachmann, P. J. In *The Immune System*; Hobart, M. J., McConnell, I., Eds.; Blackwell Scientific Publications: Philadelphia, PA, 1976.
- (46) Kuby, J. *Immunology*. W. H. Freeman & Co.: New York, 1992.
- (47) Hansch, C.; Yoshimoto, M. Structure–Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidines. *J. Med. Chem.* 1974, 17, 1160–1167.
- (48) Baker, B. R.; Erickson, E. H. Irreversible Enzyme Inhibitors. CLII. Proteolytic Enzymes. X. Inhibition of Guinea Pig Complement by Substituted Benzamidines. *J. Med. Chem.* 1969, 12, 408–414.
- (49) Baker, B. R.; Cory, M. Irreversible Enzyme Inhibitors. CLXIV. Proteolytic Enzymes. XIV. Inhibition of Guinea Pig Complement by *meta*-Substituted Benzamidines. *J. Med. Chem.* 1969, 12, 1049–1052.
- (50) Baker, B. R.; Cory, M. Irreversible Enzyme Inhibitors. CLXV. Proteolytic Enzymes. XV. Inhibition of Guinea Pig Complement by Derivatives of *m*-Phenoxypropoxybenzamide. *J. Med. Chem.* 1969, 12, 1053–1056.
- (51) Baker, B. R.; Cory, M. Irreversible Enzyme Inhibitors. 180. Irreversible Inhibitors of the C'1a Component of Complement Derived from *m*-(Phenoxypropoxy) benzamide and Phenoxyacetamide. *J. Med. Chem.* 1971, 14, 119–125.
- (52) Baker, B. R.; Cory, M. Irreversible Enzyme Inhibitors. 186. Irreversible Inhibitors of the C'1a Component of Complement Derived from *m*-(Phenoxypropoxy) benzamide by Bridging to a Terminal Sulfonyl Fluoride. *J. Med. Chem.* 1971, 14, 805–808.
- (53) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY 2.3; Copyright of the University of Minnesota, 1988.

- (54) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
- (55) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- (56) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *J. Comput. Chem.* **1984**, *5*, 581–588.
- (57) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.
- (58) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis: A Quantitative Structure–Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneim. Forsch.* **1983**, *33*, 501–503.
- (59) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the 11nd International Conference on Mathematical Modelling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri–Rolla: Rolla, MO, 1980; Vol. II, pp 851–856.
- (60) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lipais, A. I., Rodin, E. Y., Eds.; Pergamon Press: New York, 1984; pp 745–750.
- (61) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
- (62) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199–206.
- (63) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking into account Periodicities of Element Properties. *Math. Chem. (MATCH)* **1986**, *21*, 115–122.
- (64) Tripos Associates, Inc. *SYBYL Version 6.1*; Tripos Associates, Inc.: St. Louis, MO, 1994.
- (65) Tripos Associates, Inc. *CONCORD Version 3.0.1*; Tripos Associates, Inc.: St. Louis, MO, 1993.
- (66) SAS Institute Inc. In *SAS/STAT User's Guide, Release 6.03 Edition*; SAS Institute Inc.: Cary, NC, 1988; Chapters 28 and 34, pp 773–875, 949–965.
- (67) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054–1060.
- (68) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651–655.
- (69) Gute, B. D.; Basak, S. C. Predicting Acute Toxicity (LC₅₀) of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117–131.
- (70) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenicity of Chemicals, Quantitative Structure–Activity Relationships. In *Quantitative Structure–Activity Relationships in Environmental Sciences*; Chen, F., Schuurman, G., Eds.; SETAC Press: Pensacola, FL, 1997; Vol. 7, Chapter 17, p 245.
- (71) 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.* **1997**, in press.

PREDICTION OF THE DERMAL PENETRATION OF POLYCYCLIC AROMATIC HYDROCARBONS (PAHs): A HIERARCHICAL QSAR APPROACH

B. D. GUTE, G. D. GRUNWALD and S. C. BASAK*

*Natural Resources Research Institute, University of Minnesota,
5013 Miller Trunk Highway, Duluth, MN 55811, USA*

(Received 1 September 1997; In final form 2 March 1998)

Attempts were made to develop hierarchical quantitative structure-activity relationship (QSAR) models for the dermal penetration of polycyclic aromatic hydrocarbons (PAHs) using four classes of theoretical structural parameters; viz., topostructural, topochemical, geometric, and quantum chemical descriptors; and physicochemical properties such as molecular weight (MW) and lipophilicity ($\log P$ – octanol/water). The results show that topostructural, topochemical, and geometric descriptors and molecular weight are equally effective in predicting the dermal penetration of PAHs. Quantum chemical parameters did not make any improvements in the predictive power of the QSAR models.

Keywords: Hierarchical QSAR; topological indices; geometrical indices; quantum chemical parameters; dermal penetration; polycyclic aromatic hydrocarbons

INTRODUCTION

An understanding of the barrier properties of skin is important both for hazard assessment following dermal exposure to toxicants [1] as well as for the transdermal delivery of drugs [2]. Over the years transdermal delivery data on a large number of compounds have been accumulated. These compounds cover a wide range of physicochemical properties and structural types [1]. Attempts have been made to explain permeation behavior of chemicals using specific models of the permeability barrier.

*Corresponding author.

One of the contemporary interests in the field is the prediction of skin permeability from their physicochemical and structural parameters. Potts and Guy [1] and Guy [3] succeeded in predicting the permeability coefficient of diverse chemicals using molecular weight (MW), molar volume (MV) and octanol/water partition coefficient. These parameters quantify size and hydrophobicity of chemicals. Molnar and King used integrated molecular transform, FT_m , as the structural parameter for predicting skin permeability of diverse chemicals [4].

A recent interest in quantitative structure-activity relationship (QSAR) studies is the prediction of toxicological and pharmacological properties of chemicals directly from their structure [5–12]. This is particularly important for the risk assessment of chemicals where the majority of the new chemicals which have little or no available experimental data [13].

Recently we have developed a new hierarchical approach to QSAR using parameters which can be computed directly from molecular structure [14–18]. Such variables include topostructural, topochemical, geometrical and quantum chemical parameters. These parameters quantify size, shape, and stereo-electronic aspects of molecular architecture. In view of the fact that well-known molecular properties like molecular weight, octanol/water partition coefficient, molar volume and calculated molecular descriptors like integrated molecular transform have been used in predicting skin permeability of chemicals, it was of interest to investigate our hierarchical approach in estimating skin permeability. To this end, we have attempted to predict the skin permeability of a set of sixty polycyclic aromatic hydrocarbons using the hierarchical QSAR method.

THEORETICAL METHODS

Database

A data set of sixty polycyclic aromatic hydrocarbons (PAHs) was used for the development of hierarchical QSAR models. The data was taken from the work of Roy *et al.* [19]. Using equimolar concentrations for each compound, dermal penetration (%DP) was determined 24-hours after dosing. Activity was expressed as the percentage of the applied dose (40 nmoles per cm² skin surface) which penetrated the skin. The molecular structures of the PAHs were coded for evaluation using the SMILES line-notation for chemical structure [20]. This data; including compound name, Chemical Abstracts Services (CAS) registry number (when available), and measured dermal penetration; are presented in Table I.

TABLE I Sixty polycyclic aromatic hydrocarbons (PAHs) and their dermal penetration values expressed as percent of biological activity

<i>No.</i>	<i>Compound</i>	<i>CAS No.</i>	<i>Act.</i>	<i>Pred. Act.</i>	<i>Resid.</i>
1	Coronene	191-07-1	0.70	7.18	-6.48
2	Dibenzo(<i>a, l</i>)pyrene	191-30-0	2.00	7.08	-5.08
3	9,10-Diphenylanthracene	1499-10-1	6.00	-0.10	6.10
4	Perylene	198-55-0	7.00	19.68	-12.68
5	Dibenzo(<i>a, i</i>)pyrene	189-55-9	8.00	7.18	0.82
6	3-Methylcholanthrene	56-49-5	8.00	11.89	-3.89
7	Benzylhydrilindene fluorene	1836-87-9	8.00	16.93	-8.93
8	7,10-Dimethylbenzo(<i>a</i>)pyrene	63104-33-6	8.30	11.57	-3.27
9	Indeno(1,2,3- <i>c,d</i>)pyrene	193-39-5	9.00	11.52	-2.52
10	Dibenz(<i>a, h</i>)anthracene	53-70-3	9.40	13.29	-3.89
11	Benzo(<i>e</i>)pyrene	192-97-2	10.00	19.68	-9.68
12	Benzo(<i>g, h, i</i>)perylene	191-24-2	10.00	13.19	-3.19
13	9- <i>p</i> -Tolylfluorene	1815-43-0	10.00	14.97	-4.97
14	6-Ethylchrysene	2732-58-3	10.00	16.51	-6.51
15	9-Cinnamylfluorene	NA	11.00	8.08	2.92
16	6-Methylbenz(<i>a</i>)anthracene	316-14-3	14.00	22.40	-8.40
17	Benzo(<i>k</i>)fluoranthene	207-08-9	14.00	17.99	-3.99
18	Benzo(<i>a</i>)pyrene	50-32-8	15.00	19.79	-4.79
19	1-Ethylpyrene	17088-22-1	18.00	23.43	-5.43
20	1-Methyl-7-isopropylphenanthrene	483-65-8	20.00	21.95	-1.95
21	2- <i>tert</i> -Butylanthracene	18801-00-8	20.00	23.28	-3.28
22	9-Phenylanthracene	602-55-1	20.00	18.78	1.22
23	3-Methylbenzo(<i>c</i>)phenanthrene	56-49-5	20.00	11.89	8.11
24	10-Methylbenz(<i>a</i>)anthracene	2381-15-9	20.00	22.49	-2.49
25	5-Methylbenz(<i>a</i>)anthracene	2319-96-2	20.00	22.40	-2.40
26	9,10-Dihydroanthracene	613-31-0	20.00	37.63	-17.63
27	9-Phenylfluorene	789-24-2	20.00	19.07	0.93
28	1,2,3,6,7,8-Hexahydropyrene	1732-13-4	20.00	22.00	-2.00
29	<i>n</i> -Butylpyrene	35980-18-8	20.00	13.27	6.73
30	5,6-Dihydro-4H-dibenz (<i>a, k, l</i>)anthracene	7198-87-0	20.00	11.09	8.91
31	3-Ethylfluoranthene	20496-16-6	20.00	21.42	-1.42
32	Triphenylene	217-59-4	20.00	26.77	-6.77
33	7,8,9,10-Tetrahydroacephenanthrene	7468-93-1	20.00	22.03	-2.03
34	2,3-Benztriphenylene	215-58-7	20.00	13.19	6.81
35	Benzo(<i>c</i>)phenanthrene	195-19-7	20.00	26.89	-6.89
36	1-Methylpyrene	2381-21-7	22.00	29.76	-7.76
37	3,9-Dimethylbenz(<i>a</i>)anthracene	316-51-8	24.00	18.22	5.78
38	2,3-Benzofluorene	243-17-4	25.00	27.26	-2.26
39	1,2-Benzofluorene	238-84-6	25.00	27.17	-2.17
40	9-Benzylfluorene	1572-46-9	26.00	14.36	11.64
41	9- <i>m</i> -Tolylfluorene	18153-42-9	29.00	14.97	14.03
42	Pyrene	129-00-0	30.00	34.84	-4.84
43	2-Ethylanthracene	52251-71-5	30.00	31.11	-1.11
44	10-Methylbenzo(<i>a</i>)pyrene	63104-32-5	32.00	15.58	16.42
45	1-Methylanthracene	610-48-0	33.00	37.99	-4.99
46	2-Methylfluoranthene	33543-31-6	33.00	27.67	5.33
47	3,6-Dimethylphenanthrene	1576-67-6	33.00	32.78	0.22
48	Benzo(<i>a</i>)anthracene	56-55-3	35.00	27.02	7.98
49	Fluorene	86-73-7	36.00	43.80	-7.80
50	2-Methylphenanthrene	2531-84-2	38.00	37.96	0.04
51	9-Ethylfluorene	2294-82-8	38.00	31.06	6.94

TABLE I (Continued)

<i>No.</i>	<i>Compound</i>	<i>CAS No.</i>	<i>Act.</i>	<i>Pred. Act.</i>	<i>Resid.</i>
52	1-Methylphenanthrene	832-69-9	40.00	37.85	2.15
53	9,10-Dihydrophenanthrene	776-35-2	40.00	37.07	2.93
54	9-Vinyanthracene	2444-68-0	40.00	35.37	4.63
55	Anthracene	120-12-7	42.00	43.66	-1.66
56	Fluoranthene	206-44-0	42.00	32.52	9.48
57	1-Methylfluorene	1730-37-6	49.00	38.16	10.84
58	2-Methylanthracene	613-12-7	50.00	38.11	11.89
59	4H-Cyclopenta(<i>d, e, f</i>)phenanthrene	203-64-5	50.00	36.23	13.77
60	Phenanthrene	85-01-8	50.00	43.50	6.50

Computation of Indices

Five sets of parameters have been used to construct the hierarchical models presented in this study. These sets include topostructural, topochemical, geometric, quantum chemical, and physicochemical descriptors. Topostructural and topochemical indices are subsets of the set of topological indices, and the distinction between these groups will be discussed later. Geometric indices include the three-dimensional Wiener number, both hydrogen-filled and hydrogen-suppressed, and van der Waals volume. The quantum chemical parameters were calculated using four semi-empirical Hamiltonians, and the physicochemical descriptors include calculated log *P* and molecular weight. These physicochemical indices were included since they are commonly used in modeling dermal penetration. The set of indices used in this study are summarized in Table II.

TABLE II Classification of parameters used in developing models for the dermal penetration of polycyclic aromatic hydrocarbons (PAHs)

<i>Topostructural</i>	<i>Topochemical</i>	<i>Geometric</i>	<i>Quantum chemical</i>
I_D^W	I_{ORB}	V_W	E_{HOMO}
T_D^W	IC_0-IC_6	${}^{3D}W$	E_{HOMO1}
W	SIC_0-SIC_6	${}^{3D}W_H$	E_{LUMO}
I^D	CIC_0-CIC_6		E_{LUMO1}
H^V	${}^6X^b, {}^6X^b$		ΔH_f
H^D	${}^3X_c^b$ & ${}^5X_c^b$		μ
$\overline{IC}$	${}^5X_{Ch}^b$ & ${}^6X_{Ch}^b$		
O	${}^4X_{PC}^b, {}^6X_{PC}^b$		
M_1	${}^0X^v, {}^6X^v$		
M_2	${}^3X_c^v$ & ${}^5X_c^v$		
${}^0X, {}^6X$	${}^5X_{Ch}^v$ & ${}^6X_{Ch}^v$		
${}^3X_c, {}^5X_c$	${}^4X_{PC}^v, {}^6X_{PC}^v$		
${}^5X_{Ch}, {}^6X_{Ch}$	${}^jX_{PC}^v, {}^6X_{PC}^v$		
${}^4X_{PC}, {}^6X_{PC}$			
P_0-P_{10}			
J			

Topological Indices

The topological indices used in this study, both the topostructural and the topochemical, have been calculated using POLLY 2.3 [21] and software developed by the authors. These indices include Wiener index [22], connectivity indices developed by Randić [23] and higher order connectivity indices formulated by Kier and Hall [24], bonding connectivity indices defined by Basak *et al.* [25], a set of information theoretic indices defined on the distance matrices of simple molecular graphs [26, 27] and neighborhood complexity indices of hydrogen-filled molecular graphs [28,29], and Balaban's *J* indices [30–32]. Table III provides a list and brief definitions

TABLE III Symbols, definitions and classifications of topological parameters

<i>Topostructural</i>	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
IC	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
hX	Path connectivity index of order $h = 0-6$
hX_C	Cluster connectivity index of order $h = 3-5$
$^hX_{PC}$	Path-cluster connectivity index of order $h = 4-6$
$^hX_{Ch}$	Chain connectivity index of order $h = 5 \text{ \& } 6$
P_h	Number of paths of length $h = 0-10$
J	Balaban's <i>J</i> index based on distance
<i>Topochemical</i>	
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
$^hX^b$	Bond path connectivity index of order $h = 0-6$
$^hX_C^b$	Bond cluster connectivity index of order $h = 3 \text{ \& } 5$
$^hX_{Ch}^b$	Bond chain connectivity index of order $h = 5 \text{ \& } 6$
$^hX_{PC}^b$	Bond path-cluster connectivity index of order $h = 4-6$

TABLE III (Continued)

${}^hX^v$	Valence path connectivity index of order $h = 0-6$
${}^hX^v_C$	Valence cluster connectivity index of order $h = 3 \text{ \& } 5$
${}^hX^v_{Ch}$	Valence chain connectivity index of order $h = 5 \text{ \& } 6$
${}^hX^v_{PC}$	Valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
<i>Geometric</i>	
V_w	van der Waal's volume
${}^{3D}W$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

of the topostructural, topochemical, and geometrical indices included in this study.

The topological indices were divided into two subsets: topostructural and topochemical indices. Topostructural indices (TSIs) are topological indices which only encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs), irrespective of the chemical nature of the atoms involved in bonding or factors such as hybridization states and the number of core/valence electrons in individual atoms. Topochemical indices (TCIs) are parameters that quantify information regarding the topology (connectivity of atoms), as well as specific chemical properties of the atoms comprising a molecule. These indices are derived from weighted molecular graphs where each vertex (atom) or edge (bond) is properly weighted with selected chemical or physical property information. The division of the topological indices into the sets of topostructural and topochemical indices is shown in Tables II and III.

Geometrical Indices

Van der Waals volume, V_w [33–35], was calculated using *Sybyl 6.1* from Tripos Associates, Inc [36]. The 3-D Wiener numbers were calculated by *Sybyl* using an SPL (*Sybyl* Programming Language) program developed in our laboratory [37]. Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using *CONCORD 3.0.1* [38]. Two variants of the 3-D Wiener number were calculated: ${}^{3D}W_H$ and ${}^{3D}W$. For ${}^{3D}W_H$ hydrogen atoms are included in the computations and for ${}^{3D}W$, hydrogen atoms are excluded from the computations.

Quantum Chemical Parameters

Quantum chemical parameters were calculated using four semi-empirical Hamiltonian methods: modified neglect of diatomic overlap version 1 (MNDO), modified neglect of diatomic overlap Austin Model 1 (AM1), modified neglect of diatomic overlap parametric method 3 (PM3), and modified intermediate neglect of differential overlap version 3 (MINDO/3). The following quantum chemical parameters were calculated using each of the above methods: energy of the highest occupied molecular orbital (E_{HOMO}), energy of the second highest occupied molecular orbital ($E_{\text{HOMO}1}$), energy of the lowest unoccupied molecular orbital (E_{LUMO}), energy of the second lowest unoccupied molecular orbital ($E_{\text{LUMO}1}$), heat of formation (ΔH_f), dipole moment (μ), and HOMO/LUMO gap ($E_{\text{HOMO}} - E_{\text{LUMO}}$). These parameters were calculated using *MOPAC 6.00* in the *Sybyl* interface [39].

Physicochemical Descriptors

Molecular weight (MW) was calculated using *Sybyl* 6.1. Molecular weight can be thought of as a descriptor which characterizes the general size of a molecule, especially in the case a specialized set such as the PAHs. Values of $\log P$ were computed by CLOGP [40]. The calculated values of $\log P$ for the set of sixty PAHs range from approximately 4.2 to 8.3 and are presented in Table IV.

Data Reduction

Initially, all topological indices were transformed by the natural logarithm of the index plus one. This was done to scale the indices, since some may be several orders of magnitude greater than others, while other indices may equal zero. The geometric indices were also transformed by the natural logarithm of the index for consistency.

The resulting set of eighty-eight topological indices was then partitioned into two distinct sets, the topostructural indices (thirty-eight) and the topochemical indices (fifty). Further reduction of the number of independent variables available for model construction was still necessary to minimize the chance of spurious correlations. According to the guidelines described by Topliss and Edwards, for a set of sixty observations, approximately thirty-five independent variables can be used in modeling

TABLE IV Calculated values for molecular weight (MW), lipophilicity (log *P*), *P*₀, ¹*X*^b, ^{3D}*W*

<i>No</i>	<i>MW</i>	<i>log P</i>	<i>P</i> ₀	¹ <i>X</i> ^b	^{3D} <i>W</i>
1	300.360	7.044	3.2189	2.1898	7.0226
2	302.376	7.298	3.2189	2.1910	7.1475
3	330.430	8.266	3.2958	2.2821	7.3402
4	252.316	6.124	3.0445	2.0310	6.6000
5	289.357	NA	3.2189	2.1898	6.9618
6	268.359	7.067	3.0910	2.1299	6.8191
7	254.332	5.858	3.0445	2.0660	6.6916
8	280.370	7.422	3.1355	2.1339	6.8771
9	276.338	6.584	3.1355	2.1346	6.8812
10	278.354	6.838	3.1355	2.1122	6.9813
11	252.316	6.124	3.0445	2.0310	6.5945
12	276.338	6.584	3.1355	2.1135	6.8175
13	256.348	6.432	3.0445	2.0909	6.6725
14	256.348	6.842	3.0445	2.0713	6.6775
15	282.386	6.916	3.1355	2.1783	6.9571
16	242.321	6.313	2.9957	1.9965	6.5583
17	252.316	6.124	3.0445	2.0525	6.7022
18	252.316	6.124	3.0445	2.0296	6.6374
19	230.310	6.128	2.9444	1.9835	6.3486
20	234.342	6.716	2.9444	2.0023	6.4635
21	234.342	6.466	2.9444	1.9854	6.4892
22	254.332	6.378	3.0445	2.0425	6.6514
23	242.321	7.067	3.0910	2.1299	6.4636
24	242.321	6.313	2.9957	1.9954	6.5952
25	242.321	6.313	2.9957	1.9965	6.5691
26	180.250	4.674	2.7081	1.8032	5.7671
27	242.321	5.783	2.9957	2.0388	6.5159
28	208.304	5.942	2.8332	2.0016	6.0322
29	258.364	7.186	3.0445	2.1124	6.6998
30	268.359	6.977	3.0910	2.1401	6.7552
31	230.310	6.128	2.9444	2.0090	6.3900
32	228.294	5.664	2.9444	1.9410	6.3516
33	208.304	5.942	2.8332	2.0012	6.0656
34	278.354	6.838	3.1355	2.1135	6.9177
35	228.294	5.664	2.9444	1.9395	6.3531
36	216.283	5.599	2.8904	1.9032	6.1824
37	256.348	6.962	3.0445	2.0496	6.7562
38	216.283	5.399	2.8904	1.9348	6.3157
39	216.283	5.399	2.8904	1.9360	6.2906
40	256.348	6.312	3.0445	2.0986	6.5775
41	256.348	6.432	3.0445	2.0909	6.6494
42	202.256	4.950	2.8332	1.8386	6.0130
43	206.288	5.668	2.8332	1.8860	6.1723
44	266.343	6.773	3.0910	2.0831	6.7519
45	192.261	5.139	2.7726	1.7986	5.9393
46	216.283	5.599	2.8904	1.9296	6.2290
47	206.288	5.788	2.8332	1.8647	6.1080
48	228.294	5.664	2.9444	1.9379	6.4313
49	166.223	4.225	2.6391	1.7249	5.5620
50	192.261	5.139	2.7726	1.7991	5.9358
51	194.277	5.273	2.7726	1.8866	5.8875
52	192.261	5.139	2.7726	1.8004	5.9104
53	180.250	4.784	2.7081	1.8103	5.7372
54	204.272	5.214	2.8332	1.8319	6.0757

TABLE IV (Continued)

<i>No</i>	<i>MW</i>	<i>log P</i>	<i>P₀</i>	<i>¹X^b</i>	<i>^{3D}W</i>
55	178.234	4.490	2.7081	1.7267	5.7650
56	202.256	4.950	2.8332	1.8681	6.0547
57	180.250	4.874	2.7081	1.7964	5.7613
58	192.261	5.139	2.7726	1.7971	5.9752
59	190.245	4.685	2.7726	1.8210	5.8489
60	178.234	4.490	2.7081	1.7286	5.7224

while retaining a low probability of chance correlations ($P_c < 0.01$ with $R^2 \geq 0.7$) [41].

To further reduce the number of indices available, the sets of topostructural and topochemical indices were divided into subsets, or clusters, based on the correlation matrices using the SAS procedure VARCLUS [42]. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional.

From each cluster we selected the index most correlated with the cluster, as well as any indices which were poorly correlated with their cluster ($R^2 < 0.70$). These indices were then used in the modeling of the dermal penetration of the sixty PAHs. The variable clustering and selection of indices was performed independently on both the topostructural and topochemical sets of indices. This procedure resulted in a set of eight topostructural indices and nine topochemical indices.

Statistical Analysis and Hierarchical QSAR

Regression modeling of the thirteen distinct sets of indices was accomplished using the SAS procedure REG [42]. This hierarchical approach to QSAR modeling begins with the simplest parameters, the TSIs. Increasingly complex levels of parameters are then added. The indices from the best TSI model are retained and the set of TCIs are added. The indices included in the best model from this second step are then combined with the geometric indices and regression modeling is conducted again. The quantum chemical parameters from the various Hamiltonians are treated as unique sets of descriptors and are individually modeled with the other parameters, *e.g.*, the AM1 and PM3 indices are never used in the same model. The physicochemical descriptors were included in each step of the modeling process to determine how they compare with the theoretical descriptors.

In addition to the seven models developed using the hierarchical approach, seven other models were generated. These models used the

individual sets of descriptors only to determine the potential contribution of each set. Thus these models were generated using TCI indices only, geometric indices only, quantum chemical indices only, or physicochemical indices only.

RESULTS

The variable clustering of the TSIs resulted in the selection of eight indices: $\overline{IC}$, O , 3X - 5X , ${}^6X_{Ch}$, P_0 , P_3 . $\log P$ and MW were added to the set of independent variables, for this model and all subsequent models, because other studies have shown the importance of these parameters in predicting dermal penetration [1, 19]. All-possible subsets regression resulted in the selection of the following one-parameter model for the estimation of dermal penetration:

$$\begin{aligned} \%DP &= 224.1 - 67.9P_0 \\ n &= 60 \quad r^2 = 0.675 \quad s = 7.4 \quad F = 120.6 \end{aligned} \quad (1)$$

In the next step of the hierarchy, the nine TCIs selected by variable clustering (IC_0 , SIC_2 , SIC_4 , CIC_1 , ${}^1X^b$, ${}^6X^b_{Ch}$, ${}^4X^v$, ${}^5X^v_C$, J^B) were combined with P_0 , $\log P$, and MW and all-subsets regression was conducted on this set. The following model resulted:

$$\begin{aligned} \%DP &= 179.7 - 78.8{}^1X^b \\ n &= 60 \quad r^2 = 0.695 \quad s = 7.1 \quad F = 132.0 \end{aligned} \quad (2)$$

Interestingly, neither the topostructural index from the first model or either of our physicochemical descriptors were selected. Neither the geometrical nor any of the quantum chemical indices added significantly to the model produced in the second step of the hierarchy. In all cases, ${}^1X^b$ produced the best model.

To continue our comparative study of the indices, models were constructed using only geometric indices, only quantum chemical indices, and only physicochemical parameters. The use of geometric parameters alone resulted in a one-parameter model which performed as well as the TSI model:

$$\begin{aligned} \%DP &= 186.0 - 25.4{}^3PW \\ n &= 60 \quad r^2 = 0.673 \quad s = 7.4 \quad F = 119.3 \end{aligned} \quad (3)$$

The models using only quantum chemical indices were all discarded since none resulted in an explained variance (r^2) greater than 25%.

Finally, modeling was conducted using $\log P$ and MW. Molecular weight proved to be a better descriptor for modeling the dermal penetration of PAHs than was $\log P$. This step resulted in the following one-parameter model:

$$\begin{aligned}\%DP &= 90.6 - 0.3MW \\ n &= 60 \quad r^2 = 0.674 \quad s = 7.4 \quad F = 120.0\end{aligned}\tag{4}$$

The values for the parameters used in the final models (P_0 , $^1X^b$, 3DW , MW) have been provided in Table IV.

DISCUSSION

The goal of this paper was to develop models for estimating the dermal penetration of chemicals using computed molecular descriptors. To this end we used topostructural, topochemical, geometric, and quantum chemical parameters which can be computed directly from the molecular structure. We also used calculated $\log P$ (CLOGP) and molecular weight as descriptors in the development of regression equations.

Our results show that topostructural indices (P_0), topochemical parameters ($^1X^b$), geometrical descriptors (3DW) and physicochemical properties (MW) are almost equally effective in predicting the dermal penetration of the sixty PAHs studied in this paper. Additionally, we attempted to develop hierarchical QSAR models by adding selected topochemical, geometric, and quantum chemical indices to the set of topostructural parameters retained by the variable clustering method. This procedure did not result in any improvement in the models. Interestingly, $\log P$ and the quantum chemical descriptors gave QSAR models which were inferior to the predictive equations generated from topostructural, topochemical or geometric variables.

Of the four final models which were generated as part of this study, $^1X^b$, a simple bond-type connectivity index which accounts for general size and bonding patterns within the molecule, provided the best correlation with percent dermal penetration. Figure 1 shows the correlation between experimental dermal penetration and estimated dermal penetration using $^1X^b$ and Figure 2 demonstrates the scatter of the residuals. Thus, there are no apparent co-variance problems within this model.

FIGURE 1 Scatterplot of experimentally determined percent dermal penetration (%DP) vs. estimated %DP using Eq. (2) for a set of 60 polycyclic aromatic hydrocarbons.

FIGURE 2 Pattern of residual errors for the estimation of the percent dermal penetration (%DP) of 60 polycyclic aromatic hydrocarbons using Eq. (2).

QSAR models developed in this study are in line with other published models for dermal penetration of chemicals. Potts and Guy [1] developed models for dermal penetration of diverse chemicals using MW, molar

volume (MV) and $\log P$. Roy *et al.*, developed dermal penetration models for the same set of sixty PAHs analyzed in this paper [19] using $\log P$ and several molecular shape descriptors in the development of regression models ($r^2 = 64\%$). The parameters used by these authors quantify generalized shape, size, and hydrophobicity of chemicals, so it is not surprising that parameters such as P_0 , $^1X^b$, $^3D^bW$, and MW are well correlated with the dermal penetration of PAHs since these parameters also quantify general aspects of the size and shape of molecules.

Based on the results of this study, it seems that physical size and shape are more important in determining the dermal penetration of PAHs than lipophilicity. This conclusion would support the notion that larger molecules must traverse water-filled pores rather than moving across the dermal membrane. This would also account for the findings of Roy *et al.* [19] which showed an inverse relationship between the lipophilicity of PAHs and their dermal penetration. The more lipophilic the compound, the less likely it is to travel through a hydrophilic channel. Additionally, it should be noted that while these results are on par with similar studies, they also demonstrate that there is still something missing in this characterization of the dermal penetration of PAHs.

Acknowledgments

This is contribution number 215 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force and by the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

References

- [1] Potts, R. O. and Guy, R. H. (1992). Predicting skin permeability. *Pharm. Res.*, 9, 663–669.
- [2] Hirvonen, J., Rajala, R., Vihervaara, P., Laine, E., Paronen, P. and Urtti, A. (1994). Mechanism and reversibility of penetration enhancer action in the skin – a DSC study. *Eur. J. Pharm. Biopharm.*, 40, 81–85.
- [3] Guy, R. H. (1995). Percutaneous absorption: Physical chemistry meets the skin. *Curr. Prob. Dermatol.*, 22, 132–138.
- [4] Molnar, S. P. and King, J. W. (1996). Correlation of dermal transport with structure via the integrated molecular transform. *Int. J. Quantum Chem., Quantum Biol. Symp.*, 23, 1845–1849.
- [5] Basak, S. C., Grunwald, G. D. and Niemi, G. J. (1997). Use of graph-theoretic and geometrical molecular descriptors in structure-activity relationships. In, *From Chemical*

- Topology to Three-Dimensional Geometry* (A. T. Balaban, Ed.). Plenum Press, New York, pp. 73–116.
- [6] Basak, S. C., Niemi, G. J. and Veith, G. D. (1990). Optimal characterization of structure for prediction of properties. *J. Math. Chem.*, **4**, 185–205.
- [7] Basak, S. C., Niemi, G. J. and Veith, G. D. (1990). Recent developments in the characterization of chemical structure using graph-theoretic indices. In, *Computational Chemical Graph Theory and Combinatorics* (D. H. Rouvray, Ed.). Nova, New York, pp. 235–277.
- [8] Kier, L. B. and Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, Hertfordshire, U.K. p. 262.
- [9] Basak, S. C. and Grunwald, G. D. (1993). Use of graph invariants, volume and total surface area in predicting boiling point of alkanes. *Math. Model. and Sci. Comput.*, **2**, 735–740.
- [10] Basak, S. C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.*, **15**, 605–609.
- [11] Randić, M. (1997). On characterization of molecular structure. *J. Chem. Inf. Comput. Sci.*, **37**, 672–687.
- [12] Balaban, A. T., Basak, S. C., Colburn, T. and Grunwald, G. D. (1994). Correlation between structure and normal boiling points of haloalkanes C₁–C₄ using neural networks. *J. Chem. Inf. Comput. Sci.*, **34**, 1118–1121.
- [13] Auer, C. M., Nabholz, J. V. and Baetcke, K. P. (1990). 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.*, **87**, 183–197.
- [14] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1996). A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.*, **36**, 1054–1060.
- [15] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1998). The relative effectiveness of topological, geometrical and quantum chemical parameters in estimating mutagenicity of chemicals. In, *QSAR in Environmental Sciences–VII* (F. Chen, *et al.*, Eds.). SETAC Press, Pensacola, Florida, pp. 245–261.
- [16] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical approach. *J. Chem. Inf. Comput. Sci.*, **37**, 651–655.
- [17] Gute, B. D. and Basak, S. C. (1997). Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. *SAR QSAR Environ. Res.*, **7**, 117–131.
- [18] Basak, S. C. and Gute, B. D. (1997). Characterization of molecular structures using topological indices. *SAR QSAR Environ. Res.*, **7**, 1–21.
- [19] Roy, T. A., Neil, W., Yang, J. J., Krueger, A. J., Arroyo, A. M. and Mackerer, C. R. (1998). SAR models for estimating the percutaneous absorption of polynuclear aromatic hydrocarbons. *SAR QSAR Environ. Res.*, **9**, 171–185.
- [20] Anderson, E., Veith, G. D. and Weininger, D. (1987). SMILES: a line notation and computerized interpreter for chemical structures. Environmental Research Brief, EPA/600/M-87/021.
- [21] Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1988). POLLY 2.3: Copyright of the University of Minnesota.
- [22] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.
- [23] Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609–6615.
- [24] Kier, L. B. and Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, Hertfordshire, U.K. p. 262.
- [25] Basak, S. C. and Magnuson, V. R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.*, **19**, 17–44.
- [26] Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B. and Basak, S. C. (1984). Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.*, **5**, 581–588.

- [27] Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **67**, 4517–4533.
- [28] Basak, S. C., Roy, A. B. and Ghosh, J. J. (1980). Study of the structure-function relationship of pharmacological and toxicological agents using information theory. In, *Proceedings of the Second International Conference on Mathematical Modelling* (X. J. R. Avula, R. Bellman, Y. L. Luke and A. K. Rigler, Eds.). University of Missouri–Rolla, pp. 851–856.
- [29] Roy, A. B., Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications. In, *Mathematical Modelling in Science and Technology* (X. J. R. Avula, R. E. Kalman, A. I. Lapis and E. Y. Rodin, Eds.). Pergamon Press, New York, pp. 745–750.
- [30] Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399–404.
- [31] Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.*, **55**, 199–206.
- [32] Balaban, A. T. (1986). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)*, **21**, 115–122.
- [33] Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441–451.
- [34] Moriguchi, I. and Kanada, Y. (1977). Use of van der Waals volume in structure-activity studies. *Chem. Pharm. Bull.*, **25**, 926–935.
- [35] Moriguchi, I., Kanada, Y. and Komatsu, K. (1976). Van der Waals volume and the related parameters for hydrophobicity in structure-activity studies. *Chem. Pharm. Bull.*, **24**, 1799–1806.
- [36] SYBYL Version 6.1. (1994). Tripos Associates, Inc.: St. Louis, MO.
- [37] Mekenyan, O., Peitchev, D., Bonchev, D., Trinajstić, N. and Bangov, I. (1986). Modelling the interaction of small organic molecules with biomacromolecules. I. Interaction of substituted pyridines with anti-3-azopyridine antibody. *Arzneim.-Forsch./Drug Research*, **36**, 176–183.
- [38] CONCORD Version 3.0.1. (1993). Tripos Associates, Inc.: St. Louis, MO.
- [39] Stewart, J. J. P. (1990). MOPAC Version 6.00. QCPE #455. Frank J Seiler Research Laboratory: US Air Force Academy, CO.
- [40] Leo, A. and Weininger, D. (1984). *CLOGP Version 3.2 User Reference Manual*. Medicinal Chemistry Project, Pomona College, Claremont, CA.
- [41] Topliss, J. G. and Edwards, R. P. (1979). Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.*, **22**, 1238–1244.
- [42] SAS Institute Inc. (1988). In: *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc.: Cary, NC.

Use of Statistical and Neural Net Methods in Predicting Toxicity of Chemicals: A Hierarchical QSAR Approach

Subhash C. Basak

Brian D. Gute

Gregory D. Grunwald

Natural Resources Research Inst.

University of Minnesota

Duluth, MN 55811 (USA)

{sbasak, bgute, ggrunwal}@wyle.nrrri.umn.edu

David W. Opitz

Dept. of Computer Science

University of Montana

Missoula, MT 59812 (USA)

opitz@cs.umt.edu

Krishnan Balasubramanian

Dept. of Chem. and Biochem.

Arizona State University

Physical Sciences Bldg., D-106

Tempe, AZ 85287-1604 (USA)

KBalu@asu.edu

Abstract

A contemporary trend in computational toxicology is the prediction of toxicity endpoints and toxic modes of action of chemicals from parameters that can be calculated directly from their molecular structure. Topological, geometrical, substructural, and quantum chemical parameters fall into this category. We have been involved in the development of a new hierarchical quantitative structure-activity relationship (QSAR) approach in predicting physicochemical, biomedical and toxicological properties of various sets of chemicals. This approach uses increasingly more complex molecular descriptors for model building in a graduated manner. In this paper we will apply statistical and neural net methods in the development of QSAR models for predicting toxicity of chemicals using topostructural, topochemical, geometrical, and quantum chemical indices. The utility and limitations of the approach will be discussed.

Introduction

In 1998 the number of chemicals registered with the Chemical Abstract Service (CAS) rose to over 19 million (CAS 1999). This is an increase of over 3 million chemicals between 1996 and 1998. It would certainly be desirable to be able to test each of these chemicals for their effects on the environment and human health (which we refer to as *hazard assessment*); however, completing the battery of tests necessary for the proper hazard assessment of even a single compound is a costly and time-consuming process. Therefore, there is simply not enough time or money to complete these test batteries for even a tiny portion of the compounds which are registered today (Menzel 1995). An alternative to these traditional test batteries is to develop computational models for hazard assessment. Computational models are fast (milliseconds per compound), cheap (less than one cent per compound), and do not run the risk of adversely affecting the environment during testing. Thus computational models can easily process *all* registered chemicals and flag the ones that require further testing. The central problem with this approach is developing class specific models that can be considered accurate

enough to be useful. In this paper, we present computational models for hazard assessment that are indeed considered both accurate and useful.

One of the fundamental principles of biochemistry is that activity is dictated by structure (Hansch 1976). Following this principle, one can use theoretical molecular descriptors that quantify structural aspects of a molecule to quantitatively determine its activity (Basak & Grunwald 1995; Cramer, Famini, & Lowrey 1993). These theoretical descriptors can be generated directly from the known structure of the molecule and used to estimate its properties, without the need for further experimental data. This is important due to the fact that, with chemicals needing to be evaluated for hazard assessment, there is a scarcity of available experimental data that is normally required as inputs (i.e., independent variables) to traditional quantitative structure-activity relationship (QSAR) model development. A QSAR model based solely on theoretical descriptors on the other hand can process all registered chemicals for hazard assessment. Our recent studies show that hierarchical QSARs (H-QSAR) using theoretical structural descriptors give reasonable models for predicting toxicity (Basak, Gute, & Grunwald In press; Gute & Basak 1997; Gute, Grunwald, & Basak In press).

One potential problem with using our hierarchical approach is that it often gives many independent variables as compared to data points. For instance, in our case study of predicting acute toxicity (LC_{50}) of benzene derivatives, we have 95 independent variables and 69 data points. Therefore, reducing the number of independent variables is critical when attempting to model small data sets. The smaller the data set, the greater the chance of spurious error when using a large number of independent variables (descriptors). Part of our focus in this paper is attempting to reduce the size of the data set.

Hierarchical QSAR

Our recent studies have focused on the role of different classes of theoretical descriptors of in-

els of complexity and their utility in QSAR (Gute & Basak 1997; Gute, Grunwald, & Basak In press). Four distinct sets of theoretical descriptors have been used in this study: topostructural, topochemical, geometric, and quantum chemical indices. Gute and Basak 1997 provide the detailed list of the indices included in our study.

Topological Indices

The complete set of topological indices used in this study, both the topostructural and the topochemical, have been calculated using POLLY 2.3 (Basak, Harriss, & Magnuson 1988) and software developed by the authors. These indices include the Wiener index (Wiener 1947), the connectivity indices developed by Randic 1975 and higher order connectivity indices formulated by Kier and Hall 1986, bonding connectivity indices defined by Basak and Magnuson 1988, a set of information theoretic indices defined on the distance matrices of simple molecular graphs (Hansch & Leo 1995), and neighborhood complexity indices of hydrogen-filled molecular graphs, and Balaban's 1983 J indices.

Geometrical Indices

The geometrical indices are three-dimensional Wiener numbers for hydrogen-filled molecular structure, hydrogen-suppressed molecular structure, and van der Waals volume. Van der Waals volume, V_W (Bondi 1964), was calculated using Sybyl 6.1 from Tripos Associates, Inc. of St. Louis. The 3-D Wiener numbers were calculated by Sybyl using an SPL (Sybyl Programming Language) program developed in our lab (SYBYL 1998). Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using CONCORD 3.0.1 from Tripos Associates, Inc. Two variants of the 3-D Wiener number were calculated: 3DW_H and 3DW . For 3DW_H , hydrogen atoms are included in the computations and for 3DW hydrogen atoms are excluded from the computations.

Quantum Chemical Parameters

The following quantum chemical parameters were calculated using the Austin Model version one (AM1) semi-empirical Hamiltonian: energy of the highest occupied molecular orbital (E_{HOMO}), energy of the second highest occupied molecular orbital (E_{HOMO1}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), energy of the second lowest unoccupied molecular orbital (E_{LUMO1}), heat of formation (ΔH_f), and dipole moment (μ). These parameters were calculated using MOPAC 6.00 in the SYBYL interface (Stewart 1990).

Results

We tested the utility of our approach of generating numerous hierarchical theoretical descriptors of com-

pounds on the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. The data was taken from the work of Hall, Kier and Phipps 1984 where acute aquatic toxicity was measured in fathead minnow (*Pimephales promelas*). Their data was compiled from eight other sources, as well as some original work which was conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota. This set of chemicals was composed of benzene and 68 substituted benzene derivatives. According to the authors, these benzene derivatives were tested using methodologies comparable to their own 96-hour fathead minnow toxicity test system. The derivatives chosen for this study have seven different substituent groups that are present in at least six of the molecules. These groups consist of chloro, bromo, nitro, methyl, methoxyl, hydroxyl, and amino substituents.

We studied two classes of approaches for modeling toxicity: (1) giving all the descriptors to a learning algorithm (neural networks in this case), and (2) reducing the feature set before giving the (reduced) feature set to a learning algorithm. Results for our approaches are from leave-one-out experiments (i.e., 69 training/test set partitions). Leave-one-out works by leaving one data point out of the training set and giving the remaining instances (68 in this case) to the learning algorithms for training. (It is worth noting that each member of the ensemble sees the same 68 training instances for each training/test set partition and thus ensembles have no unfair advantage over other learners.) This process is repeated 69 times so that each example is a part of the test set once and only once. Leave-one-out tests *generalization* accuracy of a learner, whereas training set accuracy tests only the learner's ability to memorize. Generalization error from the test set is the true test of accuracy and is what we report here.

Table 1 gives our results. First we trained neural networks using all 95 parameters. The networks contained 15 hidden units and we trained the networks for 1000 epochs. We normalized each input parameter to a values between 0 and 1 before training. Additional parameter settings for the neural networks included a learning rate of 0.05, a momentum term of 0.1, and weights initialized randomly between -0.25 and 0.25. With these ninety-five parameters, the neural network obtained a test-set correlation coefficient between predicted toxicity and measured toxicity (explained variance) of $R^2 = 0.868$ and a standard error of 0.29. Target toxicity measurements ranged from 3.04 to 6.37.

For our next experiments, the VARCLUS method of SAS 1998 was used for selecting subsets of topostructural and topochemical parameters for QSAR model development. With this method, the set of topological indices is first partitioned into two distinct sets, the topostructural indices and the topochemical indices. To further reduce the number of independent variables for model construction, the sets of topostructural and topochemical indices were further divided into subsets,

Table 1: Relative effectiveness of statistical and neural network methods in estimating LC_{50} of 69 benzene derivatives.

Method	R^2	Standard Error
Linear regression	0.825	0.32
NN with 95 inputs	0.868	0.29
NN with VARCLUS	0.878	0.28

or clusters, based on the correlation matrix using the VARCLUS procedure. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. From each cluster we selected the index most correlated with the cluster, as well as any indices which were poorly correlated with their cluster ($R^2 < 0.70$). The variable clustering and selection of indices was performed independently for both the topostructural and topochemical indices. This procedure resulted in a set of five topostructural indices and a set of nine topochemical indices. These indices were combined with the three geometric and six quantum chemical parameters described earlier.

The linear regression approach was that described earlier by Gute and Basak 1997. This study found that an accurate linear regression model for acute aquatic toxicity required descriptors from all four levels of the hierarchy: topostructural, topochemical, geometrical and quantum chemical. This model utilized seven descriptors and obtained an explained variance (R^2) of 0.863 and a standard error of 0.30. A leave-one-out approach was then implemented to test the predictivity of the model. This testing resulted in a model with an $R^2 = 0.825$ and a standard error of 0.32.

We also trained neural networks using the 23 parameters provided by this data reduction technique. The parameter settings for these networks were the same as the settings for the other neural network experiments mentioned above. With these 23 parameters, the neural networks obtained a test-set explained variance (R^2) of 0.878 and a standard error of 0.28. Thus the inputs selected by our data reduction procedure were able to increase the accuracy of the neural network.

Discussion and Future Work

The results show that both statistical and neural network methods give acceptable estimates of toxicity. The neural network methods produced improvement over the statistical model. While the method proposed here has proven effective, there is much future work that needs to be completed. For example, though our results demonstrate that our method is able to accurately predict toxicity directly from structure, it would be interesting to know just how many compounds are needed to learn an accurate model of toxicity. Future work, then, is to empirically answer this question. We plan to run our techniques on further reduced data sets and plot leave-one-out accuracy. This would allow one to look

at a curve that plots accuracy versus training set size and decide how many compounds need to be explicitly tested for toxicity.

In the machine learning literature, the process of finding and removing the variables that are unhelpful or destructive to learning is called feature selection (Kohavi & John 1997). Previous work on feature selection has focused on finding the appropriate subset of relevant features to be used in constructing *one* inference model, such as our approach presented in this paper; however, it is appropriate to start considering feature selection with regards to ensembles. An "ensemble" is a combination of the outputs from a *set* of models that are generated from separately trained inductive learning algorithms. Ensembles have been shown, in most cases, to greatly improve generalization accuracy over a single learning model (Breiman 1996a; Maclin & Opitz 1997; Opitz & Shavlik 1996b; Shapire *et al.* 1997). Recent research has shown that an effective ensemble should consist of a set of models that are not only highly correct, but ones that make their errors on different parts of the input space as well (Hansen & Salamon 1990; Krogh & Vedelsby 1995; Opitz & Shavlik 1996a).

Varying the feature subsets used by each member of the ensemble helps promote the necessary diversity and create a more effective ensemble (Opitz submitted). Thus, this concept is particularly appropriate for large feature sets of partially correlated inputs, such as found in hazard assessment of compounds. Ensemble feature selection algorithms, then, not only have the traditional feature-selection criteria of needing to find feature subsets that are germane to the particular task and inductive-learning algorithm, but have the additional criterion of finding a *set* of features subsets that will promote disagreement among the component members of the ensemble.

The ensemble techniques we plan to test are analogous to the popular and successful ensemble approach Bagging (Breiman 1996b). Bagging is a statistical "boot-strap" (Efron & Tibshirani 1993) ensemble method that creates individuals for its ensemble by training each predictor on a random redistribution of the training set. Each predictor's training set is generated by randomly drawing, with replacement, N examples – where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual predictor in the ensemble is generated with a different random sampling of the training set. Breiman 1996a showed that Bagging is effective on "unstable" learning algorithms where small changes in the training set result in large changes in predictions. This shows that, on average, more diversity is created among the predictors by varying our training set in this manner than is lost in individual predictor accuracy by not training each predictor on the whole data set.

Bagging is not appropriate for most toxicity domains since they are data poor and one cannot afford to waste training examples; however, these domains are feature

rich and thus we can attempt to create diversity by instead varying the inputs to the learning algorithms. Thus we plan to test the approach where each predictor's feature set is generated by randomly drawing, with replacement, N features – where N is the size of the original feature set.

Acknowledgements

This work was partially supported by National Science Foundation grant IRI-9734419, a University of Montana MONTS grant, and U.S. Air Force grants F49620-94-1-0401 and F49620-96-1-0330.

References

- Balaban, A. 1983. Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* 55:199–206.
- Basak, S., and Grunwald, G. 1995. Estimation of lipophilicity from molecular structural similarity. *New Journal of Chemistry* 19:231–237.
- Basak, S., and Magnuson, V. 1988. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* 19:17–44.
- Basak, S.; Gute, B.; and Grunwald, G. In press. A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters. In Devillers, J., and Balaban, A., eds., *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach.
- Basak, S.; Harriss, D.; and Magnuson, V. 1988. Polly 2.3. Copyright of the University of Minnesota.
- Bondi, A. 1964. Van der waals volumes and radii. *J. Phys. Chem.* 68:441–451.
- Breiman, L. 1996a. Bagging predictors. *Machine Learning* 24(2):123–140.
- Breiman, L. 1996b. Stacked regressions. *Machine Learning* 24(1):49–64.
- CAS. 1999. The latest cas registry number and substance count. <http://www.cas.org/cgi-bin/regreport.pl>.
- Cramer, C.; Famini, G.; and Lowrey, A. 1993. Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationships. *Acc. Chemical Research* 26:599–605.
- Efron, B., and Tibshirani, R. 1993. *An introduction to the Bootstrap*. New York: Chapman and Hall.
- Gute, B., and Basak, S. 1997. Predicting acute toxicity (LC50) of benzen derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR and QSAR in Environmental Research* 7:117–131.
- Gute, B.; Grunwald, G.; and Basak, S. In press. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. In *SAR and QSAR in Environmental Research*.

Hall, L.; Kier, L.; and Phipps, G. 1984. Structure-activity relationship studies on the toxicities of benzene derivatives: I. an additivity model. *Environ. Toxicol. Chem.* 3:355–365.

Hansch, C., and Leo, A. 1995. Exploring QSAR: Fundamentals and applications in chemistry and biology. *American Chemical Society* 557.

Hansch, C. 1976. On the structure of medicinal chemistry. *Journal of Medicinal Chemistry* 19:1–6.

Hansen, L., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:993–1001.

Kier, L., and Hall, L. 1986. *Molecular Connectivity in Structure-Activity Analysis*. Hertfordshire, UK: Research Studies Press.

Kohavi, F., and John, G. 1997. Wrappers for feature subset selection. *Artificial Intelligence*.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, volume 7, 231–238.

Maclin, R., and Opitz, D. 1997. An empirical evaluation of bagging and boosting. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 546–551.

Menzel, D. 1995. Extrapolating the future: research trends in modeling. *Toxicology Letters* 79:299–303.

Opitz, D., and Shavlik, J. 1996a. Actively searching for an effective neural-network ensemble. *Connection Science* 8(3/4):337–353.

Opitz, D., and Shavlik, J. 1996b. Generating accurate and diverse members of a neural-network ensemble. In Touretsky, D.; Mozer, M.; and Hasselmo, M., eds., *Advances in Neural Information Processing Systems*, volume 8. Cambridge, MA: MIT Press.

Opitz, D. submitted. Feature selection for ensembles. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

Randic, M. 1975. On characterization of molecular branching. *Journal of American Chemical Society* 97:6609–6615.

SAS. 1998. Cary, NC: SAS Institute Inc. chapter SAS/STAT User's Guide, Release 6.03 Edition.

Shapire, R.; Freund, Y.; Bartlett, P.; and Lee, W. 1997. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 322–330. Nashville, TN: Morgan Kaufmann.

Stewart, J. 1990. Mopac version 6.00. qcpe #455. US Air Force Academy, CO: Frank J. Seiler Research Laboratory.

SYBYL. 1998. Sybyl version 6.1. Tripos Associates, Inc.

Wiener, H. 1947. Structural determination of paraffin boiling points. *Journal of Am. Chem. Soc.* 69:17–20.