REPORT DOCUMENTATION PAGE

ved
-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE 20 Jan 2000	3. REPORT TYPE AND DATES COVERED FINAL 01 JUL 96 - 31 DEC 99

4. TITLE AND SUBTITLE
Quantitative Characterization of Molecular Similarity Spaces: Tools for Computational Toxicology

5. FUNDING NUMBERS
F49620-96-1-0330

6. AUTHOR(S)
Subnash C. Basak

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Natural Resources Research, Institute
University of Minnesota, Duluth
5013 Miller Trunk Highway
Deluth, MN 55811

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
801 N Randolph St. Rm 732
Arlington VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

During the course of the project, most of the work focused on the first three tasks of the project; a) characterization of molecular similarity spaces, b) selection of analogs, and c) similarity-based estimation of properties; has continued. However, in the last approximately eighteen months the focus shifted to the fourth and final task of the project - the application of neural networks in property estimation.

14. SUBJECT TERMS risk assessment, molecular similarity spaces,	15. NUMBER OF PAGES 137
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UU	18. SECURITY CLASSIFICATION OF THIS PAGE UU	19. SECURITY CLASSIFICATION OF ABSTRACT UU	20. LIMITATION OF ABSTRACT

Final Technical Report
of the AFOSR
AASERT Project

Covering research period 7/1/96 to 12/31/99

Agency No: DOD/F49620-96-1-0330
U of M No: 9606754
NRRI Technical Report No: NRRI/TR-2000/02

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@nrri.umn.edu

TABLE OF CONTENTS

OBJECTIVES

The major aims of the proposed project were: a) development of quantitative methods for the characterization of structure spaces, b) application of these newly developed methods in selecting analogs, c) development of estimation methods for predicting toxicologically relevant properties of chemicals from their analogs, and d) development of neural network methods for property estimation and analog selection.

STATUS OF EFFORT

During the course of the project, most of our work focused 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, in the last approximately eighteen months the focus shifted to the fourth and final task of the project – 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 technique known as genetic ensemble feature selection (GEFS) which is based on genetic algorithms. The results show that neural networks, in general, give some improvement in modeling power over statistical methods. However, 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 physicochemical property data and *ab initio* quantum chemical parameters when available. 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, principal components analysis (PCA) and variable clustering (VC). The results of these studies has been reported in two peer-reviewed manuscripts (Pub. # 2-4, See Publications below) and in chapters in two books: a volume of the *Discrete Mathematics and Theoretical Computer Science* series (Pub. # 5) and volume 2 of the *Advances in Molecular Similarity* series (Pub. #6).

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).

Recently we have created several similarity spaces for the identified constituents of JP-8. Three similarity spaces were constructed using a variety of descriptors: topological indices, atom pairs, and physicochemical descriptors. This information was recently reported at the Air Force Office of Scientific Research's "JP-8 Jet Fuel Toxicology Workshop" that was held at University of Arizona, Tucson, AZ, Jan 11-12, 2000. These studies were conducted as part of a cluster-analysis, rather than to find analogous chemicals or for the estimation of properties (See Task 2 for further discussion).

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 six recent publications (Pub. # 7-12) and reviewed in a chapter (Pub. #13) of the book *Topological Indices and Related Descriptors in QSAR and QSPAR*. This hierarchical approach has also been employed in the development of similarity spaces (Pub. #3 & 6).

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. #6), three distinct similarity spaces were created from: a) calculated TSI only, b) calculated TCI only, and c) a combination of both TSI and TCI.

The former of these studies (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.

In addition to the selection of analogs, similarity spaces can be used in cluster-analysis. This technique assesses the molecular similarity and examines the distances between molecules within the similarity space to form clusters of related compounds. The clusters are formed around a central point (centroid) and have a set radius based on the molecular density around the centroid. The distance from the cluster centroid to any compound within that cluster can be measured, telling us which compounds are nearest the centroid and which compounds are furthest from the centroid. This type of study is useful in scanning large real or virtual chemical libraries in looking for new pharmaceutical leads or for other testing problems in which the number of compounds is simply too large, and therefore too expensive, to subject the entire set to proper toxicological screening. In this situation, representatives from each of the clusters can be tested on the assumption that since the compounds within each cluster are similar, their properties should also be similar.

Just such a study has been carried out on 194 of the isolated compounds in JP-8. The three similarity spaces (See Task 1 above) were clustered in an attempt to determine the optimal number of topological, atom pair or physicochemical clusters to be used on a set of nearly 200 chemicals and the optimal representation for this particular set of compounds.

Further research coupled with analytical testing is necessary to truly determine the optimal method for the creation of molecular similarity spaces for use with cluster-analysis. However, this technique promises to be very useful in simplifying the problems of analyzing mixture toxicity, as in the case of JP-8, and in the analysis and pre-screening of large virtual chemical libraries in search of new, novel drug leads.

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. #5). 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 pharmacodyanmic 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 (LC_{50}) in fathead minow (*Pimephales promelas*) (Pubs. #14-16). 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 clustering (VC) (Pub. #14 & 15). 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 (Pub. #16) 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

(8/1/96 – 7/31/97)
Mike Henderson – Undergraduate Research Assistant

(8/1/97 – 7/31/99)
Jennifer Maki – Undergraduate Research Assistant

(8/1/98 – 7/31/99)
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 July 1, 1996 and December 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.*, in press, 1999.

2. 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.*, **10**, 117-129, 1999.

3. 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.

4. Characterization of molecular structures using topological indices, S.C. Basak and B.D. Gute, *SAR QSAR Environ. Res.*, **7**, 1-21 1997.

5. 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.

6. 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.

7. A comparative QSAR study of benzamidines complement-inhibitory activity and benzene derivatives acute toxicity, S.C. Basak, B.D. Gute, B. Lucic, S. Nikolic, and N. Trinajstic, *Computers and Chemistry*, in press, 1999.

8. 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.

9. 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.

10. The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, S.C. Basak, B.D. Gute and G.D. Grunwald, in: *QSAR in Environmental Sciences - VII*, F. Chen and G. Schuurman, eds., SETAC Press, Pensacola, FL, 1998, Chapter 17, p 245-261.

11. Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, B.D. Gute and S.C. Basak, *SAR QSAR Environ. Res.*, **7**, 117-131, 1997.

12. Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach, S.C. Basak, B.D. Gute and G.D. Grunwald, *J. Chem. Inf. Comput. Sci.*, **37**, 651-655, 1997.

13. A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, S.C. Basak, B.D. Gute and G.D. Grunwald, in: *Topological Indices and Related Descriptors in QSAR and QSPAR*, eds. J. Devillers and A.T. Balaban, Gordon and Breach: Reading, UK, in press, 1999.

14. 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.

15. 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.

16. 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 all manuscripts have been attached as the Appendices.

INTERACTIONS/TRANSITIONS

Participation/Presentations

1. Subhash Basak gave an invited presentation "Exploring the scientific basis of Ayurvedic Medicine: A computatioal approach" at the conference "Beyond Conventional Healthcare: Understanding Alternative Choices" organized by the University of Wisconsin, Superior, November 12-13, 1999.

2. Subhash Basak gave an invited presentation on "Development of hierarchical QSAR models for predicting toxicity of chemicals: statistical and neural net approaches" at the Air Force Predictive Toxicology Conference, Wright Patterson Air Force Base, Dayton, OH, October 7, 1999.

3. Subhash Basak gave the following invited research presentations/ invited seminars during his trip to Europe and India:

   a) "A hierarchical QSAR approach for predating property/activity of chemical from structure" at the Rugjer Boskovic Institute, Zagreb, The Republic of Croatia, August 26, 1999

   b) "Predicting property/activity/toxicity of chemicals from structure: A hierarchical QSAR approach" at the National Institute of Chemistry, Slovenia, August 30, 1999

   c) "Prediction of activity/toxicity of chemicals from structure using graph invariants" at Visva Bharati University, Santiniketan, West Bengal, India, September 9, 1999

   d) "Clustering of Psoralen Derivatives using Topological Invariants: a strategy for molecular design" presented at the 13th International Biophysics Congress, New Delhi, September 19-24, 1999, authored jointly by Subhash C. Basak, Gregory D. Grunwald, Alexandru T. Balaban (Polytechnic University, Romania) and Kanika Basak (St. Xavier's Computer Center, Calcutta, India)

   e) "A Hierarchical QSAR Approach to Predicting Bioactivity of Chemicals using Theoretical Molecular Descriptors" presented at the 13th International Biophysics Congress, New Delhi, September 19-24, 1999, authored jointly by Subhash C.

Basak, Brian D. Gute, Denise Mills, Gregory D. Grunwald, David Opitz (University of Montana, Mizoulla), and Krishnan Balasubramanian (Dept. of Chemistry and Biochemistry, Arizona State University, Tempe, AZ)

f) "Modeling the Solubility of Aliphatic Alcohols in Water: Graph Connectivity Indices versus Line Graph Connectivity Indices" presented at the 13th International Biophysics Congress, New Delhi, September 19-24, 1999, authored jointly by Dragan Amic (The Rugjer Boskovic Institute, Croatia), Subhash C. Basak, Drago Beslo (Croatia), Sonja Nikolic (The Rugjer Boskovic Institute, Croatia) and Nenad Trinajstic (The Rugjer Boskovic Institute, Croatia)

g) "Design of High Quality Structure-Property Regressions" presented at the 13th International Biophysics Congress, New Delhi, September 19-24, 1999, authored jointly by Milan Randic (Drake University, IA) and Subhash C. Basak

h) "On Numerical Characterization of DNA Primary Sequences, presented at the 13th International Biophysics Congress, New Delhi, September 19-24, 1999, authored jointly by Milan Randic (Drake University), Marjan Vracko (National Institute of Chemistry, Slovenia), Ashesh Nandy (Indian Institute of Chemical Biology, Calcutta, India) and Subhash C. Basak,

i) "Predicting biomedicinal and toxicological properties of chemicals using molecular descriptors" at the University of Delhi, India, September 24, 1999

j) "The utility of Ayurvedic medicine for modern drug discovery: An exploratory analysis" at the conference organized by the East India Pharmaceutical Company, Calcutta, September 29, 1999

4. Subhash Basak presented the following papers at the QSAR Gordon Conference, July 25-30, 1999, Tilton, New Hampshire:

a) A hierarchical QSAR approach for predicting property/activity of chemicals, authored by Basak, Greg Grunwald, Brian Gute, Denise Mills, Krishnan Balasubramanian (Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona), and Alexandru Balaban (Polytechnic University, Bucharest, Romania)

b) Topological indices as molecular descriptors for QSAR, authored by Balaban and Basak

5. Subhash Basak and Milan Randic, a Distinguished Professor of Mathematics and Computer Science at Drake University, Iowa, and a Visiting Scientist at NRRI, jointly organized a one day Workshop on Applied Mathematical Chemistry: Molecular Descriptors and Their Applications in Structure-Property-Activity-Toxicity Relationship, May 3, 1999, at NRRI. Thirteen speakers from seven different countries, viz., Bulgaria, Croatia, India, Romania, Slovenia,

United Kingdom and United States, gave invited presentations on their latest research on Mathematical Chemistry, Quantitative Structure Activity Relationships (QSAR), Computational Chemistry and Predictive Toxicology. Dr. Michael J. Lalich, Director of NRRI, welcomed the guests and Dr. Vincent Magnuson, Vice Chancellor for Academic Administration at UMD inaugurated the workshop.

6. Brian Gute attended Annual American Chemical Society meeting, March 21-25, 1999, Anaheim, CA.

7. Subhash Basak gave the following invited presentations on QSAR/ predictive toxicology:

a) "A computational approach to predicting toxicity and toxic modes of action of chemicals from structure" at the International Conference 'Smarter Lead Optimization: easing the bottleneck' organized by Cambridge Health Institute, March 18-19, 1999, San Diego, CA

b) "Topological indices as molecular descriptors for lead optimization" authored jointly by Alexandru T. Balaban and Subhash C. Basak, at the International Conference 'Smarter Lead Optimization: easing the bottleneck' organized by Cambridge Health Institute, March 18-19, 1999, San Diego, CA

c) "Use of statistical and neural net methods in predicting toxicity of chemicals: a hierarchical QSAR approach" authored jointly by Subhash C. Basak, Gregory D. Grunwald, Brian D. Gute, K. Balasubramanian (Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ, and David Opitz (Department of Computer Science, University of Montana, Missoula, Montana) at the American Association of Artificial Intelligence (AAAI) conference, "Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools," Stanford University (CA), March 22-24, 1999

d) "A Graphical Technique for Preliminary Assessment of Effects on DNA Sequences from Toxic Substances" authored jointly by A. Nandy (Indian Institute of Chemical Biology, Calcutta, India), C. Raychaudhury (IICB, Calcutta, India) and Subhash Basak at the American Association of Artificial Intelligence (AAAI) conference, "Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools," Stanford University (CA), March 22-24, 1999

8. Brian Gute attended Annual Society of Toxicology meeting, March 13-17, 1999, New Orleans, LA.

9. Subhash Basak gave the following invited lectures/ presentations:

a) The first distinguished lecture in Mathematical Chemistry on "From Graph Invariants to Molecular Design: 25 years after the connectivity index" at Visva

Bharati University, Santiniketan, West Bengal, India, February 11, 1999

b) An invited seminar on "Theoretical molecular descriptors for the prediction of bioactivity, toxicity, selection of analogs, discovery and optimization of leads" at the Wockhardt Research Centre, Aurangabad, Maharashtra, India, on February 15, 1999

c) An invited lecture on "Prediction of bioactivity of chemicals from structure: a hierarchical computational approach" at Bharatiya Vidya Bhavan's Swami Prakashananda Ayurvedic Research Center, Mumbai, India, on February 18, 1999

d) An invited lecture on "Toxicology in silico: addressing the quagmire of environmental pollution and protecting public health using computational chemistry," authored jointly by Subhash C. Basak, Brian D. Gute David Opitz (Computer Science Department, University of Montana, Missoula) and Gregory D. Grunwald at the International Symposia Series: Reducing the Environmental Impacts of Toxic Chemicals in Asian Economies. The Impacts of Toxic Chemicals and Pollutants on Public Health, the Ecology and the Environment of the Bengal Basin - Bangladesh and India , Dhaka Bangladesh, on March 1, 1999

e) An invited seminar on "Novel drug discovery methods: predicting pharmacological and toxicological properties of chemicals using computational chemistry" at the School of Pharmacy, Dhaka University, Dhaka, Bangladesh on March 4, 1999

f) An invited talk on "Computational toxicology: a cost effective approach for the protection of human and environmental health" at the International Conference at Santiniketan, India, March 7, 1999

g) An invited presentation "Estimation of DNA Damage from Toxic Chemicals by Graphical Techniques" authored jointly by Ashesh Nandy (Head of the Computer Division, Institute of Chemical Biology (IICB), Calcutta, India), C. Raychaudhury and S. Ghosh, Research Scientists at IICB and Subhash Basak on March 8, 1999

10. Subhash Basak gave an invited lecture on "Novel Drug Discovery Methods: Predicting pharmacological and toxicological properties of chemicals using computational chemistry" at the Meharry Medical College, Nashville, TN, January 19, 1999.

11. Subhash Basak had a site visit to the Molecular Anatomy Laboratory, Department of Biology, Indiana University Purdue University, of Indiana, Columbus, IN, January 12-16, 1999, as part of the US Air Force Predictive toxicology program, to discuss the use of proteomics in the development of

QSAR models for JP8 jet fuel with colleagues from University of Minnesota, TC campus, University of Montana, Missoula and IUPUI.

12. Subhash Basak gave an invited presentation “Clustering of JP-8 constituents into structurally dissimilar groups: a novel computational strategy for predictive toxicology”, authored jointly by Basak and Greg Grunwald, at the Air Force Office of Scientific Research JP-8 Jet Fuel Toxicology Workshop, held at the University of Arizona, Tucson, AZ, December 2-3, 1998.

13. Brian Gute presented an invited talk “A hierarchical QSAR approach to predicting carcinogenicity of chemicals” authored jointly, by Subhash Basak, Gute and Greg Grunwald, at the 19th Annual Society of Environmental Toxicology and Chemistry meeting, Charlotte, North Carolina, November 15-19, 1998

14. Subhash Basak presented the following invited lectures:

    a) “Theoretical molecular descriptors for the prediction of bioactivity/toxicity, selection of analogs, discovery and optimization of leads” authored jointly by Basak, Brian Gute, Gregory Grunwald, and Alexandru T. Balaban (Professor of Organic Chemistry at the Polytechnic University, Bucharest, Roumania) at the Astra Symposium on “Advance in Medicinal Chemistry” organized by the Astra company, Bangalore, September 17-19, 1998.

    b) “Prediction of bioactivity of chemicals from structure: a computational approach” at the Indian Institute of Science, Bangalore, India, September 20, 1998.

    c) “Integration of traditional Indian medicine and chemoinformatics for rapid drug discovery” at the conference organized jointly by East India Pharmaceutical Company, Calcutta, October 12, 1998.

15. Subhash Basak attended the Annual American Chemical Society meeting, August 23-27, 1998, Boston, Massachusetts.

16. Dr. S.C. Basak presented the invited lecture “Use of theoretical structural descriptors in molecular design and hazard assessment of chemicals” to the scientists of the computer-aided drug design company NANODESIGN, INC, Toronto, Canada, July 6, 1998.

17. Dr. S.C. Basak presented an invited seminar “Novel Drug Design Methods: assessing activity and toxicity using computational chemistry” at the Department of Molecular Biology and Genetics, University of Guelph, Ontario, Canada, July 3, 1998.

18. Dr. S.C. Basak presented a paper "Dissimilarity-based clustering of psoralen derivatives in the topological structure space: a strategy for drug design" at the Second Annual Chemoinformatics Workshop, organized by the Cambridge

Health Institute, Boston, MA, June 15-16, 1998. The paper was co-authored by G.D Grunwald and B.D. Gute.

19. Dr. S.C. Basak presented the following papers at the International Conference "Computational Methods in Toxicology" held April 20-22, 1998, Dayton, OH:

a) "Use of computational methods in predicting potential toxicity of chemicals," authored jointly by S.C. Basak, B.D. Gute and G.D. Grunwald.

b) "On construction of optimal molecular descriptors," authored jointly by M. Randić and S.C. Basak.

c) "Predicting mode of action of chemicals from structure: a hierarchical approach," authored jointly by S.C. Basak, G.D. Grunwald and B.D. Gute.

d) "A hierarchical approach to predictive toxicology using computed molecular descriptors," authored jointly by B.D. Gute, G.D. Grunwald and S.C. Basak

20. Dr. S.C. Basak gave an invited presentation entitled "A computational approach to predicting toxicity: Possible applications to JP8 jet fuel" at the First International Conference on the Environmental Health and Safety of Jet Fuels, organized jointly by US Air Force, National Institute of Occupational Safety and Health, USEPA National Exposure Research Laboratory and American Industrial Hygiene Association, April 1-3, 1998, San Antonio, TX.

21. Dr. S.C. Basak chaired a session at the DIMACS Workshop on Discrete Mathematical Chemistry, March 23-25, 1998, held at Rugters University, New Jersey. He also presented an invited paper entitled "Use of graph invariants in QSAR and predictive toxicology" at the conference authored jointly by S.C. Basak, B.D. Gute and G.D. Grunwald.

22. Dr. S.C. Basak gave several invited lectures at various national and international symposia:

a) A distinguished lecture "Rational drug design and Ayurvedic medicine" at the conference organized by the Association of Ayurvedic Doctors of India (AADI), January 4, 1998.

b) An invited lecture on "Use of computational methods and Ayurvedic knowledge in modern drug discovery" at the conference AYURVEDA TODAY, January 8, 1998.

c) An invited seminar on "Assessment of genotoxicity of chemicals from structure: a computational approach" at the Annual Conference of the Indian Association for Cancer Congress, Calcutta, January 21-24, 1998, B.D. Gute and G.D. Grunwald.

23. Dr. S.C. Basak was the Co-Chairperson of the First Indo/US Workshop on Mathematical Chemistry, organized jointly by NRRI and Visva Bharati University, Santiniketan, West Bengal India, Jan 9-13, 1998. Basak presented the following papers at the workshop:

a) "Graph invariants, molecular similarity and QSAR" coauthored by B.D. Gute and G.D. Grunwald.

b) "Weighted paths as novel optimal molecular descriptors" authored jointly by M. Randić, President, International Society for Mathematical Chemistry and S.C. Basak.

c) "The utility of hierarchical model development in examining the structural basis of properties" authored by B.D. Gute, G.D. Grunwald and S.C. Basak.

d) "Weighted K-nearest neighbors property estimation in molecular similarity" authored by G.D. Grunwald, B.D. Gute and S.C. Basak.

e) "Dissimilarity based clustering of psoralen derivatives in the topological structure space: a strategy for drug design" authored by S.C. Basak, G.D. Grunwald, D. Panja, K. Basak and B.D. Gute.

24. Subhash C. Basak presented an invited lecture entitled "Predicting bioactivity of chemicals from structure: a hierarchical QSAR approach" to the Department of Biochemistry, University of Calcutta, Calcutta, India, July 30, 1997.

25. Subhash C. Basak presented an invited lecture entitled "Prediction of physicochemical and toxicological properties of chemicals using theoretical molecular descriptors", at Moscow State University, Moscow, Russia, June 30,1997.

26. Subhash C. Basak, Brian D. Gute, and Greg D. Grunwald presented an invited paper entitled "Use of theoretical molecular descriptors in structure-property and structure-activity studies" at the 7th International Conference on Mathematical Chemistry and 3rd Girona Seminar on Molecular Similarity, Girona, Spain, May 26-31, 1997.

27. Subhash C. Basak, Brian D. Gute and Greg D. Grunwald presented an invited paper entitled "Use of nonempirical structural descriptors in QSAR" in the session "Mathematical approaches to QSAR and predictive toxicology" of the 11th International Conference on Mathematical and Computer Modelling and Scientific Computing in Washington, DC, March 27-April 3, 1997,

28. Subhash C. Basak presented a seminar "Computational chemical graph theory and its practical applications" in the Scientific Computing Seminar Laboratory for

Intelligent Systems - ECE Dept.and CSc Dept. University of Minnesota, Duluth on January 29, 1997.

29. Subhash C. Basak gave a presentation "Development of QMSA and QSAR methods for hazard assessment of chemicals: tools for computational toxicology" at the Air Force Office of Scientific Research (AFOSR) Toxicology Program Review, December 12-13, 1996, Fairborn, Ohio.

30. Subhash C. Basak and Brian D. Gute gave an invited presentation "Quantitative Molecular Similarity Analysis (QMSA) and Toxicity Prediction" at the US Air Force Conference "Chemistry and Toxicology of Candidate Deicers" organized by the Materials Directorate of Wright Patterson Air Force Base (WPAFB), Dayton, OH.

31. Brian D. Gute, Subhash C. Basak and Greg D. Grunwald presented a paper "Development of QSARs of bioactive molecules using a hierarchical approach" at the 31st Midwest Regional meeting of the American Chemical Society, November 6-8, 1996.

32. Subhash C. Basak presented a seminar "QSAR/QMSA using nonempirical parameters: applications in predictive toxicology and drug discovery" at the Abbott Laboratories, Chicago, September 22-23, 1996.

33. Subhash C. Basak and Brian Gute presented an invited lecture at the international symposium organized for the 1995 Herman Skolnick award in chemical information. The symposium was held during the American Chemical Society meeting, Orlando, Florida, August 25-29, 1996.

Consultative and Advisor Functions

None

Transitions

1. Applied computational methods in the design a set of six anti-epileptic carbamates by Professor Alexandru T. Balaban, Vice President, Rumanian Academy of Sciences.

2. Worked with Dr. James Riviere, North Carolina State University, in the clustering of JP-8 components using dissimilarity methods developed at NRRI.

3. Worked with Professor George Mushrush, Department of Chemistry, George Mason University, Washington D.C., in the application of similarity and QSAR methods in the design of novel and benign deicing agents.

New Discoveries

1. An in-depth study of similarity space construction and analog selection resulted in the discovery that for a particular set of compounds the degree of overlap between the groups of analogs selected by theoretical descriptor spaces is relatively high. This study also revealed that a similarity space constructed from physicochemical property data provided relatively unique sets of analogs as compared to those selected from the theoretically-derived similarity spaces.

2. Hierarchical QSAR research using topostructural, topochemical, and geometrical parameters showed that the first two classes of parameters explain most of the variance in the data of toxicological and physicochemical properties.

3. It was observed that similarity spaces derived from topostructural and topochemical parameters have distinct analog selection characteristics.

Honors/Awards

1. Dr. S.C. Basak chaired a session at the DIMACS Workshop on Discrete Mathematical Chemistry, March 23-25, 1998, held at Rugters University, New Jersey.

2. Dr. S.C. Basak was the Co-Chairperson of the First Indo/US Workshop on Mathematical Chemistry, organized jointly by NRRI and Visva Bharati University, Santiniketan, West Bengal India, Jan 9-13, 1998.

3. Subhash C. Basak was invited to present a lecture on molecular similarity at the 7th International Conference on Mathematical Chemistry and 3rd Girona Seminar on Molecular Similarity, Girona, Spain, May 26-31, 1997.

4. Subhash C. Basak was invited to become a member of the Organizing and Scientific Committee of for future meetings of the International Conference on Mathematical and Computer Modelling and Scientific Computing.

5. Subhash C. Basak chaired and organized two sessions at the 11th International Conference on Mathematical and Modelling and Scientific Computing, March 31-April 3, 1997, Georgetown University, Washington, DC.

6. Subhash C. Basak was invited to become one of six invited speakers at the international symposium organized for the 1995 Herman Skolnick award in chemical information. The symposium was held during the American Chemical Society meeting, Orlando, Florida, August 25-29, 1996, to honor Milan Randic, the recipient of 1995 Herman Skolnic award.

APPENDIX 1

Appendix 1.1 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.

Appendix 1.2 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.

Appendix 1.3 A comparative QSAR study of benzamidines complement-inhibitory activity and benzene derivatives acute toxicity, S.C. Basak, B.D. Gute, B. Lucic, S. Nikolic, and N. Trinajstic, *Computers and Chemistry*, in press, 1999.

Appendix 1.4 A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, S.C. Basak, B.D. Gute and G.D. Grunwald, in: *Topological Indices and Related Descriptors in QSAR and QSPAR*, eds. J. Devillers and A.T. Balaban, Gordon and Breach: Reading, UK, in press, 1999.

Appendix 1.5 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.

Appendix 1.6 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.*, in press, 1999.

Appendix 1.7 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.

Appendix 1.8 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.*, **10**, 117-129, 1999.

Appendix 1.9 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 1.10 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.

Appendix 1.11 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 1.12 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 1.13 The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals. S.C. Basak, B.D. Gute and G.D. Grunwald, in: *QSAR in Environmental Sciences - VII*, F. Chen and G. Schüürmann, eds., SETAC Press, Pensacola, FL, 1998, Chapter 17, p 245-261.

Appendix 1.14 Characterization of molecular structures using topological indices, S.C. Basak and B.D. Gute, *SAR QSAR Environ. Res.*, **7**, 1-21, 1997.

Appendix 1.15 Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, B.D. Gute and S.C. Basak, *SAR QSAR Environ. Res.*, **7**, 117-131, 1997.

Appendix 1.16 Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach, S.C. Basak, B.D. Gute and G.D. Grunwald, *J. Chem. Inf. Comput. Sci.*, **37**, 651-655, 1997.

APPENDIX 1.1 Use of statistical and neural net approaches in predicting toxicity of chemicals

Journal of Chemical Information and Computer Sciences, submitted

Use of Statistical and Neural Net Approaches in Predicting Toxicity of Chemicals

Subhash C. Basak, Gregory D. Grunwald, Brian D. Gute, Krishnan Balasubramanian[1] and David Opitz[2]

Natural Resources Research Institute, University of Minnesota Duluth, Duluth, Minnesota 55811

[1]Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287-1604

[2]Department of Computer Science, University of Montana, Missoula, Montana 59812

We have been involved in the development of a new hierarchical quantitative structure-activity relationship (H-QSAR) approach in predicting physicochemical, biomedicinal 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 the aquatic toxicity of sixty-nine benzene derivatives using topostructural, topochemical, geometrical, and quantum chemical indices. The utility and limitations of the approach will be discussed.

1. INTRODUCTION

An important aspect of modern toxicological research is the prediction of toxicity of xenobiotics and environmental pollutants from their structure.[1-13] The potential toxicity of chemicals is usually assessed from a plethora of relevant physical and biological properties. Table 1 provides a partial list of such properties. Such toxicological indicators usually try to predict complex toxicity endpoints of chemicals to humans and the environment using simpler and relevant properties. A perusal of the combinatorics of the situation shows that the problem is astronomical. The Toxic Substances Control Act (TSCA) Inventory currently has about 80,000 structures most of which do not have data for the toxicologically relevant properties mentioned in Table 1. In fact, about 50% of these chemicals do not have any experimental property data at all.[14] Worldwide, more than 16 million chemicals are known, as is evident from the number of entries in the Chemical Abstract Service (CAS) inventory.[15] For most of these chemicals we do not have the data necessary for risk assessment. Modern combinatorial chemistry has been producing large libraries of chemicals at a very rapid rate. Most of these substances have none of the test data needed for their hazard estimation.

In recent years, there have been efforts by the chemical industry and government agencies to develop reliable databases of properties that might be used for hazard estimation.[16] This effort, although commendable, falls short of the need; and the picture will remain so in the foreseeable future. In the area of molecular biology, innovative techniques are emerging where specially engineered cell lines can be used to detect the activity or toxicity of chemicals to the genetic system.[17-19] Effects of chemicals on the pattern of cellular proteins, analyzed by proteomics technology, are being used to detect their potential toxic effects.[20-22] Such methods are faster than the traditional methods and can save large number of test animals. At present, neither the available test data nor the combination of *in vitro* toxicity testing methods provide adequate resources for hazard assessment.

Quantitative structure-activity/toxicity relationship (QSAR/QSTR) models have emerged as useful tools to handle the data gap in toxicology and pharmacology.[1-13,22-26] Such models can be used to estimate complex properties of chemicals from simpler experimental or computed properties. In view of the fact that most chemicals in commence and environmental pollutants have very little test data, it would be desirable if we could develop toxicologically-relevant QSARs from properties that can be calculated directly from a chemical's molecular structure. In some of our recent papers we have developed a novel hierarchical QSAR approach where four classes of theoretical molecular descriptors, *viz.*, topostructural, topochemical, geometrical, and quantum chemical parameters, have been used sequentially in the formulation of QSAR models for predicting physical, biomedicinal, and toxicological properties.[1,3,6,8,23-26]

Most of our hierarchical QSARs are based on linear statistical methods such as multiple linear regression, principal components analysis (PCA) and variable clustering. Such methods yield useful models; but they suffer from the limitation that in some cases the relationship between a molecular descriptor and toxicity may be intrinsically nonlinear. In such cases, the use of linear statistical methods may not result in the best models. Therefore, in this paper, we have carried out a comparative study of multiple regression *vis-a-vis* neural net methods in predicting toxicity (LC_{50}) of a set of 69 benzene derivatives.

2. METHODS

2.1 Toxicity Database. The utility of this approach of generating numerous hierarchical theoretical descriptors of compounds was tested on a set of acute aquatic toxicity (LC_{50}) data for sixty-nine benzene derivatives. The data was taken from a study by Hall, Kier and Phipps[12] who collected acute aquatic toxicity data measured in fathead minnow (*Pimephales promelas*). This data was compiled from eight other literature sources and included some original work which was conducted at the U.S. Environmental Protection Agency Environmental Research Laboratory (USEPA – ERL) in Duluth, Minnesota. This set of chemicals was composed of benzene and sixty-eight 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 (see Table 2) have seven different substituent groups that are present in at least six of the molecules: chloro-, bromo-, nitro-, methyl-, methoxyl-, hydroxyl-, and amino-.

2.2 Calculation of Topological Indices (TIs). The complete set of topological indices (TIs) used in this study, both topostructural and topochemical, have been calculated using POLLY 2.3 and software developed by the Basak *et al.* [27] These indices include Wiener index,[28] the connectivity indices developed by Randić[29] and higher order connectivity indices formulated by Kier and Hall,[30] bonding connectivity indices defined by Basak *et al*,[31] a set of information theoretic indices defined on the distance matrices of simple molecular graphs,[32,33] a set of parameters derived on the neighborhood complexity of hydrogen-filled molecular graphs,[34-36] and Balaban's *J* indices.[37-39] Table 3 provides the symbols and brief definitions of the topological indices included in this study.

The set of TIs was divided into two distinct subsets: topostructural indices (TSI) and topochemical indices (TCI). 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 that 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 3 shows the division of the topological indices into topostructural and topochemical indices.

2.3 Calculation of Geometrical Indices. The geometrical indices include three-dimensional Wiener numbers for hydrogen-filled molecular structure, hydrogen-suppressed molecular structure, and van der Waals volume. Van der Waals volume, V_W, was calculated using *SYBYL 6.4* from Tripos Associates, Inc.[40] The 3-D Wiener numbers were calculated using *SYBYL* using 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 3.2.1*.[41] The symbols and definitions of the geometrical indices are included in Table 3.

2.4 Quantum Chemical Parameters. Quantum chemical parameters were calculated using the Austin Model version one (AM1) semi-empirical Hamiltonian These parameters were calculated using *MOPAC 6.00* in the *SYBYL* interface.[42] Brief definitions and symbols for the quantum chemical parameters used in this study are included in Table 3.

2.5 Statistical Analysis and Hierarchical QSAR. 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 transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of eighty-six topological indices was then partitioned into the two distinct sets: topostructural indices (thirty-five) and topochemical indices (fifty-one). The sets of topostructural and topochemical indices were then divided into subsets, or clusters, based on the correlation matrix using the SAS variable clustering procedure (VARCLUS)[43] to further reduce the number of independent variables for use in model construction. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional.

From each cluster, the index most correlated with the cluster was selected for modeling, as well as any indices that were poorly correlated with their cluster (R^2 < 0.70). These indices were then used in the modeling of the acute aquatic toxicity of benzene derivatives in fathead minnow. 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.

Reducing the number of independent variables is critical when attempting to model small datasets using linear statistical methods. The smaller the dataset, the greater the chance of spurious error when using a large number of independent variables (descriptors). Topliss and Edwards[44] have thoroughly studied this issue of chance correlations. For a set with about seventy dependent variables (observations), to keep the probability of chance correlations less than 0.01, at most forty independent variables may be used. This number is dependent on the actual correlation achieved in the modeling process, higher correlation results in a better chance of using more variables with the same limited probability of chance correlations. In this study we are well below the cut-off of forty independent variables. In fact, the total number of descriptors which will be used for model construction and estimation is twenty-three, well within the bounds of the Topliss and Edwards criteria.[44]

Regression modeling was accomplished using the SAS procedure REG[43] on four distinct sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest parameters, the TSIs. After using the TSIs to model the activity, the next level of parameters of higher complexity are added. To the indices included in the best TSI model, we add all of the TCIs and proceed to model the activity using these parameters. Likewise, the indices included in the best model from this procedure are combined with the indices from the next complexity level, the geometrical indices and modeling is conducted once again. Finally, the best model utilizing TSIs, TCIs and geometrical indices is combined with the quantum chemical parameters to develop the final model in the hierarchy.

2.6 Neural Network Methods. Using neural networks, we studied two classes of approaches for modeling toxicity: (1) giving all the descriptors to a learning algorithm (neural network 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.*, sixty-nine training/test set partitions). Leave-one-out works by leaving one data point out of the training set and giving the remaining instances (sixty-eight in this case) to the learning algorithms for training. This process is repeated sixty-nine 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.

First we trained neural networks using all ninety-five parameters: thirty-five TSI, fifty-one TCI, three geometrical and six quantum chemical parameters. The networks contained fifteen hidden units and were trained for 1000 epochs. Each input parameter was normalized to a value 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.

For our next experiment, we used a smaller set of twenty-three independent variables. The twenty-three independent variables were the topostructural and topochemical parameters provided by the variable clustering technique (see section 4.1 for a list of the indices) combined with the three geometrical and six quantum chemical parameters described in Table 3. The parameter settings for these networks were the same as the settings for the other neural network experiments mentioned above.

3. RESULTS

3.1 Results of Statistical Regression Procedures. The variable clustering of the topostructural indices resulted in the retention of five indices: M_1, $\overline{IC}$, O, P_8, P_9. All-subsets regression resulted in the selection of a four-parameter model to estimate -log(LC_{50}) with an explained variance (R^2) of 45.3% and a standard error (s) of 0.58. While this is an unsatisfactory model, the indices were still retained and combined with the topochemical indices in the second step of model development. The second step combined the four indices used in the first tier model with the nine topochemical indices selected in the variable clustering procedure: SIC_0, SIC_1, SIC_4, CIC_0, $^{2}\chi^{b}$, $^{5}\chi^{b}_{C}$, $^{5}\chi^{v}_{C}$, $^{6}\chi^{v}_{PC}$, J^{X}. Again, all-subsets regression was conducted resulting in a four-parameter model with an explained variance (R^2) of 78.3% and a standard error (s) of 0.36. The four indices from the second tier model were combined with the three geometric parameters: $^{3D}W_H$, ^{3D}W, V_W. This resulted in a four-parameter model that replaced the topochemical index CIC_0 with the geometric parameter $^{3D}W_H$. This model had an explained variance (R^2) of 79.2% and a standard error (s) of 0.36. The final step in the hierarchical method combined the four parameters from the third tier model with the semi-empirical quantum chemical parameters: E_{HOMO}, E_{HOMO1}, E_{LUMO}, E_{LUMO1}, ΔH_f, μ. This set of ten indices led to a seven-parameter model with an explained variance (R^2) of 86.3% and a standard error (s) of 0.30. This model retained all indices from the third model and added three of the AM1 quantum chemical parameters.

The leave-one-out analysis was conducted on the final model for purposes of comparison with the results of the neural networks. This analysis resulted in a final explained variance for the model of R^2 = 0.825 and a standard error of s = 0.32.

3.2 Results of the Neural Network Procedures. The first class of approach incorporating all ninety-five parameters, 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. The second class of neural network approaches utilizing the twenty-three parameters from the data reduction step obtained a test-set explained variance (R^2) of 0.878 and a standard error (s) of 0.28. The results from the leave-one-out analysis using the linear statistical method and the neural network methods are summarized in Table 4. Table 2 presents the experimental acute aquatic toxicity (-log[LC_{50}]) values for the sixty-nine benzene derivatives as well as the values estimated by the best statistical model and the two neural network models.

4. DISCUSSION

The results show that both statistical and neural network models give acceptable estimates of the toxicity of the sixty-nine benzene derivatives studied in this paper. However, when tested using the leave-one-out approach, the statistical model falls short of the performance of the neural network models. It has to be noted, however, that statistical QSARs are based on linear models whereas the two neural network models use nonlinear methods.

It is interesting to note that the neural network model using the subset of twenty-three inputs selected in part by the VARCLUS procedure gave slightly better results as compared to the network developed using all ninety-five input variables. This could be the result of filtering out redundant, or nearly redundant, parameters from the set of independent variables.

Further work on the relative utility of statistical *vis-à-vis* neural network methods is necessary to determine which types of models are best suited to the estimation of chemical toxicity.

ACKNOWLEDGEMENT

This paper is contribution number XXX 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, grant IRI-9734419 from the National Science Foundation and a MONTS grant from the University of Montana.

REFERENCES AND NOTES

(1) Basak, S.C.; Gute, B.D.; Grunwald, G.D. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A.T., Eds.; Gordon & Breach: Reading, U.K., in press.

(2) Basak, S.C. In *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990; p 83-103.

(3) Basak, S.C.; Gute, B.D.; Grunwald, G.D. In *QSAR in Environmental Sciences – VII*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1998; p 245-261.

(4) S.C. Basak and B.D. Gute, In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Hansen, P., Paradis, N., Eds.; in press.

(5) Basak, S.C.; Gute, B.D.; Grunwald, G.D. Assessment of the Mutagenicity of Chemicals from Theoretical Structural

Parameters: A Hierarchical Approach, *SAR QSAR Environ. Res.*, in press.

(6) Gute, B.D.; Grunwald, G.D.; Basak, S.C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1-15.

(7) Basak, S.C.; Gute, B.D. Characterization of Molecular Structures using Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 1-21.

(8) Gute, B.D.; Basak, S.C. Predicting Acute Toxicity (LC_{50}) of Benzene Derivatives using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117-131.

(9) Mushrush, G.W.; Basak, S.C.; Slone, J.E.; Beal, E.J.; Basu, S.; Stalick, W.M.; Hardy, D.R. Computational Study of the Environmental Fate of Selected Aircraft Deicing Compounds. *J. Environ. Sci. Health* **1997**, *A32(8)*, 2201.

(10)Basak, S.C.; Grunwald, G.D. Predicting Mutagenicity of Chemicals using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* **1995**, *31*, 2529.

(11)Basak, S.C.; Grunwald, G.D. In *Proceeding of the XVI International Cancer Congress*, Rao, R.S., Deo, M.G., Sanghui, L.D., Eds.; Monduzzi: Bologna, Italy, 1995; p 413.

(12)Hall, L.; Kier, L.; Phipps, G. Structure-Activity Relationship Studies on the Toxicities of Benzene Derivatives: I. An Additivity Model. *Environ. Toxicol. Chem.* **1984**, *3*, 355-365.

(13)Gombar, V.K.; Enslein, K.; Blake, B.W. Assessment of Developmental Toxicity Potential of Chemicals by Quantitative Structure-Toxicity Relationship Models. *Chemosphere* **1995**, *31*, 2499-2510.

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

(15)CAS. The latest CAS registry number and substance count. http://www.cas.org/cgi-bin/regreport.pl, 1999.

(16)Johnson, J. Pact Triggers Tests: Thousands of Chemicals may be Tested under Toxicity Screening Program. *Chem. Engineer. News* **1998**, *76* (44), 19-20.

(17)Chen, J.J.; Wu, R.; Yang, P.C.; Huang, J.Y.; Sher, Y.P.; Han, M.H.; Kao, W.C.; Lee, P.J.; Chiu, T.F.; Chang, F.; Chu, Y.W.; Wu, C.W.; Peck, K. Profiling Expression Patterns and Isolating Differentially Expressed Genes by cDNA Microarray System with Colorimetry Detection. *Genomics* **1998**, *51*, 313-324.

(18)Schena, M.; Shalon, D.; Davis, R.W.; Brown, P.O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **1995**, *270*, 467-470.

(19)De Risi, J.; Penland, L.; Brown, P.O.; Bittner, M.L.; Meltzer, P.S.; Ray, M.; Chen, Y.; Su, Y.A.; Trent, J.M. Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer. *Nat. Genet.* **1996**, *14*, 457-460.

(20)Witzmann, F.A.; Fultz, C.D.; Grant, R.A.; Wright, L.S.; Kornguth, S.E.; Siegel, F.L. Differential Expression of Cytosolic Proteins in the Rat Kidney Cortex and Medulla: Preliminary Proteomics. *Electrophoresis* **1998**, *19*, 2491-2497.

(21)Anderson, N.L.; Esquer-Blasco, R.; Richardson, F.; Foxworthy, P.; Eacho, P. The Effects of Peroxisome Proliferators on Protein Abundances in Mouse Liver. *Toxicol. Appl. Pharm.* **1996**, *137*, 75-89.

(22)Lake, B.G.; Lewis, D.F.V; Gray, T.J.B.; Beamand, J.A. Structure-Activity Relationships for Induction of Peroxysomal Enzyme Activities in Primary Rat Hepatocyte Cultures. *Toxic. in Vitro* **1993**, *7*, 605-614.

(23)Basak, S.C.; Gute, B.D.; Grunwald, G.D.; Opitz, D.W.; Balasubramanian, K. 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.

(24)Basak, S.C.; Gute, B.D.; Ghatak, S. Prediction of Complement-Inhibitory Activity of Benzamidines using Topological and Geometric Parameters. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 255-260.

(25)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.

(26)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.

(27)Basak, S.; Harriss, D.; Magnuson, V. *POLLY 2.3.* Copyright of the University of Minnesota, 1988.

(28)Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.

(29)Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.

(30)Kier, L.; Hall, L. *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press: Hertfordshire, U.K., 1986.

(31)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.

(32)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.

(33)Bonchev, D.; Trinajstić, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517-4533.

(34)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; p851.

(35)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., Lapis, A.I. Rodin, E.Y., Eds.; Pergamon Press: New York, 1984; p 745.

(36)Basak, S.C.; Magnuson, V.R. Molecular Topology and Narcosis. *Arzneim-Forsch. Drug Res.* **1983**, *33*, 501-503.

(37)Balaban, A.T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399-404.

(38)Balaban, A.T. (1983). Topological Indices Based on Topological Distances in Molecular Graphs. *Pure and Appl. Chem.* **1983**, *55*, 199-206.

(39)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.

(40)Tripos Associates, Inc. *SYBYL Version 6.4.*; Tripos Associates, Inc.: St. Louis, MO, 1998.

(41)Tripos Associates, Inc. *CONCORD Version 3.2.1.*; Tripos Associates, Inc.: St. Louis, MO, 1998.

(42)Stewart, J.J.P. *MOPAC 6.00*, QCPE #455; Frank J. Seiler Research Laboratory: US Air Force Academy, CO, 1990.

(43)SAS Institute, Inc. In *SAS/STAT User's Guide*, 6.03 Edition; SAS Institute Inc.: Cary, NC, 1988; Chapters 28 and 34, pp 773-875, 949-965.

Table 1. Physicochemical and biological properties relevant to the assessment of toxicity.

Physicochemical	Biological
Molar Volume	Receptor Binding (K_D)
Boiling Point	Michaelis Constant (K_m)
Melting Point	Inhibitor Constant (K_i)
Vapor Pressure	Biodegradation
Aqueous Solubility	Bioconcentration
Dissociation Constant (pK_a)	Alkylation Profile
Partition Coefficient	Metabolic Profile
Octanol-Water (log P)	Chronic Toxicity
Air-Water	Carcinogenicity
Sediment-Water	Mutagenicity
Reactivity (Electrophile)	Acute Toxicity
	LD_{50}
	LC_{50}

Table 2. Experimental and estimated acute aquatic toxicity data for sixty-nine benzene derivatives, expressed as –log(LC_{50}) for the linear regression model (LR), the neural network model using all ninety-five parameters (NN95) and the neural network model using twenty-three parameters (NN23) selected by variable clustering.

Compound	Exp.	LR	NN95	NN23
Benzene	3.40	3.42	3.66	3.65
Bromobenzene	3.89	3.77	4.02	3.79
Chlorobenzene	3.77	3.75	3.80	3.77
Phenol	3.51	3.38	3.44	3.51
Toluene	3.32	3.66	3.50	3.62
1,2-dichlorobenzene	4.40	4.29	4.24	4.30
1,3-dichlorobenzene	4.30	4.37	4.03	4.12
1,4-dichlorobenzene	4.62	4.51	4.46	4.27
2-chlorophenol	4.02	3.79	3.82	3.91
3-chlorotoluene	3.84	3.88	3.72	3.79
4-chlorotoluene	4.33	3.87	3.78	3.76
1,3-dihydroxybenzene	3.04	3.43	3.47	3.53
3-hydroxyanisole	3.21	3.33	3.40	3.45
2-methylphenol	3.77	3.64	3.55	3.67
3-methylphenol	3.29	3.60	3.51	3.58
4-methylphenol	3.58	3.53	3.54	3.55
4-nitrophenol	3.36	3.61	3.65	3.76
1,4-dimethoxybenzene	3.07	3.28	3.79	3.51
1,2-dimethylbenzene	3.48	3.93	3.88	3.91
1,4-dimethylbenzene	4.21	3.87	3.74	3.68
2-nitrotoluene	3.57	3.66	3.78	3.81
3-nitrotoluene	3.63	3.53	3.71	3.71
4-nitrotoluene	3.76	3.49	3.68	3.68
1,2-dinitrobenzene	5.45	5.24	4.91	4.99
1,3-dinitrobenzene	4.38	4.18	4.30	4.19
1,4-dinitrobenzene	5.22	4.94	4.38	4.85
2-methyl-3-nitroaniline	3.48	3.79	3.79	3.88
2-methyl-4-nitroaniline	3.24	3.51	3.79	3.75
2-methyl-5-nitroaniline	3.35	3.68	3.82	3.86
2-methyl-6-nitroaniline	3.80	3.84	3.73	3.79
3-methyl-6-nitroaniline	3.80	3.78	3.64	3.62
4-methyl-2-nitroaniline	3.79	3.80	3.73	3.66
4-hydroxy-3-nitroaniline	3.65	3.61	3.53	3.58
4-methyl-3-nitroaniline	3.77	3.73	3.72	3.72
1,2,3-trichlorobenzene	4.89	4.89	4.85	5.04
1,2,4-trichlorobenzene	5.00	5.04	5.05	4.83
1,3,5-trichlorobenzene	4.74	5.11	4.62	4.78
2,4-dichlorophenol	4.30	4.33	4.42	4.47
3,4-dichlorotoluene	4.74	4.26	4.39	4.28
2,4-dichlorotoluene	4.54	4.36	4.47	4.44

4-chloro-3-methylphenol	4.27	3.87	3.96	4.07
2,4-dimethylphenol	3.86	3.76	3.78	3.72
2,6-dimethylphenol	3.75	3.80	3.71	3.84
3,4-dimethylphenol	3.90	3.80	3.92	3.79
2,4-dinitrophenol	4.04	4.14	4.15	4.01
1,2,4-trimethylbenzene	4.21	4.09	4.53	3.87
2,3-dinitrotoluene	5.01	5.20	5.12	5.28
2,4-dinitrotoluene	3.75	4.10	4.65	4.33
2,5-dinitrotoluene	5.15	4.84	4.71	4.72
2,6-dinitrotoluene	3.99	4.41	4.56	4.63
3,4-dinitrotoluene	5.08	5.11	5.11	5.09
3,5-dinitrotoluene	3.91	4.05	4.41	4.16
1,3,5-trinitrobenzene	5.29	5.37	5.34	5.32
2-methyl-3,5-dinitroaniline	4.12	4.13	4.30	4.23
2-methyl-3,6-dinitroaniline	5.34	4.80	4.40	4.54
3-methyl-2,4-dinitroaniline	4.26	4.28	4.14	4.20
5-methyl-2,4-dinitroaniline	4.92	4.14	4.00	4.02
4-methyl-2,6-dinitroaniline	4.21	4.67	4.57	4.58
5-methyl-2,6-dinitroaniline	4.18	4.80	4.53	4.78
4-methyl-3,5-dinitroaniline	4.46	4.34	4.32	4.43
2,4,6-tribromophenol	4.70	4.89	5.34	5.47
1,2,3,4-tetrachlorobenzene	5.43	5.62	5.50	5.56
1,2,4,5-tetrachlorobenzene	5.85	5.80	5.63	5.61
2,4,6-trichlorophenol	4.33	4.79	4.86	4.96
2-methyl-4,6-dinitrophenol	5.00	4.21	4.20	4.16
2,3,6-trinitrotoluene	6.37	6.36	5.84	5.81
2,4,6-trinitrotoluene	4.88	5.16	5.39	5.42
2,3,4,5-tetrachlorophenol	5.72	5.36	5.44	5.58
2,3,4,5,6-pentachlorophenol	6.06	6.03	5.86	5.83

Table 3. Symbols, definitions and classifications of topological, geometrical and quantum chemical parameters.

	Topostructural
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
$\overline{IC}$	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
$^h\chi$	Path connectivity index of order *h* = 0-6
$^h\chi_C$	Cluster connectivity index of order *h* = 3, 5
$^h\chi_{Ch}$	Chain connectivity index of order *h* = 6
$^h\chi_{PC}$	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
	Topochemical
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
$^h\chi^b$	Bond path connectivity index of order *h* = 0-6
$^h\chi^b_C$	Bond cluster connectivity index of order *h* = 3, 5
$^h\chi^b_{Ch}$	Bond chain connectivity index of order *h* = 6
$^h\chi^b_{PC}$	Bond path-cluster connectivity index of order *h* = 4-6
$^h\chi^v$	Valence path connectivity index of order *h* = 0-6
$^h\chi^v_C$	Valence cluster connectivity index of order *h* = 3, 5

$^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 6
$^{h}\chi^{v}_{PC}$	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^{Y}	Balaban's J index based on relative covalent radii
Geometrical	
V_{W}	van der Waals 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
Quantum Chemical	
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO1}	Energy of the second highest occupied molecular orbital
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO1}	Energy of the second lowest unoccupied molecular orbital
ΔH_{f}	Heat of formation
μ	Dipole moment

Table 4. Relative effectiveness of statistical and neural network methods in estimating the acute aquatic toxicity of 69 benzene derivatives.

Method	# Independent Variables	R^2	s
Statistical	7	0.825	0.32
Neural network	95	0.868	0.29
Neural network	23	0.878	0.28

APPENDIX 1.2 Hazard assessment modeling: An evolutionary ensemble approach

Hazard Assessment Modeling: An Evolutionary Ensemble Approach

David W. Opitz
Department of Computer Science
University of Montana
Missoula, MT 59812 (USA)
opitz@cs.umt.edu
406-243-2831

Subhash C. Basak **Brian D. Gute**
Natural Resources Research Institute
University of Minnesota
Duluth, MN 55811 (USA)
{sbasak, bgute}@wyle.nrri.umn.edu
218-720-4230

Abstract

This paper presents a novel and effective genetic algorithm approach for generating computational models for hazard assessment. With millions of proposed chemicals being registered each year, it is impossible to come even remotely close to completing the battery of tests needed for the proper understanding of the toxic effects of these chemicals. Computer models can give quick, cheap, and environmentally friendly hazard assessments of chemicals. Our approach works by first extracting a hierarchy of theoretical descriptors of the structure of a compound, then filtering these numerous descriptors with a genetic algorithm approach to ensemble feature selection. We tested the utility of our approach by modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. Our results demonstrate a very important point: that our method is able to accurately predict toxicity directly from structure.

1 INTRODUCTION

By the end of 1998 the number of chemicals registered with the Chemical Abstract Service rose to over 19 million (CAS 1999). This is an increase of over 3 million chemicals between 1996 and 1998. It is desirable 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. Additionally, these computational methods can replace or limit the amount of animal testing that is necessary. 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 a novel and effective approach for learning computational hazard assessment models by using an ensemble feature selection algorithm based on genetic algorithms (GAs) to filter numerous theoretical descriptors of chemical structure.

To better illustrate the need for effective and quick hazard assessment, we should consider the situation of the industrial chemicals "grandfathered" into continued use under the Toxic Substances Control Act (TSCA) of 1976. TSCA has required that a suite of physicochemical and toxicological screens be run on all commercial compounds (those produced or imported in volumes exceeding one million pounds annually) developed after 1976. However, there are almost 3,000 chemicals that were "grandfathered" in with the understanding that it would be the responsibility of the chemical manufacturing industry to ultimately supply information about these chemicals. Only recently, after a 20-year delay, are the chemical manufacturers talking about running 2,800 of these compounds through basic toxicity screens and while this is promising, these screens will not be completed until 2004 and at a cost of between $500 to $700 million dollars. So it will be another five years before we have basic toxicity data on compounds that have been in wide-spread use for more than twenty years (Johnson 1998).

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 that 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 hierarchical approach examines the relative contributions of theoretical descriptors of gradually increasing complexity (structural, chemical, shape, and quantum chemical descriptors). This approach is important as none of the individual classes of parameters are very effective at predicting toxicity (Gute & Basak 1997); however, we show in this paper that we can effectively predict toxicity if we combine all levels of descriptors. One potential problem with using our hierarchical approach is that it often gives many independent variables as compared to data points since having a limited number of data points in not uncommon in hazard assessment. 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). In some of our earlier QSAR studies we have used statistical methods such as principal components analysis (PCA) and variable clustering methods to reduce the number of independent variables (Basak & Grunwald 1995; Gute & Basak 1997; Gute, Grunwald, & Basak In press).

As an alternative solution, we use our previous ensemble feature selection approach (Opitz 1999) that is based on GAs. 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 to, in most cases, greatly improve generalization accuracy over a single learning model (Breiman 1996; Maclin & Opitz 1997; 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 1999). We use GAs to search through the enormous space of finding a set of feature subsets that will promote disagreement among the component members of an ensemble while still maintaining the component member's accuracy.

Combining our approach of generating hierarchical theoretical descriptors with our other approach to GA-based ensemble feature selection, we are able to generate an effective model for predicting the toxicity of benzene derivatives using only a few compounds. Our results show that our model is nearly as accurate as the battery of tests necessary for the proper hazard assessment of a single compound. Our results also confirm that our new ensemble feature selection approach is more effective than previous approaches for modeling hazard assessment.

The rest of the paper is organized as follows. First we provide background and related work for both our hierarchical QSAR approach and our GA-based ensemble feature selection approach. This is followed by results of our approach applied to benzene derivatives. Finally, we discuss these results and provide future work.

2 QSAR AND THEORETICAL METHODS

QSARs have come into widespread use for the prediction of various molecular properties, as well as biological, pharmacological and toxicological responses. Traditional QSAR techniques use empirical properties (Dearden 1990; Hansch & Leo 1995; de Waterbeemd 1995); however, due to the scarcity of available data for the majority of chemicals needing to be evaluated for hazard assessment, these physicochemical properties necessary for traditional QSAR model development may not be available. When this is the case, it is imperative that there are methods available which make use of nonempirical parameters, which we term theoretical molecular descriptors.

Topological indices (TIs) are numerical graph invariants that quantify certain aspects of molecular structure (Gute & Basak 1997; Gute, Grunwald, & Basak In press). The different classes of TIs provide us with nonempirical, quantitative descriptors that can be used in place of experimentally derived descriptors in QSARs for the prediction of properties.

Our recent studies have focused on the role of different classes of theoretical descriptors of increasing levels 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.

2.1 TOPOLOGICAL INDICES

The topostructural and topochemical indices fall into the category normally considered topological indices. Topostructural indices (TSIs) are topological indices that 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 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.

2.2 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: $^{3D}W_H$ 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.

2.3 QUANTAM 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).

3 FILTERING DESCRIPTORS

As stated above, one potential problem with including all theoretical descriptors in the hierarchy is that it gives many independent variables when compared to the limited number of data points available for hazard assessment modeling of a particular chemical derivative. Compounding this problem is that a salient descriptor for one hazard assessment model may not be a salient descriptor for another problem. That is, the relevance of a descriptor for predicting hazard assessment is often problem dependent. This section describes our approach for automatically filtering the descriptors with a GA-based approach to ensemble feature detection. Before explaining our algorithm, we briefly cover the notion of ensembles.

3.1 ENSEMBLES

Figure 1 illustrates the basic framework of a predictor ensemble. Each predictor in the ensemble (predictor 1 through predictor N in this case) is first trained using the training instances. Then, for each example, the predicted output of each of these predictors (o_i in Figure 1) is combined to produce the output of the ensemble ($\hat{o}$ in Figure 1). Many researchers (Breiman 1996; Hansen & Salamon 1990; Krogh & Vedelsby 1995;

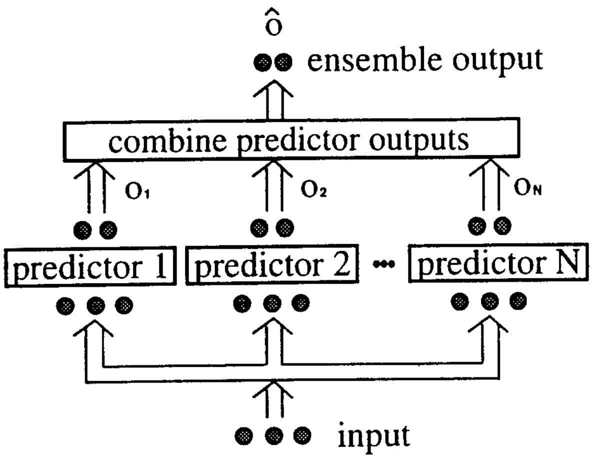


Figure 1: A predictor ensemble.

Opitz & Shavlik 1997) have demonstrated the effectiveness of combining schemes that are simply the weighted average of the predictors (i.e., $\hat{o} = \sum_{i \in N} w_i \cdot o_i$ and $\sum_{i \in N} w_i = 1$), and this is the type of ensemble on which we focus in this article.

Combining the output of several predictors is useful only if there is disagreement on some inputs. Obviously, combining several identical predictors produces no gain. Hansen and Salamon 1990 proved that for an ensemble, if the average error rate for an example is less than 50% and the predictors in the ensemble are independent in the production of their errors, the expected error for that example can be reduced to zero as the number of predictors combined goes to infinity; however, such assumptions rarely hold in practice.

Krogh and Vedelsby 1995 later proved that the ensemble error can be divided into a term measuring the average generalization error of each individual predictor and a term called diversity that measures the disagreement among the predictors. Formally, they define the diversity term, d_i, of predictor i on input x to be:

$$d_i(x) \equiv [o_i(x) - \hat{o}(x)]^2. \qquad (1)$$

The quadratic error of predictor i and of the ensemble are, respectively:

$$\epsilon_i(x) \equiv [o_i(x) - f(x)]^2, \qquad (2)$$

$$e(x) \equiv [\hat{o}(x) - f(x)]^2, \qquad (3)$$

where $f(x)$ is the target value for input x. If we define $\hat{E}$, E_i, and D_i to be the averages, over the input distribution, of $e(x)$, $\epsilon(x)$, and $d(x)$ respectively, then the ensemble's generalization error can be shown to consist of two distinct portions:

$$\hat{E} = \bar{E} - \bar{D}, \qquad (4)$$

where $\bar{E}$ $(= \sum_i w_i E_i)$ is the weighted average of the individual predictor's generalization error and $\bar{D}$ $(= \sum_i w_i D_i)$ is the weighted average of the diversity among these predictors. What the equation shows then, is that an ideal ensemble consists of highly correct predictors that disagree as much as possible. Opitz and Shavlik 1996a; 1996b empirically verified that such ensembles generalize well.

Regardless of theoretical justifications, methods for creating ensembles center around producing predictors that disagree on their predictions. Generally, these methods focus on altering the training process in the hope that the resulting predictors will produce different predictions. For example, neural network techniques that have been employed include methods for training with different topologies, different initial weights, different parameters, and training only on a portion of the training set (Alpaydin 1993; Freund & Schapire 1996; Hansen & Salamon 1990; Maclin & Shavlik 1995).

Numerous techniques try to generate disagreement among the classifiers by altering the training set each classifier sees. The two most popular techniques are Bagging (Breiman 1996) and Boosting (Freund & Schapire 1996). Bagging is a bootstrap ensemble method that trains each network in the ensemble with a different partition of the training set. It generates each partition by randomly drawing, with replacement, N examples from the training set, where N is the size of the training set. As with Bagging, Boosting also chooses a training set of size N and initially sets the probability of picking each example to be $1/N$. After the first network, however, these probabilities change to emphasize misclassified instances. A large number of extensive empirical studies have shown that these are highly successful methods that nearly always generalize better than their individual component predictors (Bauer & Kohavi 1998; Maclin & Opitz 1997; Quinlan 1996). Neither approach is appropriate for our domain since we are data poor and cannot afford to waste training examples; however, we are feature rich and can afford to create diversity by instead varying the inputs to the learning algorithms. *Varying the feature subsets to create a diverse set of accurate predictors is the focus of the next section.*

3.2 THE GEFS ALGORITHM

The goal of our algorithm is to find a set of feature subsets that creates an ensemble of classifiers (neural networks in this study) that maximize equation 1 while minimizing equation 2. The space of candidate sets is enormous and thus is particularly well suited for ge-

Table 1: The GEFS algorithm.

GOAL: Find a set of input subsets to create an accurate and diverse classifier ensemble.

1. Using varying inputs, create and train the initial population of classifiers.
2. Until a stopping criterion is reached:
   (a) Use genetic operators to create new networks.
   (b) Measure the diversity of each network with respect to the current population.
   (c) Normalize the accuracy scores and the diversity scores of the individual networks.
   (d) Calculate fitness of each population member.
   (e) Prune the population to the N fittest networks.
   (f) Adjust λ.
   (g) The current population is the ensemble.

netic algorithms. Table 1 summarizes our recent algorithm (Opitz 1999) called GEFS (for Genetic Ensemble Feature Selection) that uses GAs to generate a set of classifiers that are accurate and diverse in their predictions. GEFS starts by creating and training its initial population of networks. The representation of each individual of our population is simply a dynamic length string of integers, where each integer indexes a particular feature. We create networks from these strings by first having the input nodes match the string of integers, then creating a standard single-hidden-layer, fully connected neural network. Our algorithm then creates new networks by using the genetic operators of crossover and mutation.

GEFS trains these new individuals using backpropogation. It adds new networks to the population and then scores each population member with respect to its prediction accuracy and diversity. GEFS normalizes these scores, then defines the fitness of each population member (i) to be:

$$Fitness_i = Accuracy_i + \lambda \; Diversity_i \qquad (5)$$

where λ defines the tradeoff between accuracy and diversity. Finally, GEFS prunes the population to the N most-fit members, then repeats this process. At every point in time, the current ensemble consists of simply averaging (with equal weight) the predictions of the output of each member of the current population. Thus as the population evolves, so does the ensemble.

We define accuracy to be network i's training-set accuracy. (One may use a validation-set if there are enough training instances.) We define diversity to be the average difference between the prediction of our component classifier and the ensemble. We then separately normalize both terms so that the values range from 0 to 1. Normalizing both terms allows λ to have the same meaning across domains.

It is not always clear at what value one should set λ; therefore, we automatically adjust λ based on the discrete derivatives of the ensemble error $\hat{E}$, the average population error $\bar{E}$, and the average diversity $\bar{D}$ within the ensemble. First, we never change λ if $\hat{E}$ is decreasing; otherwise we (a) increase λ if $\bar{E}$ is not increasing and the population diversity $\bar{D}$ is decreasing; or (b) decrease λ if $\bar{E}$ is increasing and $\bar{D}$ is not decreasing. We started λ at 1.0 for the experiments in this article. The amount λ changes is 10% of its current value.

We create the initial population by randomly choosing the number of features to include in each feature subset. For classifier i, the size of each feature subset (N_i) is independently chosen from a uniform distribution between 1 and twice the number of original features in the dataset. We then randomly pick, with replacement, N_i features to include in classifier i's training set. Note that some features may be picked multiple times while others may not be picked at all; replicating inputs for a neural network may give the network a better chance to utilize that feature during training. Also, replicating a feature in a genome encoding allows that feature to better survive to future generations.

Our crossover operator uses dynamic-length, uniform crossover. In this case, we chose the feature subsets of two individuals in the current population proportional to fitness. Each feature in both parent's subset is independently considered and randomly placed in the feature set of one of the two children. Thus it is possible to have a feature set that is larger (or smaller) than the largest (or smallest) of either parent's feature subset. Our mutation operator works much like traditional genetic algorithms; we randomly replace a small percentage of a parent's feature subset with new features. With both operators, the network is trained from scratch using the new feature subset; thus no internal structure of the parents are saved during the crossover.

4 RESULTS

We tested the utility of combining our approach for generating numerous hierarchical theoretical descriptors of compounds with our approach for filtering these descriptors with GEFS by modeling 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.

Table 2 gives our results. We studied three approaches for modeling toxicity: (1) giving all theoretical descriptors to a neural network, (2) reducing the feature set in a traditional previously published (Gute & Basak 1997) manner, and (3) using our new genetic algorithm technique on the entire feature set to create a neural network ensemble. 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.

We first 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 all 95 input parameters, the neural networks 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.

Our first method for feature-set reduction follows the work of Gute and Basak 1997 on toxicity domains. Their method begins by using the VARCLUS method of SAS 1998 to select 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.

Table 2: Relative effectiveness of statistical and neural network methods in estimating LC_{50} of 69 benzene derivatives.

Method	R^2	Standard Error
NN with 95 inputs	0.868	0.29
VARCLUS	0.825	0.32
NN with GEFS	0.893	0.27

To further reduce the number of independent variables for model construction, the sets of topostructural and topochemical indices were further divided into subsets, 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. Their approach then applied linear regression to these 23 parameters. 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 on the whole data set used as a training set. Our leave-one-out experiment gave an $R^2 = 0.825$ and a standard error of 0.32.

Finally we applied our genetic algorithm technique, GEFS, using all 95 parameters. The parameter settings for the networks in the ensemble were the same as the settings for the single networks in the first experiment. Parameter settings for the genetic algorithm portion of GEFS includes a mutation rate of 50%, a population size of 20, a $\lambda = 1.0$, and a search length of 100 networks (20 networks for the initial population and 80 networks created from crossover and mutation). While the mutation rate may seem high as compared with traditional genetic algorithms, certain aspects of our approach call for a higher mutation rate (such as the criterion of generating a population that cooperates as well as our emphasis on diversity); other mutation values were tried during our pilot studies. With this approach, we obtained a test-set correlation coefficient of $R^2 = 0.893$ and a standard error of 0.27; the initial population of 20 networks obtained a test-set

$R^2 = 0.835$ and a standard error of 0.31.

5 DISCUSSION AND FUTURE WORK

The correlation coefficient between the predicted value from the computational model and the target value derived from the toxicity test is an extremely informative metric of accuracy in this case. The exact numeric value of most toxicity tests is not as important as the relative ordering and spread of these values. Thus, a perfect correlation ($R^2 = 1.0$) between the computation model and target toxicity shows the computational model is as informative as the toxicity obtained from a battery of expensive and time-consuming tests – regardless of the standard error. Note the standard error of 0.27 is fairly good, given the toxicity measurements ranged from 3.04 to 6.37.

While the neural network technique and the standard data-reduction technique obtained decent correlation with measured toxicity, our ensemble technique was about 20% closer to perfect correlation. Note that GEFS produces an accurate initial population and that running GEFS longer with our genetic operators can further increase performance. Thus our approach can be viewed as an "anytime" learning algorithm. Such a learning algorithm should produce a good concept quickly, then continue to search concept space, reporting the new "best" concept whenever one is found (Opitz & Shavlik 1997). This is important since, for most hazard assessment, an expert is willing to wait for days, or even weeks, if a learning system can produce an improved model for predicting toxicity.

Our results demonstrate a very important point: that our method is able to accurately predict toxicity directly from structure. Compared to the actual battery of tests necessary to measure toxicity, a computer model is much cheaper, much faster, and does not have a negative impact on the environment. It is important to also note that the computer model does not have to be the final measurement for hazard assessment; additional tests can be run on compounds that are either flagged by the model, or require more tests by the nature of their use (such as a benzene derivative that may become a standard fuel). Not only can good computer models become filters, they will probably be the only viable option for processing all registered chemicals.

While the method proposed here has proven effective, there is much future work that needs to be completed. For instance, we plan to test our method on other data sets of chemical derivatives; investigate other ensemble feature selection techniques; investigate variants to our genetic algorithm approach, and finally investigate the utility of other descriptors, such as bio-descriptors.

6 CONCLUSIONS

In this paper we presented a novel approach for creating a computer model for hazard assessment. Our approach works by first extracting a hierarchy of theoretical descriptors derived from the structure of a compound, then filtering the numerous possible descriptors with a genetic algorithm approach to ensemble feature selection. We tested the utility of our approach by modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. Our results demonstrate the ability of our approach to accurately predict toxicity directly from structure. Thus our new algorithm further increases the applicability of computer models to the problem of predicting chemical activity directly from its structure.

Acknowledgements

This work was partially supported by National Science Foundation grant IRI-9734419, a University of Montana MONTS grant, U.S. Air Force grant F49620-96-1-0330, and is contribution number 246 for the Center for Water and the Environment of the Natural Resources Research Institute.

References

Alpaydin, E. 1993. Multiple networks for function learning. In *Proceedings of the 1993 IEEE International Conference on Neural Networks*, volume I, 27–32. San Fransisco: IEEE Press.

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.; Harriss, D.; and Magnuson, V. 1988. Polly 2.3. Copyright of the University of Minnesota.

Bauer, E., and Kohavi, R. 1998. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*.

Bondi, A. 1964. Van der waals volumes and radii. *J. Phys. Chem.* 68:441–451.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

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.

de Waterbeemd, H. V. 1995. Discriminant analysis for activity prediction. In *Chemometric Methods in Molecular Design*, 283–294. VCH Publishers, Inc.

Dearden, J. 1990. Physico-chemical descriptors. In *Environmental Chemistry and Toxicology*, 25–59. Kluwer Academic Publisher.

Freund, Y., and Schapire, R. 1996. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156. Morgan Kaufmann.

Gute, B., and Basak, S. 1997. Predicting acute toxicity (LC50) of benzen derivatives using theoretical molecular descripors: 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.

Johnson, J. 1998. Pact triggers tests: Thousands of chemicals may be tested under toxicity screening program. *Chemical Engineering News* 76(44):19–20.

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

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. In Tesauro, G.; Touretzky, D.; and Leen, T., eds., *Advances in Neural Information Processing Systems*, volume 7, 231–238. Cambridge, MA: MIT Press.

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. Providence, RI: AAAI/MIT Press.

Maclin, R., and Shavlik, J. 1995. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*.

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., and Shavlik, J. 1997. Connectionist theory refinement: Searching for good network topologies. *Journal of Artificial Intelligence Research* 6:177–209.

Opitz, D. 1999. Feature selection for ensembles. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

Quinlan, J. R. 1996. Bagging, boosting, and c4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 725–730. AAAI/MIT Press.

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.

APPENDIX 1.3 A comparative QSAR study of benzamidines complement-inhibitory activity…

ELSEVIER

Computers and Chemistry 000 (1999) 000–000

Computers
& Chemistry

www.elsevier.com/locate/compchem

A comparative QSAR study of benzamidines complement–inhibitory activity and benzene derivatives acute toxicity

Subhash C. Basak [a,*], Brian D. Gute [a], Bono Lučić [b], Sonja Nikolić [a,b], Nenad Trinajstić [a,b]

[a] *Natural Resources Research Institute, The University of Minnesota, Duluth, MN 55811, USA*
[b] *The Rugjer Boš*ković Institute, PO Box *1016*, HR-*10001* Zagreb. Croatia

Received 5 July 1999; accepted 20 September 1999

Abstract

A novel QSAR study of benzamidines complement–inhibitory activity and benzene derivatives acute toxicity is reported and a new efficient method for selecting descriptors is used. Complement–inhibitory activity QSAR models of benzamidines contain from one to five descriptors. The best, according to fitted and cross-validated statistical parameters, is shown to be the five-descriptor model. Models with a higher number of indices did not improve over the five-descriptor model. The benzene derivatives structure–toxicity models involve up to seven linear descriptors. Multiregression models, containing up to ten nonlinear descriptors, are also reported for the sake of comparison with previously obtained additivity models. Comparison with benzamidine complement–inhibitory activity models and with benzene derivatives toxicity models from the literature favors our novel approach. © 1999 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: QSAR study: Complement–inhibitory activity; Benzene; Five-descriptor model

1. Introduction

In our recent papers a hierarchical QSAR (quantitative structure–activity relationship) approach was used to model the complement–inhibitory activity of benzamidines (Basak et al., 1999a) and the acute aquatic toxicities of benzene derivatives (Gute and Basak, 1997; Basak et al., 1999c). The hierarchical QSAR approach uses topological (partitioned into topostructural and topochemical), geometric and quantum-chemical descriptors in a stepwise fashion to build increasingly more complex structure–property–activity models (Basak et al., 1997, 1999b). Now we report the use, with the same aim, of a new efficient approach for selecting the best QSAR models using multivariate regression (MR) (Lučić and Trinajstić, 1999; Lučić et al., 1999a) and a standard approach for variable selection and model generation used in CODESSA (Katritzky et al., 1999; Lučić et al., 1999b). Sometime ago Hansch and Yoshimoto (Hansch and Yoshimoto, 1974) carried out a QSAR study on the complement–inhibitory potency of benzamidines using their own approach. After 10 years, Hall et al. (Hall et al., 1984) carried out a QSAR study on the toxicities of benzene derivatives using de novo analysis (Free and Wilson, 1964; Kubinyi and Kehrhahn, 1976), and derived an additivity model for 66 compounds (they excluded three compounds as outliers). We will analyze their models and compare to ours.

[*] Corresponding author.

PII: S0097-8485(99)00059-5

Table 1
Observed and calculated (cross-validated, CV, and fitted, FIT) complement-inhibitory activities 1/log *C* of 105 benzamidines

No.	X	1/log *C*		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
1	2-CH_3	−0.444	−0.417	−0.419
2	3.4-$(CH_3)_2$	−0.425	−0.423	−0.424
3	H	−0.418	−0.424	−0.423
4	3-OH	−0.415	−0.439	−0.434
5	3-CF_3	−0.410	−0.378	−0.382
6	3-NO_2	−0.410	−0.392	−0.395
7	3-Br	−0.405	−0.399	−0.400
8	3-CH_3	−0.398	−0.399	−0.399
9	3-OCH_3	−0.397	−0.401	−0.401
10	3-$CH_2C_6H_5$	−0.373	−0.343	−0.346
11	3,5-$(CH_3)_2$	−0.361	−0.375	−0.369
12	3-OC_3H_7	−0.355	−0.358	−0.358
13	3-*i*-C_5H_{11}	−0.355	−0.344	−0.345
14	3-OC_4H_9	−0.351	−0.340	−0.341
15	3-C_4H_9	−0.338	−0.355	−0.353
16	3-$CH{=}CHC_6H_5$	−0.339	−0.324	−0.325
17	3-$OCH_2C_6H_5$	−0.331	−0.324	−0.324
18	3-$(CH_2)_2C_6H_5$	−0.330	−0.332	−0.331
19	3-OC_6H_{13}	−0.329	−0.318	−0.319
20	3-$O(CH_2)_4OC_6H_5$	−0.325	−0.286	−0.287
21	3-$O(CH_2)_2OC_6H_5$	−0.323	−0.314	−0.315
22	3-C_6H_5	−0.323	−0.366	−0.359
23	3-$O(CH_2)_3OC_6H_4$-4-COOH	−0.321	−0.296	−0.297
24	3-OC_5H_{11}	−0.320	−0.327	−0.326
25	3-O-*i*-C_5H_{11}	−0.318	−0.338	−0.335
26	3-$O(CH_2)_2OC_{10}H_7$-α	−0.312	−0.255	−0.262
27	3-$O(CH_2)_4OC_6H_4$-4-NH_2	−0.306	−0.288	−0.289
28	3-$(CH_2)_4C_6H_5$	−0.302	−0.315	−0.313
29	3-$O(CH_2)_3OC_6H_4$-4-NO_2	−0.301	−0.282	−0.282
30	3-$O(CH_2)_3OC_6H_4$-4-NH_2	−0.300	−0.298	−0.298
31	3-$(CH_2)_2$-4-C_5H_4N	−0.299	−0.318	−0.318
32	3-$O(CH_2)_3OC_6H_5$	−0.299	−0.295	−0.295
33	3-$O(CH_2)_3C_6H_5$	−0.296	−0.290	−0.290
34	3-$(CH_2)_2$-3-C_5H_4N	−0.294	−0.298	−0.298
35	3-$(CH_2)_4C_6H_4$-4-NHAc	−0.294	−0.281	−0.282
36	3-$(CH_2)_2$-2-C_5H_4N	−0.291	−0.300	−0.299
37	3-$O(CH_2)_3OC_6H_4$-2-NH_2	−0.283	−0.288	−0.288
38	3-$O(CH_2)_3OC_6H_4$-4-NHAc	−0.278	−0.270	−0.270
39	3-$(CH_2)_4$-3-C_5H_4N	−0.276	−0.284	−0.284
40	3-$O(CH_2)_4C_6H_5$	−0.276	−0.277	−0.277
41	3-$O(CH_2)_3OC_6H_4$-3-NHAc	−0.270	−0.260	−0.260
42	3-$O(CH_2)_3OC_6H_3$-3.4-Cl_2	−0.265	−0.271	−0.271
43	3-$O(CH_2)_3OC_6H_4$-3-NH_2	−0.265	−0.283	−0.283
44	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_4$-4-SO_2F	−0.265	−0.247	−0.247
45	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_5$	−0.265	−0.258	−0.258
46	3-$O(CH_2)_3OC_6H_4$-4-OCH_3	−0.262	−0.275	−0.274
47	3-$O(CH_2)_4OC_6H_4$-4-$NHCONHC_6H_4$-4-SO_2F	−0.260	−0.236	−0.237
48	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_3$-2-OCH_3-5-SO_2F	−0.260	−0.226	−0.227
49	3-$O(CH_2)_3OC_6H_4$-4-Cl	−0.257	−0.287	−0.286
50	3-$O(CH_2)_3OC_6H_4$-2-NO_2	−0.257	−0.279	−0.279
51	3-$O(CH_2)_3OC_6H_4$-3-NO_2	−0.257	−0.268	−0.268
52	3-$O(CH_2)_3OC_6H_4$-3-OCH_3	−0.256	−0.255	−0.255
53	3-$O(CH_2)_3OC_6H_4$-2-$NHCOC_6H_3$-2-Cl-6-SO_2F	−0.255	−0.247	−0.248
54	3-$O(CH_2)_3OC_6H_4$-2-$NHCONHC_6H_5$	−0.255	−0.260	−0.259
55	3-$O(CH_2)_2OC_6H_4$-2-$NHCONHC_6H_3$-2-Cl-5-SO_2F	−0.250	−0.246	−0.246

Data Reduction and Division of the Topological Indices

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 and other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of TIs was partitioned into two distinct sets: topostructural indices and topochemical indices. Topostructural indices are 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 like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These sets of the indices are shown in Table I.

To reduce the number of independent variables that were used for model construction in the smaller sets of compounds, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [38]. The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional. From each cluster we select the index most correlated with the cluster, as well as any indices which are poorly correlated with the cluster ($r < 0.70$). These indices are then used in model construction. The variable clustering and selection of indices is performed independently for both the topostructural and topochemical subsets.

III. DEVELOPMENT OF HIERARCHICAL QSAR MODELS

In the development of hierarchical QSAR models, between two and four sets of indices have been used. A schematic of this method is given in figure 1 and the SAS procedure REG is used to conduct the all-subsets regression analyses [38]. Final model selection from the all-subsets regression is based on the results for both RSQUARE and CP (Mallow's C_p statistic). The hierarchy begins with the simplest indices, the topostructural. After developing our initial model utilizing the topostructural indices, the level of complexity is increased one step. To the indices included in the best topostructural model, all of the topochemical indices are added and modeling is conducted using the combined set of parameters. Likewise, the indices included in the best model from this procedure are combined with the geometrical indices and modeling is conducted once again. Finally, in some studies we have included quantum chemical parameters calculated by MOPAC. The parameters are added to the best model selected from modeling with the combination of topostructural, topochemical and geometrical parameters, and all subsets regression is used to find the best-fit model. In some of our studies we have also used each level of the hierarchy individually to compare the results of using only one higher-level set, *e.g*, geometrical indices, alone to

[Insert Figure 1 here]

Table 2
69 benzene derivatives and their observed and calculated (cross-validated, CV, and fitted, FIT) fathead minnow toxicities, expressed as $-\log(LC_{50})$

No.	Compound	$-\log(LC_{50})$		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
1	Benzene	3.40	3.29	3.32
2	Bromobenzene	3.89	4.04	4.01
3	Chlorobenzene	3.77	3.75	3.75
4	Phenol	3.51	3.31	3.35
5	Toluene	3.32	3.51	3.49
6	1,2-Dichlorobenzene	4.40	4.33	4.33
7	1,3-Dichlorobenzene	4.30	4.10	4.12
8	1,4-Dichlorobenzene	4.62	4.80	4.77
9	2-Chlorophenol	4.02	4.01	4.01
10	3-Chlorotoluene	3.84	3.72	3.73
11	4-Chlorotoluene	4.33	4.11	4.13
12	1,3-Dihydroxybenzene	3.04	3.31	3.28
13	3-Hydroxyanisole	3.21	3.13	3.14
14	2-Methylphenol	3.77	3.62	3.62
15	3-Methylphenol	3.29	3.52	3.51
16	4-Methylphenol	3.58	3.64	3.64
17	4-Nitrophenol	3.36	3.68	3.66
18	1,4-Dimethoxybenzene	3.07	3.01	3.01
19	1,2-Dimethylbenzene	3.48	3.84	3.81
20	1,4-Dimethylbenzene	4.21	3.94	3.97
21	2-Nitrotoluene	3.57	3.70	3.69
22	3-Nitrotoluene	3.63	3.67	3.66
23	4-nitrotoluene	3.76	3.71	3.71
24	1,2-Dinitrobenzene	5.45	4.95	5.09
25	1,3-Dinitrobenzene	4.38	4.12	4.15
26	1,4-Dinitrobenzene	5.22	4.83	4.91
27	2-Methyl-3-nitroaniline	3.48	3.74	3.73
28	2-Methyl-4-nitroaniline	3.24	3.50	3.47
29	2-Methyl-5-nitroaniline	3.35	3.80	3.77
30	2-Methyl-6-nitroaniline	3.80	3.76	3.76
31	3-Methyl-6-nitroaniline	3.80	3.61	3.62
32	4-Methyl-2-nitroaniline	3.79	3.78	3.78
33	4-Hydroxy-3-nitroaniline	3.65	3.51	3.52
34	4-Methyl-3-nitroaniline	3.77	3.78	3.78
35	1,2,3-Trichlorobenzene	4.89	4.84	4.84
36	1,2,4-Trichlorobenzene	5.00	5.02	5.02
37	1,3,5-Trichlorobenzene	4.74	4.36	4.45
38	2,4-Dichlorophenol	4.30	4.53	4.52
39	3,4-Dichlorotoluene	4.74	4.46	4.48
40	2,4-Dichlorotoluene	4.54	4.57	4.56
41	4-Chloro-3-methylphenol	4.27	4.27	4.27
42	2,4-Dimethylphenol	3.86	3.74	3.76
43	2,6-Dimethylphenol	3.75	3.75	3.75
44	3,4-Dimethylphenol	3.90	3.90	3.90
45	2,4-Dinitrophenol	4.04	4.03	4.04
46	1,2,4-Trimethylbenzene	4.21	4.07	4.09
47	2,3-Dinitrotoluene	5.01	5.29	5.21
48	2,4-Dinitrotoluene	3.75	4.29	4.27
49	2,5-Dinitrotoluene	5.15	4.89	4.93
50	2,6-Dinitrotoluene	3.99	4.43	4.41
51	3,4-Dinitrotoluene	5.08	5.29	5.23
52	3,5-Dinitrotoluene	3.91	4.25	4.23
53	1,3,5-Trinitrobenzene	5.29	5.29	5.29
54	2-Methyl-3,5-dinitroaniline	4.12	4.23	4.22

Table 2 (Continued)

No.	Compound	$-\log(LC_{50})$		
		Observed	Calculated (CV)[a]	Calculated (FIT)[a]
55	2-Methyl-3,6-dinitroaniline	5.34	4.59	4.64
56	3-Methyl-2,4-dinitroaniline	4.26	3.97	4.00
57	5-Methyl-2,4-dinitroaniline	4.92	3.88	3.97
58	4-Methyl-2,6-dinitroaniline	4.21	4.76	4.72
59	5-Methyl-2,6-dinitroaniline	4.18	4.64	4.61
60	4-Methyl-3,5-dinitroaniline	4.46	4.33	4.34
61	2,4,6-Tribromophenol	4.70	4.98	4.82
62	1,2,3,4-Tetrachlorobenzene	5.43	5.55	5.53
63	1,2,4,5-Tetrachlorobenzene	5.85	5.76	5.77
64	2,4,6-Trichlorophenol	4.33	4.68	4.64
65	2-Methyl-4,6-dinitrophenol	5.00	4.45	4.48
66	2,3,6-Trinitrotoluene	6.37	6.39	6.38
67	2,4,6-Trinitrotoluene	4.88	5.32	5.26
68	2,3,4,5-Tetrachlorophenol	5.72	5.64	5.65
69	2,3,4,5,6-Pentachlorophenol	6.06	6.01	6.03

[a] CV and FIT values are calculated using Eq. (10).

which was achieved by the orthogonalization of descriptors, because in the orthogonal basis the computation of R is much faster and simpler (Lučić et al., 1995a,b,c; Lučić, 1997). Namely, in the case one has the MR model based on the set of I orthogonalized descriptors di ($i = 1, \ldots, I$), the correlation coefficient between the experimental values of modeled activity A and the values estimated by the model A^{est} can be calculated in a very simple way (Eq. (1)):

$$R = \left[\sum_{i=1}^{I} R_i^2 \right]^{1/2} \quad (1)$$

where Ri is the correlation coefficient between each orthogonalized descriptor di and the modeled activity A. For example, using this procedure it takes 28 CPU min on Hewlett-Packard 9000/E55 computer, which is configured as a server, to select the best MR model with five out of 104 descriptors among ~ 10^8 possible models.

3. Results and discussion

3.1. QSAR of benzamidines

The best one-descriptor structure–complement–inhibitory activity model of benzamidines obtained is:

$$1/\log C = -0.9332(\pm 0.0229) + 0.4395(\pm 0.0152)H^{V}$$
$$n = 105 \quad R = 0.943 \quad R_{cv} = 0.941 \quad S = 0.0195$$
$$S_{cv} = 0.0199 \quad F = 832 \quad (2)$$

where H^{V} is the graph-vertex complexity (Basak, 1987), n is the number of benzamidine derivatives considered, R is the correlation coefficient, R_{cv} is the leave-one-out (cross-validated) correlation coefficient, F is F-value, S is the standard error and S_{cv} is the cross-validated (leave-one-out) standard error of estimate (root-mean-square error), both with N-2 in the denominator. This model is only slightly better than the earlier obtained one-descriptor model, but with a different descriptor (Basak et al., 1999a):

$$1/\log C = -0.6428(\pm 0.0129) + 0.0490(\pm 0.0017)^{3D}W$$
$$n = 105 \quad R = 0.943 \quad R_{cv} = 0.940 \quad S = 0.0196$$
$$S_{cv} = 0.0200 \quad F = 824 \quad (3)$$

where ${}^{3D}W$ is the 3-D Wiener number for the hydrogen-suppressed structures computed using their geometric distance matrices (Bogdanov et al., 1989). Close to this model is a model with 3-D Wiener number computed for structures containing all atoms including hydrogens (Bosnjak et al., 1989) ($n = 105$, $R = 0.941$, $R_{cv} = 0.939$, $S = 0.0199$ $S_{cv} = 0.0203$).

The best two-descriptor model of the benzamidine structure-complement-inhibitory activity is:

$$1/\log C = -0.6878(\pm 0.0175) + 0.1327(\pm 0.0367)W + 0.1864(\pm 0.0380)^{3D}W$$
$$n = 105 \quad R = 0.950 \quad R_{cv} = 0.947 \quad S = 0.0184$$
$$S_{cv} = 0.0189 \quad F = 467 \quad (4)$$

where W is the 2-D Wiener number (Wiener, 1947). The best three-descriptor model is given by:

$$1/\log C = -0.6400(\pm 0.0239) + 0.1273(\pm 0.0355)W + 0.0103(\pm 0.0037)P_9 + 0.1698(\pm 0.0372)^{3D}W$$

Table 3
Descriptions of all considered descriptors and symbols of only those descriptors involved in the models

	Information index for the magnitude of distances between all possible pairs of vertices of a graph
	Mean information index for the magnitude of distance
W	Wiener index, the half-sum of the off-diagonal elements of the molecular distance matrix
	Degree complexity
H^{V}	Graph vertex complexity
	Graph distance complexity
	Information content of the distance matrix partitioned by frequency of occurrences of distance l
	Information content of the hydrogen-suppressed graph at its maximum neighborhood of vertices
	Order of neighborhood when ICr reaches its maximum value for the hydrogen-filled graph
	A Zagreb group parameter, the sum of square of degree over all vertices
	A Zagreb group parameter, the sum of cross-product of degrees over all neighboring (connected) vertices
IC$_r$	Mean information content of a graph based on the rth ($r = 0$–6) order neighborhood of vertices in a hydrogen-filled graph
SIC$_r$	Structural information content for rth ($r = 0$–6) order neighborhood of vertices in a hydrogen-filled graph
CIC$_r$	Complementary information content for rth ($r =$ 0–6) order neighborhood of vertices in a hydrogen-filled graph
	Path connectivity index of order $h = 0$–6
	Cluster connectivity index of order $h = 3$–6
	Chain connectivity index of order $h = 6$
	Path-cluster connectivity index of order $h = 4$–6
	Bond path connectivity index of order $h = 0$–6
$^{hb}\chi_{c}$	Bond cluster connectivity index of order $h = 3$–6
$^{hb}\chi_{ch}$	Bond chain connectivity index of order $h = 6$
	Bond path-cluster connectivity index of order $h = 4$–6
$^{h}\chi^{v}$	Valence path connectivity index of order $h = 0$–6
$^{hv}\chi_{c}$	Valence cluster connectivity index of order $h =$ 3–6
$^{hv}\chi_{ch}$	Valence chain connectivity index of order $h = 6$
$^{h}\chi^{v}_{pc}$	Valence path-cluster connectivity index of order $h = 4$–6
P_l	Number of paths of length $l = 0$–10
	Balaban's J index based on distance
	Balaban's J index based on relative electronegativities
	Balaban's J index based on relative covalent radii
	Balaban's J index based on bond types
	Energy of the highest occupied molecular orbital
	Energy of the second highest occupied molecular orbital
E_{lumo}	Energy of the lowest unoccupied molecular orbital
	Energy of the second lowest unoccupied molecular orbital
ΔH_{f}	Heat of formation
μ	Dipole moment
	van der WaalSs volume
$^{3D}W_{H}$	3-D Wiener index for the hydrogen-filled geometric distance matrix
^{3D}W	3-D Wiener index for the hydrogen-suppressed geometric distance matrix

$$n = 105 \quad R = 0.954 \quad R_{cv} = 0.949 \quad S = 0.0177$$

$$S_{cv} = 0.0185 \quad F = 335 \tag{5}$$

where P_9 is the path of length nine. P_9 could be omitted from Eq. (5) because the related value of error of regression coefficient is relatively large comparing to the value of regression coefficient. Then Eq. (5) simply converts into Eq. (4). The best four-descriptor model is:

$$\begin{aligned}1/\log C = & -0.6999(\pm 0.0194) + 0.1327(\pm 0.0354)W \\ & + 5.0332(\pm 1.2285)^{6}\chi^{b}_{ch} \\ & - 5.1120\ (\pm 1.2486)^{6}\chi^{v}_{ch} \\ & + 0.1885(\pm 0.0359)^{3D}W\end{aligned}$$

$$n = 105 \quad R = 0.957 \quad R_{cv} = 0.953 \quad S = 0.0170$$

$$S_{cv} = 0.0177 \quad F = 272 \tag{6}$$

where $^{6}\chi^{b}_{ch}$ and $^{6}\chi^{v}_{ch}$ denote the bond-chain and valence-chain connectivity indices of order six, respectively.

Hansch and Yoshimoto (Hansch and Yoshimoto, 1974) published, 25 years ago, the following four-descriptor model for benzamidine derivatives inhibiting complement (the model is given in their notation):

$$\begin{aligned}\log(1/C) = & \ 0.15(\pm 0.03)(\mathrm{MR} - 1.2) \\ & + 1.07(\pm 0.13)(\text{D-1}) + 0.52(\pm 0.28)(\text{D-2}) \\ & + 0.43(\pm 0.14)(\text{D-3}) + 2.425(\pm 0.12)\end{aligned}$$

$$n = 108 \quad R = 0.935 \quad S = 0.258 \tag{7}$$

where MR is the molar refractivity of substituents at positions 1 and 2, taken from the compilation by Hansch et al. (Hansch et al., 1973) or computed, while D-1, D-2, and D-3 are indicator variables for the presence or absence of three kinds of the substructural units in a given benzadimine. To compare fitted statistical parameters of our four-descriptor model (Eq. (6)) with those of model given by Eq. (7), we retransformed our results into a log $(1/C)$ scale used by Hansch and Yoshimoto. Thus, we obtained statistical parameters ($R = 0.941$ and $S = 0.237$) that are comparable with their result. However, Hansch and Yoshimoto considered 108 benzamidine derivatives and we only considered 105. This discrepancy is caused by problematic data for three compounds which in our case are discarded from the set of benzamidine derivatives (Basak et al., 1999a). But, the nature of descriptors used in these two types of models is different. Descriptors used by us are calculated solely from the structures of studied molecules while the Hansch–Yoshimoto parameters (molar refractivities of substituents) are experimentally-based.

Finally, the five-descriptor model is:

$$\begin{aligned}1/\log C &= 1.5264(\pm 0.3534) + 0.6323(\pm 0.0936)(\mathrm{IC})_2\\ &\quad - 1.6788(\pm 0.2720)(\mathrm{IC})_6 - 1.4540(\pm 0.2043)\\ &(\mathrm{SIC})_1 - 0.4239(\pm 0.0680)(\mathrm{CIC})_6 + 0.1286\\ &(\pm 0.0149)^{3\mathrm{D}}W\\ &n = 105 \quad R = 0.963 \quad R_{\mathrm{cv}} = 0.957 \quad S = 0.0158\\ &S_{\mathrm{cv}} = 0.0170 \quad F = 253 \qquad (8)\end{aligned}$$

where $(\mathrm{IC})_2$ and $(\mathrm{IC})_6$ denote the mean information content of structure based on the second- and sixth-order neighborhood of atoms, including hydrogens, in the structure, respectively, $(\mathrm{SIC})_1$ and $(\mathrm{CIC})_6$ are, respectively, the structural information content for the first order neighborhood and complementary information content for the sixth order neighborhood of atoms, including hydrogens, in the structure. $(\mathrm{IC})_r$, $(\mathrm{SIC})_r$ and $(\mathrm{CIC})_r$ are molecular complexity indices introduced some times ago by one of us (Basak, 1987) for use in predictive pharmacology and toxicology.

It is interesting to note that the 3-D Wiener number is present in all models given here, except in the very best model with a single descriptor, although is present in the next best single-descriptor model. This is not surprising because this descriptor has shown to be very useful in the structure–property–activity modeling (Bogdanov et al., 1989; Bosnjak et al., 1991; Mihalić and Trinajstić, 1991; Nikolić et al., 1991; Trinajstić, 1992).

The models containing more decriptors did not outperform the above five-descriptor model. Thus, the model with five-descriptors (Eq. (8)), selected from the initial set of descriptors, is the best QSAR model, according to the calculated cross-validated statistical parameters, for predicting the benzamidine structure–complement-inhibitory activity. This model is better than one-descriptor model previously obtained using hierarchical approach (Basak et al., 1999a). However, according to F-values one-descriptor models selected in this paper and our previous work (Basak et al., 1999a) appear to be better models than the model with five-descriptors. But, the F-value is calculated only from the fitted correlation coefficient R and taking into account the number of parameters optimized in the model. Because it is accepted (Ortiz et al., 1997) that the cross-validated statistical parameters give better evidence into the model quality than fitted statistical parameters, our final conclusions are based on cross-validated statistical parameters, although the prediction for compounds from an external data set would be the best way of model quality testing. A plot between the experimental and predicted values, calculated in the cross-validation procedure using Eq. (8), of $1/\log C$ is given in Fig. 2. Computed (fitted and leave-one-out cross-validated) $1/\log C$ values are given in Table 1.

3.2. *QSAR of benzene derivatives*

The best linear five-descriptor structure–toxicity model of benzene derivatives selected by CROMRsel program is:

$$\begin{aligned}&-\log(\mathrm{LC}_{50})\\ &= 5.2032(\pm 0.546) + 0.8488(\pm 0.106)P_9\\ &\quad + 1.7979(\pm 0.183)^4\chi^{\mathrm{v}}_{\mathrm{Pc}} - 0.4439(\pm 0.0523)E_{\mathrm{lumo}}\\ &\quad - 0.1379(\pm 0.0195)\mu - 0.2961(\pm 0.0927)^{3\mathrm{D}}W_{\mathrm{H}}\\ &n = 69 \quad R = 0.927 \quad R_{\mathrm{cv}} = 0.914 \quad S = 0.287 \quad S_{\mathrm{cv}} = 0.312\\ &F = 77 \qquad (9)\end{aligned}$$

where P_9 is the path of length nine, ${}^4\chi^{\mathrm{v}}_{\mathrm{Pc}}$ valence path-cluster connectivity index of order four, Elumo is the energy of the lowest unoccupied molecular orbital, μ is dipole moment, and ${}^{3\mathrm{D}}W$H is the 3-D Wiener number for the hydrogen-filled structures computed using their geometric distance matrices (Bogdanov et al., 1989). This model has two descriptors fewer than the best model obtained by hierarchical approach (see Gute and Basak, 1997) and ~~posses~~ almost the same statistical parameters.

possesses

The best linear seven-descriptor model is:

$$\begin{aligned}&-\log(\mathrm{LC}_{50})\\ &= 4.4100(\pm 0.809) + 0.8637(\pm 0.0988)P_9\\ &\quad + 2.5278(\pm 0.833)^2\chi^{\mathrm{v}} - 3.1248(\pm 0.655)^4\chi^{\mathrm{v}}\\ &\quad + 1.5628(\pm 0.372)^6\chi^{\mathrm{v}}_{\mathrm{Pc}} - 0.44157(\pm 0.051)E_{\mathrm{lumo}}\\ &\quad - 0.1364(\pm 0.018) - 0.34054(\pm 0.087)^{3\mathrm{D}}W_{\mathrm{H}}\\ &n = 69 \quad R = 0.940 \quad R_{\mathrm{cv}} = 0.925 \quad S = 0.262 \quad S_{\mathrm{cv}} = 0.291\\ &F = 66 \qquad (10)\end{aligned}$$

where ${}^2\chi^{\mathrm{v}}$ and ${}^4\chi^{\mathrm{v}}$ denote valence path connectivity indices of order two and four, respectively, and ${}^6\chi^{\mathrm{v}}_{\mathrm{Pc}}$ is the valence path-cluster connectivity index of order six. Other descriptors are the same as those from five-descriptor model (Eq. (9)). This model ($R^2 = 0.884$, $F = 66$, $S = 0.26$) is better than the seven-descriptor model obtained by hierarchical procedure (see Gute and Basak, 1997) ($R^2 = 0.863$, $F = 50$, $S = 0.30$), and one can see that these two models contain three identical descriptors: $P9$, ${}^{3\mathrm{D}}W$H, and μ. Fitted and cross-validated predicted values for all benzene derivatives obtained using Eq. (10) are given in Table 2. A plot between the experimental and predicted values, calcu-

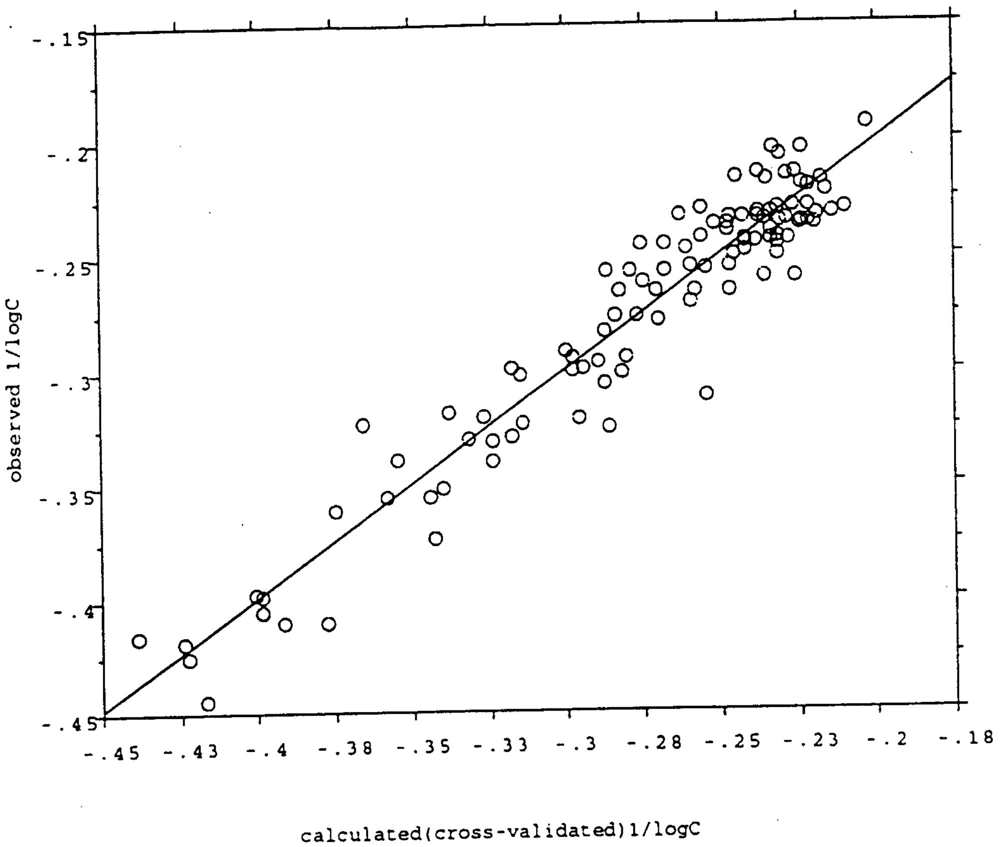


Fig. 2. A plot of observed versus calculated (cross-validated) 1/log *C* complement-inhibitory activity of benzamidines.

lated in the cross-validation procedure using Eq. (10), of $-\log(LC_{50})$ is given in Fig. 3.

We also found several seven-descriptor linear multiregression models with better statistical prameter than the best seven-descriptor model of Gute and Basak (see Gute and Basak, 1997). One of them is very similar to the model given as Eq. (10) and involving the following set of descriptors H^{v}, $P9$, ${}^{3}\chi_{c}^{b}$, $5\chi_{c}^{v}$, ΔH_{f}, μ, ${}^{3D}W_{H}$ (see Table 3 for description of descriptors), and possessing the following statistical parameters $R = 0.9398$, $R_{cv} = 0.9245$, $S = 0.262$, $S_{cv} = 0.292$, $F = 66$).

In addition, we perform modeling in order to compare our seven-descriptor model with the additivity model (using eight terms, i.e. eight optimized parameters) derived by Hall et al. (Hall et al., 1984). To do this we omitted from the data set compounds 53, 57 and 65, which were identified in by Hall et al. as outliers. For 66 compounds statistical parameters of seven-descriptor model (Eq. (10)) are: $R = 0.955$, $R_{cv} = 0.943$, $S = 0.225$, $S_{cv} = 0.255$ $F = 87$). This parameters are better than those for additivity models obtained by Hall et al. ($R = 0.951$, $S = 0.249$, $F = 67$).

4. Concluding remark

Presented results show that the optimum way to carry out QSAR modeling is by selecting the best descriptors in (linear, as was the case here, or nolinear (Lučić and Trinajstić, 1999) multiregression models.

Acknowledgements

This work was supported, in part, by Grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force and 00980606 from the Ministry of Science and Technology of the Republic of Croatia. We thank reviewers for helpful comments.

References

Basak, S.C., 1987. Use of molecular complexity indices in predictive pharmacology and toxicology. Med. Sci. Res. 15, 605–609.

Basak, S.C., Gute, B.D., Ghatak, S., 1999a. Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. J. Chem. Inf. Comput. Sci. 39, 255–260.

Basak, S.C., Gute, B.D., Grunwald. G.D., 1997. Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach. J. Chem. Inf. Comput. Sci. 37, 651–655.

Basak, S.C., Gute, B.D., Grunwald, G.D., 1999b. In: Devillers, J., Balaban, A.T. (Eds.), Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Reading, pp. 245–261.

Basak, S.C., Gute, B.D., Opitz, D.W., Balasubramanian, K., 1999c. Use of Statistical and Neural Net Methods in Predicting Toxicity of Chemicals: A Hierarchical QSAR Approach. Reported at the American Association of Artificial Intelligence (AAAI) Conference — Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools, Stanford University, March 22–24.

Bogdanov, B., Nikolić, S., Trinajstić, N., 1989. On the three-dimensional Wiener number. J. Math. Chem. 3, 299–309.

Bosnjak, N., Mihalić, Z., Trinajstić, N., 1991. Application of topographic indices to chromatographic data: calculation of the retention indices of alkanes. J. Chromatogr. 540, 430–440.

Free, S.M., Wilson, J.W., 1964. A mathematical contribution to structure-activity studies. J. Med. Chem. 1, 395–399.

Gute, B.D., Basak, S.C., 1997. Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. SAR QSAR Environ. Res. 7, 117–131.

Hall, L.H., Kier, L.B., 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., Leo, A., Unger, S.H., Kim, K.H., Nikaitani, D., Lien, E.J., 1973. Aromatic substituent constants for structure-activity correlations. J. Med. Chem. 16, 1207–1216

Hansch, C., Yoshimoto, M., 1974. Structure-activity relationships in immunochemistry. 2. Inhibition of complement by benzamidines. J. Med. Chem. 17, 1160–1167.

Katritzky, A.R., Chen, K., Wang, Y., Karelson, M., Lučić, B., Trinajstić, N., Suzuki, T., Schüürmann, G., 1999. Prediction of liquid viscosity for organic compounds by a quantitative structure-property relationship. J. Phys. Org. Chem. (in press).

Kubinyi, H., Kehrhahn, O.H., 1976. Quantitative structure-activity relationships: a comparison of different free-Wilson models. J. Med. Chem. 19, 1040–1045.

Kuby, J., 1992. Immunology. Freeman, New York.

Lučić, B., 1997. Ph. Dissertation, University of Zagreb. Zagreb. CROMRsel.f (CROatian MultiRegression selection of descriptors) is a computer program for the selection of descriptors for the best MR models.

Lučić, B., Amić, D., Trinajstić, N., 1999a. Nonlinear multivariate regression outperforms several concisely designed neural networks in QSAR modeling. J. Chem. Inf. Comput. Sci. (in press).

Lučić, B., Nikolić, S., Trinajstić, N., Juretić, D., 1995a. The structure-property models can be improved using the orthogonalized descriptors. J. Chem. Inf. Comput. Sci. 35, 532–538.

Lučić, B., Nikolić, S., Trinajstić, N., Juretić, D., Jurić, A., 1995b. A novel QSPR approach to physicochemical properties of the α-amino acids. Croat. Chem. Acta 68, 435–450.

Lučić, B., Nikolić, S., Trinajstić, N., Jurić, A., Mihalić, Z., 1995c. A structure-property study of the solubility of aliphatic alcohols in water. Croat. Chem. Acta 68, 417–434.

Lučić, B., Trinajstić, N., 1999. Multivariate regression outperforms several robust architectures of neural networks in QSAR modeling. J. Chem. Inf. Comput. Sci. 39, 121–132.

Lučić, B., Trinajstić, N., Sild, S., Karelson, M., Katritzky, A.R., 1999b. A new efficient approach for variable selection based on multiregression: rediction of gas chromatographic retention times and response factors. J. Chem. Inf. Comput. Sci. 39, 610–621.

Mihalić, Z., Trinajstić, N., 1991. The algebraic modelling of chemical structures: on the development of three-dimensional molecular descriptors. J. Mol. Struct. (Theochem.) 232, 65–78.

Nikolić, S., Trinajstić, N., Mihalić, Z., Carter, S., 1991. On the geometric distance matrix and the corresponding structural invariants of molecular systems. Chem. Phys. Lett. 179, 21–28.

Ortiz, A.R., Pastor, M., Palomer, A., Cruciani, G., Gago, F., Wade, R.C., 1997. Reliability of comparative molecular field analysis models: effect of data scaling and variable selection using a set of human synovial fluid phospholipase A_2 inhibitors. J. Med. Chem. 40, 1136–1148.

Trinajstić, N., 1992. Chemical Graph Theory. second revised. CRC, Boca Raton, FL, pp. 262–269.

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

APPENDIX 1.4 A hierarchical approach to the development of QSAR models using topological, geometrical...

Topological Indices and Related Descriptors in QSAR and QSPAR, Gordon and Breach, in press

A Hierarchical Approach to the Development of QSAR Models Using Topological, Geometrical and Quantum Chemical Parameters

Subhash C. Basak*, Brian D. Gute and Gregory D. Grunwald
Natural Resources Research Institute, University of Minnesota Duluth, Duluth, Minnesota 55811

Abstract

A current trend in quantitative structure-property/activity relationship studies (QSPR/QSAR) studies is the use of theoretical molecular descriptors that can be calculated directly from molecular structure. One advantage of such descriptors is that they can be calculated for any chemical structure, real or hypothetical. Topological indices (TIs) or numerical graph invariants constitute an important subset of these theoretical descriptors. TIs are derived from different classes of weighted graphs, representing various levels of chemical structural information. They are numerical quantifiers of molecular topology and encode information regarding the size, shape, branching pattern, cyclicity, and symmetry of molecular graphs. The Wiener index, different types of connectivity indices, and complexity or information theoretic topological indices have been widely used in QSAR/QSPR research.

We have been involved in the use of TIs in QSAR/QSPR model development to estimate pharmacological, physicochemical, and toxicological properties of diverse sets of molecules. More recently, we have developed a hierarchical approach in the use of theoretical descriptors where topological, geometrical, and quantum chemical indices are used. The goal of this approach has been to use the simplest descriptors first and to only use more complex descriptors if necessary. For this reason, the TIs have been divided into two subsets: a) topostructural indices (TSIs), the topological indices which are defined on the skeletal molecular graph and which do not distinguish among the various atoms or bonds present in the molecule, and b) topochemical indices (TCIs), which explicitly encode information regarding atom and bond types.

In this chapter we will discuss the utility of TIs, geometrical indices, and quantum chemical parameters in hierarchical QSAR studies. The results of studies where the various levels of indices are used in estimating physicochemical, biological, and toxicological properties of different sets of molecules will be presented.

I. INTRODUCTION

A recent interest in pharmaceutical drug design and hazard assessment of chemicals is the prediction of environmental, physicochemical, toxicological, and pharmacological properties of chemicals directly from their structure [1-11]. Early quantitative structure-structure activity relationship (QSAR) studies by Hansch and others used physical properties and physicochemical substituent constants for the prediction of other more complex physicochemical, biomedicinal and toxicological properties [12]. Such property-property correlation is useful only when properties necessary for prediction are available for all chemicals under consideration. In the field of environmental risk assessment, most chemicals do not have the data required for proper hazard estimation [13]. In contemporary drug design, one can produce large (real or virtual) combinatorial libraries of chemicals for screening. Most of these chemicals will have no physicochemical data and predictive methods based on experimental data are of no use in this situation. Therefore, there is a need for the development of QSAR methods using nonempirical parameters, *i.e.*, parameters that can be calculated from the molecular structure. Topological indices (TIs), the various molecular size and shape indices as well as quantum chemical parameters fall in this category.

Recently we have developed a new hierarchical approach to QSAR using parameters which are algorithmically defined, *i.e.*, which can be computed from structure using computer software [14-19]. We have successfully used four classes of computed parameters, *viz.*, topostructural, topochemical, geometrical, and quantum chemical parameters, in the development of QSAR models using a hierarchical approach (*vide infra*). This approach was found to be quite useful in the estimation of different properties.

In this chapter we will review the results of our hierarchical QSAR studies pertaining to the prediction of physicochemical, biological, and toxicological properties of different groups of chemicals.

II. CALCULATION OF PARAMETERS

Computation of Topological Indices

Topological indices used in this study have been calculated by POLLY 2.3 [20] which calculates a total of 102 indices. These indices include the Wiener index [21], the connectivity indices of Kier and Hall [2], and Randic [22], information theoretic indices defined on distance matrices of graphs [23,24], a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [25-27], and Balaban's J indices [28-30]. Table I provides brief definitions for the indices included in this study.

Computation of Geometrical Indices

Van der Waal's volume, V_W, [31-33] was calculated using *Sybyl 6.2* [34]. The 3-D Wiener numbers [35] were calculated by *Sybyl* using an SPL (Sybyl Programming Language) program developed in our laboratory. 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 3.2.1* [36]. 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.

Computation of Quantum Chemical Parameters

The quantum chemical parameters E_{HOMO}, E_{HOMO1}, E_{LUMO}, E_{LUMO1}, ΔH_f, and m were calculated for all of the following semi-empirical Hamiltonians: AM1, PM3, MNDO, MINDO/3. These parameters were calculated by *MOPAC 6.00* in the *SYBYL* interface [37]. One difficulty was encountered in using the MINDO/3 Hamiltonian.

Data Reduction and Division of the Topological Indices

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 and other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of TIs was partitioned into two distinct sets: topostructural indices and topochemical indices. Topostructural indices are 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 like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These sets of the indices are shown in Table I.

To reduce the number of independent variables that were used for model construction in the smaller sets of compounds, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [38]. The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional. From each cluster we select the index most correlated with the cluster, as well as any indices which are poorly correlated with the cluster ($r < 0.70$). These indices are then used in model construction. The variable clustering and selection of indices is performed independently for both the topostructural and topochemical subsets.

III. DEVELOPMENT OF HIERARCHICAL QSAR MODELS

In the development of hierarchical QSAR models, between two and four sets of indices have been used. A schematic of this method is given in figure 1 and the SAS procedure REG is used to conduct the all-subsets regression analyses [38]. Final model selection from the all-subsets regression is based on the results for both RSQUARE and CP (Mallow's C_p statistic). The hierarchy begins with the simplest indices, the topostructural. After developing our initial model utilizing the topostructural indices, the level of complexity is increased one step. To the indices included in the best topostructural model, all of the topochemical indices are added and modeling is conducted using the combined set of parameters. Likewise, the indices included in the best model from this procedure are combined with the geometrical indices and modeling is conducted once again. Finally, in some studies we have included quantum chemical parameters calculated by MOPAC. The parameters are added to the best model selected from modeling with the combination of topostructural, topochemical and geometrical parameters, and all subsets regression is used to find the best-fit model. In some of our studies we have also used each level of the hierarchy individually to compare the results of using only one higher-level set, *e.g*, geometrical indices, alone to

[Insert Figure 1 here]

determine the degree of contribution to modeling from the given set. Thus, there may be as many as seven final models in a hierarchical study to illustrate the individual contributions of the three higher-level sets of indices, as well as the four model from the stepwise procedure of the hierarchical modeling.

IV. HIERARCHICAL QSAR/QSPR STUDIES

The hierarchical method has been used in developing QSAR models for predicting a wide variety of properties. The following are examples from our previous studies employing the hierarchical approach in the construction of useful models.

Physicochemical Properties

Three large sets of chemicals have been used to model physicochemical properties, *viz.*, normal boiling point, lipophilicity (log*P*), and normal vapor pressure. The normal boiling point data was a subset of the Toxic Substances Control Act (TSCA) Inventory [13] for which measured normal boiling point data were available and where HB_1, a simple measure of the hydrogen bonding potential of a chemical, was equal to zero. This resulted in a set of 1023 diverse chemicals [14]. For this particular set, only the first three levels of the hierarchical approach were used, mainly due to the large amount of computational time necessary to generate quantum chemical parameters for a set of over 1000 chemicals. Eight topostructural indices were selected for the first model (Eq. 1). The second level of the hierarchy resulted in the retention of two of those topostructural indices and the addition of six topochemical indices (Eq. 2). Finally, the addition of geometric indices resulted in a ten parameter model using the two topostructural indices, the six topochemical indices, and two of the geometric indices (Eq.3). The results of this modeling are presented below (Eq. 1-3):

$$\begin{aligned} BP = {} & -21.9 + 30.6(W) - 21.5(O) + 69.9({}^{3}\chi) + 35.8({}^{6}\chi) - 106.5({}^{6}\chi_{C}) - 96.1({}^{5}\chi_{Ch}) \\ & - 17.7({}^{5}\chi_{PC}) + 19.5(P_{10}) \qquad \text{Eq.1} \\ & n = 1023,\ r^2 = 0.812,\ s = 39.7^\circ C,\ F = 547 \end{aligned}$$

$$\begin{aligned} BP = {} & -332.9 + 134.6({}^{6}\chi) + 10.9(P_{10}) + 110.0(IC_0) - 133.8({}^{6}\chi^{b}) - 80.2({}^{3}\chi^{b}{}_{c}) \\ & + 176.5({}^{0}\chi^{v}) + 44.8({}^{2}\chi^{v}) + 16.8({}^{5}\chi^{v}{}_{PC}) \qquad \text{Eq.2} \\ & n = 1023,\ r^2 = 0.961,\ s = 18.0^\circ C,\ F = 3151 \end{aligned}$$

$$\begin{aligned} BP = {} & -285.7 + 125.3({}^{6}\chi) + 10.6(P_{10}) + 74.5(IC_0) - 125.0({}^{6}\chi^{b}) - 86.3({}^{3}\chi^{b}{}_{c}) \\ & + 175.3({}^{0}\chi^{v}) + 49.1({}^{2}\chi^{v}) + 18.7({}^{5}\chi^{v}{}_{PC}) - 9.1({}^{3D}W_{H}) + 8.1({}^{3D}W) \qquad \text{Eq.3} \\ & n = 1023,\ r^2 = 0.963,\ s = 17.6^\circ C,\ F = 2650 \end{aligned}$$

From the three equations presented, it is clear that the replacement of six topostructural indices with six topochemical indices greatly enhanced the predictive power of the model, while the addition of the geometric parameters did not add much to the model. A scatterplot of experimental versus predicted boiling point from equation 3 is shown in figure 2.

[Insert Figure 2 here]

The lipophilicity data are a subset of 219 chemicals derived from the STARLIST set with log$P$ values between –2 to 5.5 obtained from CLOGP [39] and HB_1 equal to zero [14]. This subset was chosen to examine the effectiveness of model based on topological indices in the prediction of lipophilicity for compounds that do not have explicit hydrogen-bonding centers. Compounds were chosen within the range of log$P$ values described to avoid the problematic nature of compounds having exceptionally high values for lipophilicity. As with the boiling point models, only the first three levels of the hierarchy were applied to modeling lipophilicity. Seven topostructural indices were initially selected (Eq.4), and again, only two were retained with the addition of eight topochemical indices (Eq. 5). In equation 6, with the addition of two geometric parameters, an additional topostructural index is removed from the model. These equations (4-7) are presented below:

$$\log P = -1.42 + 1.08(\mathrm{W}) - 1.58({}^{2}\chi) + 1.51({}^{6}\chi) - 0.92({}^{6}\chi\mathrm{c}) - 0.32(\mathrm{P}_7) + 0.20(\mathrm{P}_{10}) + 1.97(\mathrm{J}) \quad \text{Eq. 4}$$

$$n = 219,\ r^2 = 0.789,\ s = 0.54,\ F = 112$$

$$\log P = -2.13 - 0.20({}^{2}\chi) + 0.18(\mathrm{P}_{10}) - 1.86(\mathrm{IC}_0) + 1.33(\mathrm{CIC}_2) - 0.92(\mathrm{CIC}_3) - 1.36({}^{6}\chi^{b}) + 5.76({}^{0}\chi^{v}) - 2.98({}^{1}\chi^{v}) + 0.54({}^{4}\chi^{v}) - 0.39({}^{3}\chi^{v}\mathrm{c}) \quad \text{Eq. 5}$$

$$n = 219,\ r^2 = 0.908,\ s = 0.36,\ F = 206$$

$$\log P = -5.60 + 0.19(\mathrm{P}_{10}) - 1.46(\mathrm{IC}_0) + 1.09(\mathrm{CIC}_2) - 0.77(\mathrm{CIC}_3) - 1.36({}^{6}\chi^{b}) + 5.34({}^{0}\chi^{v}) - 3.41({}^{1}\chi^{v}) + 0.55({}^{4}\chi^{v}) - 0.41({}^{3}\chi^{v}\mathrm{c}) + 1.10(\mathrm{V_W}) - 0.17({}^{3D}\mathrm{W}) \quad \text{Eq. 6}$$

$$n = 219,\ r^2 = 0.912,\ s = 0.35,\ F = 194$$

These three equations show similar results as those for the modeling of normal boiling point. The replacement of topostructural indices with an equal or greater number of topochemical indices results in marked improvement in the predictive power of the model, while the addition of geometric indices resulted in only a minor improvement. Figure 3 presents a plot of the experimental log$P$ values versus the log$P$ values predicted from equation 6. The 219 chemicals and their observed and predicted values for log$P$ have been presented previously in the literature [14].

[Insert Figure 3 here]

The 476 chemicals in the normal vapor pressure data [16] are a subset of the TSCA inventory taken from the ASTER (Assessment Tools for the Evaluation of Risk) database [40]. This is a diverse subset of chemicals all have vapor pressure (p_{vap}) data measured at 25°C and ranging between 3-10,000 mmHg.

The first three levels of the hierarchical method have been employed; however, the addition of geometric parameters to the modeling process did not result in the selection of a novel model and so there is no geometric model reported.

$$\log_{10}(p_{vap}) = 4.88 + 0.20(\mathrm{O}) - 2.56({}^{1}\chi) + 0.49({}^{4}\chi\mathrm{c}) + 0.79({}^{6}\chi\mathrm{c}) + 0.98(\mathrm{P}_{10}) \quad \text{Eq. 7}$$

$$n = 476,\ r^2 = 0.515,\ s = 0.53,\ F = 99.7$$

$$\log_{10}(p_{vap}) = 8.44 - 1.77({}^{1}\chi) + 1.25(\mathrm{P}_{10}) - 5.69(\mathrm{IC}_1) + 3.91(\mathrm{IC}_2) - 1.24(\mathrm{IC}_5) + 1.41({}^{3}\chi^{b}\mathrm{c}) - 1.70({}^{1}\chi^{v}) \quad \text{Eq. 8}$$

$$n = 476, r^2 = 0.793, s = 0.34, F = 224.0$$

As can be seen from equation 7, five topostructural indices were initially selected to model normal vapor pressure. The addition of the topochemical indices resulted in the retention of two topostructural indices and the addition of five topochemical indices (Eq. 8). As was seen for the other two physicochemical properties, *viz.*, normal boiling point and lipophilicity, the predictive power of the model is greatly enhanced by the addition of the topochemical indices. A scatterplot of experimental versus predicted normal vapor pressure, based on equation 8, is shown in Figure 4. These results are adequate, however, as can be seen from Figure 5 while the residuals show fairly uniform scatter when plotted against the dependent variable there are some significant outliers and the data tends to be somewhat skewed to the lower end of the vapor pressure range.

[Insert Figure 4 here]

[Insert Figure 5 here]

Biological Properties

Two smaller sets of congeneric chemicals have been used in the study of biological properties. The smaller of the two sets [19] consisted of sixty polycyclic aromatic hydrocarbons for which 24-hour dermal penetration (DP) data were available from the work of Roy *et al* [41]. For the purposes of this study, all four levels of the hierarchical method were employed. Only two equations are being presented since the addition of geometric and quantum chemical parameters to the modeling procedure did not result in the formulation of improved QSAR equations.

$$DP = 224.1 - 67.9(P_0) \qquad \text{Eq. 9}$$
$$n = 60, r^2 = 0.675, s = 7.4, F = 120.6$$

$$DP = 179.7 - 78.8({}^1\chi^b) \qquad \text{Eq. 10}$$
$$n = 60, r^2 = 0.695, s = 7.1, F = 132.0$$

Equation 9 shows the model resulting from the topostructural modeling. A one parameter model which explains 67.5% of the variance was generated. A small improvement is seen in the model resulting from the addition of the topochemical indices (Eq. 10), in which the topostructural index is replaced by the topochemical index, ${}^1\chi^b$. Figure 6 presents a scatterplot of experimental dermal penetration versus the predicted results from equation 10.

[Insert Figure 6 here]

The second set of biological data studied using the hierarchical method was a set of 107 benzamidines [18] that act as inhibitors of the complement system, collected from the literature by Hansch and Yoshimoto [42]. The base structure for the benzamidines is presented in figure 6 and the side-chains and activity values have been published previously [18]. The large size of these molecules made the calculation of quantum chemical indices prohibitively time consuming. As a result, the first three levels of the hierarchical modeling procedure were used for this study.

[Insert Figure 7 here]

$$1/\log C = 1.1245 + 0.4989(I^D) \qquad \text{Eq. 11}$$

$$n = 105,\ r^2 = 0.884,\ s = 0.0200,\ F = 785$$

$$1/\log C = -0.6428 + 0.0490(^{3D}W) \qquad \text{Eq. 12}$$
$$n = 105,\ r^2 = 0.889,\ s = 0.0196,\ F = 824$$

A single topostructural index provided a strong correlation with the inhibitory activity of these large compounds (Eq. 11). This one index modeled the activity so well, that the addition of topochemical indices did not add significantly to the predictive power of the model. Finally, with the addition of geometric parameters to the modeling of inhibitory activity, it was found that one geometric parameter provided a slightly better correlation with activity than did the topostructural index (Eq. 12), explaining 89% of the variance in the data. The results of this final model (Eq. 12) are shown in Figure 8 as a scatterplot of experimental versus predicted activity.

[Insert Figure 8 here]

Toxicological Properties

Two sets of compounds have been studied using the hierarchical modeling for toxicological properties. The first set consists of acute aquatic toxicity data for 69 benzene derivatives determined by the 96-hour fathead minnow toxicity test system [17]. This data was compiled by Hall, Kier, and Phipps [43] from eight literature sources and was supplemented by some original work conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota.

$$LC_{50} = -7.50 + 3.50(M_1) - 1.72(\overline{IC}) - 0.52(P_8) + 0.68(P_9) \qquad \text{Eq. 13}$$
$$n = 69,\ r^2 = 0.453,\ s = 0.58,\ F = 13.3$$

$$LC_{50} = 23.68 + 5.04(M_1) + 0.55(P_9) - 43.27(SIC_0) - 20.04(CIC_0) \qquad \text{Eq. 14}$$
$$n = 69,\ r^2 = 0.783,\ s = 0.36,\ F = 57.9$$

$$LC_{50} = 0.59 + 5.82(M_1) + 0.55(P_9) - 14.23(SIC_0) - 2.36(^{3D}W_H) \qquad \text{Eq. 15}$$
$$n = 69,\ r^2 = 0.792,\ s = 0.36,\ F = 61.1$$

$$LC_{50} = -3.83 + 5.97(M_1) + 0.77(P_9) - 8.26(SIC_0) - 1.98(^{3D}W_H) + 0.41(E_{LUMO1})$$
$$+ 0.01(\Delta H_f) - 0.12(\mu) \qquad \text{Eq. 16}$$
$$n = 69,\ r^2 = 0.863,\ s = 0.30,\ F = 55.0$$

Equation 13 shows the results of the initial modeling using topostructural indices. Even using four indices, the topostructural set did a poor job of modeling acute toxicity. The addition of topochemical indices led to a significant improvement in predictive power, with the replacement of two topostructural indices with topochemical indices (Eq. 14). The geometrical indices slightly improved the QSAR modeling (Eq. 15); however, it was the addition of quantum chemical indices which drastically improved the predictive power of our model (Eq. 16). The addition of quantum chemical indices increased the variance explained by 7.1% over the model including geometrical indices, resulting in an overall explanation of 86.3% of the variance. Figure 9 presents the scatterplot of experimental versus predicted toxicity for these 69 compounds based on the results of equation 16.

[Insert Figure 9 here]

A set of 520 compounds, 260 mutagens and 260 non-mutagens, was taken from the literature [44] as a source of mutagenicity data. These data provided qualitative assessments of mutagenicity based on a positive or negative result in the Ames' mutagenicity assay. A discriminant function analysis (DFA) was conducted on this set using the SAS procedure DISCRIM [38] to create a function capable of classifying the compounds as active or inactive. Based on the results of a previous study and the amount of time required for the calculations, the quantum chemical parameters were excluded and indicators of molecular fragments associated with mutagenic activity were included [15]. See the original manuscript for a further discussion of the data used in this study and the molecular fragments keyed for the analysis. These classification results, the indices used in each case, and brief notes on the fragment groups included in the final models are presented in Table II.

[Insert Table II here]

As can be seen in Table II, the topostructural indices alone correctly classify over 75% of the mutagens; however, they only correctly classify 57.3% of the non-mutagens. This leaves over 40% of the non-mutagens incorrectly classified. The combination of topostructural and topochemical indices results in a comparable classification rate for mutagens (74.6%) and a significant increase (5.8%) in the classification of non-mutagens. The addition of information regarding the presence or absence of known structural fragments associated with mutagenic activity results in a significant decrease (5.4%) in classification rate for mutagens, from 74.6% down to 69.2%. However, the addition of these structural fragments also increases the correct classification rate for non-mutagens increasing it from 63.1% to 71.9%, and overall increase of 8.7%. As a result of this dramatic increase in classification rate for non-mutagens, this model was retained and supplemented by the geometrical indices. Addition of the geometric indices brought the classification rate for mutagens up to 71.5% (an overall decrease of 4.7% from the topostructural model) and retained the classification rate for non-mutagens at 71.9% (an overall increase of 14.6% over the initial model). While these results are by no means spectacular, it is a reasonably accurate model for the prediction of mutagenic activity.

V. DISCUSSION

The goal of hierarchical QSAR studies is to investigate the relative roles of different classes of parameters, *viz.*, topostructural and topochemical indices, 3-D parameters and calculated quantum chemical parameters in predicting different types of molecular properties. It is clear from the results presented here that topostructural and topochemical indices explain most of the variance in the data for physicochemical, biological and toxicological properties. In most cases geometrical and quantum chemical indices make only marginal improvements in the predictive power of the models. This indicates that the easily calculable topostructural and topochemical indices will be an effective first choice in QSAR studies.

It is evident from these studies that the expanded levels of the hierarchical method are extremely useful for large, diverse sets of chemicals where there are many factors influencing the variation of properties between chemical structures. They are also useful in modeling the more complex biological interactions involving the modulation of toxicants. It is interesting to note that studies involving the inhibition of a specific enzymatic system or the passage of large compounds through the skin are modeled well using simply shape and size descriptors, and do not seem to benefit significantly from the addition of more complex indices. There is still a need for better descriptors that will help us to more accurately model complex biological and toxicological systems.

ACKNOWLEDGEMENT

This paper is contribution number 236 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

[1] Randic, M. Nonempirical Approaches to Structure-Activity Studies. *Int. J. Quantum Chem: Quant. Biol. Symp.* **1984**, *11*, 137-153.

[2] Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure-Activity Analysis. Research Studies Press: Letchworth, Hertfordshire, U.K, 1986.

[3] Rouvray, D. H.; Pandey, R. B. The Fractal Nature, Graph Invariants and Physicochemical Properties of Normal Alkanes. *J. Chem. Phys.* **1986**, *85*, 2286-2290.

[4] Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605-609.

[5] 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.

[6] Basak, S. C. Binding of Barbiturates to Cytochrome P_{450}: A QSAR Study Using Log P and Topological Indices. *Med. Sci. Res.* **1988**, *16*, 281-282.

[7] Basak, S. C. In *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology,* Karcher, W. and Devillers, J., Eds.; Kluwer Academic: Dordrecht/Boston/London, 1990; pp. 83-103.

[8] Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243-272.

[9] Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. Correlation Between Structure and Normal Boiling Points of Haloalkanes C_1-C_4 using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118-1121.

[10] 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.

[11] 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.

[12] Hansch, C. and Leo, A. (1995). *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology.* American Chemical Society, Washington, D.C., p. 557.

[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. and Grunwald, G.D. (1995). Predicting genotoxicity of chemicals using nonempirical parameters. In, *Proceedings of the XVI International Cancer Congress* (R.S. Rao, M.G. Deo, and L.D. Sanghui, Eds.). Monduzzi, Bologna, Italy, Vol. 7, pp 413-416.

[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 QSAR approach. *J. Chem. Inf. Comput. Sci.* **37**, 651-655.

[17] Gute, B.D. and Basak, S.C. (1997). Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.* **7**, 117-131.

[18] Basak, S.C., Gute, B.D., and Grunwald, G.D. (1998). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.* In press.

[19] Gute, B.D., Grunwald, G.D., and Basak, S.C. (1998). Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.* In press.

[20] Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1988). POLLY 2.3: Copyright of the University of Minnesota.

[21] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **69**, 17-20.

[22] Randic, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.* **97**, 6609-6615.

[23] 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.

[24] Bonchev, D., and Trinajstic, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **67**, 4517-4533.

[25] 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.

[26] Basak, S.C. and Magnuson, V.R. (1983). Molecular topology and narcosis: A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim. Forsch.* **33**, 501-503.

[27] 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.

[28] Balaban, A.T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89**, 399-404.

[29] Balaban, A.T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* **55**, 199-206.

[30] 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.

[31] Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem.* **68**, 441-451.

[32] 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.

[33] Moriguchi, I., and Kanada, Y. (1977). Use of van der Waals volume in structure-activity studies. *Chem. Pharm. Bull.* **25**, 926-935.

[34] *SYBYL Version 6.1.* (1994). Tripos Associates, Inc.: St. Louis, MO.

[35] Mekenyan, O., Peitchev, D., Bonchev, D., Trinajstic, 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.

[36] *CONCORD Version 3.0.1.* (1993). Tripos Associates, Inc.: St. Louis, MO.

[37] Stewart, J.J.P. (1990). MOPAC Version 6.00. QCPE #455. Frank J Seiler Research Laboratory: US Air Force Academy, CO.

[38] SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition.* SAS Institute Inc.: Cary, NC.

[39] Leo, A. and Weininger, D. (1984). *CLOGP Version 3.2 User Reference Manual.* Medicinal Chemistry Project, Pomona College, Claremont, CA.

[40] Russom, C.L.; Anderson, E.B.; Greenwood, B.E.; Pilli, A. (1991). ASTER: An integration of the AQUIRE data base and the QSAR system for use in ecological risk assessments. *Sci. Total Environ.* **109/110**, 667-670.

[41] 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.*, in press.

[42] Hansch, C.; Yoshimoto, M. (1974). Structure-activity relationships in immunochemistry. 2. Inhibition of complement by benzamidines. *J. Med. Chem.* **17**, 1160-1167.

[43] Hall, L.H., Kier, L.B., and Phipps, G. (1984). Structure-activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem.* **3**, 355-365.

[44] Soderman, J.V. (Ed.). (1982). *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*, CRC Press, Inc., Boca Raton, FL, Volume I, p 655.

Table I. Symbols, definitions and classifications of topostructural, topochemical, geometrical and quantum chemical descriptors.

	Topostructural
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
$\overline{IC}$	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
$^h\chi$	Path connectivity index of order h = 0-6
$^h\chi$	Cluster connectivity index of order h = 3-6
$^h\chi_{Ch}$	Chain connectivity index of order h = 3-6
$^h\chi_{PC}$	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
	Topochemical
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
$^h\chi^b$	Bond path connectivity index of order h = 0-6
$^h\chi_C^b$	Bond cluster connectivity index of order h = 3-6

${}^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3-6
${}^{h}\chi^{b}_{PC}$	Bond path-cluster connectivity index of order h = 4-6
${}^{h}\chi^{v}$	Valence path connectivity index of order h = 0-6
${}^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3-6
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3-6
${}^{h}\chi^{v}_{PC}$	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^{Y}	Balaban's J index based on relative covalent radii
Geometrical	
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
Quantum Chemical	
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO1}	Energy of the second highest occupied molecular orbital
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO1}	Energy of the second lowest unoccupied molecular orbital
ΔH_{f}	Heat of formation
μ	Dipole moment

Table II. Classification results for 520 mutagens/non-mutagens from DFA.

Model type	Indices included	% Mutagens correct	% Non-mutagens correct
topostructural	W, H^V, H^D, IC, $\overline{M_1}$, $^2\chi$, $^3\chi$, $^4\chi$, $^6\chi_C$, $^6\chi_{PC}$, P_{10}	76.2	57.3
topostructural + topochemical	H^D, M_1, $^2\chi$, P_{10}, IC_5, $^6\chi^b{}_{Ch}$, $^0\chi^v$, $^2\chi^v$, $^3\chi^v{}_{Ch}$, $^6\chi^v{}_{Ch}$, $^6\chi^v{}_{PC}$, J^X, J^B	74.6	63.1
topostructural + topochemical + fragments	H^D, M_1, $^2\chi$, P_{10}, IC_5, $^0\chi^v$, $^3\chi^v{}_{Ch}$, $^6\chi^v{}_{PC}$, J^B, nitroso[1], mustard[2], sulf[3], benz[4]	69.2	71.9
topostructural + topochemical + fragments + geometrical	H^D, M_1, $^2\chi$, P_{10}, IC_5, $^0\chi^v$, $^3\chi^v{}_{Ch}$, $^6\chi^v{}_{PC}$, J^B, nitroso[1], mustard[2], sulf[3], benz[4], V_W	71.5	71.9

[1] Nitroso- compounds.

[2] Halogenated substituted mustard, sulfur mustard or oxygen mustard.

[3] Organic sulfates or sulfonates.

[4] Biphenyl amine, benzidine or 4,4'-methylenedianiline derivatives.

Figure Legend:

Figure 1 Diagramatic representation of the first two stages in hierarchical QSAR model development from topological indices.

Figure 2 Scatterplot of experimental normal boiling point *vs* estimated normal boiling point using equation 3 for 1023 diverse chemicals.

Figure 3 Scatterplot of experimental log*P* *vs* estimated log*P* using equation 6 for 219 diverse chemicals.

Figure 4 Scatterplot of experimental normal vapor pressure *vs* estimated normal vapor pressure using equation 8 for 476 diverse chemicals.

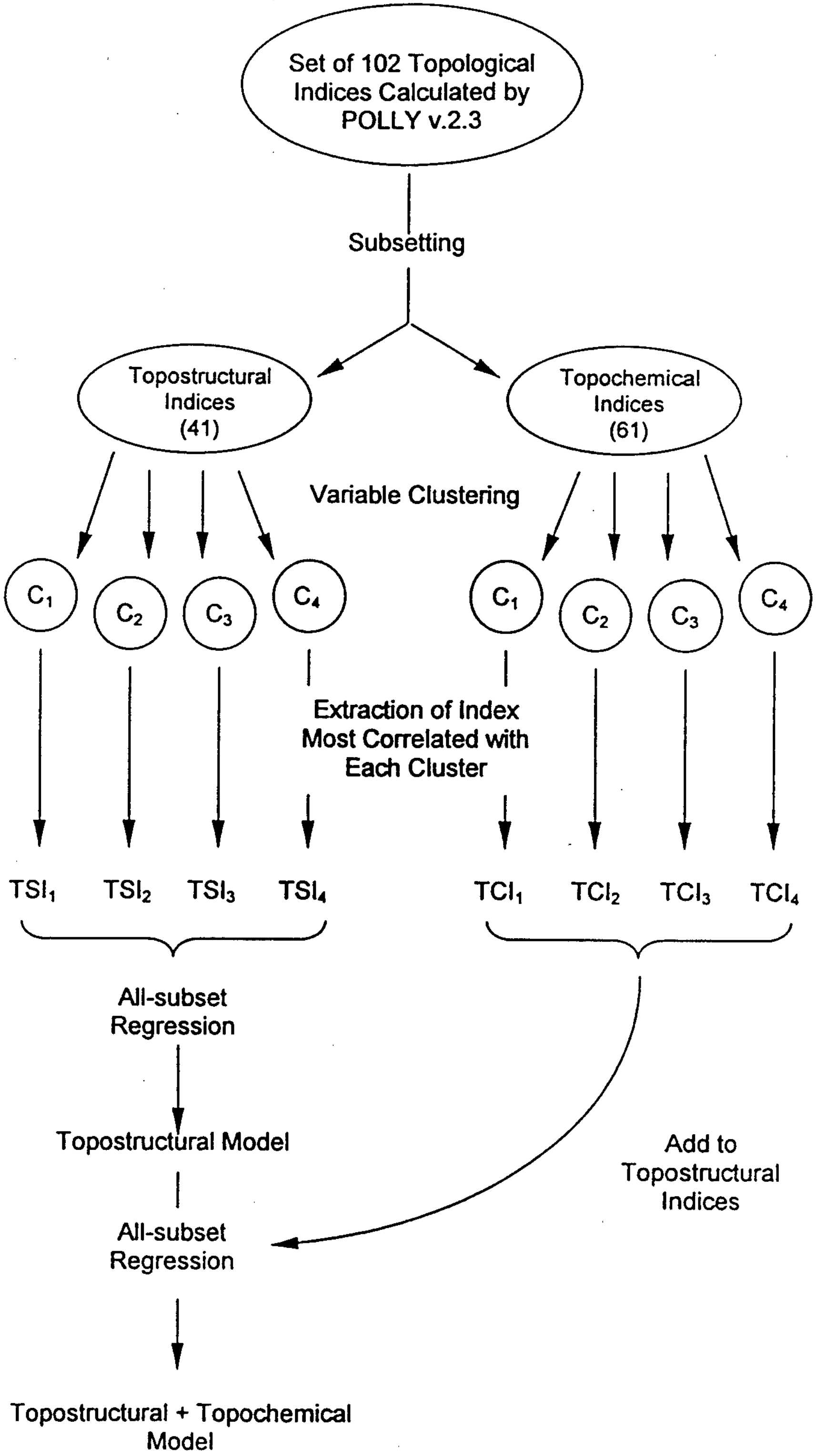
Figure 5 Scatterplot of the residual *vs* experimental normal vapor pressure from equation 8 for 476 diverse chemicals.

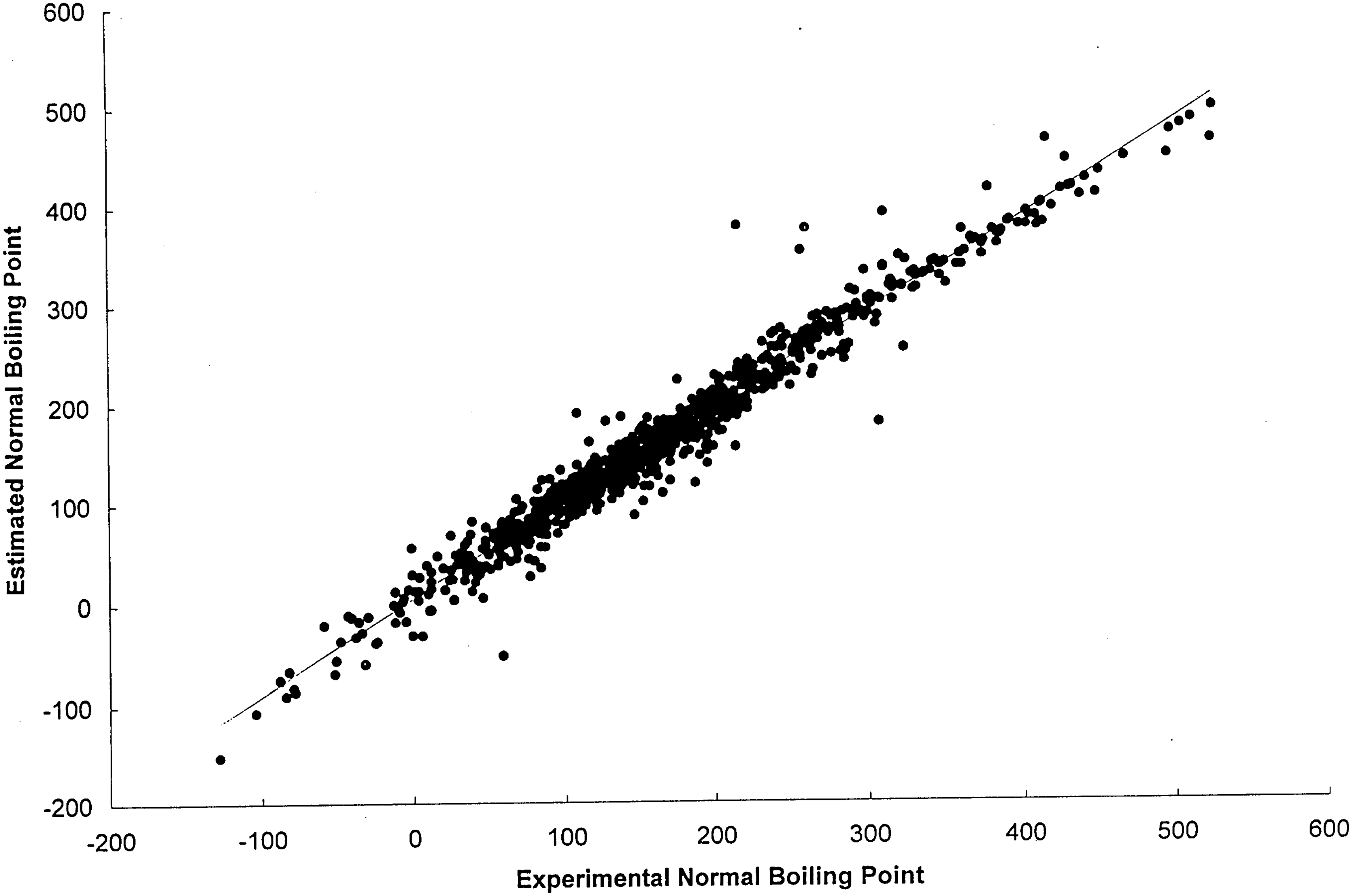
Figure 6 Scatterplot of experimental percent dermal penetration *vs* estimated percent dermal penetration using equation 10 for 60 polycyclic aromatic hydrocarbons.

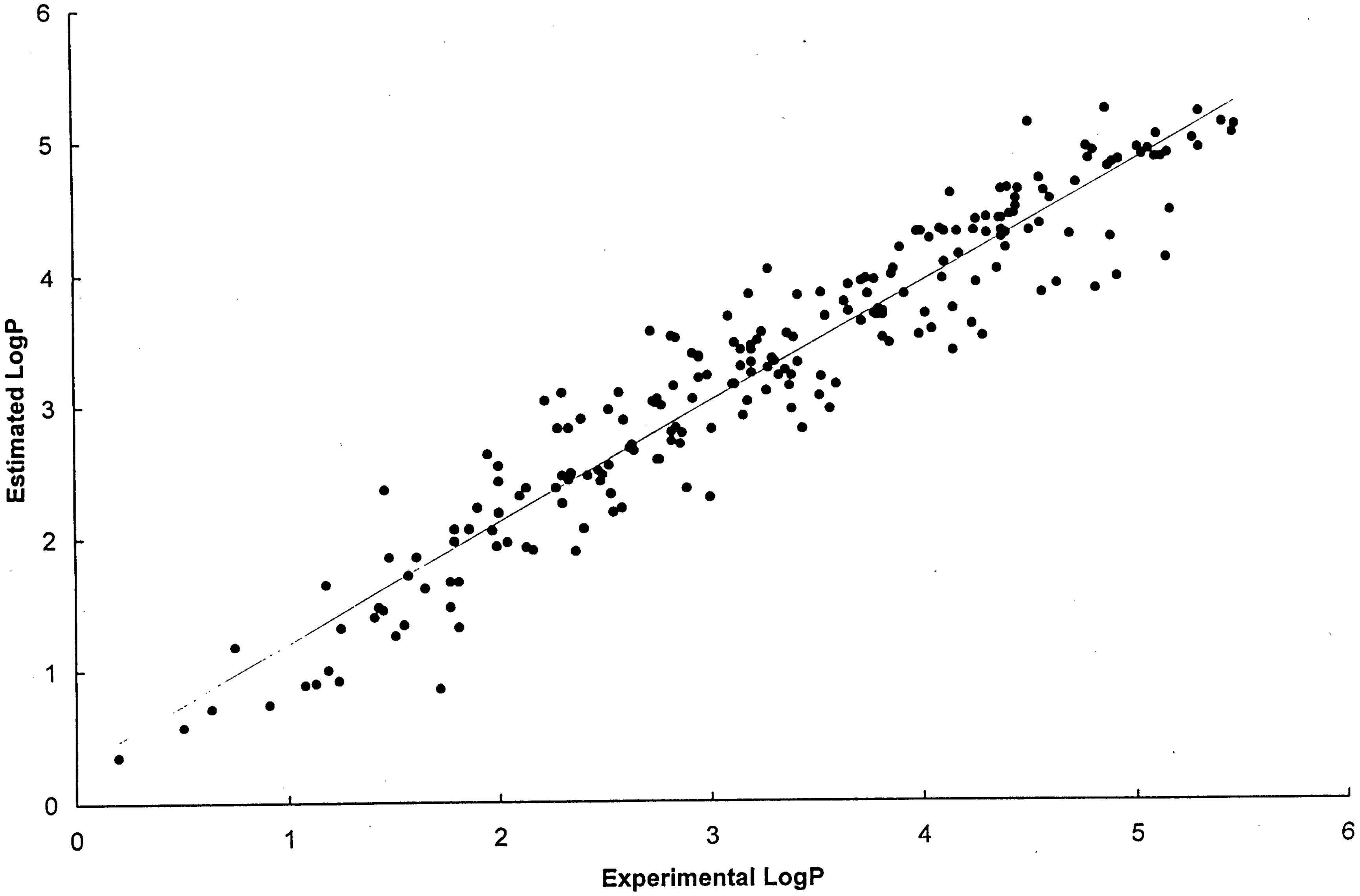
Figure 7 Neutral base structure for the 107 benzamidines.

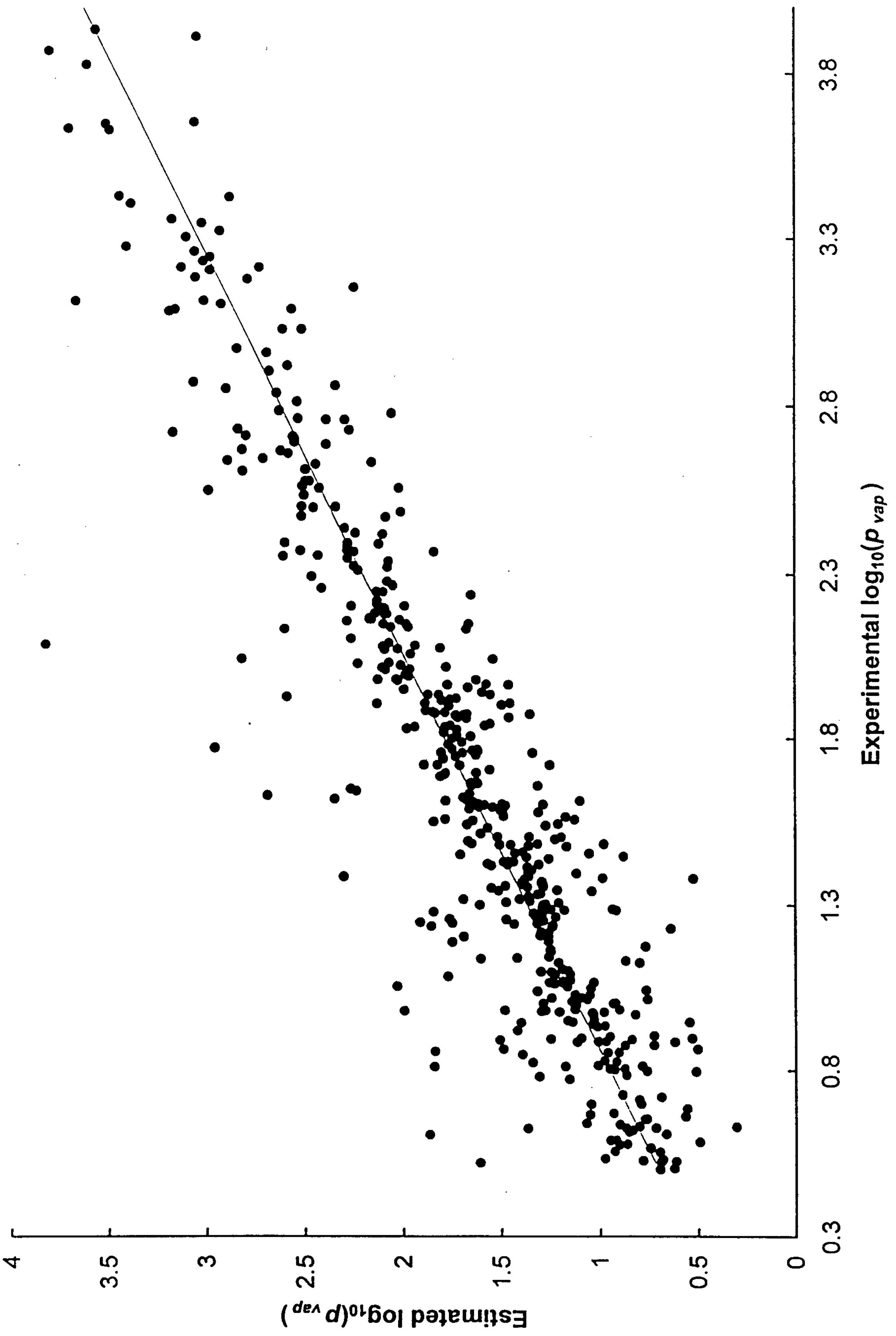
Figure 8 Scatterplot of experimental complement inhibition *vs* estimated complement inhibition using equation 12 for 105 benzamidines.

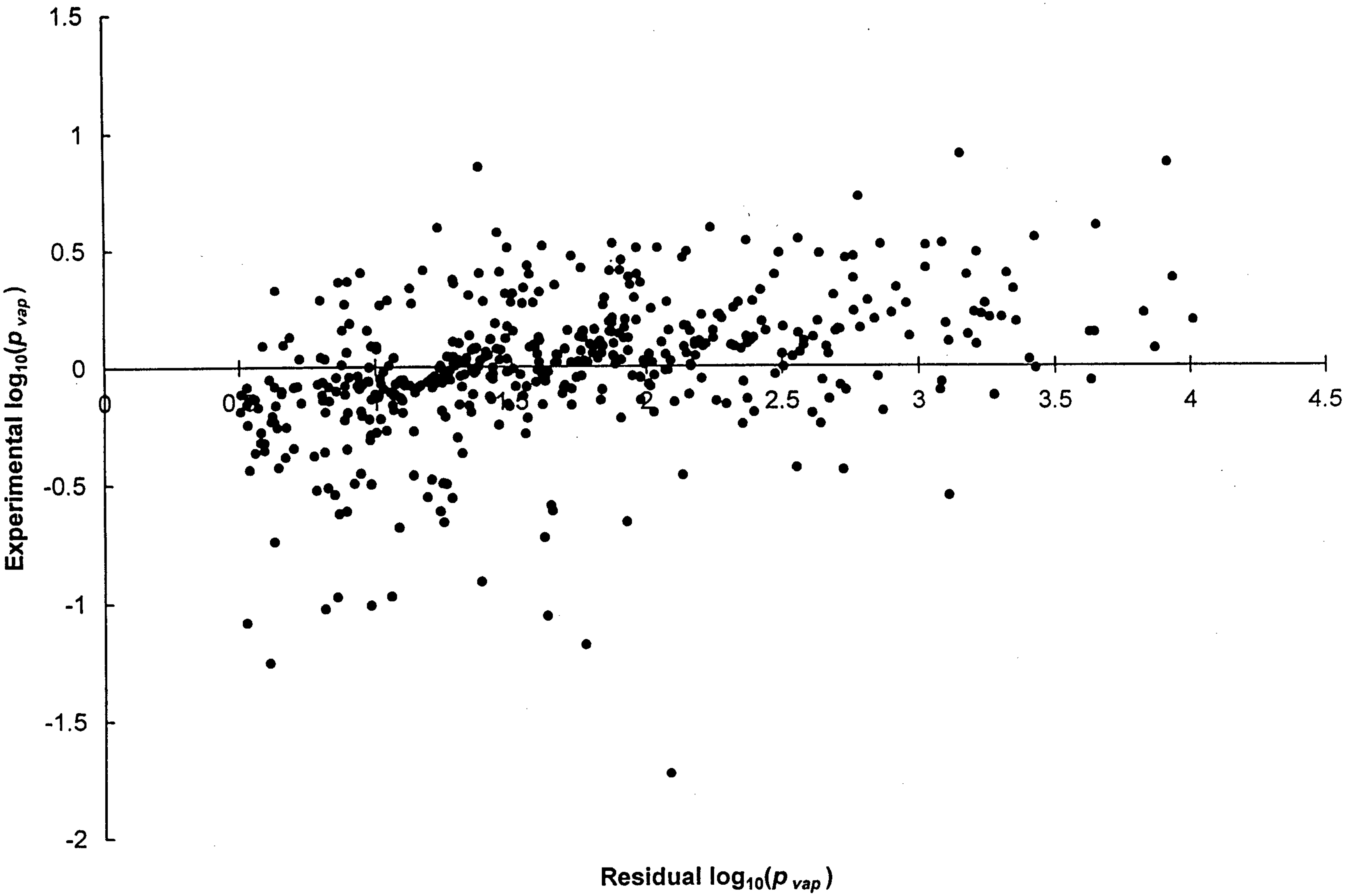
Figure 9 Scatterplot of experimental acute aquatic toxicity (LC_{50}) *vs* estimated acute aquatic toxicity using equation 16 for 69 benzene derivatives.

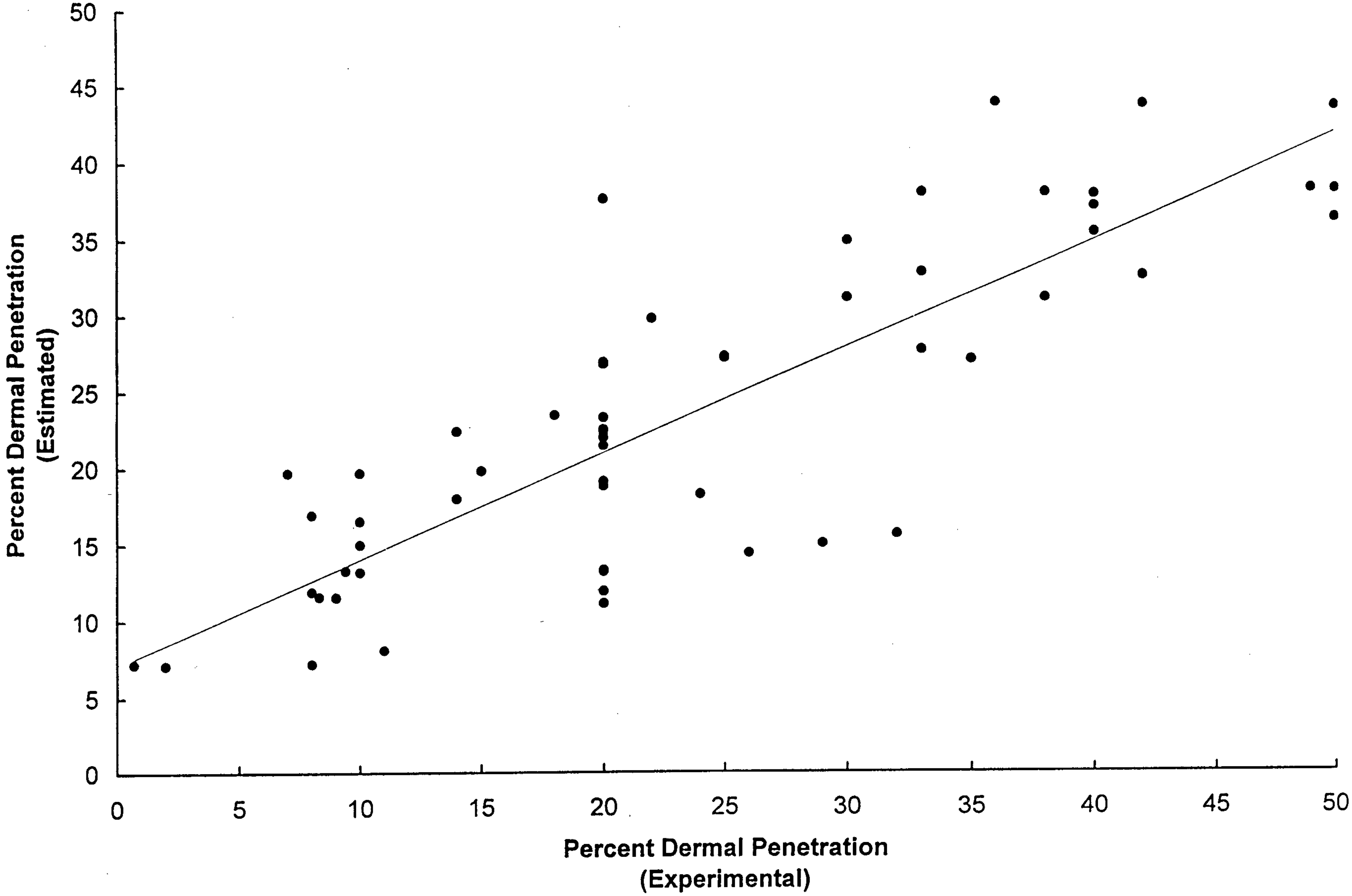

Set of 102 Topological Indices Calculated by POLLY v.2.3
Subsetting
Topostructural Indices (41)
Topochemical Indices (61)
Variable Clustering
C_1
C_2
C_3
C_4
C_1
C_2
C_3
C_4
Extraction of Index Most Correlated with Each Cluster
TSI_1
TSI_2
TSI_3
TSI_4
TCI_1
TCI_2
TCI_3
TCI_4
All-subset Regression
Topostructural Model
All-subset Regression
Add to Topostructural Indices
Topostructural + Topochemical Model

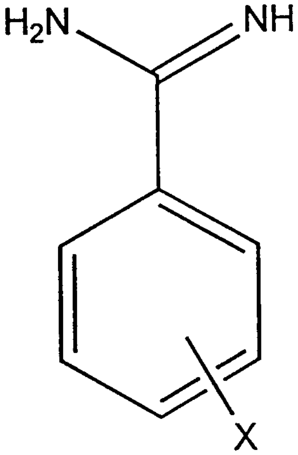

600
500
400
300
200
100
0
-100
-200
Estimated Normal Boiling Point
-200
-100
0
100
200
300
400
500
600
Experimental Normal Boiling Point

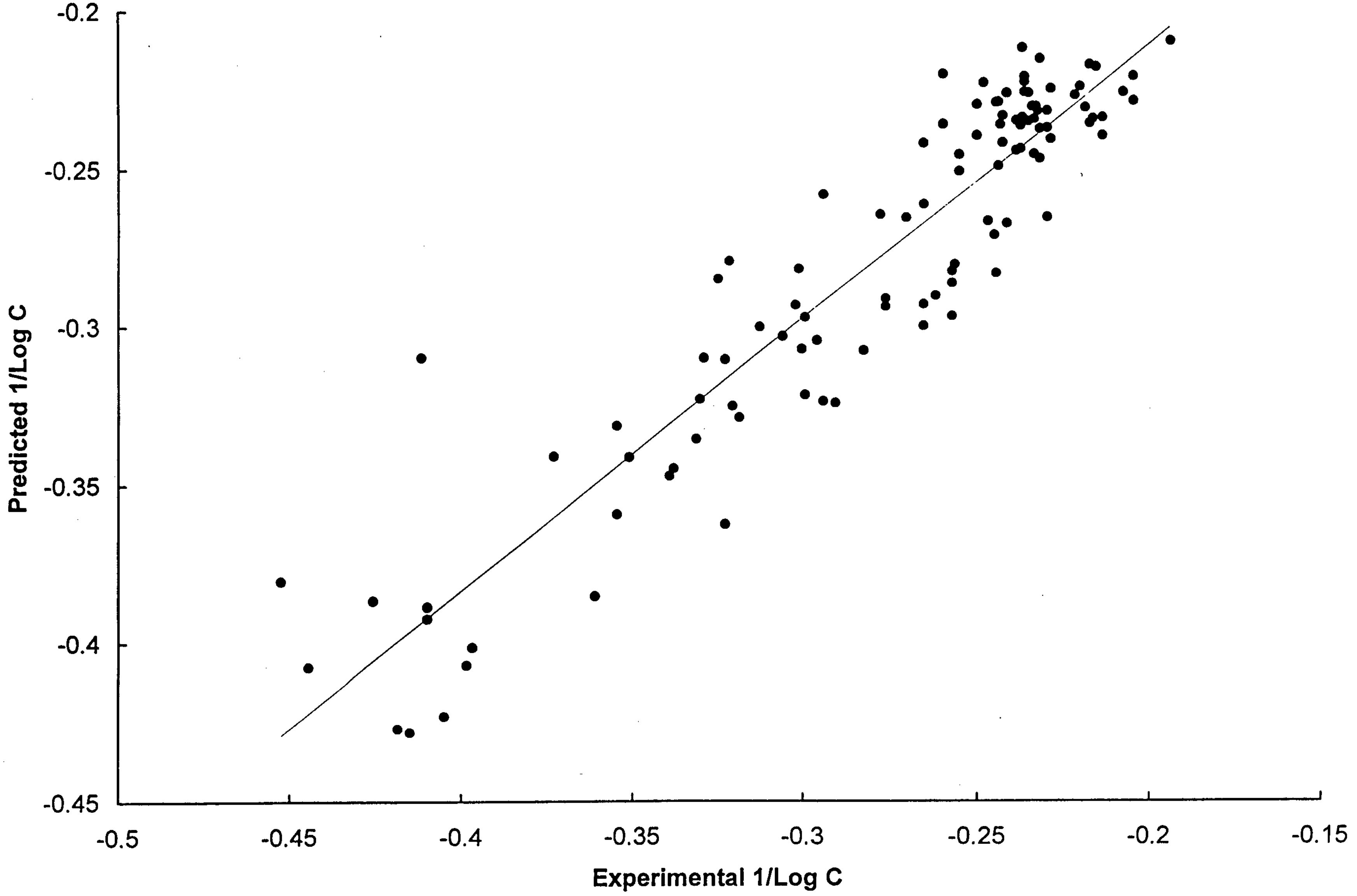

Estimated LogP
0
1
2
3
4
5
6
Experimental LogP
0
1
2
3
4
5
6


Estimated $\log_{10}(p_{vap})$
0
0.5
1
1.5
2
2.5
3
3.5
4
Experimental $\log_{10}(p_{vap})$
0.3
0.8
1.3
1.8
2.3
2.8
3.3
3.8


1.5
1
0.5
0
-0.5
-1
-1.5
-2
Experimental log$_{10}$(p_{vap})
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Residual log$_{10}$(p_{vap})


Percent Dermal Penetration
(Estimated)
50
45
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35
40
45
50
Percent Dermal Penetration
(Experimental)


H2N
NH
X


-0.2
-0.25
-0.3
-0.35
-0.4
-0.45
Predicted 1/Log C
-0.5
-0.45
-0.4
-0.35
-0.3
-0.25
-0.2
-0.15
Experimental 1/Log C

APPENDIX 1.5 Quantitative comparison of five molecular structure spaces in selecting analogs...

Math. Model. Comput. Sci
vol. 8, 1997

QUANTITATIVE COMPARISON OF FIVE MOLECULAR STRUCTURE SPACES IN SELECTING ANALOGS OF CHEMICALS

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

Natural Resources Research Institute, University of Minnesota - Duluth,
5013 Miller Trunk Highway, Duluth, MN 55811, USA
Phone: (218) 720-4230 E-Mail: sbasak@wyle.nrri.umn.edu

ABSTRACT

Five methods for characterizing intermolecular similarity have been used in the selection of analogs for a diverse set of seventy-six compounds. These methods include an atom pair (AP) based similarity measure, three principal component spaces derived from topostructural indices, topochemical indices, the combined set of all (topostructural and topochemical) indices, as well as one structure space consisting of principal components calculated from physicochemical properties. Each method has been used in the selection of sets analogs, ranging from five to forty in number in increments of five, for each of the seventy-six compounds. The degree of overlap of the sets of analogs selected by the five separate methods was analyzed.

KEYWORDS

molecular graph, atom pairs, principal components, analog selection, molecular similarity

INTRODUCTION

Molecular similarity is an intuitive concept which is subjectively understood by the chemist. In the realm of mathematical and computational chemistry, intermolecular similarity can be objectively quantified in terms of descriptors derived from the molecular structure (Basak et al, 1988b; Basak et al, 1997; Carbó et al, 1980; Fisanick et al, 1992; Fisanick et al, 1994; Johnson et al, 1988; Maggiora and Johnson, 1990; Randić, 1992; Willet and Winterman, 1986). Chemical structures can be represented by various types of models, *e.g.*, simple molecular graphs, multigraphs, pseudographs, 3-D models, and quantum chemical hamiltonian functions. Similarity, being context specific, is quantified in terms of a user-defined set of parameters or properties of molecules. Consequently, there are a potentially endless number of methods that one can define to quantify intermolecular similarity.

In recent years molecular similarity methods based on topological and substructural descriptors have become popular. Such methods are based on different types of graph invariants such as topological indices, atom pairs, and fragments (Basak and Grunwald, 1994, 1995c; Basak and Gute, 1997; Basak et al, 1988b; Carbó et al, 1980; Carhart et al, 1985; Fisanick et al, 1992; Johnson et al, 1988; Randić, 1992; Willet and Winterman, 1986). Similarity/dissimilarity methods have been used in the clustering of large sets of chemicals (Lajiness, 1990), the selection of analogs for toxicological risk assessment (Basak and Grunwald, 1994; Basak et al, 1995), and the estimation of the physicochemical and biomedicinal properties of chemicals (Basak and Grunwald, 1995a, 1995c; Basak et al, 1996a; Basak and Gute, 1997). Usually some number, *n*, of descriptors is used to define the structure space of chemicals and either Euclidean distance in the *n*-dimensional space or some association coefficient is used to quantify intermolecular similarity. The basic paradigm underlying molecular similarity analysis is "similar structures have similar properties." However, it has been shown that different molecular similarity methods select quite different sets of analogs from a specific database for the same set of query chemicals (Basak and Grunwald, 1995c). In the case of the automated selection of analogs for testing chemicals in drug design protocols or toxicological hazard assessment one would like to select analogs by reasonably non-redundant molecular similarity methods. Therefore, it is of interest to investigate the degree to which various similarity methods differ from each other. In a previous study we analyzed the analog selection profiles for topologically-based *vis-a-vis* empirical property-based molecular similarity techniques in the selection of nearest neighbors of molecules (Basak and Grunwald, 1995c). In this paper we have compared the analog selection profile of five different molecular similarity methods, four of which are based on graph invariants and one is derived from physicochemical property data.

DATABASE AND PARAMETERS

Development of the database

The data used in this study is a subset of the U.S. EPA ASTER system (Russom, 1992) which met the following criteria. These compounds have experimental values for:

1. Log $K_{o/w}$ Logarithm of the octanol/water partition coefficient (hydrophobicity).
2. BP Boiling point at 760 Torr.
3. MP Melting point.

within the ASTER database. Kamlet (1987) provided the remaining physicochemical properties used in this study. These four solvatochromic parameters are:

1. *V*/100 The molar volume of a molecule calculated as its molecular weight divided by the liquid density at 20° C.
2. α A measure of the hydrogen bond donor acidity of a compound in forming a hydrogen bond.
3. β A scale of the hydrogen bond acceptor basicity of a compound in forming a hydrogen bond.
4. π^* A measure of solute or solvent dipolarity or polarizability that quantifies the ability of a compound to stabilize a neighboring charge or dipole by virtue of its dielectric effect.

Kamlet et al (1988) describe in detail the methods used in the determination of these solvatochromic parameters.

Calculation of Atom Pairs

Atom pairs (APs) were calculated using the method of Carhart *et al* (1985). An atom pair is defined as a substructure which consists of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<descriptor$_i$>-<separation>-<descriptor$_j$>

where <descriptor> contains information about the element type, number of non-hydrogen neighbors, and the number of π electrons for each atom. The interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms. These calculations were conducted using the APProbe software developed by Basak and Grunwald (1993).

Calculation of Topological Indices

The topological indices used in this study have been calculated using the program POLLY 2.3 (Basak et al, 1988a) and software developed by the authors to calculate Balaban's *J* indices. A complete listing of these indices, along with examples of their calculation have been given in detail previously (Basak and Gute, 1997; Basak et al, 1997).

The topological indices were further divided into two subsets, topostructural and topochemical indices. Topostructural indices are topological indices which only encode information about the adjacency and distances of the vertices (atoms) within a graph (molecular structure), irrespective of the chemical nature of the atoms involved. The topochemical indices are parameters which quantify information regarding the topology of the graph (molecule), as well as specific chemical properties of the atoms and bonds comprising the molecule. These indices are derived from weighted graphs where each vertex (atom) or edge (bond) is properly weighted with selected chemical information. The division of the topological indices into these distinct sets has been discussed in previous studies (Basak et al, 1996b, 1997).

Similarity Measures

Two measures of intermolecular similarity were used in this study. The methods have been described in detail previously (Basak and Grunwald, 1995b) and include an associative measure using atom pairs (AP) and Euclidean distance (ED) within an *n*-dimensional principal component (PC) space. The Euclidean distance method was used in conjunction with the topological indices and the physicochemical property data.

ANALOG SELECTION

Following the quantification of intermolecular similarity for the five similarity spaces, the *K*-nearest neighbors or analogs ($K = 5, 10, 15, 20, 25, 30, 35, 40$) were determined on the basis of the associative measure used in conjunction with the AP method or based on ED within a principal component space.

RESULTS AND DISCUSSION

In generating the principal components for the sets of topological indices, only the principal components with eigenvalues greater than 1.0 were retained. This left six PCs for the set of topostructural indices which cumulatively explained 94.1% of the variance in the indices, eight PCs for the set of topochemical indices which explained 93.5% of the variance in these indices, and ten PCs for the set of all topological indices which cumulatively explained 95.2% of the variance in the topological indices. These formed the final sets of PCs which were used in creation of the similarity spaces and selection of analogs for these three methods.

Each similarity method was used to select sets of analogs for each of the seventy-six compounds in the dataset. The analogs selected by each set were compared with the analogs selected by every other method to examine the overlap between the sets of analogs. The results of this comparison are presented in Table 1 below as the arithmetic mean of the cardinalities of the intersection of subsets of analogs chosen by a particular pair of similarity methods for a specific value of *K*. For example, the topostructural and topochemical similarity methods selected an average of 2.2 identical analogs out of five for the entire set of seventy-six chemicals. Thus, slightly under half of the analogs selected by the two methods were identical.

It is clear from the data in Table 1 that the five molecular similarity methods studied in this paper are not radically different from one another because they have a substantial degree of overlap in the profile of selected neighbors. This is an interesting observation in view of the fact that the structure spaces are constructed from such diverse, independent variables as experimentally determined physicochemical properties and calculated graph invariants.

A perusal of the data also shows that the property-based similarity method is distinct from the group of methods based on topological indices and atom pairs. For $K = 20$, for example, the average number of common neighbors for the property-based methods *vis-a-vis* the topostructural, topochemical, all index and atom pair-based methods are 8.7, 8.9, 8.6 and 8.9, respectively. For the same value of *K*, the number of common analogs for the topostructural method with atom pair, topochemical and all index methods are 12.3, 12.2 and 13.1, respectively.

Table 1. Comparisons of the overlap in analog selection for five distinct similarity methods.

K	S vs C	S vs T	C vs T	S vs P	C vs P	T vs P	S vs A	C vs A	T vs A	P vs A
5	2.2	2.5	3.5	1.2	1.6	1.6	2.2	2.1	2.3	1.9
10	5.0	5.4	7.1	3.1	3.4	3.5	4.8	4.7	5.0	4.1
15	8.6	9.2	11.3	5.6	5.7	5.7	8.2	7.8	8.1	6.3
20	12.2	13.1	15.1	8.7	8.9	8.6	12.3	10.7	11.0	8.9
25	15.7	16.7	19.5	12.1	12.3	11.9	16.3	14.3	14.3	12.1
30	20.0	20.9	23.8	16.0	16.6	15.8	19.5	17.4	17.4	15.7
35	24.7	25.6	28.9	20.5	21.1	20.0	22.9	21.4	21.1	20.4
40	30.4	30.9	33.9	25.1	25.9	25.0	26.6	25.9	25.5	24.6

S = topostructural indices
C = topochemical indices
T = all topological indices
P = physicochemical parameters
A = atom pairs

For the three similarity methods calculated from the topological indices, the topochemical indices seem to have more influence on the selection of neighbors when they are used along with topostructural parameters as independent variables. This is clear from the fact that for almost all values of *K* the topochemical and all index methods have a uniformly higher degree of overlap as compared to that between the topostructural and all index methods.

In conclusion, if one is interested in selecting only two candidates from the set of five methods studied here for analog selection, the property-based method and any one of the theoretically-based methods would be the choice. There is no criteria to decide which of the four topologically-based methods should be selected for a particular occasion. Further studies of the analog selection and property prediction profile of these methods are necessary to guide the selection of a specific method for a particular practical situation.

ACKNOWLEDGMENTS

This paper is contribution number 204 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

Basak, S. C., S. Bertelsen and G. D. Grunwald (1995). Use of graph theoretic parameters in risk assessment of chemicals. Toxicol. Lett., 79, 239-250.

Basak, S. C. and G. D. Grunwald (1993). APProbe: Copyright of the University of Minnesota.

Basak, S. C. and G. D. Grunwald (1994). Molecular similarity and risk assessment: analog selection and property estimation using graph invariants, SAR QSAR Environ. Res., 2, 289-307.

Basak, S. C. and G. D. Grunwald (1995a). Estimation of lipophilicity from molecular structural similarity. New J. Chem., 19, 231-237.

Basak, S. C. and G. D. Grunwald (1995b). Molecular similarity and estimation of molecular properties. J. Chem. Inf. Comput. Sci., 35, 366-372.

Basak, S. C. and G. D. Grunwald (1995c). Use of topological space and property space in selecting structural analogs. Mathl. Model. Sci. Comput., in press.

Basak, S. C., B. D. Gute and G. D. Grunwald (1996a). Estimation of normal boiling points of haloalkanes using molecular similarity. Croat. Chem. Acta, 69, 1159-1173.

Basak, S. C., B. D. Gute and G. D. Grunwald (1996b). 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.

Basak, S. C. and B. D. Gute (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal *p*-hydroxylation of aniline by alcohols: a molecular similarity approach. In: Proceedings of the 2nd International Congress on Hazardous Waste: Impact on Human and Ecological Health (B.L. Johnson, C. Xintaras and J.S. Andrews, Jr., eds.), pp. 492-504, Princeton Scientific Publishing Co., Inc., New Jersey.

Basak, S. C., B. D. Gute and G. D. Grunwald (1997). Characterization of the molecular similarity of chemicals using topological invariants. In: Advances in Molecular Similarity: Highlights of the 3rd Girona Seminar on Molecular Similarity (P.G. Mezey, ed.), in press, JAI Press Inc, Greenwich, Connecticut.

Basak, S. C., D. K. Harriss and V. R. Magnuson (1988a). POLLY 2.3: Copyright of the University of Minnesota.

Basak, S. C., V. R. Magnuson, G. J. Niemi and R. R. Regal (1988b). Determining structural similarity of chemicals using graph theoretic indices. Discrete Appl. Math., 19, 17-44.

Carbó, R., L. Leyda and M. Arnau (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. Int. J. Quant. Chem., 17, 1185-1189.

Carhart, R. E., D. H. Smith and R. Venkataraghavan (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. J. Chem. Inf. Comput. Sci., 25, 64-73.

Fisanick, W., K. P. Cross and A. Rusinko, III (1992). Similarity searching on CAS registry substances. 1. global molecular property and generic atom triangle geometric searching. J. Chem. Inf. Comput. Sci., 32, 664-674.

Fisanick, W., A. H. Lipkus and A. Rusinko III (1994). Similarity searching on CAS registry substances. 2. 2D structural similarity. J. Chem. Inf. Comput. Sci., 34, 130-140.

Johnson, M., S. C. Basak and G. Maggiora (1988). A characterization of molecular similarity methods for property prediction. Mathl. Comp. Model., II, 630-634.

Kamlet, M. J. (1987). Personal communication.

Kamlet, M. J., R. M. Doherty, M. H. Abraham, Y. Marcus and R. W. Taft (1988). Linear solvation energy relationships. 46. An improved equation for correlation and prediction of octanol/water partition coefficients of organic nonelectrolytes (including strong hydrogen bond donor solutes). J. Phys. Chem., 92, 5244-5255.

Lajiness, M. (1990). Molecular similarity-based methods for selecting compounds for screening. In: Computational Chemical Graph Theory (D.H. Rouvray, ed.), pp. 299-316, Nova, New York.

Maggiora, G. M. and M. A. Johnson (1990). Introduction to molecular similarity. In: Concepts and Applications of Molecular Similarity (M. A. Johnson and G. M. Maggiora, eds.), pp. 1-13, John Wiley & Sons, Inc., New York.

Randić, M. (1992). Similarity based on extended basis descriptors. J. Chem. Inf. Comput. Sci., 32, 686-692.

Russom, C. L. (1992). Assessment Tools for the Evaluation of Risk, v. 1.0. U.S. Environmental Protection Agency.

Willett, P. and V. Winterman (1986). A comparison of some measures for the determination of inter-molecular structural similarity. Quant. Struct. -Act. Relat., 5, 18-25.

APPENDIX 1.6 Topological indices: Their nature and mutual relatedness

Journal of Chemical Information and Computer Sciences, submitted

Topological Indices: Their Nature and Mutual Relatedness

Subhash C. Basak,[a] Alexandru T. Balaban,[b] Gregory D. Grunwald,[a]
and Brian D. Gute [a]

[a] Natural Resources Research Institute, University of Minnesota Duluth, Duluth, Minnesota 55811
[b] Organic Chemistry Department, Polytechnic University Bucharest, 77206 Bucharest, Romania

We calculated 202 molecular descriptors (topological indices, TIs) for two chemical databases (a set of 139 hydrocarbons and another set of 1037 diverse chemicals). Variable cluster analysis of these TIs grouped these structures into 14 clusters for the first set and into 18 clusters for the second set. Correspondences between the same TIs in the two sets reveal how and why the various classes of TIs are mutually related and provide insight into what aspects of chemical structure they are expressing.

1. INTRODUCTION

A major part of the current research in mathematical chemistry, chemical graph theory, and quantitative structure-activity/property relationship studies involves topological indices. Topological indices (TIs) are numerical graph invariants that quantitatively characterize molecular structure. A graph G = (V, E) is an ordered pair of two sets V and E, the former representing a nonempty set and the latter representing unordered pairs of elements of the set V. When V represents the atoms of a molecule and elements of E symbolize covalent bonds between pairs of atoms, then G becomes a *molecular graph* (or *constitutional graph*, because there is no stereochemical information). Such a graph depicts the topology of the chemical species. A graph is characterized using graph invariants. An invariant may be a polynomial, a sequence of numbers, or a single number. A numerical graph invariant (i. e. a single number) which characterizes the molecular structure is called a topological index.

2. OVERVIEW OF TOPOLOGICAL INDICES USED IN THE PRESENT STUDY

A large number of topological indices have been defined and used.[1-11] The majority of TIs are derived from the various matrices corresponding to molecular graphs. The adjacency matrix **A**(G) and the distance matrix **D**(G) of the molecular graph G have been most widely used in the formulation of TIs. Integer-number local vertex invariants (LOVIs) are the vertex degrees (v_i) and the distance sums (distasums, d_i) resulting from summation over rows or columns of entries in the adjacency and distance matrices, respectively. By mathematical operations performed on such LOVIs, one can obtain a molecular descriptor, *i. e.*, a topological index. Wiener's index W (eq. 1),[2] the Zagreb group index M_1 (eq. 2),[11] Randić's connectivity index, χ (eq. 3),[4] the higher order connectivity indices, $^n\chi$, for paths of length n defined by Kier and Hall,[5] and the J index (eq. 4),[6] fall in this category.

$$W = (\Sigma_i d_i)/2 \qquad (1)$$

$$M_1 = \Sigma_i v_i^2 \qquad (2)$$

$$\chi = \Sigma_{ij} (v_i v_j)^{-1/2} \qquad (3)$$

$$J = [q / (\mu + 1)]\Sigma_{ij} (d_i d_j)^{-1/2} \qquad (4)$$

The summations in formulas (3) and (4) are over all edges i–j in the hydrogen-depleted graph. The numbers q of graph edges, and μ of cycles in the graph are introduced into formula (4) in order to avoid the automatic increase of J with graph size and cyclicity. Indeed, for an infinite linear carbon chain it was demonstrated that $J = \pi = 3.14159$. The nature of atoms can be taken into account by means of parameters based on their relative atomic numbers, electronegativities, or covalent radii, with respect to those of carbon atoms, multiplying the corresponding distasum in the formula (4) for J.

The mean square root distance D derived from all topological distances (denoted by i in the next formula) is defined as:[6b]

$$D = [(\Sigma_i i^2) / (\Sigma_i i)]^{1/2} \qquad (5)$$

For taking into account the chemical nature of atoms symbolized by vertices, Kier and Hall advocated the use of "valence connectivity indices".[5a,b] These are calculated with formulas similar to Randić's (eq. 3) but products of edge endpoint (or path vertex) invariants are no longer of vertex degrees but of weights (valence delta values δ_i) given by formula (5):

$$\delta_i = (Z_i^v - H_i)/(Z_i - Z_i^v - 1) \qquad (6)$$

where Z_i^v stands for the number of valence electrons in atom i, Z_i is its atomic number, and H_i is the number of hydrogen atoms attached to atom i.

The most recent additions to the Kier-Hall armamentary of TIs are electrotopological state indices.[5c]

Another class of molecular descriptors, the information-theoretic indices, are derived from an entirely different reasoning. In this case, the complexity or mode of partitioning of structural features is decomposed into disjoint subsets using an equivalence relation; a molecular complexity index is then computed using Shannon's idea of information content or complexity.[12] Real-number local vertex invariants (LOVIs), on the other hand, may also be defined starting from different matrices other than **A**(G) or **D**(G), or by applying information theory at the vertex level. Thus, topological indices U, V, X, and Y were defined. [13] Bonchev and Trinajstić described several information-theoretic TIs reviewed thoroughly in Bonchev's book. [7] The information-theoretic indices developed by Basak and coworkers take into account all atoms in the constitutional formula (hydrogens also being included), and one considers the information content provided by various classes of atoms based on their topological neighborhood.. There are three main types of informational indices developed by Basak et al: IC (mean information content or complexity of a hydrogen-filled graph, with vertices grouped into equivalence classes having r vertices; the equivalence is based on the nature of atoms and bonds, in successive neighborhood groups); CIC (complementary information content); and SIC (structural information content), and they are not intercorrelated with other TIs. In the following formula, the summation spans the range from i = 1 to i = r:

$$IC_r = - \Sigma_i p_i \log_2 p_I \qquad (10)$$

$$SIC_r = IC_r / \log_2 N \qquad (11)$$

$$CIC_r = \log_2 N - IC_r \qquad (12)$$

The probability that a randomly selected vertex occurs in the i-th equivalence class is denoted by p_i. The IC_r, SIC_r and CIC_r indices can be calculated for different orders of neighborhoods, r (r = 0, 1, 2,……ρ) where ρ is the radius of the molecular graph G. At the 0th order level, the atom set is partitioned based solely on their chemical nature; at the level of the first-order topological neighborhood, the atoms are partitioned into disjoint subsets based on their chemical nature and their first-order bonding topology. At the next level, the atom set is decomposed into equivalence classes using their chemical nature and bonding pattern up to the second-order bonded neighbors. The process is continued until consideration of higher-order neighbors does not yield further increase in the number or composition of disjoint subsets.

A large variety of real-number local vertex invariants, and thence a larger variety of TIs, were described based on converting a matrix (**A** or **D** for instance) into a system of linear equations. This is done by means of two column vectors that can convey topological, chemical, or numerical information. One non-zero vector is the free term of the system of equations. The other one (which may be zero, but this restricts the choices on available supplementary information) becomes the main diagonal of the matrix (if both vectors would be zero, then some negative LOVIs would result with difficulties of interpretation). These vectors may be the following integers: Z (atomic number of the atom corresponding to each vertex), V (vertex degree), I (identity), N (number of non-hydrogen atoms, or order of the graph), N^k (power k of N). Less frequently, one may use for periodicity of chemical properties real numbers: S (electronegativity) or R (covalent radius) of the atom corresponding to each vertex. The resulting matrix with the vector for the main diagonal constitutes the set of coefficients for the N unknowns which represent the real-number LOVIs of the N vertices. The triplet (matrix, vector for the main diagonal and vector for the free term) also serves as notation for LOVIs and for the derived TIs. After solving the system of N linear equations, the LOVIs (x_i) are assembled into a "triplet TI" based on one of the following operations:

1. Summation, $\Sigma_i x_i$;
2. Summation of squares, $\Sigma_i x_i^2$;
3. Summation of square roots, $\Sigma_i x_i^{1/2}$;
4. Sum of inverse square root of cross-product over edges ij, $\Sigma_{ij} x_i x_j)^{-1/2}$;
5. Product, $N[\Pi_i x_i]^{1/N}$.

Numbers 1 through 5 of the above operation after the triplet complete the notation of the triplet TIs.[14]

To conclude this brief review of TIs, one should mention recent progress that includes other matrices such as the reciprocal distance matrix which yields Harary indices,[15] the regressive distance matrices,[16] the Szeged matrix,[17] and the resistance distance matrix which affords Kirchhoff indices.[18] So-called optimal structural descriptors can be obtained from some TIs by varying some parameters and thereby adapting them to the data base;[19] alternatively, in Randić-type formulas (eqs. 3, 4) the exponent is allowed [20] to differ from ½. Three-dimensional molecular descriptors can be derived from geometrical and topological structural features of molecules.[21]

Each of the indices above discussed is a "global" parameter, *i.e.*, it quantifies certain aspects of the entire molecular structure using a single number.

It is clear from the above discussion that the set of TIs is a group of heterogeneous entities. They have been defined to characterize molecular structure based on distinct objectives and motivations. In spite of their distinctive characteristics, TIs share certain common features. A topological index maps a set of chemicals C into the set R of real or integer numbers. Therefore, TIs quantify some general aspects of molecular architecture like size, shape, symmetry, bonding type, cyclicity, branching pattern, etc.

Topological indices have been used for isomer discrimination, quantification of the structural similarity/ dissimilarity of molecules, and prediction of property/ activity from structure.[19] The widespread use of TIs obviously encourages one to ask some fundamental questions about them: What is the fundamental nature of TIs? To what degree are they intercorrelated? How does one extract orthogonal information from TIs?

The intercorrelation of TIs was studied earlier with a limited set of invariants. Thus, Motoc and Balaban [22] described graphically the intercorrelations of the few TIs known till 1981. These aspects were reviewed in the early 1980s.[23] Basak *et al.* studied the mutual relatedness of a set of ninety TIs calculated for a set of 3,692 diverse chemicals.[24] A third study by Todeschini *et al.* will be discussed in the last section of this paper.

All such studies were limited in the sense that they analyzed data on a smaller and less diverse group of TIs. Therefore, in this paper, we have studied the mutual relatedness of a set of 202 TIs. We have also tried to extract useful and orthogonal structural information from the calculated TIs. This study also reports, for the first time, a comprehensive discussion of Basak's information content indices (IC_r, SIC_r, CIC_r), the triplet indices (proposed by one of the present authors), and Balaban's average distance-based connectivity index J as compared to the traditional and more widely-used indices.

The goal of this paper is two-fold: (a) to study the degree of intercorrelation among the various types of topological indices, and (b) to extract mutually uncorrelated (orthogonal) topological parameters which can be used for QSAR/QSPR studies, quantitation of intermolecular similarity/ dissimilarity as well as characterization of real and virtual combinatorial libraries. To this end, we studied the mutual relatedness of a set of over two hundred topological indices in this paper.

3. METHODS

3.1 Chemical Databases. There were two sets of chemicals analyzed in this study: a set of 139 hydrocarbons to represent a moderately homogeneous set of chemicals and a set of 1037 diverse chemicals. The hydrocarbons consisted of 73 C3-C9 alkanes, 29 alkylbenzenes, and 37 polycyclic aromatic hydrocarbons.[25] The diverse set of 1037 compounds consists of those chemicals from the US EPA ASTER system [26] for which a measured boiling point was available and hydrogen bonding potential (as measured by HB1 = 0). did not exist. The composition of these data sets is indicated in Table 1. Table 2 presents the list of all 202 parameters calculated in this study.

Tables 1 and 2 around here

3.2 Calculation of TIs. The TIs calculated for this study (some of which are included in Table 1) include Wiener number W,[2] molecular connectivity indices as calculated by Randić [4] and Kier and Hall,[5] frequency of path lengths of varying size,[5] information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić,[7] Roy et al.,[27] Basak et al.,[28-31] as well as those of Raychaudhury et al.,[32] parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,[28-32] and Balaban's J indices [6] as well as triplet indices.[14] The majority of the TIs were calculated using the program POLLY 2.3.[33] The J indices and triplet indices were calculated using software developed in-house by the authors.

4. STATISTICAL ANALYSIS

For both sets of chemicals, the computed TIs were transformed by the natural logarithm of the index plus a constant, generally one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices.

For each set, a technique known as variable clustering was performed using SAS procedure VARCLUS.[34] The variable clustering procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. This is accomplished by a repeated principal components analysis of the sets of indices. The initial principal component analysis examines all indices and defines two principal components or eigenvectors. If the eigenvalue for the second component is > 1.0, the indices are split into separate clusters by correlating the indices with the first and second principal component. Those indices most correlated with the first component form one cluster and those indices most correlated with the second component form another cluster, thus forming two disjoint clusters. A principal component analysis is then performed for each cluster of indices, with the cluster being split if the eigenvalue for the second component is > 1.0. The procedure is repeated until the second eigenvalue is < 1.0 for all clusters.

5. RESULTS AND DISCUSSION

The first database (denoted by A) consists of 139 hydrocarbons (alkanes, alkylbenzenes and polycyclic aromatics) and 162 TIs. The number of indices examined was reduced from the original 202 by removing all but one of the degenerate (i.e. correlation of 1.0) indices and those indices that were constant (0.0) for all chemicals. The second database (denoted by B) is a diverse one and contains 1037 chemical structures and 176 non-degenerate, non-constant indices.

The results of the variable cluster analysis will be presented, discussing first how the descriptors (variables) for database A become clustered, and then surveying the descriptor clustering for database B, as well as the correspondence between these clusters. Inter-cluster correlation will then be described.

The clusters have been ordered according to decreasing numbers of descriptors in each cluster; when clusters contain the same number of descriptors, the numbering of the corresponding clusters is arbitrary.

In Fig. 1, one can see, in graphical form, on the left-hand side the points denoting the clusters that group together the descriptors for the hydrocarbon database A, and on the right-hand side those corresponding to the diverse database B. Each cluster is denoted by a letter (A or B) and a number. The total number of variables in each cluster is written under each point. Full lines connect A-type with B-type clusters, having inscribed on them the numbers of descriptors common to each pair of clusters; when no number is inscribed, this indicates a single common descriptor. Dashed side-lines denote the descriptors that do not have counterparts in the other set of clusters, and the associated numbers on these side-lines indicate the numbers of such "orphan" descriptors. Because the two data sets differ both in the numbers of compounds and in their structures, it is normal to expect that clusters for one data set will have counterparts in several clusters in the other data set. This is indeed what was found to happen, as will be shown below when the diverse data set will be analyzed.

Fig. 1 around here

Only in a single case have we found a one-to-one correspondence between clusters of descriptors corresponding to the two data sets (A12 and B14). Nevertheless, in several instances (A6, A11; B4, B9, B15, B16, and B17), a cluster for one data set (say, A) was found to have all its descriptors in common with only one cluster of the other data set (say, B); however, this latter cluster also contains descriptors found in more than one cluster of the other set.

5.1 Clustering of descriptors for hydrocarbons. The descriptors for database A are grouped in 14 clusters summarized in Table 3. Cluster A1 has 54 from the total of 162 descriptors, therefore it groups together about one third of all variables. These descriptors depend both on the shape and the size (magnitude) of the molecular graph; such descriptors include the Randić connectivity index, the Kier-Hall simple path connectivity indices, the Zagreb group indices, and many triplet indices having as the main diagonal column vector the atomic numbers Z or the total number N of vertices.

Table 3 around here

Cluster A2 with about 1/8 of the total number of descriptors includes molecular connectivity indices of order higher than five, the J indices, as well as two closely similar triplet indices. Cluster A3 contains mainly valence/bond-corrected molecular connectivity indices. The next cluster, A4, consists mainly of the information-based indices IC (information content), SIC (structural information content) and CIC (complementary information content) for the hydrogen-filled graphs of order higher than 2 for IC and higher than 3 for SIC and CIC. Cluster A5 is composed mainly of triplet indices having as main diagonal unit vectors either distance sums or total number N of vertices.

Each of the remaining clusters have less than 10 descriptors. Clusters A6 and A7 contain mostly triplet descriptors: A6 with the distance sum S, and A7 with the order N of the hydrogen-depleted graph, as the main diagonal unit vector; cluster A7 also includes two simple path-cluster molecular connectivity indices. Cluster A8 contains simple cluster- and bond/valence-corrected cluster connectivities of high order (4 through 6). Cluster A9 again consists exclusively of triplet indices, and they are based on summing squares of LOVIs based mainly on distance sum unit vectors on the main diagonal.

Cluster A10 includes three information-theoretic indices IC and SIC of low order (0 and 1) as well as two triplet indices having in common the two unit vectors (distance sum S for the main diagonal, vertex degree V for the free term) and the operation for assembling LOVIs into an index (summation of LOVI square roots).

Interestingly, the four smallest clusters having four descriptors each are pairwise similar in type: A11 with A13, and A12 with A14. Cluster A11 consists of *information TIs* (IC, SIC, CIC) of low order (0 through 2) whereas A13 includes the same TIs of slightly higher order (2 and 3). Clusters A12 and A14 group together *molecular connectivity indices* based on simple cluster and simple cycle, respectively.

A general remark for the triplet indices is that what groups them together is not the matrix on which they are based (adjacency matrix or distance matrix) but the two unit vectors that convert such matrices into systems of linear equations.

5.2 Clustering of descriptors for the diverse set of compounds. There are 18 variable clusters grouping together 176 variables for the database of 1037 diverse compounds (Table 4). Cluster B1, with 49 descriptors, includes 28 % of all variables; 35 of these descriptors are common to cluster A1. Some of these indices, e.g. W (Wiener number), P_0 (number of non-hydrogen atoms), P_1 (number of bonds in the hydrogen-depleted graph), express molecular size. It is interesting that most of the triplet variables (AZVi, AZNi and ANNi with i = 1 through 5 as well as several other ones) are found to be common to clusters A1 and B1. Five other descriptors (${}^0\chi^b$, ${}^2\chi^b$, ${}^3\chi^b$, ${}^0\chi^v$ and ${}^3\chi^v$) also appear in both clusters A1 and B1.

Table 4 around here

Cluster B2 has nine variables in common with cluster A1; most of these (3χ, 4χ, P_2 through P_4) are path connectivities of intermediate order. A couple of triplet indices (ANV1 and ANV5 are also in common with cluster A1; another pair of triplet indices (ASN3 and ASN4) are in common with cluster A7.

Cluster B3 contains triplet indices with distance sums as main-diagonal vector; they occur in clusters A5 and A9. In addition, two descriptors (MIDW and H^D) appear also in cluster A1.

Cluster B4 is uniquely associated with cluster A2, and consists in indices 5χ, 6χ, $5\chi b$, $6\chi b$; $5\chi v$, $6\chi v$, and P_6 through P_{10}. These descriptors are based on long paths, therefore these variables appear only when large molecules are involved.

Seven of the 11 variables of cluster B5 form exclusively cluster A6; they are related to molecular shape via vertex complexity and graph radius. Five triplet indices such as ASN1, ASN5, DSN1, DSN5 and ANV2 also are common to these two clusters.

Very interesting correspondences are manifested by cluster B6, which is mainly associated with two clusters involving the hydrocarbon database, namely A4 and A13 (plus one descriptor in B6 which appears in A10). All variables are of information theoretic type. These higher-order variables (SIC_3 through SIC_6 and CIC_3 through CIC_6) are common to clusters B6 and A4 and represent a true measure of molecular complexity. The lower- and intermediate-order indices such as IC_1 or SIC_2 which appear in clusters B6 and A10 or B6 and A13, respectively, provide information on lower-order complexity that may be more degenerate than that furnished by the higher-order information indices. One should stress here that information content indices form clusters that are separate from clusters with other descriptors, meaning that such variables convey unique information relative to structure and molecular complexity.

Cluster B7 consists only of path-cluster molecular connectivity descriptors which were included in clusters A3, A7 and A8 for the hydrocarbons.

Cluster B8 includes triplet indices, all of which have the atomic number Z for the free term vector in the system of linear equations. Most of these descriptors appear in clusters A1, A5, A9.

Cluster B9 consists of high-order connectivity-cluster terms all contained in cluster A8. For hydrocarbons, descriptors ${}^6\chi^b{}_C$ and ${}^6\chi^v{}_C$ are perfectly correlated with descriptor ${}^6\chi_C$, therefore, the former variables did not appear in the hydrocarbon cluster A8. For the diverse-compound database, such a correlation is not perfect because of differences in atom types.

An interesting observation concerns cluster B10: all six variables are absent from the hydrocarbon database because this database does not contain any 3- or 4-membered rings, unlike the diverse compound database. This is why indices ${}^{3/4}\chi_{Ch}$, ${}^{3/4}\chi^b{}_{Ch}$ and ${}^{3/4}\chi^v{}_{Ch}$ appear only in cluster B10.

Cluster B11 has all but one of its descriptors contained in cluster A4; these information content indices, IC_2 through IC_6, measure a high degree of non-redundancy of topological neighborhoods.

Cluster B12 has four of its variables contained in cluster A11; these descriptors (SIC_0, CIC_0 through CIC_2) express lower-order redundancy of topological neighborhoods. This is true of indices IC_0 and SIC_1 as well, which are present in cluster A10.

From cluster B13, the six descriptors (simple, bond and valence corrected chain molecular connectivity indices) are partitioned equally between clusters A2 and A14, according to the 6- versus 5-membered ring size, respectively; in the hydrocarbon data base A, six-membered chain (or rings) predominate.

Cluster B14 is exclusively associated in a one-to-one relationship with cluster A12. The corresponding descriptors ${}^3\chi_C$, ${}^4\chi_C$, as well as their bond and valence corrected counterparts represent connectivity indices on three- and four-vertex structural clusters. For the hydrocarbon database, we have again a case in which the two indices $4\chi b_C$ and ${}^4\chi v_C$, are perfectly correlated with ${}^4\chi_C$, do not appear explicitly in cluster A12.

Half of the variables (J-type indices) in cluster B15 are contained in cluster A2. These J indices again form a cluster apart from all other ones in the case of the diverse data base, proving that when heteroatoms are taken into account, the information provided by such J-type indices is unique.

Clusters B16, B17 and B18 have each a small number of triplet-type descriptors; the three descriptors of cluster B17 are all contained in cluster A7.

5.3 Inter-cluster correlations. From each cluster we select 15-25% of the descriptors according to the maximal value of the correlation coefficient with their own cluster. In most cases, the first selected descriptor also has the minimal value of the correlation with the next closest cluster, expressed by the $1\text{-}r^2$ value. When choosing more than one index from the same cluster, after the first one was selected as indicated above, the next one must also fulfill a third criterion, namely a low intercorrelation with the previously selected indices of the same cluster.

There were four inter-cluster correlations within the hydrocarbon data set that were greater than 0.9 and all involved cluster A1. Cluster A1 was positively correlated with A2, A3, and A7. Cluster A1 was correlated negatively with A5. Each of the clusters characterizes some aspect of molecular size and shape.

Cluster B1 showed an inter-cluster correlation of 0.92 with cluster B2 and –0.90 with cluster B3. These were the only inter-cluster correlations greater than 0.9. These clusters are the three largest clusters in set B. Like cluster A1, cluster B1 groups TIs expressing molecular size and shape. Interestingly, in set A cluster A1 also had a negative inter-cluster correlation with cluster A5; it is therefore not surprising that clusters A5 and B3 have the most abundantly populated line connecting them in Fig. 1.

In summary, for the hydrocarbon data base there are four inter-cluster correlations with r>0.90 all involving on one hand the first cluster A1, and on the other hand clusters A2, A3, A5, and A7. For the diverse compound data base there are only two such inter-cluster correlations with r>0.90, namely B1 with B2 and B3. This is not unexpected, as the combination of the first three clusters in each case contain more descriptors than the parameters remaining in all the remaining ones together.

In this context, one should mention that Todeschini and coworkers published an interesting study [35] on 23 TIs for a set of 667 diverse chemicals, 20% of which were hydrocarbons; the above authors excluded 10 of these TIs because they were degenerate, redundant, or had intercorrelation factors higher than 0.90. A graph depicting highly intercorrelated indices using data published by these authors is presented in Fig. 2, which is similar to a graph published earlier.[22]

Fig. 2 around here

Ten TIs were then selected by Todeschini *et al.*,[35] namely the molecular weight (MW), J, IC, CIC, the bonding information content (BIC), mean Randić connectivity (χ) the information content on atomic composition (I_{AC}), the mean Wiener index (W), and the mean information indices on equality of distance degree and on the magnitude of distance degree ($I^E_{D,\ deg}$ and

$I^{W}_{D, deg}$, respectively). Then, using principal component analysis for the above 10 TIs, Todeschini *et al.* analyzed the composition of the first six principal components. They found that the first PC is mainly composed of indices that express the size of molecules (MW, W, IC, $I^{E}_{D, deg}$ and $I^{W}_{D, deg}$). This is in agreement with the earlier finding of Basak et al. for a diverse set of 3,692 diverse chemicals that the first PC is related to molecular size.[29] Further, Todeschini et al. found that the second PC is dominated by indices expressing information on bonds (IC, CIC, and BIC). This is also analogous to the results reported by Basak et al. [29] that the second axis represents molecular complexity as encoded by higher order neighborhood complexity indices (IC_2, IC_3, SIC_2, SIC_3, CIC_2, CIC_3, etc.). The IC, CIC and BIC indices used by Todeschini et al. are based solely on first-order topological bonding/neighborhoods, and slightly different equivalence relations as compared to the IC_r, SIC_r, and CIC_r indices defined by Roy et al. [27] In studies by Basak et al.,[29] the first-order complexity indices (IC_1, SIC_1, CIC_1) were usually most correlated with the first PC. Each of the next four PCs in Todeschini et al.'s study [35] are dominated by a single TI,viz.,: χ, I_{AC}, J (indicating branching), and $I^{E}_{D, deg}$ (connected with the position of substituents on the molecular scaffold), respectively.

ACKNOWLEDGEMENT

This is contribution number 251 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force.

REFERENCES AND NOTES

1. Devillers, J.; Balaban, A. T., editors. *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, The Netherlands, 1999. (a) Chapter 2 by Balaban, A. T.; Ivanciuc, O. Historical Development of Topological Indices; (b) Chapter 7 by Kier,L. H.; Hall, L. B. Molecular Connectivity Chi Indices for Database Analysis and Structure-Property Modeling; (c) Chapter 10 by Kier, L. H.; Hall, L. B. The Kappa Indices for Molecular Modeling of Molecular Shape and Flexibility; (d) Chapter 11 by Kier, L. H.; Hall, L. B. The Electrotopological State: Structure Modeling for QSAR and Database Analysis; (e) Chapter 12 by Basak, S. C. Information-Theoretic Indices of Neighborhood Complexity and Their Application; (f) Chapter 14 by Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Hierarchical Approach to the Development of QSAR Models Using Topological, Geometrical and Quantum Chemical Parameters.
2. Wiener, H. Structural Determination of Paraffin Boiling Points, *J. Am. Chem. Soc.* **1947**, *69*, 17-20.
3. (a) Balaban, A. T. Chemical Graphs. Part 35. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta*, **1979**,. *5*, 239-261; (b) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Centre Concept and Derived Topological Indices. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 196-213; (c) 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. (MATCH)*, **1994**, 30, 55-72.
4. Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.
5. (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research.* Academic Press, New York, 1976; (b) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Studies.* Research Studies Press, Letchworth, 1986; (c) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State.* Academic Press, New York, 1999.
6. (a) Balaban, A. T. Highly Discriminating Distance-Based Topological Index, *Chem. Phys. Lett.* **1982**, *80*, 399-404; (b) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199-206; (c) Balaban, A. T. Chemical Graphs. 48. Topological Index J for Heteroatom-Containing

Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH),* **1986**, *21*, 115-122; (d) Balaban, A. T.; Filip, P. Computer Program for Topological Index J (Average Distance Sum Connectivity). *Math. Chem. (MATCH),* **1984**, *16*, 163-190

7. Bonchev, D. *Information-Theoretic Indices for Characterization of Chemical Structure.* Research Studies Press, Letchworth, 1993.
8. Trinajstić, N. *Chemical Graph Theory*, 2nd Ed. CRC Press, Boca Raton, Florida, 1992, pp. 225-273.
9. Balaban, A. T. Using Real Numbers as Vertex Invariants for Third-Generation Topological Indices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 23-28.
10. Basak, S. C.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological Indices: Their Nature, Mutual Relatedness, and Applications. *Math. Modelling*, **1987**, *8*, 300-305.
11. Gutman, I.; Ruscic, B; Trinajstić, N.; Wilcox Jr., C. F. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399-3405.
12. Shannon, C., A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379-423.
13. Balaban, A. T.; Balaban, T. S.; New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8*, 383-397.
14. Filip, P. A.; Balaban, T. S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61-83.
15. (a) Ivanciuc, O.; Balaban, T. S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. J. *Math. Chem.* **1993**, *12*, 309-318; (b) Plavsic, D.; Nikolić, S.; Trinajstić, N.; Mihalic, Z. On the Harary Index for Characterization of Chemical Graphs. *J. Math. Chem.* **1993**, *12*, 235-250.
16. (a) Balaban, A. T.; Diudea, M. V. Real Number Vertex Invariants: Regressive Distance Sums and Related Topological Indices. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 421-428; (b) Diudea, M. V.; Minailiuc, O.; Balaban, A. T. Molecular Topology. Part 4. Regessive Vertex Degrees (New Graph Invariants) and Derived Topological Indices", *J. Comput. Chem.* **1991**, *12*, 527-535.
17. Diudea, M. V.; Minailiuc, O.; Katona, G.; Gutman, I. Szeged Matrices and Related Numbers. *Math. Chem. (MATCH),* **1997**, *35*, 129-143.
18. Klein, D. J.; Randić, M. Resistance Distance. *J. Math. Chem.* **1993**, *17*, 147-154.
19. (a) Randić, M.; Basak, S. C. Optimal Molecular Descriptors Based on Weighted Path Numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261-266; (b) Basak, S. C.; Gute, B. D. Characterization of Molecular Structures Using Topological Indices, *SAR QSAR Environ. Res.* **1997**, *7*, 1-21; (c) Gute, B. D.; Basak, S. C. Predicting Acute Toxicity of Benzene Derivatives Using Theoretical Molecular Descriptors: a Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7,* 117-131; (d) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243-272; (e) Basak, S. C.; Grunwald, G. D. Use of Graph Invariants, Volume and Total Surface Area in Predicting Boiling Point of Alkanes. *Math. Modelling Sci. Comput.* **1993**, *2*, 735-740; (f) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modelling Sci. Comput.* in press; (g) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Math. Model. Comput. Sci.* in press; (h) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modelling Sci. Comput.* in press; (i) 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; (j) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals, *Toxicology Lett.* **1995**, *79*, 239-250; (k) Basak, S. C.; Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: 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. Eds., Princeton Scientific Publishing Co. Inc. pp. 492-504 (1997); (l) Basak, S. C.; Grunwald, G. D. Predicting Genotoxicity of Chemicals Using Nonempirical Parameters. In *Proceeding of XVI International Cancer Congress*, pp. 413-416 (1995); (m) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical and Neural

Network Methods for Predicting Toxic Modes of Action of Chemicals, *Environ. Toxicol. Chem.* **1998**, 17, 1056-1064; (n) Basak, S. C.; Veith, G. D.; Grunwald, G. D. Prediction of Octanol-Water Partition Coefficient (K_{ow}) Using Algorithmically-Derived Variables G J. *Environ. Toxicol. Chem.* **1992**,*11*, 893-900; (o) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood-Brain Transport of Drugs: a Computational Approach., *Pharm. Res. 13*, **1996**, 775-778. (p) 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*, 1054-106; (q) 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 Environmental Res.* **1999**, *10*, 1-15.

20. Ivanciuc, O.; Balaban, A. T. Investigation of Alkanė Branching with Topological Indices. *Math. Chem. (MATCH)*, in press.

21. Balaban, A. T. (ed.), *From Chemical Topology to Three-Dimensional Geometry.* Plenum Press, New York, 1998.

22. Motoc, I.; Balaban, A. T. Topological Indices: Intercorrelations, Physical Meaning, Correlational Ability. *Rev. Roum. Chim.* **1981**,*26*, 593-600; Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. Topological Indices: Inter-Relations and Composition. *Math. Chem. (MATCH),* **1982**, *13*, 369-404;

23. Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological Indices for Structure-Activity Correlations, in *Steric Effects in Drug Design* (editors Charton, M.; Motoc, I.) *Top. Curr. Chem.* **1983**, *114*, 21-55; Balaban, A. T.; Niculescu-Duvaz, I;. Simon, Z. Topological Aspects in QSAR for Biologically-Active Molecules. *Acta Pharm. Jugosl.* **1987**, *37*, 7-36; Voiculetz, N.; Balaban, A. T. Niculescu-Duvaz, I.; Simon, Z. *Modeling of Cancer Genesis and Prevention.* CRC Press, Boca Raton, Florida, 1990.

24. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the Molecular Similarity of Chemicals Using Topological Indices. In *Advances in Molecular Similarity*, **1998**, *2*, 169-183.

25. Needham, D.E.; Wei, I.C.; Seybold, P.G. Molecular Modelling of the Physical Properties of Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186-4194; Mekenyan, O.; Bonchev, D.; Trinajstić, N. Chemical Graph Theory: Modelling the Thermodynamic Properties of Molecules. *Int. J. Quantum Chem.* **1980**, *18*, 369-380; Karcher, W. *Spectral Atlas of Polycyclic Aromatic Hydrocarbons, Vol. 2.* Kluwer Academic Press, Dordrecht 1988, pp. 16-19.

26. Russom, C. L. Assessment Tools for the Evaluation of Risk (ASTER) v. 1.0; U.S. Environmental Protection Agency, 1992.

27. Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In: *Mathematical Modelling in Science and Technology.* 4th Internat. Conf. Zurich. Eds. Avula, X. J. R.; Kalman, R. E.; Liapis, A. I.; Rodin, E. Y. Pergamon Press, New York, 1983, pp. 745-750.

28. Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: a QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605-609; Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. A Quantitative Structure Activity Relationship Study of Tumor Inhibitory Triazenes Using Bonding Information Content and Lipophilicity. *ICRS Med. Sci.* **1982**, *10*, 933-934; Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. A Quantitative Structure-Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneimitt.-Forsch. Drug Res.* **1983**, *33*, 501-503.

29. Basak, S. C.; Magnuson, V. R. Niemi, G. J.; Regal, R.R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discr. Appl. Math.* **1988**, *19*, 17-44.

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

31. (a) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Math. Modelling Sci. Comput.* (in press); (b) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Ibid.* (in press); (c); Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals. *Ibid.* (in press).

32. Raychaudhury, C.; Ray, S. K.; Roy, A. B.; Ghosh, J. J.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Indices. *J. Comput. Chem.* **1984**, *5*, 581-588.

33. Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY 2.3. Copyright of the University of Minnesota, 1988.

34. SAS Institute Inc. The VARCLUS Procedure. In SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 2, Cary, NC. SAS Institute Inc., 1989, 846 pp.

35. Todeschini, R.; Cazar, R.; Collina, E. The Chemical Meaning of Topological Indices. *Chemometrics Intell. Lab. Syst.* **1992**, *15*, 51-59.

Table 1. Summary of Chemical Classes or Features in Databases Analyzed.

Chemical Classes or features	Database A (Hydrocarbons)	Database B (Diverse)
Total Number of Compounds	139	1037
Hydrocarbons	139	565
◆ Alkanes, Cyclic Alkanes	73	206
◆ Aromatics	66	288
– Alkyl Benzenes	29	80
– Fused Rings	37	56
– Polycyclic Aromatics	37	49
Non-hydrocarbons	0	472
◆ Halogen containing compounds		359
◆ Heteroatom containing compounds (Sulphur or Phosphorous)		101
◆ Compounds containing both halogens & heteroatoms		12
– Organosulfides		105
– Organophosphorous		8

Table 2. Symbols and definitions of topological parameters

Index	Definition
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
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
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
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
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
${}^h\chi$	Path connectivity index of order h = 0-6
${}^h\chi$	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 = 3-6
${}^h\chi^b$	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 = 3-6
${}^{h}\chi_{PC}^{b}$	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 = 3-6
${}^{h}\chi_{PC}^{v}$	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^Y	Balaban's J index based on relative covalent radii
Triplet	Global invariants based on solutions of linear equation systems using the adjacency matrix (A), distance matrix (D), and column/row vectors: distance sums (S), atomic number (Z), number of non-hydrogen atoms (N and N^2), vertex degree (V), or numerical constants (1). Notation is described by triplets (e.g. AZV). Results are weightings for each atom in a molecule. These weights are combined by 5 possible formulas: 1 = Sum of weights: $\Sigma_i x_i$; 2 = Sum of squared weights $\Sigma_i x_i^2$; 3 = Sum of square root of weights $\Sigma_i x_i^{1/2}$; 4 = Sum of cross-product $\Sigma_i(xi \cdot xj)^{-1/2}$; and 5 = product of weights $N \cdot [\Sigma_i x_i]^{1/N}$

Table 3. Summary of Variable Clustering for 139 Hydrocarbons

Cluster	Number of Variables	Representative Variables (max. 25% of total listed)
A1	54	DN^2Z_4, DN^2N_4, P0, AZV4, ASZ4, ANN3, ANN5, AZN3
A2	19	${}^6\chi$, $P_{7,}$ ${}^5\chi$, $6\chi b$, $6\chi v$
A3	13	$0\chi b$, $0\chi v$, ANZ1
A4	13	SIC_6, SIC_5, IC_6
A5	12	DSZ_1, DSZ_5, ASZ_1
A6	9	DSZ_3, DSN_5
A7	9	DSN_3, DN^2N_1
A8	6	${}^5\chi^v{}_C$, ${}^5\chi^b{}_C$
A9	6	DSZ_2, ASZ_2
A10	5	SIC_1
A11	4	CIC_1
A12	4	${}^3\chi^v{}_C$
A13	4	SIC_3
A14	4	${}^5\chi_{Ch}$

Table 4. Summary of Variable Clustering for 1037 Diverse Chemicals.

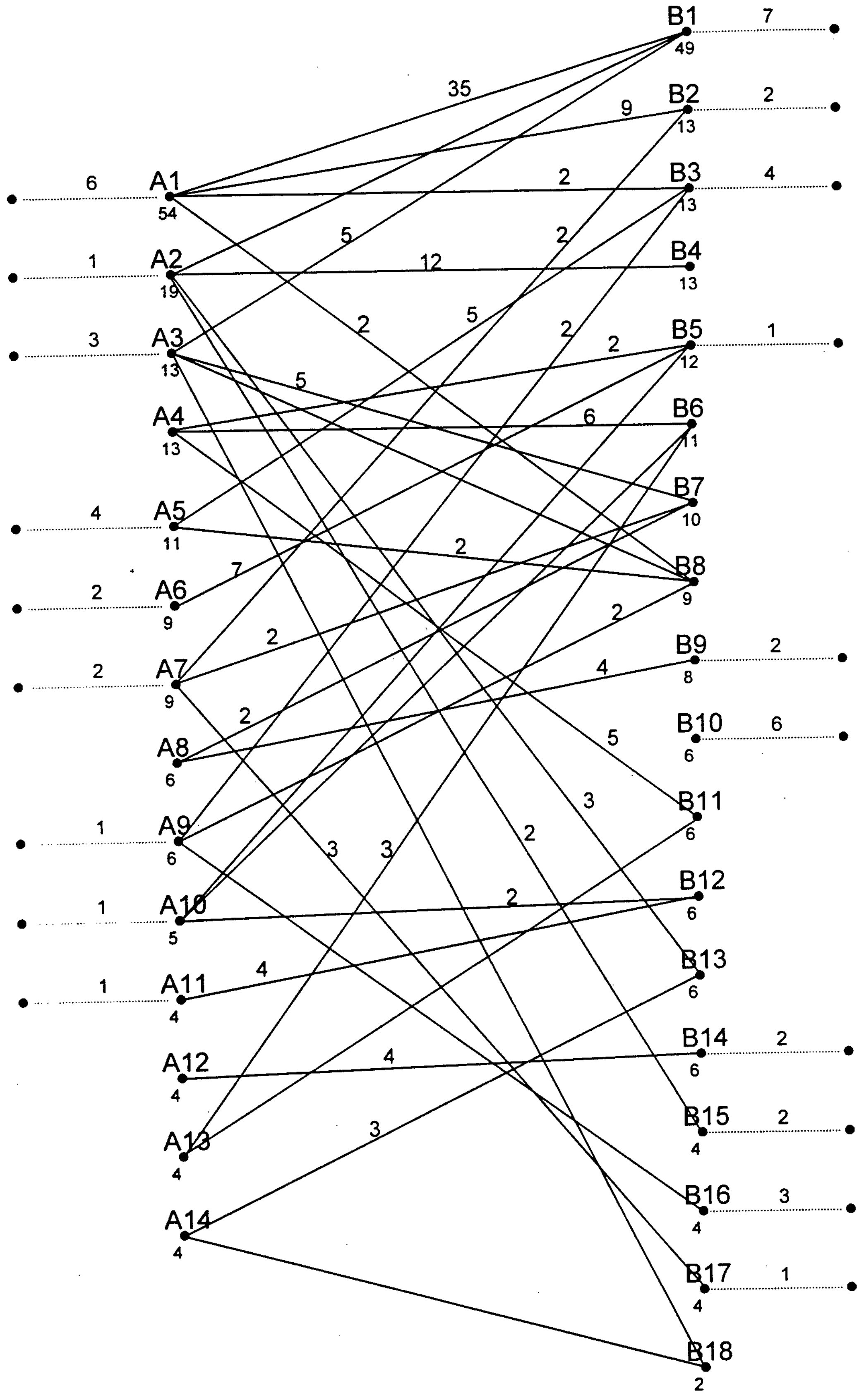
Cluster	Number of Variables	Representative Variables (max. 25% of total listed)
B1	49	P0, ANN3, ANN5, AN13, ANN1, ANV4, AS14, $DN^2$14
B2	13	ANV1, P3, M2
B3	13	AS11, AS15, DS11
B4	13	6χ, $6\chi b$, P7
B5	11	ASN5, AS13, ASN1
B6	10	SIC3, SIC4, CIC4
B7	9	$5\chi b_{PC}$, $5\chi_{PC}$
B8	8	ASZ2, ASZ1
B9	6	$5\chi b_{C}$, $5\chi_{C}$
B10	6	$3\chi_{Ch}$, $3\chi b_{Ch}$
B11	6	IC_4, IC_5
B12	6	CIC_1, SIC_1
B13	6	$6\chi v_{Ch}$, $6\chi b_{Ch}$
B14	6	$3\chi b_{C}$, $4\chi_{C}$
B15	4	J^B
B16	4	AS12
B17	4	DN^2N1
B18	2	ANS1

Legends of figures

Fig. 1. Associations between clusters of descriptors for the hydrocarbon database (A-type clusters) and the database with diverse compounds (B-type clusters). Solid lines connect A-type descriptors with B type descriptors, and the numbers of common descriptors are indicated on such lines (when no number is indiceted, there is just one common descriptor). Dashed lateral lines indicate descriptors that have no correspondence for the other type.

Fig. 2. Graph of highly correlated topological indices (TIs) according to Todeschini et al. (notation of TIs as in Tab. 3 of ref.[31]). Lines connect TIs with $r > 0.90$.

Figure 1


A1
54
6
A2
19
1
A3
13
3
A4
13
A5
11
4
A6
9
2
A7
9
2
A8
6
A9
6
1
A10
5
1
A11
4
1
A12
4
A13
4
A14
4
B1
49
7
B2
13
2
B3
13
4
B4
13
B5
12
1
B6
11
B7
10
B8
9
B9
8
2
B10
6
6
B11
6
B12
6
B13
6
B14
6
2
B15
4
2
B16
4
3
B17
4
1
B18
2
35
9
2
5
12
2
2
5
2
2
5
6
2
7
2
2
2
4
5
3
3
3
2
2
4
4
3

APPENDIX 1.7 Use of graph invariants in QMSA and predictive toxicology

DIMACS Series in Discrete Mathematics and
Theoretical Computer Science, in press

Use of Graph Invariants in QMSA and Predictive Toxicology

S.C. Basak and B.D. Gute
Natural Resources Research Institute,
5013 Miller Trunk Hwy., Duluth, MN, 55811 USA

I. INTRODUCTION

A contemporary interest in mathematical chemistry is the characterization of molecular structure using graph theoretic formalism [1-11]. A graph $G = [V,E]$ consists of an ordered pair of two sets V and E, representing the vertices and edges, respectively. G becomes a molecular graph when the set V represents the set of atoms in a molecule and the set E symbolizes chemical bonds between adjacent atoms [8].

Mathematical characterization of molecular graphs (structures) may be accomplished using graph invariants. An invariant may be a polynomial, a sequence of numbers, or a real number. A real number characterizing a molecular graph is called a topological index (TI). TIs quantify different aspects of molecular architecture, viz., size, shape, cyclicity, branching, symmetry, etc [8].

TIs have been used extensively in quantitative structure-property/activity relationships (QSPR and QSAR respectively) and the quantification of intermolecular similarity/dissimilarity of chemicals [10-24]. In quantitative molecular similarity analysis (QMSA) studies, TIs have been used to derive high dimensional structure spaces where the Euclidean distance D_{ij} between a pair of molecules i and j is used to quantify the similarity between them. Similarity measures can be used either for the selection of analogs of chemicals or in the prediction of the property/activity of a molecule from the property of its selected neighbor(s).

In some of our recent QSAR/QMSA studies we have used different similarity measures derived from TIs in the selection of analogs and prediction of properties/activities for diverse sets of chemicals. We have also used orthogonal descriptors derived from a set of over 100 graph invariants to estimate bioactivity/toxicity of different graphs of molecules. In this paper we have used similarity measures derived from TIs in: a) selecting analogs of an isospectral graph from a diverse set of 221 compounds, and b) predicting the mutagenicity of a set of 113 mutagens and non-mutagens using QMSA methods.

II. METHODS

Databases

A set of 19 pairs of isospectral graphs from the work of Balasubramanian and Basak [25] were added to a set of 107 benzamidines [26] and a composite set of 76 diverse compounds used in an earlier study by Basak and Grunwald [23] to create a varied library of 221 compounds. This composite library was created to provide a large set containing both congeneric and non-congeneric sets to test analog selection methods. The chemical structures for the 19 pairs of isospectral graphs have been presented in a previously [25].

A second data set, representing a subset of the set of 277 chemicals presented by Yamaguchi *et al.* [27] was also used in the current study. This subset consisted of all the chemicals in the set of 277 chemicals that had reported results for mutagenicity in the Ames test, mutagenicity in the medium term liver carcinogenesis bioassay, and carcinogenicity in the two-year rodent bioassay in rat and/or mouse. This subseting resulted in a set of 113 chemicals, 68 of which are classified as non-mutagens and 45 of which are classified as mutagens in the Ames test. This set of chemicals and their observed mutagenicity are reported in Table 1.

Calculation of Topological Indices

The TIs calculated for this study are listed in Table 2 and include Wiener number [28], molecular connectivity indices as calculated by Randić [29] and Kier and Hall [4], frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić [30] as well as those of Raychaudhury *et al.* [31], parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [32-34], and Balaban's *J* indices [35-37]. The majority of the TIs were calculated using POLLY 2.3 [38]. The *J* indices were calculated using software developed by the authors.

The Wiener index (*W*) [28], the first topological index reported in the chemical literature, 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 x 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 thioacetamide *(Fig.1)*:

$$D(G_1) = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 2 & 1 & 0 & 2 \\ 4 & 2 & 1 & 2 & 0 \end{array}$$

W is calculated as:

$$W = ½\sum_{ij} d_{ij} = \sum_{h} h \cdot g_h \qquad (1)$$

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 nine.

[Insert Fig. 1 here]

Randić's connectivity index [29], and higher-order connectivity path, cluster, path-cluster and chain types of simple, bond and valence connectivity parameters were calculated using the method of Kier and Hall [4]. The generalized form of the simple path connectivity index is as follows:

$$^{h}\chi = \sum \left(v_i v_j \ldots v_{h+1}\right)^{-½} \qquad (2)$$

where $v_i, v_j, \ldots, v_{h+1}$ 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,...,10) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Information-theoretic topological indices 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 upon 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, ..., $h;_{i}\, n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$\begin{array}{c} A_1, A_2, \ldots\ldots, A_h \\ p_1, p_2, \ldots\ldots, p_h \end{array}$$

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 [39]:

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

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 x IC. Figure 2 provides a sample calculation for IC_1.

[Insert Fig. 2 here]

It is to be noted that the information content of a graph G is not uniquely defined. It depends on how the set A is derived from G as well as on the equivalence relation which partitions A into disjoint subsets A_i. For example, when A constitutes the vertex set of a chemical graph G, two methods of partitioning have been widely used: a) chromatic-number coloring of G where two vertices of the same color are considered equivalent, and b) determination of the orbits of the automorphism group of G thereafter vertices belonging to the same orbit are considered equivalent.

Rashevsky was the first to calculate the information content of graphs where "topologically equivalent" vertices were placed in the same equivalence class [40]. In Rashevsky's approach, two vertices u and v of a graph are said to be topologically equivalent if and only if for each neighboring vertex u_i (i = 1, 2, ..., k) of the vertex u, there is a distinct neighboring vertex v_i of the same degree for the vertex v. While Rashevsky used simple linear graphs with indistinguishable vertices to symbolize molecular structure, weighted linear graphs or multigraphs are better models for conjugated or aromatic molecules because they more properly reflect the actual bonding patterns, *i.e.*, electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.* [41] calculated 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 stereo-electronic 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 non-negative 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, $S(v, 0) =$, $S(v, r) = v$ 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 which 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 [34].

In this method, chemicals are symbolized by weighted linear graphs. Two vertices u_o and v_o 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_o, u_1, ..., u_r of length r, there is a distinct path v_o, v_1, ..., v_r of the same length such that the paths have similar edge weights, and both u_o and v_o 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 [34,42].

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by Eq. 2. Basak *et al.* [32] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r / \log_2 n \tag{4}$$

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

Another information-theoretic invariant, complementary information content (CIC_r) [43], is defined as:

$$CIC_r = \log_2 n - IC_r \tag{5}$$

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*$_r$. Sample calculations for *SIC*$_1$ and *CIC*$_1$ have been included in Figure 2.

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 [30]:

$$I_D^W = W\log_2 W - \sum_h g_h \cdot h\log_2 h \tag{6}$$

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* [31].

Balaban defined a series of indices based upon distance sums within the distance matrix for a chemical graph that he designated as *J* indices [35-37]. These indices are highly discriminating with low degeneracy. Unlike *W*, the *J* indices range of values are independent of molecular size. The general form of the *J* index calculation is as follows:

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

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 [35]. Variants were proposed by Balaban for incorporating information on bond type, relative electronegativities, and relative covalent radii [36,37].

Calculation of Atom Pairs

Atom pairs (APs) were calculated using the method of Carhart *et al* [3]. An atom pair is defined as a substructure consisting of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<atom descriptor$_i$> – <separation> – <atom descriptor$_j$>

where <atom descriptor> contains information about the atomic type, number of non-hydrogen neighbors and the number of π electrons. The interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms. APs used in this study were calculated by the APProbe software [43].

III. STATISTICAL METHODS AND COMPUTATION OF INTERMOLECULAR 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 the intercorrelation of indices. The PCA was conducted using the SAS procedure PRINCOMP [44]. 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* [13]. These PCs were subsequently used to determine similarity scores as described below.

Similarity Measures

Intermolecular similarity was measured using two distinct methods. The AP method uses an associative measure described by Carhart *et al.* [3] and is based on atom pair descriptors. The measurement is the ratio of the number of shared atom pairs between two molecules over the total number of atom pairs present in the two molecules. Similarity (S) between molecules i and j is defined as:

$$S_{ij} = 2C/\left(T_i + T_j\right) \tag{8}$$

where C is the number of atom pairs common to molecule i and j. T_i and T_j are the total number of atom pairs in molecule i and j, respectively. The numerator is multiplied by a factor of 2 to reflect the presence of shared atom pairs in both compounds.

The second similarity method, Euclidean distance (ED) within an n-dimensional PC space derived from TIs was used. ED between molecules i and j is defined as:

$$ED_{ij} = \left[\sum_{k=1}^{n}\left(D_{ik} - D_{jk}\right)^2\right]^{1/2} \tag{9}$$

where n equals 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 molecules i and j, respectively.

Analog / K-Nearest Neighbor Selection

Following the quantification of intermolecular similarity of the molecules, analogs or nearest neighbors are determined on the basis of both S and ED. In the case of the AP method, two molecules are considered identical if S=1, while they have no atom pairs in common if S=0. The ED method measures a distance between molecules, thus the lower the value of ED the greater the similarity between two molecules.

Property Estimation

Since the data presented in the work of Yamaguchi *et al.* [27] represented mutagenicity as non-mutagen (-) or mutagen (+) this data was treated as a zero-one relationship, where non-mutagens have a value of zero and mutagens have a value of one. In estimating the mutagenicity of the probe compound, the mean of the observed mutagenicity of the K-nearest neighbors was used as the estimate. Thus, if the mean resulted in a value greater than 0.5, the compound was classified as a mutagen.

However, if the mean was equal to 0.5, the compound was not classified as the results were inconclusive.

IV. RESULTS

Principal Component Analysis

From the PCA of the 102 TIs, eight PCs with eigenvalues greater than one were retained. These eight PCs explained, cumulatively, 95.2% of the total variance within the TI data. Table 3 lists the eigenvalues of the eight PCs, the proportion of variance explained by each PC, the cumulative variance explained, and the two TIs most correlated with each individual PC.

Analog Selection

Figure 3 shows the results of the analog selection for isospectral graph 10.1.1 using atom pairs to derive a similarity space and PCs to derive a Euclidean distance space. The first five analogs (neighbors) for the probe compound, 10.1.1, are presented for each of the similarity methods.

[Insert Fig. 3 here]

K-Nearest Neighbor Estimation

Table 4 presents the results for the prediction of mutagenicity for the 113 molecules over a range of *K* values (*K* = 1-5) for both the *AP* and *ED* methods. The results are presented as percent correctly classified and over-all percent correct prediction rates are provided as a means of comparing the efficacy of the individual models. The variability between the *K* levels is easily explained by the problematic nature of using a binary relationship such as this one in estimation. When the number of neighbors was even, the potential for unclassified compounds led to lower prediction rates than in the case of an odd number of neighbors.

V. DISCUSSION

The major objective of this paper was to study the effectiveness of mathematical invariants in the characterization of molecular structure and the estimation of the toxicity of chemicals. An invariant maps a chemical structure into the set *R* of real numbers. A specific invariant may be used for the ordering or partial ordering of sets of molecules or in structure-activity relationship studies [45]. A particular structural invariant quantifies distinct aspects of molecular structure. Therefore, a combination of such indices might be more powerful in the mathematical characterization of molecular structure as compared to the use of one specific invariant. The problem arises out of the fact that often the various graph theoretic indices of molecular structures are strongly correlated. We have attempted to resolve this problem through the implementation of a PCA to derive orthogonal variables from a large set of calculated TIs, and using the orthogonal parameters in the characterization of structure [10,12,15,17,18,22,23].

In the present study we have used calculated atom pairs and principal components derived from TIs to select structural analogs for a probe compound from a diverse set containing closely related structures. The result of this analog selection, depicted in Figure 3, shows that the five neighbors selected by each of the methods exhibit sufficient power to reject dissimilar structures. In other words, we may conclude that both the atom pair and Euclidean distance methods are capable of choosing similar molecules from a collection of structurally diverse structures. This is in line with our earlier studies with various diverse sets of molecules [10,12,15,17,18,22,23].

The central paradigm of QSAR holds that similar structures usually have similar properties. To test this idea, we selected *K*-nearest neighbors (*K*=1-5) for each molecule from a set of 113 mutagens and non-mutagens using the *ED* and *AP* methods and used the selected nearest neighbors in estimating mutagenicity. The results in Table 4 show that both methods lead to reasonably good estimates, although the *AP* method was superior to the *ED* method.

In conclusion, both the ED and AP methods, based on calculated graph theoretic structural invariants, did reasonably well in the selection of structural analogs and in the estimation of chemical properties based on nearest neighbors.

ACKNOWLEDGEMENTS

This is contribution number 239 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force. The authors would like to extend their thanks to Greg Grunwald for technical support.

References

1. Narumi, H.; Hosoya, H. Topological Index and Thermodynamic Properties. II. Analysis of the Topological Factors on the Absolute Entropy of Acyclic Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1980**, *53*, 1228-1237.

2. Randić, M. Nonempirical Approaches to Structure-Activity Studies. *Int. J. Quantum Chem: Quant. Biol. Symp.* **1984**, *11*, 137-153.

3. 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.

4. Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure-Activity Analysis. Research Studies Press: Letchworth, Hertfordshire, U.K, 1986.

5. Rouvray, D. H.; Pandey, R. B. The Fractal Nature, Graph Invariants and Physicochemical Properties of Normal Alkanes. *J. Chem. Phys.* **1986**, *85*, 2286-2290.

6. Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605-609.

7. Basak, S. C.; Niemi, G. J.; Veith, G. D. In *Computational Chemical Graph Theory*, D.H. Rouvray, Ed.; NOVA: New York, 1990, pp. 235-277.

8. Trinajstić, N. *Chemical Graph Theory*, Klein, D. J., and Randić, M., Eds.; CRC Press: Boca Raton, 1992.

9. 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.

10. 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.

11. 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.

12. Johnson, M.; Basak, S. C.; Maggiora, G. A Characterization of Molecular Similarity Methods for Property Prediction. *Mathematical and Computer Modelling* **1988**, *II*, 630-635.

13. 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.

14. Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243-272.

15. Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity from Structural Similarity. *New J. Chem.* **1995**, *19*, 231-237.

16. 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.

17. 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.

18. 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.

19. 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, pp 245.

20. 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.

21. Gute, B. D.; Basak, S. C. Predicting Acute Toxicity (LC_{50}) of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach, *SAR QSAR Environ. Res.* **1997**, 7, 117-131.

22. Basak, S. C., Gute, B. D. and Grunwald, G. D. Development and Applications of Molecular Similarity Methods using Nonempirical Parameters. *Mathl. Modelling Sci. Computing*, in press, 1998.

23. Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathl Modelling Sci. Computing*, in press, 1998.

24. 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.*, in press, 1998.

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

26. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Prediction of Complement-Inhibitory Activity of Benzamidines Using Topological and Geometric Parameters. *J. Chem. Inf. Comput. Sci.*, accepted, 1998.

27. Yamaguchi, T.; Hasegawa, R.; Hagiwara, A.; Hirose, M.; Imaida, K.; Ito, N.; Shirai, T. Results for 277 Chemicals in the Medium Term Liver Carcinogenesis Bioassay of Rats.

28. Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, 69, 17-20.

29. Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.

30. Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517-4533.

31. 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.

32. Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the Structure-Function Relationship of Pharmacological and Toxicological Agents Using Information Theory. In *Proceedings of the 2nd International Conference on Mathematical Modelling,* Avula, X. J. R., Bellman, R., Luke, Y. L., and Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, Missouri, 1980; Vol. II, pp. 851-856.

33. 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.

34. Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology,* Avula, X. J. R., Kalman, R. E., Lipais, A. I., and Rodin, E. Y., Eds.; Pergamon Press: New York, 1984, pp. 745-750.

35. Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399-404.

36. Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure & Appl. Chem.* **1983**, *55*, 199-206.

37. 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.

38. Basak, S. C.; Harriss, D. K.; Magnuson, V. R. POLLY 2.3: Copyright of the University of Minnesota, 1988.

39. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379-423.

40. Rashevsky, N. Life, Information Theory and Topology. *Bull. Math. Biophys.* **1955**, *17*, 229-235.

41. Sarkar, R.; Roy, A. B.; Sarkar, R. K. Topological Information Content of Genetic Molecules - I. *Math. Biosci.* **1978**, *39*, 299-312.

42. Magnuson, V. R.; Harriss, D. K.; Basak, S. C. Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications. In *Studies in*

Physical and Theoretical Chemistry, King, R. B., Ed.; Elsevier: Amsterdam, 1983, pp. 178-191.

43. Basak, S. C.; Grunwald, G. D. APProbe: Copyright of the University of Minnesota, 1993.

44. SAS Institute Inc, in: *SAS/STAT User's Guide, Release 6.03 Edition* (SAS Institute Inc., Cary, NC, 1988) p. 751.

45. Wilkins, C. L.; Randić, M. A Graph Theoretical Approach to Structure-Property and Structure-Activity Correlations. *Theor. Chim. Acta* **1980**, *58*, 45-68.

Table 1. Mutagenicity in the Ames test for 113 chemicals.

No.[a]	Compound Name	Obs. Ames Mutagenicity
1.5	butylated hydroxyanisole (BHA)	0
1.6	caffeic acid	0
1.7	catechol	0
1.8	clofibrate	0
1.9	di(2-ethylhexyl)phthalate (DEHP)	0
1.10	hydroquinone	0
1.11	p-methoxyphenol	0
1.12	sesamol	0
1.13	tamoxifen	0
1.14	acetaminophen	0
1.15	benzoin	0
1.16	EPN	0
1.17	gallic acid	0
1.18	a-tocopherol	0
2.2	2-acethylaminofluorene (AAF)	1
2.3	adriamycin	1
2.4	aflatoxin B1	1
2.5	benzo[a]pyrene	1
2.7	captafol	1
2.8	captan	1
2.9	carbazole	1
2.10	dibutylnitrosamine (DBN)	1
2.11	diethylnitrosamine (DEN)	1
2.12	3,2'-dimethyl-4-aminobiphenyl (DMAB)	1
2.14	dimethylnitrosamine (DMN)	1
2.15	N-ethyl-N-hydroxyethylnitrosamine (EHEN)	1
2.16	N-ethyl-N-nitrosourea (ENU)	1
2.20	hydrazobenzene	1
2.22	laciocarpine	1
2.26	3'-methyl-4-dimethylaminoazobenzene (3'-Me-DAB)	1
2.27	3-amino-9-ethylcarbazole	1
2.28	N-nitrosooxazolidine	1
2.29	N-nitrosodi-n-propylamine (NDPA)	1
2.30	N-nitrosomorpholine	1
2.31	N-nitrosopiperidine	1
2.32	N-nitrosopyrrolidine	1
2.33	quinoline	1
2.34	sterigmatocystin	1
2.35	4,4'-thiodianiline	1
2.42	alachlor	0
2.43	aldrin	0
2.44	auramine O	0
2.45	barbital	0

2.46	chlordane	0
2.47	chlorendic acid	0
2.48	chlorobenzilate	0
2.49	DDT	0
2.50	dieldrin	0
2.51	diethylstilbestrol	0
2.53	ethenzamide	0
2.54	17α-ethinyl estradiol	0
2.55	DL-ethionine	0
2.56	hexachlorobenzene (HCB)	0
2.57	a-hexachlorocyclohexane (a-HCH)	0
2.58	d-limonene	0
2.59	monoclotaline	0
2.60	N-nitrosodiethanolamine	0
2.61	phenobarbital	0
2.64	safrole	0
2.66	thioacetamide	0
2.67	triadimefon	0
2.68	trifluralin	0
2.69	urethane	0
2.70	polychlorinated biphenyl (PCB)	0
2.71	malathion	0
2.72	vinclozolin	0
3.1	acetophenetidine (phenacetin)	1
3.2	azathioprine	1
3.3	N-butyl-N-(4-hydroxybutyl)nitrosamine (BBN)	1
3.4	chrysazin (danthron)	1
3.5	4,4'-diaminodiphenylmethane (DDPM)	1
3.6	7,12-dimethylbenz[a]anthracene (DMBA)	1
3.7	N-ethyl-N-(4-hydroxybutyl)nitrosamine (EHBN)	1
3.8	folpet	1
3.9	hydrogen peroxide	1
3.11	3-methylcholanthrene (3-MC)	1
3.12	N-methyl-N'-nitro-N-nitrosoguanidine (MNNG)	1
3.13	N-methyl-N-nitrosourea (MNU)	1
3.14	8-nitroquinoline	1
3.17	streptozotocin	1
3.18	o-toluidine	1
3.20	6-methylquinoline	1
3.21	8-methylquinoline	1
3.22	nitrofrantoln	1
3.23	6-nitroquinoline	1
3.24	quercetin	1
3.32	acetaldehyde	0
3.33	atrazine	0
3.34	di(2-ethylhexyl)adipate (DEHA)	0

3.35	1,1-dimethylhydrazine	0
3.39	trichloroacetic acid	0
3.42	4-acethylaminofluorene (AAF)	0
3.43	aspirin	0
3.44	butylated hydroxytoluene (BHT)	0
3.45	caffeine	0
3.46	caprolactam	0
3.47	chenodeoxicholic acid	0
3.49	cypermethrin	0
3.50	deltamethrin	0
3.51	diltiazem	0
3.52	dimethylsulfoxide (DMSO)	0
3.53	diazinon	0
3.54	fenvalerate	0
3.55	glutathione	0
3.56	4-o-hexyl-2,3,6-trimethylhydroquinone (HTHQ)	0
3.58	lithocolic acid	0
3.59	d-mannitol	0
3.61	phenol	0
3.64	propyl galiate	0
3.65	propylparaben	0
3.66	pyrene	0
3.67	resorcinol	0
3.71	trimorphamide	0

[a]The numbering scheme refers to the enumeration of the chemicals in the presentation by Yamaguchi *et al.* [27] where the numeral before the decimal place refers to the table in which the compound was listed (see below) and the numerals after the decimal refer to the compounds location within the table.

Table 1 – Association between inhibitory results in the medium-term liver bioassay (Ito test) and reported mutagenicity and carcinogenicity.

Table 2 – Association between positive results in the medium-term liver bioassay (Ito test) and reported mutagenicity and carcinogenicity.

Table 3 – Association between negative results in the medium-term liver bioassay (Ito test) and reported mutagenicity and carcinogenicity.

Table 2. Symbols and brief definitions for 101 topological indices.

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
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
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
IC_r	Mean information content or complexity of a graph based on the rth (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for rth (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for rth (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
${}^h\chi_{Ch}$	Chain connectivity index of order h = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^h\chi^b$	Bond path connectivity index of order h = 0-6
${}^h\chi_C^b$	Bond cluster connectivity index of order h = 3-6

${}^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3-6
${}^{h}\chi^{b}_{PC}$	Bond path-cluster connectivity index of order h = 4-6
${}^{h}\chi^{v}$	Valence path connectivity index of order h = 0-6
${}^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3-6
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3-6
${}^{h}\chi^{v}_{PC}$	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^Y	Balaban's J index based on relative covalent radii

Table 3. Eigenvalues, variance explained and two TIs most correlated with the eight principal components.

PC	Eigenvalue	Percent variance explained	Cumulative variance explained	First most correlated TI	Second most correlated TI
PC_1	55.52	54.97	54.97	$^{4}\chi^{b}$ (96.5%)	$^{3}\chi$ (96.4%)
PC_2	12.38	12.26	67.23	SIC_3 (86.4%)	SIC_4 (85.5%)
PC_3	11.73	11.61	78.84	$^{5}\chi^{b}{}_{Ch}$ (77.3%)	$^{5}\chi^{v}{}_{Ch}$ (76.1%)
PC_4	6.78	6.71	85.55	IC_0 (55.0%)	$^{4}\chi^{v}{}_{Ch}$ (49.7%)
PC_5	4.60	4.55	90.10	J (68.9%)	J^{Y} (62.4%)
PC_6	2.35	2.32	92.43	IC_0 (-47.2%)	SIC_0 (-36.4%)
PC_7	1.65	1.63	94.06	$^{4}\chi b_{C}$ (44.4%)	$^{4}\chi^{v}{}_{C}$ (43.5%)
PC_8	1.16	1.14	95.21	$^{4}\chi^{v}{}_{C}$ (-34.6%)	$^{6}\chi^{b}{}_{C}$ (23.0%)

Table 4. KNN results for the prediction of mutagenicity for 113 chemicals.

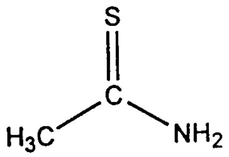
	Percent Negative Correct		**Percent Positive Correct**		**Total Percent Correct**	
K	**AP**	**ED**	**AP**	**ED**	**AP**	**ED**
1	73.5	75.0	84.1	66.7	77.7	71.7
2	66.2	64.7	72.7	33.3	68.8	52.2
3	77.9	80.9	88.6	53.3	82.1	69.9
4	70.6	69.1	77.3	42.2	73.2	58.4
5	79.4	77.9	86.4	53.3	82.1	68.1

Figure Captions

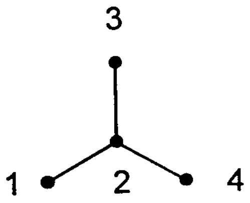
Figure 1 – Unlabeled, hydrogen-suppressed graph of thioacetamide (G_1).

Figure 2 – Labeled, hydrogen-filled graph of thioacetamide (G_2) and sample calculations for IC_1, SIC_1 and CIC_1.

Figure 2 – Analogs selected for isospectral graph 10.1.1.

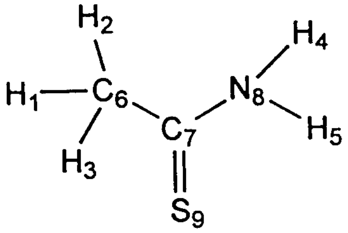


Thioacetamide

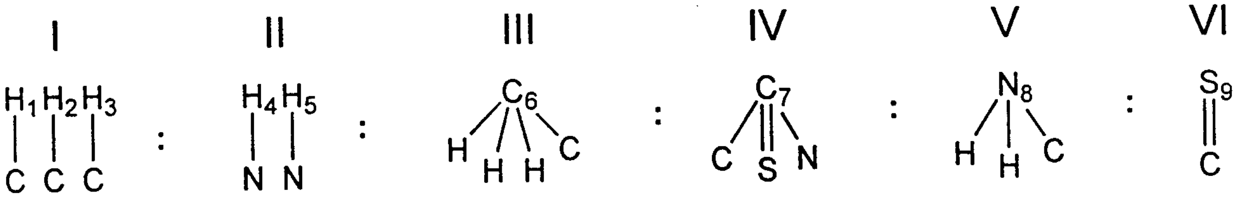


G_1

G_2: thioacetamide



First order neighbors:



Subsets:

I	II	III	IV	V	VI
$(H_1\text{-}H_3)$	$(H_4\text{-}H_5)$	C_6	C_7	N_8	S_9

Probability:

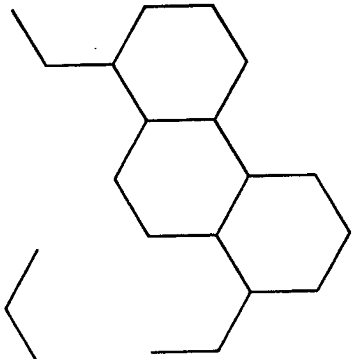
I	II	III	IV	V	VI
3/9	2/9	1/9	1/9	1/9	1/9

$IC_1 = 4 * 1/9 * \mathrm{Log}_2\ 9 + 2/9 * \mathrm{Log}_2\ 9/2 + 3/9 * \mathrm{Log}_2\ 9/3$ = 2.419 bits

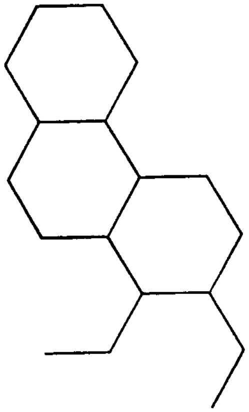
$SIC_1 = IC_1/\mathrm{Log}_2\ 9$ = 0.763 bits

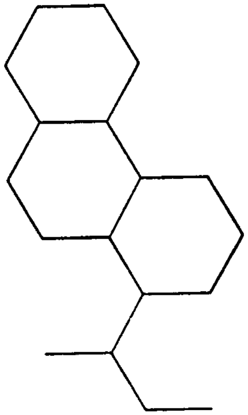
$CIC_1 = \mathrm{Log}_2\ 12 - IC_2$ = 0.751 bits

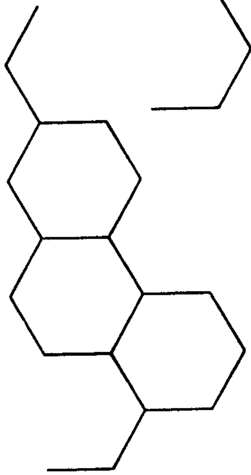
Probe

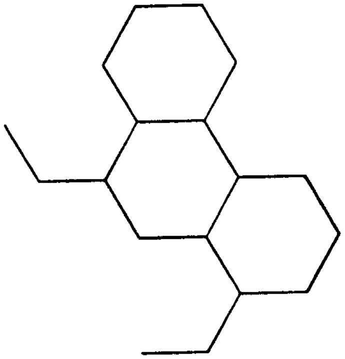


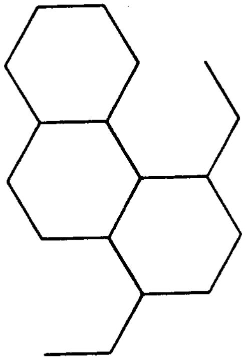
Atom Pair Method











Similarity Score

S=0.95

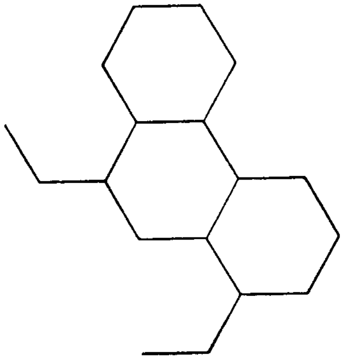
S=0.93

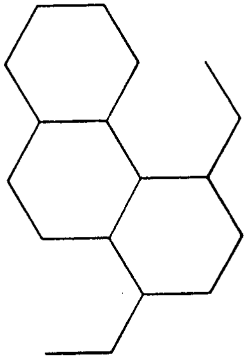
S=0.88

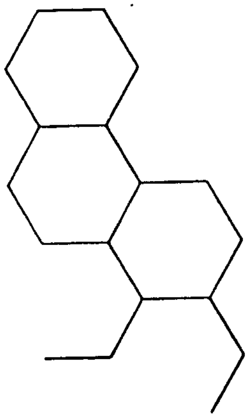
S=0.86

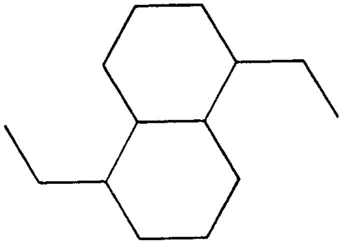
S=0.86

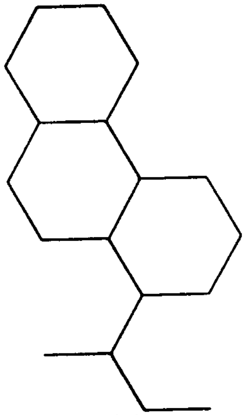
Euclidean Distance Method











Euclidean Distance

ED=0.19

ED=0.20

ED=0.20

ED=0.20

ED=0.21

Math. Model. Comput. Sci
Vol. 8, 1997

QUANTITATIVE COMPARISON OF FIVE MOLECULAR STRUCTURE SPACES IN SELECTING ANALOGS OF CHEMICALS

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

Natural Resources Research Institute, University of Minnesota - Duluth,
5013 Miller Trunk Highway, Duluth, MN 55811, USA
Phone: (218) 720-4230 E-Mail: sbasak@wyle.nrri.umn.edu

ABSTRACT

Five methods for characterizing intermolecular similarity have been used in the selection of analogs for a diverse set of seventy-six compounds. These methods include an atom pair (AP) based similarity measure, three principal component spaces derived from topostructural indices, topochemical indices, the combined set of all (topostructural and topochemical) indices, as well as one structure space consisting of principal components calculated from physicochemical properties. Each method has been used in the selection of sets analogs, ranging from five to forty in number in increments of five, for each of the seventy-six compounds. The degree of overlap of the sets of analogs selected by the five separate methods was analyzed.

KEYWORDS

molecular graph, atom pairs, principal components, analog selection, molecular similarity

INTRODUCTION

Molecular similarity is an intuitive concept which is subjectively understood by the chemist. In the realm of mathematical and computational chemistry, intermolecular similarity can be objectively quantified in terms of descriptors derived from the molecular structure (Basak et al, 1988b; Basak et al, 1997; Carbó et al, 1980; Fisanick et al, 1992; Fisanick et al, 1994; Johnson et al, 1988; Maggiora and Johnson, 1990; Randić, 1992; Willet and Winterman, 1986). Chemical structures can be represented by various types of models, *e.g.*, simple molecular graphs, multigraphs, pseudographs, 3-D models, and quantum chemical hamiltonian functions. Similarity, being context specific, is quantified in terms of a user-defined set of parameters or properties of molecules. Consequently, there are a potentially endless number of methods that one can define to quantify intermolecular similarity.

In recent years molecular similarity methods based on topological and substructural descriptors have become popular. Such methods are based on different types of graph invariants such as topological indices, atom pairs, and fragments (Basak and Grunwald, 1994, 1995c; Basak and Gute, 1997; Basak et al, 1988b; Carbó et al, 1980; Carhart et al, 1985; Fisanick et al, 1992; Johnson et al, 1988; Randić, 1992; Willet and Winterman, 1986). Similarity/dissimilarity methods have been used in the clustering of large sets of chemicals (Lajiness, 1990), the selection of analogs for toxicological risk assessment (Basak and Grunwald, 1994; Basak et al, 1995), and the estimation of the physicochemical and biomedicinal properties of chemicals (Basak and Grunwald, 1995a, 1995c; Basak et al, 1996a; Basak and Gute, 1997). Usually some number, *n*, of descriptors is used to define the structure space of chemicals and either Euclidean distance in the *n*-dimensional space or some association coefficient is used to quantify intermolecular similarity. The basic paradigm underlying molecular similarity analysis is "similar structures have similar properties." However, it has been shown that different molecular similarity methods select quite different sets of analogs from a specific database for the same set of query chemicals (Basak and Grunwald, 1995c). In the case of the automated selection of analogs for testing chemicals in drug design protocols or toxicological hazard assessment one would like to select analogs by reasonably non-redundant molecular similarity methods. Therefore, it is of interest to investigate the degree to which various similarity methods differ from each other. In a previous study we analyzed the analog selection profiles for topologically-based *vis-a-vis* empirical property-based molecular similarity techniques in the selection of nearest neighbors of molecules (Basak and Grunwald, 1995c). In this paper we have compared the analog selection profile of five different molecular similarity methods, four of which are based on graph invariants and one is derived from physicochemical property data.

DATABASE AND PARAMETERS

Development of the database

The data used in this study is a subset of the U.S. EPA ASTER system (Russom, 1992) which met the following criteria. These compounds have experimental values for:

1.	Log $K_{o/w}$	Logarithm of the octanol/water partition coefficient (hydrophobicity).
2.	BP	Boiling point at 760 Torr.
3.	MP	Melting point.

within the ASTER database. Kamlet (1987) provided the remaining physicochemical properties used in this study. These four solvatochromic parameters are:

1. *V*/100 — The molar volume of a molecule calculated as its molecular weight divided by the liquid density at 20° C.
2. α — A measure of the hydrogen bond donor acidity of a compound in forming a hydrogen bond.
3. β — A scale of the hydrogen bond acceptor basicity of a compound in forming a hydrogen bond.
4. π^* — A measure of solute or solvent dipolarity or polarizability that quantifies the ability of a compound to stabilize a neighboring charge or dipole by virtue of its dielectric effect.

Kamlet et al (1988) describe in detail the methods used in the determination of these solvatochromic parameters.

Calculation of Atom Pairs

Atom pairs (APs) were calculated using the method of Carhart *et al* (1985). An atom pair is defined as a substructure which consists of two non-hydrogen atoms *i* and *j* and their interatomic separation:

$$<\text{descriptor}_i>-<\text{separation}>-<\text{descriptor}_j>$$

where <descriptor> contains information about the element type, number of non-hydrogen neighbors, and the number of π electrons for each atom. The interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms. These calculations were conducted using the APProbe software developed by Basak and Grunwald (1993).

Calculation of Topological Indices

The topological indices used in this study have been calculated using the program POLLY 2.3 (Basak et al, 1988a) and software developed by the authors to calculate Balaban's *J* indices. A complete listing of these indices, along with examples of their calculation have been given in detail previously (Basak and Gute, 1997; Basak et al, 1997).

The topological indices were further divided into two subsets, topostructural and topochemical indices. Topostructural indices are topological indices which only encode information about the adjacency and distances of the vertices (atoms) within a graph (molecular structure), irrespective of the chemical nature of the atoms involved. The topochemical indices are parameters which quantify information regarding the topology of the graph (molecule), as well as specific chemical properties of the atoms and bonds comprising the molecule. These indices are derived from weighted graphs where each vertex (atom) or edge (bond) is properly weighted with selected chemical information. The division of the topological indices into these distinct sets has been discussed in previous studies (Basak et al, 1996b, 1997).

Similarity Measures

Two measures of intermolecular similarity were used in this study. The methods have been described in detail previously (Basak and Grunwald, 1995b) and include an associative measure using atom pairs (AP) and Euclidean distance (ED) within an *n*-dimensional principal component (PC) space. The Euclidean distance method was used in conjunction with the topological indices and the physicochemical property data.

ANALOG SELECTION

Following the quantification of intermolecular similarity for the five similarity spaces, the *K*-nearest neighbors or analogs ($K = 5, 10, 15, 20, 25, 30, 35, 40$) were determined on the basis of the associative measure used in conjunction with the AP method or based on ED within a principal component space.

RESULTS AND DISCUSSION

In generating the principal components for the sets of topological indices, only the principal components with eigenvalues greater than 1.0 were retained. This left six PCs for the set of topostructural indices which cumulatively explained 94.1% of the variance in the indices, eight PCs for the set of topochemical indices which explained 93.5% of the variance in these indices, and ten PCs for the set of all topological indices which cumulatively explained 95.2% of the variance in the topological indices. These formed the final sets of PCs which were used in creation of the similarity spaces and selection of analogs for these three methods.

Each similarity method was used to select sets of analogs for each of the seventy-six compounds in the dataset. The analogs selected by each set were compared with the analogs selected by every other method to examine the overlap between the sets of analogs. The results of this comparison are presented in Table 1 below as the arithmetic mean of the cardinalities of the intersection of subsets of analogs chosen by a particular pair of similarity methods for a specific value of *K*. For example, the topostructural and topochemical similarity methods selected an average of 2.2 identical analogs out of five for the entire set of seventy-six chemicals. Thus, slightly under half of the analogs selected by the two methods were identical.

It is clear from the data in Table 1 that the five molecular similarity methods studied in this paper are not radically different from one another because they have a substantial degree of overlap in the profile of selected neighbors. This is an interesting observation in view of the fact that the structure spaces are constructed from such diverse, independent variables as experimentally determined physicochemical properties and calculated graph invariants.

A perusal of the data also shows that the property-based similarity method is distinct from the group of methods based on topological indices and atom pairs. For $K = 20$, for example, the average number of common neighbors for the property-based methods *vis-a-vis* the topostructural, topochemical, all index and atom pair-based methods are 8.7, 8.9, 8.6 and 8.9, respectively. For the same value of *K*, the number of common analogs for the topostructural method with atom pair, topochemical and all index methods are 12.3, 12.2 and 13.1, respectively.

Table 1. Comparisons of the overlap in analog selection for five distinct similarity methods.

K	S vs C	S vs T	C vs T	S vs P	C vs P	T vs P	S vs A	C vs A	T vs A	P vs A
5	2.2	2.5	3.5	1.2	1.6	1.6	2.2	2.1	2.3	1.9
10	5.0	5.4	7.1	3.1	3.4	3.5	4.8	4.7	5.0	4.1
15	8.6	9.2	11.3	5.6	5.7	5.7	8.2	7.8	8.1	6.3
20	12.2	13.1	15.1	8.7	8.9	8.6	12.3	10.7	11.0	8.9
25	15.7	16.7	19.5	12.1	12.3	11.9	16.3	14.3	14.3	12.1
30	20.0	20.9	23.8	16.0	16.6	15.8	19.5	17.4	17.4	15.7
35	24.7	25.6	28.9	20.5	21.1	20.0	22.9	21.4	21.1	20.4
40	30.4	30.9	33.9	25.1	25.9	25.0	26.6	25.9	25.5	24.6

S = topostructural indices
C = topochemical indices
T = all topological indices
P = physicochemical parameters
A = atom pairs

For the three similarity methods calculated from the topological indices, the topochemical indices seem to have more influence on the selection of neighbors when they are used along with topostructural parameters as independent variables. This is clear from the fact that for almost all values of *K* the topochemical and all index methods have a uniformly higher degree of overlap as compared to that between the topostructural and all index methods.

In conclusion, if one is interested in selecting only two candidates from the set of five methods studied here for analog selection, the property-based method and any one of the theoretically-based methods would be the choice. There is no criteria to decide which of the four topologically-based methods should be selected for a particular occasion. Further studies of the analog selection and property prediction profile of these methods are necessary to guide the selection of a specific method for a particular practical situation.

ACKNOWLEDGMENTS

This paper is contribution number 204 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

Basak, S. C., S. Bertelsen and G. D. Grunwald (1995). Use of graph theoretic parameters in risk assessment of chemicals. Toxicol. Lett., 79, 239-250.

Basak, S. C. and G. D. Grunwald (1993). APProbe: Copyright of the University of Minnesota.

Basak, S. C. and G. D. Grunwald (1994). Molecular similarity and risk assessment: analog selection and property estimation using graph invariants, SAR QSAR Environ. Res., 2, 289-307.

Basak, S. C. and G. D. Grunwald (1995a). Estimation of lipophilicity from molecular structural similarity. New J. Chem., 19, 231-237.

Basak, S. C. and G. D. Grunwald (1995b). Molecular similarity and estimation of molecular properties. J. Chem. Inf. Comput. Sci., 35, 366-372.

Basak, S. C. and G. D. Grunwald (1995c). Use of topological space and property space in selecting structural analogs. Mathl. Model. Sci. Comput., in press.

Basak, S. C., B. D. Gute and G. D. Grunwald (1996a). Estimation of normal boiling points of haloalkanes using molecular similarity. Croat. Chem. Acta, 69, 1159-1173.

Basak, S. C., B. D. Gute and G. D. Grunwald (1996b). 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.

Basak, S. C. and B. D. Gute (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal *p*-hydroxylation of aniline by alcohols: a molecular similarity approach. In: Proceedings of the 2nd International Congress on Hazardous Waste: Impact on Human and Ecological Health (B.L. Johnson, C. Xintaras and J.S. Andrews, Jr., eds.), pp. 492-504, Princeton Scientific Publishing Co., Inc., New Jersey.

Basak, S. C., B. D. Gute and G. D. Grunwald (1997). Characterization of the molecular similarity of chemicals using topological invariants. In: Advances in Molecular Similarity: Highlights of the 3rd Girona Seminar on Molecular Similarity (P.G. Mezey, ed.), in press, JAI Press Inc, Greenwich, Connecticut.

Basak, S. C., D. K. Harriss and V. R. Magnuson (1988a). POLLY 2.3: Copyright of the University of Minnesota.

Basak, S. C., V. R. Magnuson, G. J. Niemi and R. R. Regal (1988b). Determining structural similarity of chemicals using graph theoretic indices. Discrete Appl. Math., 19, 17-44.

Carb6, R., L. Leyda and M. Arnau (1980). How similar is a molecule to another? An electron density measure of similarity between two molecular structures. Int. J. Quant. Chem., 17, 1185-1189.

Carhart, R. E., D. H. Smith and R. Venkataraghavan (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. J. Chem. Inf. Comput. Sci., 25, 64-73.

Fisanick, W., K. P. Cross and A. Rusinko, III (1992). Similarity searching on CAS registry substances. 1. global molecular property and generic atom triangle geometric searching. J. Chem. Inf. Comput. Sci., 32, 664-674.

Fisanick, W., A. H. Lipkus and A. Rusinko III (1994). Similarity searching on CAS registry substances. 2. 2D structural similarity. J. Chem. Inf. Comput. Sci., 34, 130-140.

Johnson, M., S. C. Basak and G. Maggiora (1988). A characterization of molecular similarity methods for property prediction. Mathl. Comp. Model., II, 630-634.

Kamlet, M. J. (1987). Personal communication.

Kamlet, M. J., R. M. Doherty, M. H. Abraham, Y. Marcus and R. W. Taft (1988). Linear solvation energy relationships. 46. An improved equation for correlation and prediction of octanol/water partition coefficients of organic nonelectrolytes (including strong hydrogen bond donor solutes). J. Phys. Chem., 92, 5244-5255.

Lajiness, M. (1990). Molecular similarity-based methods for selecting compounds for screening. In: Computational Chemical Graph Theory (D.H. Rouvray, ed.), pp. 299-316, Nova, New York.

Maggiora, G. M. and M. A. Johnson (1990). Introduction to molecular similarity. In: Concepts and Applications of Molecular Similarity (M. A. Johnson and G. M. Maggiora, eds.), pp. 1-13, John Wiley & Sons, Inc., New York.

Randić, M. (1992). Similarity based on extended basis descriptors. J. Chem. Inf. Comput. Sci., 32, 686-692.

Russom, C. L. (1992). Assessment Tools for the Evaluation of Risk, v. 1.0. U.S. Environmental Protection Agency.

Willett, P. and V. Winterman (1986). A comparison of some measures for the determination of intermolecular structural similarity. Quant. Struct. -Act. Relat., 5, 18-25.

APPENDIX 1.8 Assessment of the mutagenicity of chemicals from theoretical structural parameters

SAR and QSAR in Environmental Research
1999, Vol. 10, pp. 117 - 129
Reprints available directly from the publisher
Photocopying permitted by license only

ASSESSMENT OF THE MUTAGENICITY OF AROMATIC AMINES FROM THEORETICAL STRUCTURAL PARAMETERS: A HIERARCHICAL APPROACH*

S. C. BASAK[†], B. D. GUTE and G. D. GRUNWALD

Natural Resources Research Institute, 5013 Miller Trunk Hwy., Duluth, MN 55811, USA

(Received 8 July 1998; In final form 16 October 1998)

A hierarchical approach has been used in this paper in predicting the mutagenicity/non-mutagenicity of a set of 127 chemicals from their molecular descriptors. The set of descriptors consisted of topostructural and topochemical parameters, experimental properties like log *P*, and quantum chemical indices calculated using a semi-empirical method. The results show that a combination of topostructural and topochemical molecular descriptors explain most of the variance in the experimental data The addition of physical properties or quantum chemical parameters did not make any significant improvement in the predictive power of the models.

Keywords: Aromatic amines; hierarchical similarity; mutagenicity; quantum chemical descriptors; topological indices

INTRODUCTION

A current interest in the fields of chemistry, toxicology and biomedical sciences is the prediction of the property/activity of chemicals from calculated molecular descriptors [1 – 6]. In both environmental hazard assessment and pharmaceutical drug design, one has to deal with thousands, sometimes millions, of real or hypothetical chemical structures. Most of these compounds have very little of the experimental data necessary for the

*Presented at the workshop *Computational Methods in Toxicology*, April 20 – 22, 1998, Dayton, OH, USA.

[†]Corresponding author

estimation of their toxicity or efficacy. In this age of combinatorial chemistry, one can synthesize thousands of chemicals very quickly. However, experimental testing of these large numbers of chemicals would not be cost effective. Also, it is possible to create virtual libraries consisting of billions of structures. In this case one would like to know the toxic, as well as therapeutic, potential of such a vast collection of chemicals. The experimental data necessary for the prediction of the toxicity/activity of these large and diverse sets of chemicals will not be available to us in the near future.

This pervasive lack of experimental data demonstrates the need for the development of predictive models based on parameters that can be calculated directly from a chemical's molecular structure. Recently, our research group has been involved in the development of a hierarchical approach to quantitative structure-activity relationship (QSAR) model development for predicting physicochemical, toxicological and pharmacological properties of chemicals using theoretical molecular descriptors [3, 6 – 10]. Various topological indices (TIs) fall in this category of molecular descriptors [11 – 23]. Balaban has classified TIs into three generations based on whether they are integers, real numbers or a sequence of numbers [24]. Different classes of TIs quantify various aspects of molecular structure. We have shown in the past that various indices, *viz.*, connectivity indices and complexity indices developed and used by Basak *et al.* [15 – 18] quantify distinctly different types of molecular structural information. Such indices can be calculated very rapidly. On the other hand, geometrical and quantum chemical parameters encode information regarding the stereo-electronic aspects of molecules. These classes of parameters are also algorithmically derived, *i.e.*, they can be calculated for any real or hypothetical molecular structure without any input of experimental data.

One of our recent interests has been to test the relative effectiveness of the four classes of theoretical molecular descriptors mentioned above in the development of QSARs for predicting property/activity/toxicity of chemicals [3, 6 – 10]. In this paper we have used these parameters in the development of models for predicting mutagenicity/non-mutagenicity of a set of 127 aromatic amines.

METHODS

Datasets

A set of 127 aromatic and heteroaromatic amines, previously collected from the literature by Debnath *et al.* [25], were used to study mutagenicity. The mutagenicity of these compounds in *S. Typhimurium* TA98 + S9 microsomal preparation has been expressed as positive or negative mutagenicity by Benigni [26]. Compounds included in this study and their mutagenic classification based on experimentally determined mutagenic potency are given in Table I. Of the compounds used in this study, 106 were classified as mutagens while twenty-one were determined to be non-mutagens.

TABLE I Aromatic and heteroaromatic amines[1]

Chemicals	*TA98* (*Expt.*)	*TA98* (*Pred.*)[2]
2-Bromo-7-aminofluorene	1	1
2-Methoxy-5-methylaniline (*p*-cresidine)	1	1
5-Aminoquinoline	1	1
4-Ethoxyaniline (*p*-phenetidine)	1	1
1-Aminonaphthalene	1	1
4-Aminofluorene	1	1
2-Aminoanthracene	1	1
7-Aminofluoranthene	1	1
8-Aminoquinoline	1	1
1,7-Diaminophenazine	1	1
2-Aminonaphthalene	1	1
4-Aminopyrene	1	1
3-Amino-3′-nitrobiphenyl	1	1
2,4,5-Trimethylaniline	1	1
3-Aminofluorene	1	1
3,3′-Dichlorobenzidine	1	1
2,4-Dimethylaniline (2,4-xylidine)	1	1
2,7-Diaminofluorene	1	1
3-Aminofluoranthene	1	1
2-Aminofluorene	1	1
2-Amino-4′-nitrobiphenyl	1	1
4-Aminobiphenyl	1	1
3-Methoxy-4-methylaniline (*o*-cresidine)	1	0
2-Aminocarbazole	1	1
2-Amino-5-nitrophenol	1	1
2,2′-Diaminobiphenyl	1	1
2-Hydroxy-7-aminofluorene	1	1
1-Aminophenanthrene	1	1
2,5-Dimethylaniline (2,5-xylidine)	1	1
4-Amino-2′-nitrobiphenyl	1	1
2-Amino-4-methylphenol	1	1
2-Aminophenazine	1	1
4-Aminophenylsulfide	1	1
2,4-Dinitroaniline	1	1
2,4-Diaminoisopropylbenzene	1	1
2,4-Difluoroaniline	1	1
4,4′-Methylenedianiline	1	1
3,3′-Dimethylbenzidine	1	1
2-Aminofluoranthene	1	1
2-Amino-3′-nitrobiphenyl	1	1
1-Aminofluoranthene	1	1

TABLE I (Continued)

Chemicals	*TA*98 (*Expt.*)	*TA*98 (*Pred.*)[a]
4,4′-Ethylenebis(aniline)	1	1
4-Chloroaniline	1	1
2-Aminophenanthrene	1	1
4-Fluoroaniline	1	1
9-Aminophenanthrene	1	1
3,3′-Diaminobiphenyl	1	1
2-Aminopyrene	1	1
2,6-Dichloro-1,4-phenylenediamine	1	1
2-Amino-7-acetamidofluorene	1	1
2,8-Diaminophenazine	1	1
6-Aminoquinoline	1	1
4-Methoxy-2-methylaniline (*m*-cresidine)	1	1
3-Amino-2′-nitrobiphenyl	1	1
2,4′-Diamino-biphenyl	1	1
1,6-Diaminophenazine	1	1
4-Aminophenyldisulfide	1	1
2-Bromo-4,6-dinitroaniline	1	1
2,4-Diamino-*n*-butylbenzene	1	0
4-Aminophenylether	1	1
2-Aminobiphenyl	1	1
1,9-Diaminophenazine	1	1
1-Aminofluorene	1	1
8-Aminofluoranthene	1	1
2-Chloroaniline	1	0
2-Amino-aaa-trifluorotoluene	1	1
2-Amino-1-nitronaphthalene	1	1
3-Amino-4′-nitrobiphenyl	1	1
4-Bromoaniline	1	1
2-Amino-4-chlorophenol	1	1
3,3′-Dimethoxybenzidine	1	1
4-Cyclohexylaniline	1	1
4-Phenoxyaniline	1	1
4,4′-Methylenebis (*o*-ethylaniline)	1	0
2-Amino-7-Nitrofluorene	1	1
Benzidine	1	1
1-Amino-4-Nitronaphthalene	1	1
4-Amino-3′-Nitrobiphenyl	1	1
4-Amino-4′-Nitrobiphenyl	1	1
1-Aminophenazine	1	1
4,4′-Methylenebis (*o*-fluoroaniline)	1	1
4-Chloro-2-nitroaniline	1	1
3-Aminoquinoline	1	1
3-Aminocarbazole	1	1
4-Chloro-1,2-phenylenediamine	1	1
3-Aminophenanthrene	1	1
3,4′-Diaminobiphenyl	1	1
1-Aminoanthracene	1	1
1-Aminocarbazole	1	1
9-Aminoanthracene	1	1
4-Aminocarbazole	1	1
6-Aminochrysene	1	1
1-Aminopyrene	1	1
4-4′-Methylenebis(*o*-isopropyl-aniline)	1	0

TABLE I (Continued)

Chemicals	*TA98* *(Expt.)*	*TA98* *(Pred.)*[2]
2,7-Diaminophenazine	1	1
4-Aminophenanthrene	0	1
2,4-Diaminotoluene	1	1
3,3′-Diaminobenzidine	1	1
1,3-Phenylenediamine	1	0
3,4-Diaminotoluene	1	1
1,2-Phenylenediamine	1	0
3-Amino-6-methylphenol	1	1
2,4-Diaminoethylbenzene	1	1
4,4′-Methylenebis (2,6-diisopropylaniline)	0	0
4,4′-Methylenebis (2,6-diethylaniline)	0	0
4,4′-Methylenebis (2-methyl-6-*t*-butylaniline)	0	0
4,4′-Methylenebis (2-methyl-6-isopropylaniline)	0	0
4,4′-Methylenebis (2-methyl-6-ethylaniline)	0	0
4,4′-Methylenebis (2,6-dimethylaniline)	0	1
3-Aminobiphenyl	0	1
2,3-Diaminobiphenyl	0	1
2-Methyl-4-chloroaniline	0	1
2-Chloro-4-methylaniline	0	1
4-Methoxyaniline	0	1
3-Methoxyaniline	0	1
Aniline	0	0
3-Chloroaniline	0	0
3-Ethoxyaniline	0	1
2-Ethoxyaniline	0	1
4-Aminophenol	0	1
3-Aminophenol	0	0
2-Aminophenol	0	0
2-Methoxyaniline	0	1
4-Chloro-1,3-phenylenediamine	1	1
2-Nitro-1,4-phenylenediamine	1	1
4-Nitro-1,3-phenylenediamine	1	1
4-Nitro-1,2-phenylenediamine	1	1

[1] The table reports the mutagenicity of the aromatic and heteroaromatic amines as. 0 = negative; 1 = positive

[2] TA98 results predicted using topostructural and topochemical indices

Computation of Indices

Topological indices used in this study have been calculated by POLLY 2.3 [27] which can calculate a total of 102 indices. These indices include Wiener index [28], connectivity indices [11, 12], information theoretic indices defined on distance matrices of graphs [13, 14], a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [15 – 18], as well as Balaban's *J* indices [19 – 21]. Table II provides brief definitions for the topological indices included in this study.

TABLE II Symbols, definitions and classifications of topological parameters

	Topostructural
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
$\overline{\mathrm{IC}}$	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
${}^h\chi$	Path connectivity index of order $h = 0-6$
${}^h\chi_C$	Cluster connectivity index of order $h = 3-6$
${}^h\chi_{Ch}$	Chain connectivity index of order $h = 3-6$
${}^h\chi_{PC}$	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
	Topochemical
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
${}^h\chi^b$	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 $h = 3-6$
${}^h\chi_{PC}^b$	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 $h = 3-6$
${}^h\chi_{PC}^v$	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^Y	Balaban's J index based on relative covalent radii

Values for $\log P$ and the quantum chemical parameters ϵ_{HOMO} and ϵ_{LUMO} were taken from the work of Debnath *et al.* [25]. Octanol/water partition coefficients ($\log P$) were determined experimentally for a set of 67 aromatic and heteroaromatic amines and, when these values were determined to be in agreement with values calculated using the CLOGP program (release

3.54), the remainder of the log *P* values were calculated using CLOGP [29]. The quantum chemical parameters provided by Debnath *et al.*, ϵ_{HOMO} and ϵ_{LUMO} were calculated using the semi-empirical AM1 of MOPAC 4.10 (Quantum Chemistry Program Exchange No. 455) [30].

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 and other indices may equal zero.

The set of 95 TIs was partitioned into two distinct sets: 38 topostructural indices and 57 topochemical indices. Topostructural indices are 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 like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. The categorization of the 95 TIs into these sets is shown in Table II.

To further reduce the number of independent variables to be used for model construction, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [31]. This variable clustering procedure divides the set of indices into disjoint clusters such that each cluster is essentially unidimensional. The index most correlated with each cluster, as well as any indices which were poorly correlated with the cluster ($r < 0.70$), were selected for model development. Variable clustering was performed independently for both the topostructural and topochemical subsets.

Statistical Analysis and Hierarchical DFA

Selection of indices for the final models was conducted using all subsets regression on the sets of indices chosen through variable cluster analysis in the SAS procedure REG [32]. This all subsets procedure was performed on four distinct sets of indices: (1) the topostructural indices selected by variable clustering, (2) the topostructural indices selected in all subsets regression and the topochemical indices selected during variable clustering, (3) the topostructural and topochemical indices selected in all subsets regression and $\log P$, and 4) the model chosen for topostructural and topochemical indices with $\log P$ and with the addition of ϵ_{HOMO} and ϵ_{LUMO}. These sets of indices were then used to develop and crossvalidate discriminant function models for classifying the mutagenicity/non-mutagenicity of the 127 aromatic and heteroaromatic amines. Figure 1 illustrates the process for the selection of indices and formulation of DFA models.

RESULTS AND DISCUSSION

In the first step of our hierarchical modeling, 38 topostructural parameters were subjected to variable clustering procedure. The following indices were retained from the five clusters generated: $\mathrm{I}_D^W, \overline{\mathrm{IC}}, \mathrm{O}, {}^4\chi_C, {}^6\chi_{\mathrm{Ch}}, {}^4\chi_{\mathrm{PC}}, P_3, J$. These five clusters explained a total variation of 35.29 and the proportion of the variance explained was equal to 92.86%. Of the 57 topochemical indices, the following ten indices were selected from eight clusters: $\mathrm{IC}_0, \mathrm{IC}_2, \mathrm{IC}_4, \mathrm{SIC}_2, \mathrm{SIC}_4, {}^4\chi_C^b, {}^6\chi_{\mathrm{Ch}}^b, {}^4\chi_{\mathrm{PC}}^b, {}^2\chi^v, J^Y$. The eight clusters generated from the topochemical indices resulted in a total variation explained of 51.65 and the proportion of the variance explained was equal to 90.61%. These indices were then included in the all subsets regression procedure for the selection of final indices for discriminant function analysis. In all cases, the RSQUARE and ADJRSQ values were examined as indicators of model fit, however the final models were selected based on the Mallow's *Cp* statistic (CP). Statistics for the cluster analysis and the inter-correlation of the clusters for the topostructural indices are presented in Tables III and IV, respectively. Similar statistics for the variable clustering of the topochemical indices can be found in Tables V and VI.

The all subsets regression of the eight topostructural indices resulted in the selection of the following indices for model development: $\mathrm{I}_D^W, \overline{\mathrm{IC}}, P_3$. These indices were used to create the topostructural DFA model, the simplest model in the hierarchy, and were also combined with the ten topochemical indices to create the second model in the hierarchy. All subsets regression of the thirteen topostructural and topochemical indices resulted in the selection of the following indices for modeling: $\mathrm{I}_D^W, \overline{\mathrm{IC}}, P_3, \mathrm{IC}_0, \mathrm{SIC}_2$. These indices were combined with $\log P$ and resulted in a six parameter model with $\log P$ added to the complete set of descriptors from the second model. Finally, the quantum chemical descriptors, ϵ_{HOMO} and ϵ_{LUMO}, were combined with the set of six indices and all subsets regression was used again

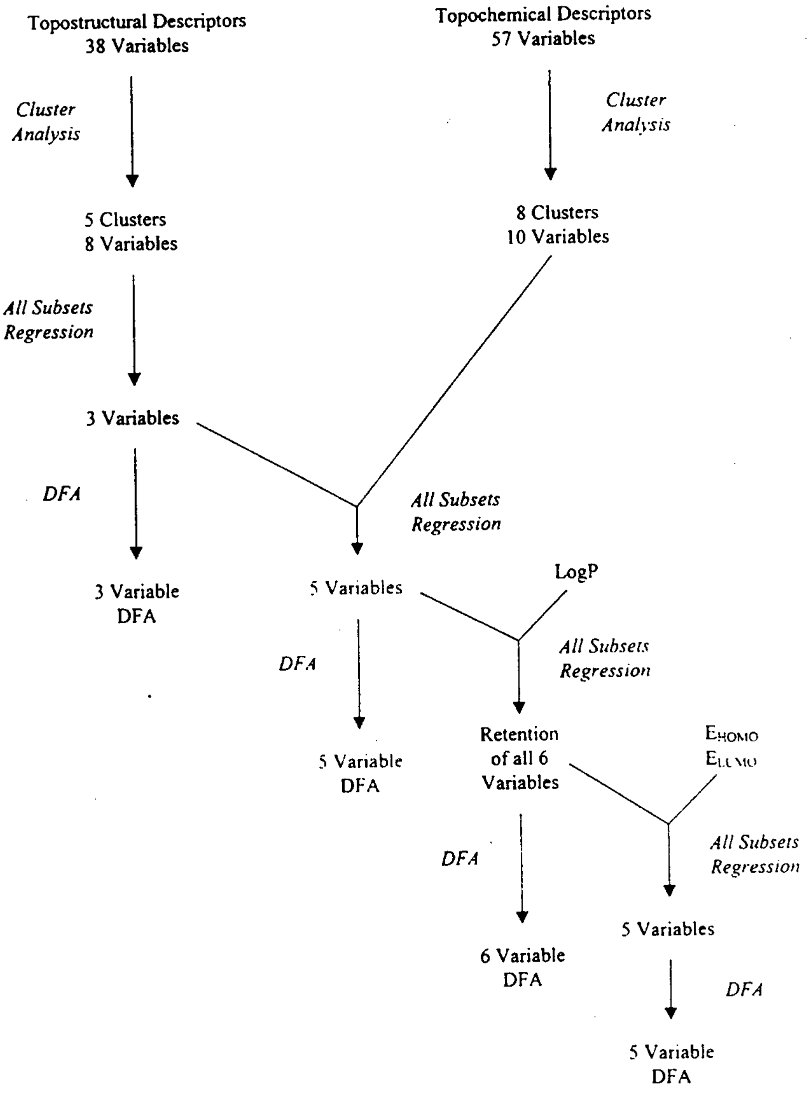


FIGURE 1 Illustration of the hierarchical method of index selection and discriminant function analysis.

to select the best parameters for model construction. This procedure resulted in the selection of the following model: I_D^W, $\overline{\mathrm{IC}}$, P_3, log P, $\in_{\mathrm{LUMO}}$.

Discriminant function analysis, using the SAS procedure DISCRIM [33], was used to develop models for predicting mutagenicity/non-mutagenicity

TABLE III Statistics for the variable cluster analysis of the topostructural indices

Cluster	*Members*	*Variation explained*	*Proportion explained*	*Second eigenvalue*	*Index most correlated*	*Correlation*
1	18	16.99	0.94	0.71	P_3	0.9918
2	2	2.00	1.00	0.00	$^{4}\chi_{C}$	0.9992
3	3	2.15	0.71	0.72	$^{6}\chi_{Ch}$	0.9104
4	12	11.41	0.95	0.45	I_D^W	0.9977
5	3	2.73	0.91	0.18	$^{4}\chi_{PC}$	0.9474

TABLE IV Intercorrelation of the clusters generated in the variable cluster analysis of the topostructural indices

Cluster	1	2	3	4	5
1	1.0000				
2	0.0735	1.0000			
3	0.6317	−0.0707	1.0000		
4	0.9327	0.1389	0.3922	1.0000	
5	0.7131	0.4006	0.2275	0.7793	1.0000

TABLE V Statistics for the variable cluster analysis of the topochemical indices

Cluster	*Members*	*Variation explained*	*Proportion explained*	*Second eigenvalue*	*Index most correlated*	*Correlation*
1	19	17.61	0.93	0.58	$^{2}\chi'$	0.9686
2	8	7.52	0.94	0.42	SIC_4	0.9876
3	4	3.76	0.94	0.24	$^{4}\chi_{C}^{b}$	0.9484
4	6	5.11	0.85	0.80	J^Y	0.8889
5	5	4.72	0.94	0.23	IC_4	0.9880
6	4	3.72	0.93	0.27	$^{6}\chi_{Ch}^{b}$	0.9419
7	6	4.68	0.78	0.79	SIC_2	0.9079
8	5	4.52	0.90	0.21	$^{4}\chi_{PC}^{b}$	0.9225

TABLE VI Intercorrelation of the clusters generated in the variable cluster analysis of the topochemical indices

Cluster	1	2	3	4	5	6	7	8
1	1.0000							
2	−0.4121	1.0000						
3	0.2311	−0.2150	1.0000					
4	−0.8162	0.4459	−0.0885	1.0000				
5	0.3407	0.6649	−0.0641	−0.2594	1.0000			
6	0.4739	0.2192	−0.0509	−0.4812	0.5033	1.0000		
7	−0.5604	0.4636	−0.1072	0.7565	−0.0130	−0.2089	1.0000	
8	0.7805	−0.5046	0.5542	−0.4287	0.0484	0.1481	−0.2913	1.0000

TABLE VII Results of the cross-validated discriminant function analyses

Hierarchical classes	*Indices*	*% Correct (non-mutagens)*	*% Correct (mutagens)*
Topostructural	$I_D^W, \overline{IC}, P_3$	28.6	95.3
Topostructural + Topochemical	$I_D^W, \overline{IC}, P_3, IC_0, SIC_2$	42.9	93.4
Topological + log P	$I_D^W, \overline{IC}, P_3, IC_0, SIC_2, \log P$	38.1	95.3
Topological + log P + Quantum chemical	$I_D^W, \overline{IC}, P_3, \log P, \epsilon_{LUMO}$	33.3	95.3

of chemicals in the Ames test. Four distinct models were developed using the indices selected from the all subsets regression procedure as described above. The results in Table VII shows that all four models could predict the mutagenicity of chemicals 93% to 95% of the time whereas they were less effective in predicting non-mutagenicity (29% to 43%).

The addition of topochemical to the set of topostructural indices, resulting in the best predictive model, are shown in Table VII. It is clear from the results that the addition of topochemical indices to the set of topostructural indices did slightly decrease the prediction of mutagenicity. However, there was a significant improvement in the prediction of non-mutagenicity by the addition of topochemical indices to the set of independent variables.

Finally, the addition of log P and quantum chemical indices did not make any improvement in the models. This is in line with our earlier work with physical and biochemical properties which showed that topostructural and topochemical indices explain most of the variance in the data [3, 6–10].

Acknowledgments

This paper is contribution number 232 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

[1] Hall, L. H. and Story, C. T. (1997). Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological states indices using artificial neural networks. *SAR QSAR Environ. Res.*, 6, 139–161.

[2] Trinajstić, N., Nikolić, S., Lucić, B., Amić, D. and Mihalić, Z. (1997). The detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.*, 37, 631–638.

[3] Gute, B. D. and Basak, S. C. (1997). Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, **7**, 117-131.

[4] Todeschini, R., Vighi, M., Finizio, A. and Gramatica, P. (1997). 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR QSAR Environ. Res.*, **7**, 173-193.

[5] Guo, M., Xu, L., Hu, C. Y. and Yu, S. M. (1997). Study on structure-activity relationship of organic compounds - Applications of a new highly discriminating topological index. *Math. Chem. (MATCH)*, **35**, 185-197.

[6] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1998). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, In press.

[7] 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 QSAR approach. *J. Chem. Inf. Comput. Sci.*, **37**, 651-655.

[8] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals. In: *Quantitative Structure-Activity Relationships in Environmental Sciences-7* (Chen, F. and Schüürman, G., Eds.). SETAC Press: Pensacola, FL, pp. 245-261.

[9] Gute, B. D., Grunwald, G. D. and Basak, S. C. (1999). Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, In press.

[10] 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.

[11] Kier, L. B. and Hall, L. H. (1986) *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press: Letchworth, Hertfordshire, U.K.

[12] Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609-6615.

[13] 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.

[14] Bonchev, D. and Trinajstić, N. (1977) Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **67**, 4517-4533.

[15] 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* (Avula, X. J. R., Bellman, R., Luke, Y. L. and Rigler, A. K., Eds.). University of Missouri-Rolla: Rolla, Missouri, pp. 851-856

[16] Basak, S. C. and Magnuson, V. R. (1983). Molecular topology and narcosis: A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim. Forsch.*, **33**, 501-503.

[17] 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* (Avula, X. J. R., Kalman, R. E., Lipais, A. I. and Rodin, E. Y., Eds.) Pergamon Press: New York, pp. 745-750.

[18] Basak, S. C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Med. Sci. Res.*, **15**, 605-609.

[19] Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399-404

[20] Balaban, A. T. (1983) Topological indices based on topological distances in molecular graphs. *Pure & Appl. Chem.*, **55**, 199-206.

[21] 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

[22] Kier, L. B. and Hall, L. H. (1990). An electrotopological-state index for atoms in molecules. *Pharm. Res.*, **8**, 801-807.

[23] Kier, L. B., Hall, L. H. and Frazer, J. W. (1991). An index of electrotopological state for atoms in molecules. *J. Math. Chem.*, 7, 229-241.

[24] Balaban, A. T. (1992). Using real numbers as vertex invariants for third-generation topological indices. *J. Chem. Inf. Comput. Sci.*, **32**, 23-28.

[25] Debnath, A. K., Debnath, G., Shusterman, A. J. and Hansch, C. (1992). A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.*, **19**, 37-52.

[26] Benigni, R., Andreoli, C. and Giuliani, A. (1994). QSAR models for both mutagenic potency and activity: Application to nitroarenes and aromatic amines. *Environ. Mol. Mutagen.*, **24**, 208-219.

[27] Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1988). POLLY 2.3: Copyright of the University of Minnesota.

[28] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17-20.

[29] Leo, A. (1988). CLOGP 3.54. Medicinal Chemistry Project, Pomona College, Claremont, CA.

[30] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. and Stewart, J. J. P. (1985). AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, **107**, 3902-3909.

[31] SAS Institute Inc. (1988). The VARCLUS procedure. In: *SAS/STAT User's Guide, Release* 6.03 *Edition*, SAS Institute Inc.: Cary, NC, Chapter 34, pp. 949-965.

[32] SAS Institute Inc. (1988). The REG procedure. In: *SAS/STAT User's Guide, Release* 6.03 *Edition*, SAS Institute Inc.: Cary, NC, Chapter 28, pp. 773-875.

[33] SAS Institute Inc. (1988). The DISCRIM procedure. In: *SAS/STAT User's Guide, Release* 6.03 *Edition*, SAS Institute Inc.: Cary, NC, Chapter 16, pp. 359-447.

APPENDIX 1.9 Prediction of complement-inhibitory activity of benzamidines using topological and...

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 the 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 I^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 biomedicinal 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 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.

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

no.	X	obsd log 1/*C*
77	3-O$(CH_2)_3OC_6H_4$-3-NHCONHC_6H_4-3[a]-SO_2F	4.23
95	3-O$(CH_2)_3OC_6H_4$-3-NHCONHC_6H_4-3-SO_2F	4.51
97	3-O$(CH_2)_3OC_6H_4$-3-NHCOC_6H_4-4-SO_2F	4.57
108	3-O$(CH_2)_3OC_6H_4$-3-NHCOC_6H_4-4-SO_2F	5.21

[a] This SO_2F group should be *meta*- instead of *para*-.

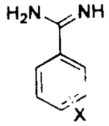


Figure 1. Neutral base structure for the 107 benzamidines.

Hansch and Yoshimoto[47] 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.nrri.umn.edu.

[†] The University of Minnesota.

[‡] Tufts University.

10.1021/ci9800716 CCC: $18.00
Published on Web 12/04/1998

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

		1/log *C*					1/log *C*		
no.	X	obsd	predict.^{a}	resid	no.	X	obsd	predict.^{a}	resid
1	3,5-(OCH$_3$)$_2$	−0.452	−0.367^b	−0.085	55	3-O(CH$_2$)$_3$OC$_6$H$_4$-2-NHCOC$_6$H$_3$-2-Cl-6-SO$_2$F	−0.255	−0.237	−0.018
2	2-CH$_3$	−0.444	−0.405	−0.040	56	3-O(CH$_2$)$_3$OC$_6$H$_4$-2-NHCONHC$_6$H$_5$	−0.255	−0.249	−0.006
3	3,4-(CH$_3$)$_2$	−0.425	−0.389	−0.036	57	3-O(CH$_2$)$_2$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$_2$)$_3$OC$_6$H$_4$-2-NHCONHCH$_2$C$_6$H$_4$-4-SO$_2$F	−0.250	−0.228	−0.022
5	3-OH	−0.415	−0.402	−0.012	59	3-O(CH$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$OC$_6$H$_4$-2-NHCOCH$_2$C$_6$H$_4$-4-SO$_2$F	−0.243	−0.233	−0.010
15	3-*i*-C$_5$H$_{11}$	−0.355	−0.353	−0.002	69	3-O(CH$_2$)$_3$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$_2$)$_3$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$_4$H$_9$	−0.338	−0.362	0.024	71	3-O(CH$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$OC$_6$H$_4$-2-NHCO(CH$_2$)$_4$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$OC$_6$H$_4$-4-COOH	−0.321	−0.277	−0.044	79	3-O(CH$_2$)$_4$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$_2$)$_3$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*-C$_5$H$_{11}$	−0.318	−0.341	0.022	81	3-O(CH$_2$)$_4$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$_{10}$H$_7$-α	−0.312	−0.283	−0.030	82	3-O(CH$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_5$H$_4$N	−0.299	−0.326	0.026	87	3-O(CH$_2$)$_3$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$_2$)$_3$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$_2$)$_4$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_2$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$OC$_6$H$_4$-3-NHCOCH$_2$C$_6$H$_4$-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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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$_2$)$_2$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$_2$)$_4$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$_2$)$_3$OC$_6$H$_4$-3-NHCOC$_6$H$_4$-4-NO$_2$	−0.214	−0.233	0.019
49	3-O(CH$_2$)$_4$OC$_6$H$_4$-4-NHCONHC$_6$H$_4$-4-SO$_2$F	−0.260	−0.219	−0.040	103	3-O(CH$_2$)$_2$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$_2$)$_4$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$_2$)$_3$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$_2$)$_3$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$_2$)$_3$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 concentration for 50% inhibition of complement (I_{50}), for 108 benzamidines. The numbered ordering used by Hansch and Yoshimoto will be used in this manuscript as well for 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 benzamidine derivatives. The base structure of the benzamidines is presented in Figure 1, while their side chains and biological activities, both measured and estimated, are presented in Table 2.

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
$\overline{IC}$	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 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
IC_r	mean information content or complexity of a graph based on the rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) 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^Y	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 *Sybyl*[64] 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.

The set of 92 TIs was divided into two distinct sets: topostructural indices (TSI) and topochemical indices (TCI). 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 accomplished using the SAS procedure REG.[66] During the initial stages of statistical analysis it became apparent that it

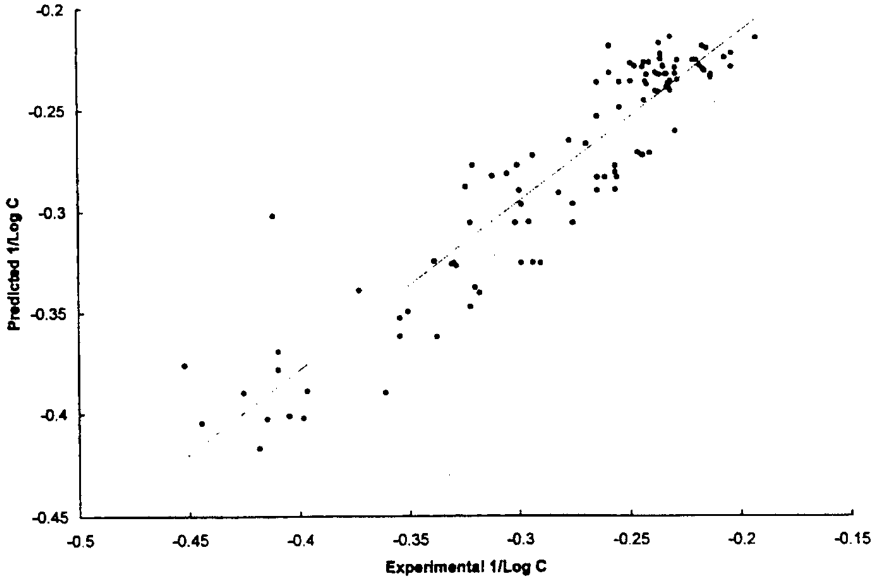


Figure 2. Scatterplot for observed 1/log C versus predicted 1/log C using eq 2 for the set of 107 benzamidines.

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

topological	topochemical	geometric
I_D^W	I_{ORB}	V_w
$\bar{I}_D^W$	$IC_0 - IC_6$	${}^{3D}W$
W	$SIC_0 - SIC_6$	${}^{3D}W_H$
I^D	$CIC_0 - CIC_6$	
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}:

$$1/\log C = -1.1245 + 0.4989(I^D) \qquad (1)$$

$$n = 105,\ r = 0.940,\ r_c = 0.938,\ s = 0.0200,\ F = 785$$

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:

$$1/\log C = -0.6428 + 0.0490({}^{3D}W) \qquad (2)$$

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

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]

In comparing our study to the work of Hanch and Yoshimoto,[47] it must be pointed out that our models did little to improve on their QSAR analysis as can be seen from examining our retransformed results with the results of their best equation.

	n	*r*	*s*
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 P_{450}: 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.

(8) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. Correlation Between Structure and Normal Boiling Points of Haloalkanes C_1-C_4 using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118–1121.

(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.

(34) Johnson, M.; Basak, S. C.; Maggiora, G. A Characterization of Molecular Similarity Methods for Property Prediction. *Mathematical Comput. Modelling* **1988**, *11*, 630–635.

(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.; Gute, 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. 1. 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 Triazenes: 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-Phenoxypropoxybenzamidine. *J. Med. Chem.* **1969**, *12*, 1053−1056.

(51) Baker, B. R ; Cory, M. Irreversible Enzyme Inhibitors. 180. Irreversible Inhibitors of the C'la Component of Complement Derived from m-(Phenoxypropoxy) benzamidine and Phenoxyacetamide. *J. Med. Chem.* **1971**, *14*, 119−125.

(52) Baker, B. R.; Cory, M. Irreversible Enzyme Inhibitors. 186. Irreversible Inhibitors of the C'la Component of Complement Derived from m-(Phenoxypropoxy) benzamidine 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 IInd 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 (LC_{50}) 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.

CI9800716

APPENDIX 1.10 Use of statistical and neural net methods in predicting toxicity of chemicals

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.nrri.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, biomedicinal 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 increasing levels 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: ${}^{3D}W_H$ 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

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 compounds 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,

Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach

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

Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received December 31, 1996[⊗]

Numerous quantitative structure−activity relationships (QSARs) have been developed using topostructural, topochemical, and geometrical molecular descriptors. However, few systematic studies have been carried out on the relative effectiveness of these three classes of parameters in predicting properties. We have carried out a systematic analysis of the relative utility of the three types of structural descriptors in developing QSAR models for predicting vapor pressure at STP for a set of 476 diverse chemicals. The hierarchical technique has proven to be useful in illuminating the relationships of different types of molecular description information to physicochemical property and is a useful tool for limiting the number of independent variables in linear regression modeling to avoid the problems of chance correlations.

1. INTRODUCTION

A large number of quantitative structure−activity relationship (QSAR) studies have been reported in recent literature using theoretical molecular descriptors in predicting physicochemical, pharmacological, and toxicological properties of molecules.[1−15] Such descriptors comprise graph invariants, geometrical or 3-D parameters, and quantum chemical indices. One of the reasons for the current upsurge of interest is the fact that such descriptors can be derived algorithmically, i.e., can be computed for any molecule, real or hypothetical, using standard software. Both in pharmaceutical drug design and in risk assessment of chemicals, one has to evaluate potential biological effects of chemicals. Evaluation schemes based on property−property correlation paradigms are not very useful in practical situations, because, for most of the candidate structures, the experimental data necessary for proper evaluation are not available. This is especially true for the thousands of chemicals rapidly produced by methods of combinatoric chemistry[16] as well as for the large number of chemicals present in the Toxic Substances Control Act (TSCA) Inventory.[17]

A large number of physicochemical and biological endpoints are necessary for estimating the ecotoxicological fate, transport, and effects of environmental pollutants.[17−19] The vapor pressure of chemicals is important in determining the partitioning of chemicals among different phases once they are released in the environment. Many QSARs have been reported for predicting normal vapor pressure of chemicals. Such studies are usually carried out on small sets of congeneric chemicals. Also, many QSARs use experimental data as inputs in the model. Therefore, it becomes necessary to develop QSARs based on nonempirical parameters which can predict the vapor pressure for a heterogeneous collection of chemicals so that such models are generally applicable. With this end in mind, in the current paper we have carried out a QSAR study of 476 diverse chemicals using three types of nonempirical molecular descriptors.

2. MATERIALS AND METHODS

2.1. Normal Vapor Pressure Database. Measured values for a subset of the Toxic Substances Control Act (TSCA) Inventory[17] were obtained from the ASTER (Assessment Tools for the Evaluation of Risk) database.[20] This subset consisted of a diverse set of chemicals where vapor pressure (p_{vap}) was measured at 25 °C and over a pressure range of approximately 3−10 000 mmHg. Due to the size of the dataset being used in this study, data for these chemicals will not be listed in this paper. An electronic copy of the data may be obtained by contacting the authors.

2.2. Computation of Topological Indices. The majority of the topological indices (TIs) used in this study have been calculated by the computer program POLLY 2.3.[21] These indices include Wiener index,[22] the molecular connectivity indices developed by Randić and Kier and Hall,[1,23] information theoretic indices defined on distance matrices of graphs,[24,25] and a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs.[2,26−28] Balaban's *J* indices[29−31] were calculated using software developed by the authors.

van der Waal's volume (V_w)[32−34] was calculated using Sybyl 6.2.[35] The 3-D Wiener numbers[36] were calculated by Sybyl using an SPL (Sybyl Programming Language) program developed by the authors. Calculation of 3-D Wiener numbers consists of the summation of the 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 3.2.1.[37] Two variants of the 3-D Wiener number were calculated, $^{3D}W_H$ and ^{3D}W, where hydrogen atoms are included and excluded from the computations, respectively.

Table 1 provides a complete listing of all of the topological and geometrical parameters which have been used in this study. The listing includes the symbols used to represent the parameters and brief definitions for each of the parameters.

Two additional parameters were used in modeling normal vapor pressure, HB_1, and dipole moment (μ). HB_1 is a simple hydrogen bonding parameter calculated using a program developed by Basak,[38] which is based on the ideas

* All correspondence should be addressed to Dr. Subhash C. Basak, Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811.

[⊗] Abstract published in *Advance ACS Abstracts,* June 1, 1997.

S0095-2338(96)00176-X CCC: $14.00

Table 1. Symbols and Definitions of Topological and Geometrical Parameters

I^{W}_{D}	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I^{W}_{D}}$	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
$\overline{\mathrm{IC}}$	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 IC_r reaches it 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
IC_r	mean information content or complexity of a graph based on the r^{th}($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complmentary information content for rth ($r = 0-5$) 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 $h = 5, 6$
${}^{h}\chi^{b}$	bond path connectivity index of order $h = 0-6$
${}^{h}\chi^{b}_{C}$	bond cluster connectivity index of order $h = 3-6$
${}^{h}\chi^{b}_{Ch}$	bond chain connectivity index of order $h = 5, 6$
${}^{h}\chi^{b}_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
${}^{h}\chi^{v}$	valence path connectivity index of order $h = 0-6$
${}^{h}\chi^{v}_{C}$	valence cluster connectivity index of order $h = 3-6$
${}^{h}\chi^{v}_{Ch}$	valence chain connectivity index of order $h = 5, 6$
${}^{h}\chi^{v}_{PC}$	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^{Y}	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

of Ou *et al.*[39] Dipole moment was calculated using Sybyl 6.2.[35]

2.3. Data Reduction. The set of 92 TIs was partitioned into two distinct subsets: topostructural indices and topochemical indices. The distinction was made as follows: topostructural indices 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 like hybridization states of atoms and number of core/valence electrons in individual atoms, while topochemical indices quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. 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 2.

The partitioning of the indices left 38 topostructural indices and 54 topochemical indices. At this point no further data reduction is called for, since the ratio of the number of observations in the training set (342) to the total number of variables (92 maximum) falls well within the condition limits suggested by Topliss and Edwards[40] for reducing the probability of spurious correlations even at the more conservative $R^2 \geq 0.7$ level.

Table 2. Classification of Parameters used in Modeling Normal Vapor Pressure [$\log_{10}(p_{\mathrm{vap}})$]

topological	topochemical	geometric	other parameters
I_{D}^{W}	I_{ORB}	V_{W}	HB_1
$\overline{I_{D}^{W}}$	$\mathrm{IC}_0-\mathrm{IC}_5$	${}^{3D}W$	μ
W	$\mathrm{SIC}_0-\mathrm{SIC}_5$	${}^{3D}W_{H}$	
I^{D}	$\mathrm{CIC}_0-\mathrm{CIC}_5$		
H^{V}	${}^{0}\chi^{b}-{}^{6}\chi^{b}$		
H^{D}	${}^{3}\chi^{b}_{C}-{}^{6}\chi^{b}_{C}$		
$\overline{\mathrm{IC}}$	${}^{5}\chi^{b}_{Ch}$ and ${}^{6}\chi^{b}_{Ch}$		
O	${}^{4}\chi^{b}_{PC}-{}^{6}\chi^{b}_{PC}$		
M_1	${}^{0}\chi^{v}-{}^{6}\chi^{v}$		
M_2	${}^{3}\chi^{v}_{C}-{}^{6}\chi^{v}_{C}$		
${}^{0}\chi-{}^{6}\chi$	${}^{5}\chi^{b}_{Ch}$ and ${}^{6}\chi^{b}_{Ch}$		
${}^{3}\chi_{C}-{}^{6}\chi_{C}$	${}^{4}\chi^{b}_{PC}-{}^{6}\chi^{b}_{PC}$		
${}^{5}\chi_{Ch}$ and ${}^{6}\chi_{Ch}$	J^{B}		
${}^{4}\chi_{PC}-{}^{6}\chi_{PC}$	J^{X}		
P_0-P_{10}	J^{Y}		
J			

2.4. Statistical Analysis and Hierarchical QSAR. 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. The geometric parameters were transformed by the natural logarithm of the parameter.

Two regression procedures were used in developing the linear models. When the number of independent variables was high, typically greater than 25, a stepwise regression procedure was used to maximize the improvement of the explained variance (R^2). When the number of independent variables was smaller, all possible subsets regression was used. Models were then optimized to reduce problems of variance inflation and collinearity. Regression modeling was conducted using the REG procedure of the statistical package SAS.[41]

The vapor pressure data (p_{vap}) was split into a training set (342 compounds) and a test set (134 compounds), an approximately 75/25 split. Models were developed using the training set of chemicals and then used to predict the p_{vap} values of the test chemicals. Final models were then developed using the combined training and test set of chemicals.

Five sets of indices were used in model development. These sets were constructed as part of a hierarchical approach to QSAR modeling. The hierarchy begins with the simplest indices, the topostructural. After developing our initial model utilizing the topostructural indices, we increase the level of complexity. To the indices included in the best topostructural model, we add all of the topochemical indices and proceed to model p_{vap} using these parameters. Likewise, the indices included in the best model from this procedure are combined with the geometrical indices and modeling is conducted once again. In addition to this hierarchical approach, models were also constructed using the topochemical indices alone and the geometrical indices alone for purposes of comparison.

3. RESULTS

Stepwise regression analyses for $\log_{10}(p_{\mathrm{vap}})$ of the training set of chemicals is summarized in Table 3. As shown in

Table 3. Summary of the Regression Results for the Training Set and the Prediction Results for the Test Set for the Hierarchical Analysis of $\log_{10}(p_{\mathrm{vap}})$

	training set ($N = 342$)				test set ($N = 134$)	
parameter class	variables included	F	R^2	s	R^2	s
topostructural	${}^1\chi$, ${}^6\chi_C$, P_9	104.6	48.1	0.56	57.9	0.46
topochemical	SIC_0, SIC_2, SIC_3, CIC_0, CIC_1, ${}^3\chi^b{}_C$, ${}^1\chi^v$, ${}^5\chi^v$, ${}^3\chi^v{}_C$, J^Y	126.3	79.2	0.36	85.8	0.27
geometrical	${}^{3D}W$, ${}^{3D}W_H$, V_W	168.9	51.8	0.53	62.2	0.44
topostructural + topochemical	${}^1\chi$, P_9, IC_1, SIC_2, CIC_1, ${}^3\chi^b{}_C$, ${}^1\chi^v$, ${}^3\chi^v$, ${}^6\chi^v$, ${}^3\chi^v{}_C$, ${}^5\chi^v{}_{Ch}$	112.5	80.4	0.35	84.7	0.28
all indices	H^V, SIC_1, SIC_2, CIC_0, CIC_3, ${}^6\chi_C$, ${}^1\chi^v$, ${}^3\chi^v$, ${}^6\chi^v{}_C$, P_6, P_{10}	117.4	79.6	0.35	84.2	0.28
ttg + HB_1 + μ	${}^1\chi$, P_3, P_9, IC_0, ${}^1\chi^b$, ${}^3\chi^b{}_C$, ${}^1\chi^v$, ${}^3\chi^v$, ${}^3\chi^v{}_C$, HB_1	160.8	82.9	0.32	83.1	0.29

Table 3, the topostructural model using three parameters resulted in an explained variance (R^2) of 48.1% and a standard error (s) of 0.56. Addition of the topochemical parameters to the three topostructural parameters led to a significant increase in the effectiveness of the model. The resulting model used 12 parameters, two topostructural and ten topochemical. This model had an R^2 of 80.4% and s of 0.35. All subsets regression of the two topostructural and ten topochemical indices retained thus far and the three geometrical indices resulted in the selection of the same 12 parameter model, thus the geometrical indices did not contribute significantly to model development. Several other models were constructed for comparative purposes. Using topochemical indices only, a ten parameter model was developed which had an R^2 of 79.2% and s of 0.36. A geometrical model was developed which utilized all three geometrical indices and resulted in an R^2 of 51.8% and s of 0.53. Finally, two additional stepwise models were developed. One model simply used all indices for a comparison between a simple stepwise analysis of the data and the results of the hierarchical procedure. This resulted in an 11 parameter model with R^2 of 79.6% and s of 0.35. The second model added two new parameters, HB_1 and μ. We thought that it might be possible to improve our modeling by adding in some other nonempirical parameters which could be important to the determination of normal vapor pressure. We selected the parameters HB_1 and μ, since they would be important in intermolecular interactions which could have a dramatic effect on vapor pressure. To look at the addition of these parameters, we conducted a stepwise regression analysis using all topostructural, topochemical, and geometric indices so that we would be able to optimize our model, just as we had done with the previous models. The addition of these parameters led to the selection of a ten parameter model which included three topostructural indices, nine topochemical indices, and HB_1. This was the best model yet, with an R^2 of 82.9% and s of 0.32.

Application of these six models to the test set of chemicals resulted in comparable R^2 and s; actually all models improved slightly on their predictions of the test set, and these values are also listed in Table 3. Based on these results, we decided that it was pointless to develop further models using only geometrical parameters. Also, based on the findings that the geometrical indices did not contribute significantly to any of the training models, they were dropped from the development of final models for the full set of 476 chemicals. However, even though the topostructural indices did not perform well in modeling vapor pressure by themselves, they will be used in model development since they did contribute significantly to most of the models.

Regression analyses of the combined set of 476 chemicals showed similar results for estimating $\log_{10}(p_{\mathrm{vap}})$ as analysis of the training set. Using only the topostructural indices, stepwise regression analysis resulted in a five parameter model to estimate vapor pressure:

$$\log_{10}(p_{\mathrm{vap}}) = 4.88 + 0.20(\mathrm{O}) - 2.56({}^1\chi) + 0.49({}^4\chi_C) + 0.79({}^6\chi_C) + 0.98(\mathrm{P}_{10}) \quad (1)$$

$$n = 476, \quad R^2 = 51.5\%, \quad s = 0.53, \quad F = 99.7$$

Stepwise regression using the five topostructural parameters and all topochemical parameters resulted in the selection of the following seven parameter model:

$$\log_{10}(p_{\mathrm{vap}}) = 8.44 - 1.77({}^1\chi) + 1.25(\mathrm{P}_{10}) - 5.69(\mathrm{IC}_1) + 3.91(\mathrm{IC}_2) - 1.24(\mathrm{IC}_5) + 1.41({}^3\chi^b{}_C) - 1.70({}^1\chi^v) \quad (2)$$

$$n = 476, \quad R^2 = 79.3\%, \quad s = 0.34, \quad F = 224.0$$

Only two of the topostructural indices used in eq 1 were retained by the stepwise regression procedure used to produce eq 2: ${}^1\chi$ and P_{10}. The improvement in R^2 was significant, increasing from 51.5% for eq 1 to 79.3% for eq 2. Also, the model error decreased significantly, dropping by 0.19 logarithmic units. Since we have dropped the geometrical indices, this becomes our final hierarchical model.

The stepwise regression analysis of only topochemical parameters resulted in a 12 parameter model:

$$\log_{10}(p_{\mathrm{vap}}) = 6.65 - 3.44(\mathrm{IC}_0) - 1.33(\mathrm{IC}_5) + 3.47(\mathrm{SIC}_2) + 0.87(\mathrm{CIC}_1) - 0.48({}^4\chi^b) + 1.44({}^3\chi^b{}_C) - 1.00({}^1\chi^v) - 0.41({}^3\chi^v) - 0.70({}^5\chi^v) - 1.08({}^3\chi^v{}_C) + 1.42({}^6\chi^v{}_{Ch}) - 1.23(\mathrm{J}^Y) \quad (3)$$

$$n = 476, \quad R^2 = 75.8\%, \quad s = 0.38, \quad F = 120.5$$

This model which is inferior to the topostructural + topochemical model (eq 2), because its variance explained is lower and, more importantly, it requires more independent variables (parameters) to achieve this explanation of variance.

Stepwise regression of all indices resulted in the selection of an 11 parameter model. This approach selected three topostructural indices and eight topochemical indices to arrive at the following model:

$$\log_{10}(p_{\mathrm{vap}}) = 7.85 - 2.56(\mathrm{H}^V) + 1.17({}^6\chi_C) - 5.01(\mathrm{IC}_1) + 3.65(\mathrm{IC}_2) - 0.99(\mathrm{IC}_5) + 0.51(\mathrm{CIC}_1) - 1.54({}^1\chi^v) - 0.36({}^3\chi^v) - 0.36({}^4\chi^v) - 1.40({}^6\chi^v{}_C) \quad (4)$$

$$n = 476, \quad R^2 = 80.4\%, \quad s = 0.33, \quad F = 173.4$$

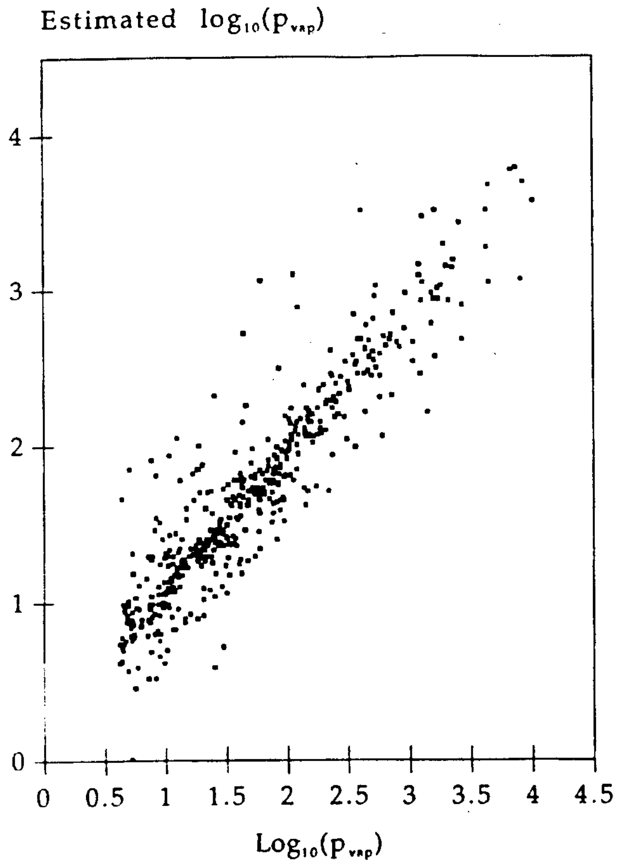


Figure 1. Scatterplot of observed $\log_{10}(p_{vap})$ *vs* estimated $\log_{10}(p_{vap})$ using eq 5 for 476 diverse compounds.

While eq 4 shows some slight improvements over eq 2, the hierarchical model, eq 2 is preferred since it is a simpler model using seven indices instead of 11 and based on a comparison of *F* values it is a more robust model than that in eq 4.

Finally, we conducted the stepwise regression modeling using all topostructural and topochemical indices with HB_1 and *u* for the complete set of 476 chemicals. The resulting ten parameter model used three topostructural indices, six topochemical indices, and HB_1:

$$\log_{10}(p_{vap}) = 9.67 - 3.66(^1\chi) + 0.35(P_3) + 0.74(P_9) - 1.78(IC_0) - 3.33(SIC_1) - 0.81(CIC_2) + 2.05(^2\chi^b) - 1.73(^2\chi^v) - 0.79(^3\chi^v) - 0.29(HB_1) \quad (5)$$

$$n = 476,\quad R^2 = 84.3\%,\quad s = 0.29,\quad F = 249.5$$

Equation 5 shows marked improvement over eq 2, justifying the addition of indices to the model. Also, it meets the criteria on which eq 4 was judged to be lacking. Overall, there is an improvement in variance explained of 5%, with a comparable decrease in standard deviation. A scatter plot of observed $\log_{10}(p_{vap})$ versus estimated $\log_{10}(p_{vap})$ using eq 5 is presented in Figure 1.

4. DISCUSSION

The purpose of this paper was 2-fold: (a) to study the utility of algorithmically-derived molecular descriptors in developing QSAR models for predicting the vapor pressure of chemicals from structure and b) to investigate the relative roles of topostructural, topochemical, and geometrical indices in the estimation of standard vapor pressure.

Table 4. Summary of the Chemical Class Composition of the Normal Vapor Pressure Dataset

compd classification	no. of compds	pure	substituted
total normal vapor pressure dataset	476		
hydrocarbons	253		
non-hydrocarbons[a]	223		
nitro compounds	4	3	1
amines	20	17	3
nitriles	7	6	1
ketones	7	7	0
halogens	100	95	5
anhydrides	1	1	0
esters	18	16	2
carboxylic acids	2	2	0
alcohols	10	6	4
sulfides	39	38	1
thiols	4	4	0
imines	2	2	0
epoxides	1	1	0
aromatic compounds[b]	15	10	4
fused-ring compounds[c]	1	1	0

[a] The non-hydrocarbons are further broken down into the following groups. [b] The 15 aromatic compounds are a mixture of 11 aromatic hydrocarbons and four aromatic halides. [c] The only fused-ring compound was a polycyclic aromatic hydrocarbon.

Results described in this paper (eqs 1–5) show that nonempirical parameters derived predominantly from graph theoretic models of molecules can estimate normal vapor pressure of diverse chemicals reasonably well. The explained variance of data ($R^2 = 84.3\%$) is excellent in view of the fact that the database of chemicals analyzed in this paper is very diverse (see Table 4). It should be mentioned that most published QSAR models for the estimation of vapor pressure have dealt with much smaller data sets with limited structural variety.[42,43]

The relative effectiveness of topostructural, topochemical, and geometrical indices in predicting normal vapor pressure of chemicals is evident from the result presented above. Equation 1 explains over 51% of variance in the data. All parameters used to derive eq 1 are topostructural, *i.e.*, they are parameters which encode information about the adjacency and distance of vertices in skeletal molecular graphs without quantifying any explicit information about such chemical aspects like bond order, electronic character of atoms, etc. Yet, the high explained variance of the property indicates that adjacency and distance in chemical graphs, being general descriptors of molecular size, shape, and branching, are important in predicting properties. This may explain the success of parameters like simple connectivity indices in estimating many diverse properties.[1]

Equation 3 is derived only from topochemical indices. The explained variance of vapor pressure (75.8%) shows that topochemical parameters, as a class, explain a larger fraction of the variance as compared to models derived from only topostructural indices (eq 1). Geometrical parameters were dropped from the set of descriptors after their limited success in prediction for the training and test sets. This is in line with our earlier studies with normal boiling point and hydrophobicity, where it was reported that the addition of geometrical indices could not significantly improve the predictive power of QSAR models derived from a combined set of topostructural and topochemical parameters.[15] It would be interesting to see whether this pattern holds good for other properties as well. Finally, the addition of the simple nonempirical parameter, HB_1, which contains information relevant to intermolecular interactions further improves our ability to estimate normal vapor pressure resulting in an explained variance of 84.3% (eq 5).

ACKNOWLEDGMENT

This is contribution number 209 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, a grant from Exxon Corporation and the Structure–Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to dedicate this paper to Professor Milan Randić in appreciation of his contributions in chemical information, quantitative structure–activity relationships, and chemical graph theory.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986

(2) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res* **1987**, *15*, 605–609.

(3) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. Correlation Between Structure and Normal Boiling Points of Haloalkanes C_1-C_4 Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118–1121

(4) Basak, S. C. A Nonempirical Approach to Predicting Molecular Properties Using Graph-Theoretic Invariants. In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht/Boston/London, 1990, pp 83–103

(5) Basak, S. C.; Bertelsen, S.; Grunwald, G. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270–276

(6) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett* **1995**, *79*, 239–250

(7) Basak, S. C., Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathematical Modelling and Scientific Computing*. In press.

(8) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR and QSAR in Environ. Res* **1994**, 2, 289–307.

(9) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity From Molecular Structural Similarity. *New J. Chem* **1995**, *19*, 231–237

(10) 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 Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73–116

(11) Basak, S. C., Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: 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

(12) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood-Brain Transport of Drugs: A Computational Approach. *Pharm Res* **1996**, *13*, 775–778

(13) Basak, S. C., Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math Modelling Sci. Computing* In press.

(14) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Estimation of Normal Boiling Points of Haloalkanes Using Molecular Similarity. *Croat Chem Acta* **1996**, *69*, 1159–1173

(15) 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.

(16) Martin, Y. C. Opportunities for Computational Chemists Afforded by the New Strategies in Drug Discovery: An Opinion. *Network Science* **1996**.

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

(18) NRC. *Toxicity Testing: Strategies to Determine Needs and Priorities*; National Academy Press: Washington, DC, 1984; p 84.

(19) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243–272.

(20) Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. ASTER: An Integration of the AQUIRE Data Base and the QSAR System for Use in Ecological Risk Assessments. *Sci. Total Environ.* **1991**, *109/110*, 667–670.

(21) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY Version 2.3*; Copyright of the University of Minnesota, 1988.

(22) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(23) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(24) 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.

(25) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.

(26) 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.

(27) 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: Rolla, MO, 1980; p 745.

(28) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon: New York, 1984; p 745.

(29) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.

(30) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199–206.

(31) 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.

(32) Bondi, A. van der Waal's Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(33) Moriguchi, I.; Kanada, Y. Use of van der Waal's Volume in Structure–Activity Studies. *Chem. Pharm. Bull.* **1977**, *25*, 926–935.

(34) Moriguchi, I.; Kanada, Y.; Komatsu, K. van der Waal's Volume and the Related Parameters for Hydrophobicity in Structure–Activity Studies. *Chem. Pharm. Bull.* **1976**, *24*, 1799–1806.

(35) Tripos Associates, Inc. *SYBYL Version 6.2*; Tripos Associates, Inc.: St. Louis, MO, 1994.

(36) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the Three-Dimensional Wiener Number. *J. Math. Chem.* **1989**, *3*, 299–309.

(37) Tripos Associates, Inc. *CONCORD Version 3.2.7*; Tripos Associates, Inc.: St. Louis, MO, 1995.

(38) Basak, S. C. *H-Bond*; Copyright of the University of Minnesota, 1988.

(39) Ou, Y. C.; Ouyang, Y.; Lien, E. J. *J. Mol. Sci.* **1986**, *4*, 89.

(40) Topliss, J. G.; Edwards, R. P. Chance Factor in Studies of Quantitative Structure–Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(41) 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.

(42) Drefahl, A.; Reinhard, M. *Handbook for Estimating Physico-Chemical Properties of Organic Compounds*; Stanford University Bookstore, Stanford, CA, 1995.

(43) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*; McGraw-Hill Book Company: New York, 1982.

CI960176D

APPENDIX 1.11 Characterization of the molecular similarity of chemicals using topological invariants

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 descripors: 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.

CHARACTERIZATION OF THE MOLECULAR SIMILARITY OF CHEMICALS USING TOPOLOGICAL INVARIANTS

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

Advances in Molecular Similarity,
Volume 2, pages 171–185.

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 ASTER[18] 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 Randić[20] 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 Trinajstić[22] 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 $n \times n$ 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):

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

W is calculated as

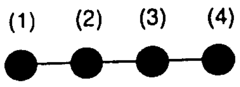
$$W = 1/2\sum_{ij} d_{ij} = \sum_{h} h \bullet g_h \tag{1}$$

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

	Topostructural
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
$^{h}\chi$	Path connectivity index of order h = 0–6
$^{h}\chi$	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 = 3–6
P_h	Number of paths of length h = 0–10
J	Balaban's J index based on distance
	Topochemical
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 rth (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for rth (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for rth (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
$^{h}\chi^{b}$	Bond path connectivity index of order h = 0–6
$^{h}\chi^{b}_{C}$	Bond cluster connectivity index of order h = 3–6
$^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3–6
$^{h}\chi^{b}_{PC}$	Bond path–cluster connectivity index of order h = 4–6
$^{h}\chi^{v}$	Valence path connectivity index of order h = 0–6
$^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3–6
$^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3–6
$^{h}\chi^{v}_{PC}$	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^Y	Balaban's J index based on relative covalent radii



G₁

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:

$$^{h}X = \sum_{\text{paths}} (v_i v_j \ldots v_{h+1})^{-1/2} \tag{2}$$

where $v_i, v_j, \ldots, v_{h+1}$ are the degrees of the vertices in the path of length h. The path length parameters (P_h), number of paths of length h ($h = 0, 1, \ldots, 10$) 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 ($i = 1, 2, \ldots, h$; $\Sigma_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, \ldots, A_h$$

$$p_1, p_2, \ldots, p_h$$

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

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

$$IC = -\sum_{i=1}^{h} p_i \log_2 p_i \tag{3}$$

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 rth-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 $u_0, u_1, \ldots, u_r$ of length r, there is a distinct path $v_0, v_1, \ldots, v_r$ 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 rth-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

$$SIC_r = IC_r/\log_2 n \tag{4}$$

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

$$CIC_r = \log_2 n - IC_r \tag{5}$$

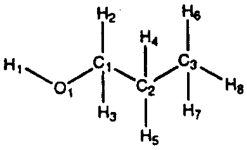
CIC_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r.[25]

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.

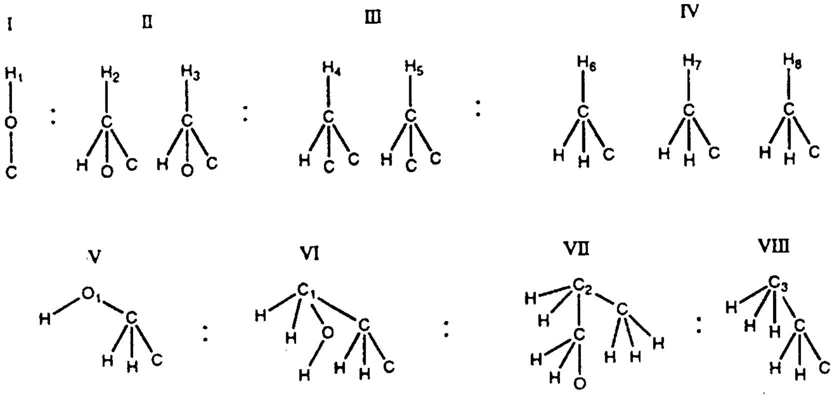
The information-theoretic index on graph distance, I_D^W, is calculated from the distance matrix $D(G)$ of a chemical graph G as follows:[22]

$$I_D^W = W\log_2 W - \sum_h g_h \cdot h\log_2 h \tag{6}$$

G_2: n-propanol



Second order neighbors:



Subsets:

I	II	III	IV	V	VI	VII	VIII
(H_1)	$(H_2$-$H_3)$	$(H_4$-$H_5)$	$(H_6$-$H_8)$	(O_1)	(C_1)	(C_2)	(C_3)

Probability:

I	II	III	IV	V	VI	VII	VIII
1/12	2/12	2/12	3/12	1/12	1/12	1/12	1/12

$IC_2 = 5*1/12*\log_2 12 + 2*2/12*\log_2 12/2 + 3/12*\log_2 12/3 = 2.855$ bits

$SIC_2 = IC_1/\log_2 12 = 0.796$ bits

$CIC_2 = \log_2 12 - IC_2 = 0.730$ bits

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, $\overline{I_D^W}$, 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:

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

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

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

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, ${}^{1}\chi$
2	11.0	18.0	64.3	${}^{4}\chi_{PC}$, ${}^{5}\chi_{PC}$, ${}^{6}\chi_{PC}$
3	5.9	9.6	73.9	${}^{3}\chi_{C}$, ${}^{5}\chi_{C}$, ${}^{4}\chi_{PC}$
4	4.1	6.7	80.6	J, ${}^{6}\chi_{Ch}$, ${}^{4}\chi_{C}$
5	2.8	4.6	85.2	${}^{4}\chi_{Ch}$, ${}^{5}\chi_{Ch}$, ${}^{3}\chi_{Ch}$
6	1.9	3.1	88.3	${}^{3}\chi_{Ch}$, ${}^{4}\chi_{Ch}$, ${}^{5}\chi_{Ch}$
7	1.5	2.4	90.8	${}^{6}\chi_{C}$, 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	${}^{1}\chi^{b}$, ${}^{2}\chi^{b}$, ${}^{3}\chi^{b}$
2	10.8	17.8	51.2	SIC_4, SIC_3, SIC_5
3	8.1	13.3	64.6	${}^{3}\chi^{b}_{C}$, ${}^{4}\chi^{b}_{C}$, ${}^{4}\chi^{b}_{PC}$
4	6.1	9.9	74.5	${}^{5}\chi^{b}_{Ch}$, ${}^{5}\chi^{v}_{Ch}$, ${}^{4}\chi^{b}_{Ch}$
5	3.0	5.0	79.5	${}^{3}\chi^{b}_{Ch}$, ${}^{3}\chi^{v}_{Ch}$, ${}^{4}\chi^{b}_{Ch}$
6	2.4	3.9	83.4	IC_0, SIC_0, IC_1
7	1.7	2.8	86.2	${}^{6}\chi^{b}_{C}$, ${}^{5}\chi^{b}_{C}$, ${}^{6}\chi^{v}_{C}$
8	1.4	2.2	88.4	${}^{4}\chi^{v}_{C}$, ${}^{2}\chi^{v}$, ${}^{6}\chi^{v}_{C}$
9	1.2	2.0	90.4	${}^{5}\chi^{v}_{C}$, ${}^{6}\chi^{v}_{C}$, ${}^{4}\chi^{b}_{C}$
10	1.1	1.8	92.1	${}^{4}\chi^{b}_{C}$, ${}^{4}\chi^{v}_{C}$, ${}^{6}\chi^{v}_{PC}$

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, ${}^{1}\chi$
2	13.3	13.0	54.7	${}^{4}\chi^{b}_{PC}$, ${}^{4}\chi_{PC}$, ${}^{3}\chi_{C}$
3	11.4	11.1	65.8	SIC_5, SIC_6, CIC_6
4	8.9	8.7	74.5	${}^{5}\chi^{b}_{Ch}$, ${}^{5}\chi_{Ch}$, ${}^{5}\chi^{v}_{Ch}$
5	5.1	5.0	79.6	J, ${}^{4}\chi_{Ch}$, ${}^{4}\chi^{b}_{Ch}$
6	3.7	3.6	83.2	IC_0, SIC_0, SIC_1
7	2.6	2.6	85.8	${}^{6}\chi^{b}_{C}$, ${}^{6}\chi_{C}$, ${}^{5}\chi^{b}_{C}$
8	2.0	1.9	87.7	${}^{4}\chi^{v}_{C}$, ${}^{5}\chi^{v}_{Ch}$, ${}^{6}\chi^{v}_{Ch}$
9	1.7	1.7	89.4	${}^{4}\chi^{v}_{C}$, IC_0, SIC_0
10	1.4	1.4	90.8	${}^{5}\chi^{v}_{C}$, ${}^{4}\chi^{v}_{PC}$, ${}^{6}\chi^{v}_{C}$
11	1.1	1.1	91.9	IC_1, J^{X}, 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.

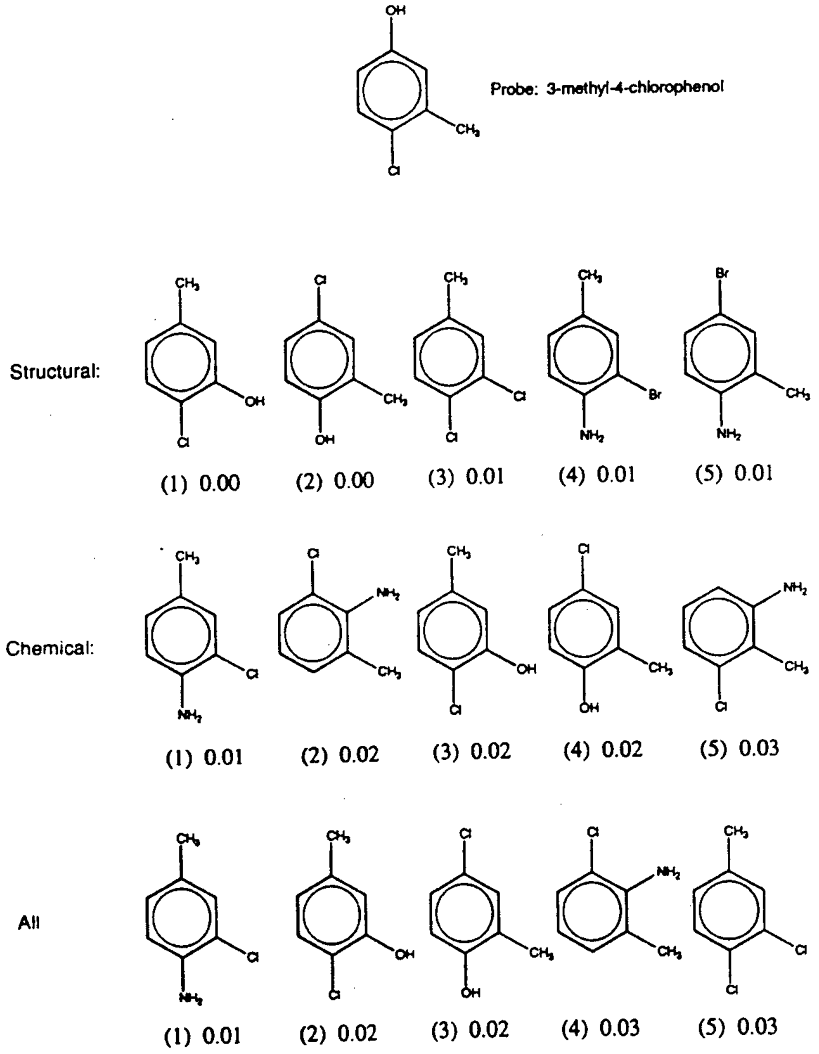


Figure 3. The five analogues selected for the probe 3-methyl-4-chlorophenol using three molecular similarity spaces: topostructural, topochemical, and all indices. The numbers under the structures indicate the ranking of the analogues and the Euclidean distance to the probe.

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)

Indices	*K*	*r*	*s*
Topostructural	10	0.881	39.0
Topochemical	6	0.883	38.6
Topostructural + topochemical	8	0.896	36.6

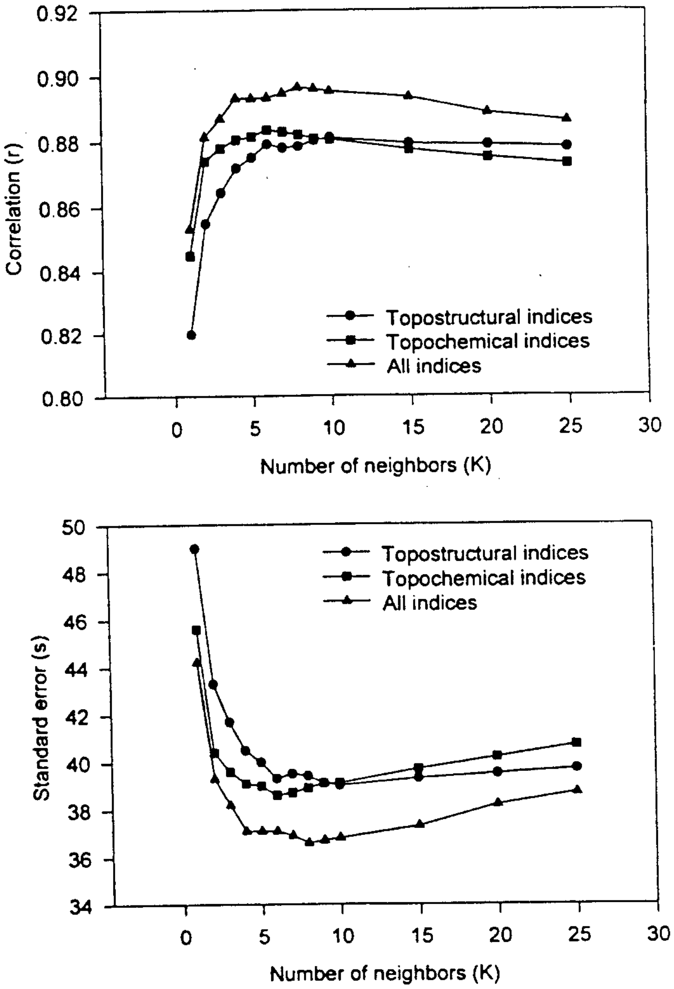


Figure 4. Pattern of (top) correlation (*r*) and (bottom) standard error (*s*) of the estimates according to the *K*-nearest-neighbor selection for 2926 normal boiling points using three molecular similarity spaces.

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.

APPENDIX 1.12 Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs)

SAR and QSAR in Environmental Research
1999, Vol. 10, pp. 1 – 15
Reprints available directly from the publisher
Photocopying permitted by license only

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 cm^2 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

No.	*Compound*	*CAS No.*	*Act.*	*Pred. Act.*	*Resid.*
1	Coronene	191-07-1	0.70	7.18	−6.48
2	Dibenzo(*a*, *l*)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(*a*, *i*)pyrene	189-55-9	8.00	7.18	0.82
6	3-Methylcholanthene	56-49-5	8.00	11.89	−3.89
7	Benzylhydrilindenefluorene	1836-87-9	8.00	16.93	−8.93
8	7,10-Dimethylbenzo(*a*)pyrene	63104-33-6	8.30	11.57	−3.27
9	Indeno(1,2,3:*c*,*d*)pyrene	193-39-5	9.00	11.52	−2.52
10	Dibenz(*a*, *h*)anthracene	53-70-3	9.40	13.29	−3.89
11	Benzo(*e*)pyrene	192-97-2	10.00	19.68	−9.68
12	Benzo(*g*, *h*, *i*)perylene	191-24-2	10.00	13.19	−3.19
13	9-*p*-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(*a*)anthracene	316-14-3	14.00	22.40	−8.40
17	Benzo(*k*)fluoranthene	207-08-9	14.00	17.99	−3.99
18	Benzo(*a*)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-*tert*-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(*c*)phenanthrene	56-49-5	20.00	11.89	8.11
24	10-Methylbenz(*a*)anthracene	2381-15-9	20.00	22.49	−2.49
25	5-Methylbenz(*a*)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	*n*-Butylpyrene	35980-18-8	20.00	13.27	6.73
30	5,6-Dihydro-4H-dibenz (*a*, *k*, *l*)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(*c*)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(*a*)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-*m*-Toylfluorene	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(*a*)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(*a*)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)

No.	Compound	CAS No.	Act.	Pred. Act.	Resid.
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-Vinylanthracene	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(*d, e, f*)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)

Topostructural	*Topochemical*	*Geometric*	*Quantum chemical*
I_D^W	I_{ORB}	V_W	E_{HOMO}
$\bar{I}_D^W$	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	${}^0X^b$-${}^6X^b$		ΔH_f
H^D	${}^3X^b_c$ & ${}^5X^b_c$		μ
$\overline{IC}$	${}^5X^b_{Ch}$ & ${}^6X^b_{Ch}$		
O	${}^4X^b_{PC}$-${}^6X^b_{PC}$		
M_1	${}^0X^v$-${}^6X^v$		
M_2	${}^3X^v_c$ & ${}^5X^v_c$		
${}^0X$-${}^6X$	${}^5X^v_{Ch}$ & ${}^6X^v_{Ch}$		
${}^3X_c$-${}^5X_c$	${}^4X^v_{PC}$-${}^6X^v_{PC}$		
${}^5X_{Ch}$ & ${}^6X_{Ch}$	J^B		
${}^4X_{PC}$-${}^6X_{PC}$			
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

	Topostructural
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
$\overline{IC}$	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
${}^{h}\chi$	Path connectivity index of order $h = 0-6$
${}^{h}\chi_C$	Cluster connectivity index of order $h = 3-5$
${}^{h}\chi_{PC}$	Path-cluster connectivity index of order $h = 4-6$
${}^{h}\chi_{Ch}$	Chain connectivity index of order $h = 5$ & 6
P_h	Number of paths of length $h = 0-10$
J	Balaban's *J* index based on distance
	Topochemical
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
${}^{h}\chi^b$	Bond path connectivity index of order h =0-6
${}^{h}\chi_C^b$	Bond cluster connectivity index of order $h = 3$ & 5
${}^{h}\chi_{Ch}^b$	Bond chain connectivity index of order $h = 5$ & 6
${}^{h}\chi_{PC}^b$	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 $h = 3$ & 5
${}^hX^v_{Ch}$	Valence chain connectivity index of order $h = 5$ & 6
${}^hX^v_{PC}$	Valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
	Geometric
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 chemicals 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_{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), dipole moment (μ), and HOMO/LUMO gap ($E_{HOMO}-E_{LUMO}$). 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, ${}^1X^b$, ${}^{3D}W$

No	*MW*	*log P*	P_0	${}^1\chi^b$	${}^{3D}W$
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)

No	*MW*	*log P*	P_0	$^1\chi^b$	^{3D}W
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: $\overline{\mathrm{IC}}$, O, $^3\mathrm{X}$-$^5\mathrm{X}$, $^6\mathrm{X}_{\mathrm{Ch}}$, 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:

$$\begin{gathered}\%\mathrm{DP} = 224.1 - 67.9P_0 \\ n = 60 \; r^2 = 0.675 \; s = 7.4 \; F = 120.6\end{gathered} \tag{1}$$

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

$$\begin{gathered}\%\mathrm{DP} = 179.7 - 78.8\,^1\mathrm{X}^{\mathrm{b}} \\ n = 60 \; r^2 = 0.695 \; s = 7.1 \; F = 132.0\end{gathered} \tag{2}$$

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, $^1\mathrm{X}^{\mathrm{b}}$ 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:

$$\begin{gathered}\%\mathrm{DP} = 186.0 - 25.4\,^{3\mathrm{D}}\mathrm{W} \\ n = 60 \; r^2 = 0.673 \; s = 7.4 \; F = 119.3\end{gathered} \tag{3}$$

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:

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

The values for the parameters used in the final models (P_0, ${}^1\mathrm{X}^\mathrm{b}$, ${}^{3\mathrm{D}}\mathrm{W}$, 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 (${}^1\mathrm{X}^\mathrm{b}$), geometrical descriptors (${}^{3\mathrm{D}}\mathrm{W}$) 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, ${}^1\mathrm{X}^\mathrm{b}$, 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 ${}^1\mathrm{X}^\mathrm{b}$ and Figure 2 demonstrates the scatter of the residuals. Thus, there are no apparent co-variance problems within this model.

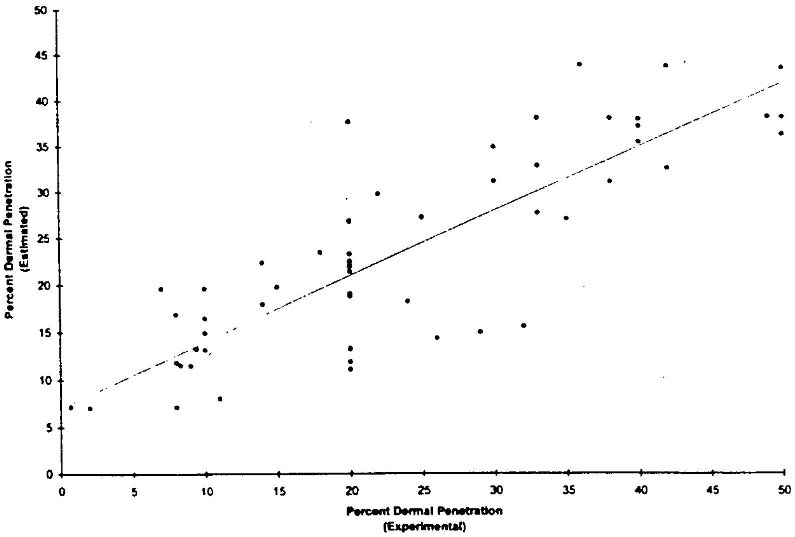


FIGURE 1 Scatterplot of experimentally determined percent dermal penetration (%DP) *vs.* estimated %DP using Eq. (2) for a set of 60 polycyclic aromatic hydrocarbons.

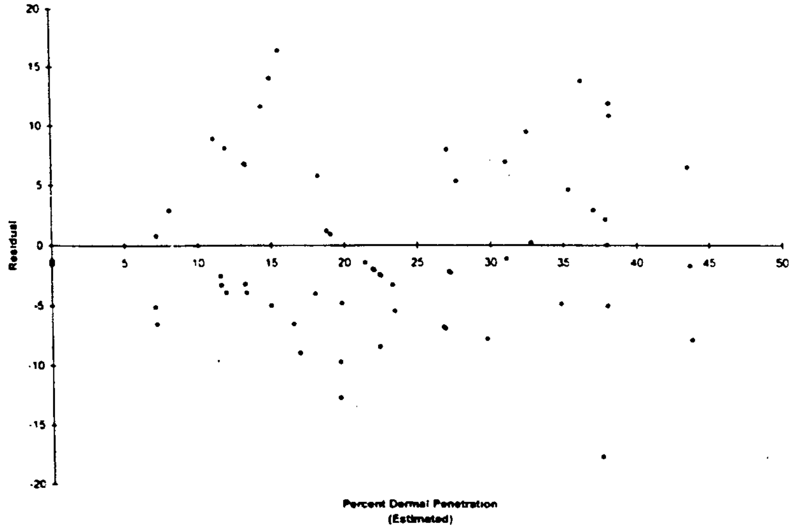


FIGURE 2 Pattern of residual errors for the estimation of the percent dermal penetration (%DP) of 60 polycyclic aromatic hydrocarbons using Eq. (2).

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, $^{1}X^{b}$, ^{3D}W, 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 characterizatiȯn 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 C_1-C_4 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., Trinajstic, 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.

APPENDIX 1.13 The relative effectiveness of topological, geometrical, and quantum chemical...

Chapter 17

Relative Effectiveness of Topological, Geometrical, and Quantum Chemical Parameters in Estimating Mutagenicity of Chemicals

Subhash C. Basak, Brian D. Gute, Gregory D. Grunwald
Natural Resources Research Institute, University of Minnesota,
5013 Miller Trunk Highway, Duluth MN 55811, USA

Abstract – Adequate experimental data necessary for hazard assessment is not available for the majority of environmental pollutants and chemicals in commerce. This has led to the increasing use of theoretical structural parameters in the hazard estimation of such chemicals. In this paper we have used a hierarchical quantitative structure-activity relationship (QSAR) approach involving topological indices, geometrical 3-dimensional (3D) indices, and quantum chemical indices to estimate the mutagenicity of a set of 95 aromatic and heteroaromatic amines. The results show that topological indices explain the major part of the variance in mutagenicity. The addition of quantum chemical indices to the set of descriptors makes some improvement in the predictive models.

The assessment of the environmental and human health hazard posed by chemicals is frequently carried out using insufficient experimental data. This is true for industrial chemicals as well as for substances identified in industrial effluent, hazardous waste sites, and environmental monitoring surveys (Auer et al. 1990). In 1984, the National Research Council (NRC) studied the availability of toxicity data on industrial chemicals and found that many of these chemicals have very little or no test data (1984). About 15 million distinct chemical entities have been registered with the Chemical Abstract Service (CAS), and the list is growing by nearly 750,000 per year. Out of these chemicals, about 1,000 enter into societal use every year (Arcos 1987). Very few of these chemicals have empirical properties needed for hazard assessment. In the United States, the Toxic Substances Control Act (TSCA) inventory has over 72,000 entries, and the list is growing by nearly 3,000 per year (U.S. General Accounting Office [GAO] 1993). Of the some 3,000 chemicals submitted yearly to the United States Environmental Protection Agency (USEPA) for the premanufacture notification (PMN) process, less than 50% have any experimental data at all, less than 15% have empirical mutagenicity data, and only about 6% have ecotoxicological and environmental fate data. The Superfund list of hazardous substances has only limited data for many of the more than 700 chemicals as well (Auer et al. 1990).

This pervasive lack of empirical data shows the real need for the development of methods that can estimate environmental and toxic properties of chemicals using parameters that can be calculated directly from molecular structure. In recent years we have been involved in the development of such models (Basak and Magnuson 1983; Basak 1987, 1990; Basak et al. 1988, 1994; Balaban et al. 1994; Basak and Grunwald 1994a, 1994b, 1995a–1995e, 1996; Basak, Bertelsen, and Grunwald 1995; Basak, Gute, and Grunwald 1995, 1996a, 1996b; Basak, Gute, and Drewes 1996; Basak, Grunwald and Niemi 1997; Basak and Gute 1997). Specifically, we have used graph theoretic indices, geometrical (3-dimensional [3D]) parameters, and semiempirical quantum chemical indices in the development of quantitative structure-activity relationship (QSAR) models pertinent to biomedicinal chemistry and toxicology. In this chapter, we have used a hierarchical approach in the development of QSARs for a group of 95 aromatic and heteroaromatic amines using topological indices, 3D parameters, and a set of quantum chemical descriptors.

The purpose in using a hierarchical approach is to begin to look at the importance of the contribution of different classes of parameters to modeling physicochemical or biologically relevant properties. To this end we ask these questions: What nonempirical molecular information is adequate for the estimation of mutagenic potency? Is specific chemical or quantum chemical information necessary, or do simple structural descriptors do an adequate job? These questions should lead us to a deeper understanding of the principles and molecular basis for determining mutagenic potency.

Theoretical Methods

Database

A set of 95 aromatic and heteroaromatic amines previously collected from the literature by Debnath et al. (1992) were used to study mutagenic potency. The mutagenic activities of these compounds in *S. typhimurium* TA98 + S9 microsomal preparation are expressed as the mutation rate, ln(R), in natural logarithm (revertants/nanomole). Table 17-1 lists the compounds used in this study and their experimentally measured mutation rates.

Computation of topological indices

Topological indices (TIs) used in this study have been calculated by POLLY 2.3 (Basak et al. 1988), which can calculate a total of 102 indices. These indices include Wiener index (Wiener 1947), connectivity indices (Randic 1975; Kier and Hall 1986), information theoretic indices defined on distance matrices of graphs (Bonchev and Trinajstic 1977; Raychaudhury et al. 1984), a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs (Basak et al. 1980; Basak and Magnuson 1983; Roy et al. 1984; Basak 1987), as well as Balaban's J indices (Balaban 1982, 1983, 1986). Table 17-2 provides brief definitions for the topological indices included in this study.

Table 17-1 Observed and estimated mutagenic potency [ln(revertants/nmol)] for 95 aromatic and heteroaromatic amines

			Est. ln(R)
Nr.	Compound	Exp. ln(R)	(Equation 17-10)
1	2-bromo-7-aminofluorene	2.62	1.10
2	2-methoxy-5-methylaniline (p-cresidine)	–2.05	–3.13
3	5-aminoquinoline	–2.00	–2.30
4	4-ethoxyaniline (p-phenetidine)	–2.30	–3.76
5	1-aminonaphthalene	–0.60	–0.32
6	4-aminofluorene	1.13	0.44
7	2-aminoanthracene	2.62	1.61
8	7-aminofluoranthene	2.88	2.54
9	8-aminoquinoline	–1.14	–1.66
10	1,7-diaminophenazine	0.75	1.36
11	2-aminonaphthalene	–0.67	–0.80
12	4-aminopyrene	3.16	3.10
13	3-amino-3'-nitrobiphenyl	–0.55	–0.19
14	2,4,5-trimethylaniline	–1.32	–0.74
15	3-aminofluorene	0.89	0.74
16	3,3'-dichlorobenzidine	0.81	0.24
17	2,4-dimethylaniline (2,4-xylidine)	–2.22	–1.63
18	2,7-diaminofluorene	0.48	0.97
19	3-aminofluoranthene	3.31	2.57
20	2-aminofluorene	1.93	1.08
21	2-amino-4'-nitrobiphenyl	–0.62	0.37
22	4–aminobiphenyl	–0.14	0.06
23	3-methoxy-4-methylaniline (o-cresidine)	–1.96	–3.27
24	2-aminocarbazole	0.60	0.60
25	2-amino-5-nitrophenol	–2.52	–2.01
26	2,2'-diaminobiphenyl	–1.52	–1.24
27	2-hydroxy-7-aminofluorene	0.41	1.61
28	1-aminophenanthrene	2.38	1.80
29	2,5-dimethylaniline (2,5-xylidine)	–2.40	–1.55
30	4-amino-2'-nitrobiphenyl	–0.92	–0.50
31	2-amino-4-methylphenol	–2.10	–2.43
32	2-aminophenazine	0.55	1.32
33	4-aminophenylsulfide	0.31	–0.47
34	2,4-dinitroaniline	–2.00	–0.75
35	2,4-diaminoisopropylbenzene	–3.00	–3.36
36	2,4-difluoroaniline	–2.70	–1.29
37	4,4'-methylenedianiline	–1.60	–0.97
38	3,3'-dimethylbenzidine	0.01	–0.23
39	2-aminofluoranthene	3.23	2.66
40	2-amino-3'-nitrobiphenyl	–0.89	–0.42
41	1-aminofluoranthene	3.35	2.23
42	4,4'-ethylenebis (aniline)	–2.15	–0.92
43	4-chloroaniline	–2.52	–2.94

Table 17-1 *continued*

Nr.	Compound	Exp. ln(R)	Est. ln(R) (Equation 17-10)
44	2-aminophenanthrene	2.46	1.96
45	4-fluoroaniline	−3.32	−2.57
46	9-aminophenanthrene	2.98	1.13
47	3,3'-diaminobiphenyl	−1.30	−0.20
48	2-aminopyrene	3.50	2.58
49	2,6-dichloro-1,4-phenylenediamine	−0.69	−1.46
50	2-amino-7-acetamidofluorene	1.18	0.89
51	2,8-diaminophenazine	1.12	1.55
52	6-aminoquinoline	−2.67	−2.31
53	4-methoxy-2-methylaniline (m-Cresidine)	−3.00	−2.44
54	3-amino-2'-nitrobiphenyl	−1.30	−0.90
55	2,4'-diaminobiphenyl	−0.92	−0.40
56	1,6-diaminophenazine	0.20	0.20
57	4-aminophenyldisulfide	−1.03	−1.00
58	2-bromo-4,6-dinitroaniline	−0.54	−1.25
59	2,4-diamino-n-butylbenzene	−2.70	−3.72
60	4-aminophenylether	−1.14	−0.76
61	2-aminobiphenyl	−1.49	−0.77
62	1,9-diaminophenazine	0.04	0.09
63	1-aminofluorene	0.43	0.28
64	8-aminofluoranthene	3.80	2.69
65	2-chloroaniline	−3.00	−2.37
66	2-amino-α,α,α-trifluorotoluene	−0.80	−1.63
67	2-amino-1-nitronaphthalene	−1.17	−0.90
68	3-amino-4'-nitrobiphenyl	0.69	0.14
69	4-bromoaniline	−2.70	−3.08
70	2-amino-4-chlorophenol	−3.00	−2.39
71	3,3'-dimethoxybenzidine	0.15	0.05
72	4-cyclohexylaniline	−1.24	−0.73
73	4-phenoxyaniline	0.38	−0.50
74	4,4'-methylenebis (o-ethylaniline)	−0.99	−0.51
75	2-amino-7-nitrofluorene	3.00	1.19
76	benzidine	−0.39	−0.52
77	1-amino-4-nitronaphthalene	−1.77	−0.95
78	4-amino-3'-nitrobiphenyl	1.02	0.47
79	4-amino-4'-nitrobiphenyl	1.04	0.73
80	1-aminophenazine	−0.01	1.28
81	4,4'-methylenebis (o-fluoroaniline)	0.23	0.41
82	4-chloro-2-nitroaniline	−2.22	−2.06
83	3-aminoquinoline	−3.14	−2.22
84	3-aminocarbazole	−0.48	0.60
85	4-chloro-1,2-phenylenediamine	−0.49	−2.01
86	3-aminophenanthrene	3.77	1.79
87	3,4'-diaminobiphenyl	0.20	−0.34

Table 17-1 *continued*

Nr.	Compound	Exp. ln(R)	Est. ln(R) (Equation 17-10)
88	1-aminoanthracene	1.18	1.86
89	1-aminocarbazole	−1.04	0.65
90	9-aminoanthracene	0.87	1.15
91	4-aminocarbazole	−1.42	0.38
92	6-aminochrysene	1.83	3.41
93	1-aminopyrene	1.43	3.51
94	4-4'-methylenebis (o-isopropyl-aniline)	−1.77	−1.13
95	2,7-diaminophenazine	3.97	1.93

Computation of geometrical indices

Van der Waal's volume, V_W (Bondi 1964; Moriguchi et al. 1975; Moriguchi and Kanada 1977) was calculated using SYBYL 6.2 (Tripos Associates, Inc. 1994). The 3D Wiener numbers (Bogdanov et al. 1989) were calculated by SYBYL using an SPL (SYBYL Programming Language) program developed in our laboratory. Calculation of 3D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3D coordinates for the atoms were determined using CONCORD 3.2.1 (Tripos 1993). Two variants of the 3D 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.

Computation of quantum chemical parameters

The quantum chemical parameters E_{HOMO}, E_{HOMO1}, E_{LUMO}, E_{LUMO1}, ΔHf, and μ were calculated for all of the following semiempirical Hamiltonians: AM1, PM3, MNDO, MINDO/3. These parameters were calculated by MOPAC 6.00 in the SYBYL interface (Stewart 1990). One difficulty was encountered in using the MINDO/3 Hamiltonian. This particular interface does not include the information necessary for handling bromine, present in 3 of the 95 molecules. To avoid omitting any compounds from one of the models, we accounted for the bromine by substituting dummy atoms which were assigned the Gasteiger-Huckel charges calculated for the original bromine atoms. These molecules containing the dummy atoms with assigned charges were then entered into MOPAC for calculation.

Data reduction

Initially, all TIs were transformed by the natural logarithm of the index plus 1. This was done because the scale of some indices may be several orders of magnitude greater than that of other indices, and other indices may equal 0. The geometric indices were trans-

Table 17-2 Symbols and definitions of topological and geometrical parameters

Symbol	Definition
I_D^W	Information index for magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	Mean information index for magnitude of distance
W	Wiener index = half-sum of off-diagonal elements of distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
$\underline{H}^D$	Graph distance complexity
IC	Information content of distance matrix partitioned by frequency of occurrences of distance h
I_{ORB}	Information content or complexity of hydrogen-suppressed graph at its maximum neighborhood of vertices
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
IC_r	Mean information content or complexity of a graph based on 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 rth (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 = 3-5
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 5, 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, 5
${}^h\chi_{Ch}^b$	Bond chain connectivity index of order h = 5, 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, 5
${}^h\chi_{Ch}^v$	Valence chain connectivity index of order h = 5, 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^Y	Balaban's J index based on relative covalent radii
V_W	van der Waal's volume
${}^{3D}W$	3D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3D Wiener number for the hydrogen-filled geometric distance matrix

formed by the natural logarithm of the index for consistency; the addition of 1 was unnecessary.

The set of 91 TIs was partitioned into 2 distinct sets: topostructural indices and topochemical indices. Topostructural indices are indices that 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 like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters that quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These sets of the indices are shown in Table 17-3.

Table 17-3 Classification of parameters used in developing models for mutagenic potency (ln(R))

Topostructural	Topochemical	Geometric	Quantum chemical: AM1, PM3, MNDO, MINDO/3
I_D^W	I_{ORB}	V_w	E_{HOMO}
$\overline{I_D^W}$	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	${}^{0}\chi^{b}$ - ${}^{6}\chi^{b}$		ΔHf
$\underline{H^D}$	${}^{3}\chi_{C}^{b}$ and ${}^{5}\chi_{C}^{b}$		μ
IC	${}^{5}\chi_{Ch}^{b}$ and ${}^{6}\chi_{Ch}^{b}$		
O	${}^{4}\chi_{PC}^{b}$ - ${}^{6}\chi_{PC}^{b}$		
M_1	${}^{0}\chi^{v}$ - ${}^{6}\chi^{v}$		
M_2	${}^{3}\chi_{C}^{v}$ and ${}^{5}\chi_{C}^{v}$		
${}^{0}\chi$ - ${}^{6}\chi$	${}^{5}\chi_{Ch}^{b}$ and ${}^{6}\chi_{Ch}^{b}$		
${}^{3}\chi_{C}$ and ${}^{5}\chi_{C}$	${}^{4}\chi_{PC}^{b}$ - ${}^{6}\chi_{PC}^{b}$		
${}^{5}\chi_{Ch}$ and ${}^{6}\chi_{Ch}$	J^B		
${}^{4}\chi_{PC}$ - ${}^{6}\chi_{PC}$	J^X		
P_0 - P_{10}	J^Y		
J			

According to Topliss and Edwards (1979), in conducting QSAR studies it is important to bear in mind that the indiscriminate use of too many independent variables can lead to spurious (chance) correlations. Using their findings, we have determined that, for a set of 95 compounds, no more than 60 independent variables can be used in generating regression analyses with explained variance (R^2) of 0.7 or greater. It must be kept in mind that this is the total number of variables initially used in modeling, not the final number of variables used in the model. This number of independent variables should keep the probability of chance correlations below the 0.01 level.

To reduce the number of independent variables that we would use for model construction, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS (SAS 1988). The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional.

From each cluster, we selected the index most correlated with the cluster, as well as any indices that were poorly correlated with the cluster ($r < 0.70$). These indices were then used in the modeling of mutagenic potency of aromatic and heteroaromatic amines. The variable clustering and selection of indices were performed independently for both the topostructural and topochemical subsets.

Statistical analysis and hierarchical QSAR

Regression modeling was accomplished using the SAS procedure REG on 13 sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest indices, the topostructural. After using the topostructural indices to model the activity, we then proceed to add the next level of complexity, the topochemical indices from the clustering procedure, and proceed to model the activity using these parameters. Likewise, the indices included in the model selected from this procedure are combined with the indices from the next level, the geometrical indices, and modeling is conducted once again. Finally, the best model utilizing topostructural, topochemical, and geometrical indices is combined with the quantum chemical parameters and modeling is conducted. This final step was repeated 4 times, each time using quantum chemical parameters from a different semiempirical Hamiltonian, namely, AM1, PM3, MNDO, MINDO/3. Thus quantum chemical models are developed individually, one using the AM1 parameters, one using the MNDO parameters, one using the PM3 parameters, and one using the MINDO/3 parameters. The regression analysis resulted in the final selection of indices for each of the models.

Results and Discussion

The variable clustering of topostructural and topochemical indices resulted in 8 topostructural and 13 topochemical indices being retained for model construction (see Table 17-3). The results for the all possible subsets' regression analyses have been summarized in Table 17-4. Because all sets were well under 25 parameters, all possible subsets' regressions were used for all analyses.

Table 17-4 Summary of regression results for all classes of parameters

Equation	Parameter class	Variables included	*F*	R^2	*s*
17-1	topostructural	O, ${}^4\chi_{PC}$, P_0, J	58.1	0.721	1.04
17-2	topochemical	IC_4, SIC_2, SIC_4, ${}^4\chi^v$, ${}^5\chi^b_C$, ${}^4\chi^b_{PC}$	41.1	0.737	1.02
17-3	geometric	${}^{3D}W$	61.8	0.399	1.50
17-4	Q_C: AM1	E_{HOMO1}, E_{LUMO}, μ	31.8	0.512	1.37
17-5	Q_C: MNDO	E_{HOMO1}, E_{LUMO}	54.7	0.543	1.31
17-6	Q_C: MINDO/3	E_{HOMO}, E_{LUMO}, Δ*Hf*	32.4	0.517	1.36
17-7	Q_C: PM3	E_{HOMO}, E_{HOMO1}, E_{LUMO}	30.0	0.497	1.39
17-8	topostructural + topochemical	${}^4\chi_{PC}$, P_0, J, SIC_2, SIC4, ${}^5\chi^b_C$	44.5	0.752	0.99
17-9	topostructural + topochemical + geometric	${}^4\chi_{PC}$, J, SIC_2, SIC_4, ${}^5\chi^b_C$, ${}^{3D}W$	42.9	0.746	1.00
17-10	topostructural + topochemical + geometric + AM1	${}^4\chi_{PC}$, P_0, J, SIC_2, SIC_4, ${}^5\chi^b_C$, E_{HOMO1}, Δ*Hf*, μ	35.8	0.791	0.92
17-11	topostructural + topochemical + geometric + MNDO	${}^4\chi_{PC}$, P_0, J, SIC_2, SIC_4, ${}^5\chi^b_C$, Δ*Hf*	40.4	0.765	0.97
17-12	topostructural + topochemical + geometric + MINDO/3	${}^4\chi_{PC}$, P_0, J, SIC_2, SIC_4, E_{LUMO}	45.8	0.758	0.98
17-13	topostructural + topochemical + geometric + PM3	${}^4\chi_{PC}$, P_0, J, SIC_2, SIC_4, ${}^5\chi^b_C$, Δ*Hf*	39.7	0.761	0.98

As can be seen from Table 17-4, using only the topostructural class of indices resulted in a 4 parameter model to estimate ln(R) with a variance explained (R^2) of 72.1% and a standard error (s) of 1.04 (Equation 17-1). The P_0 and J indices are related to the size and shape of molecular graphs; the ${}^4\chi_{PC}$ encodes information about the degree of branching of molecular graphs; the O parameter is related to the degree of symmetry of graphs (Basak et al. 1987). Therefore, size, branching, and symmetry (or complexity) of skeletal graphs corresponding to molecular structures seem to be the predominant factors in determining mutagenic potency of the set of 95 aromatic amines.

The second step of the hierarchical method combined the 4 topostructural parameters from Equation 17-1 with the set of 13 topochemical parameters. The resulting model for estimation of ln(R) included 6 parameters (Equation 17-8), which had an R^2 of 75.2% and an s of 0.99. Thus we see that the addition of topochemical information does lead to an increase in the explained variance, improving our model without greatly increasing the number of independent variables. The independent variables of Equation 17-8 quantify 1) shape and size of molecular graphs (J, P_0), 2) branching (${}^4\chi_{PC}$), 3) molecular complexity / redundancy (SIC_2, SIC_4), and 4) degree of cyclicity (${}^5\chi^b_C$). It may be mentioned that we have found very similar sets of topostructural and topochemical parameters useful in estimating normal boiling point, octanol/water partition coefficient (Basak, Gute, and Grunewald 1996b), and vapor pressure (Basak, Gute, and Grunewald 1997) of diverse sets of molecules.

The next step of the hierarchical method takes this topostructural + topochemical model and adds the 3 geometric indices; however, this actually led to a decrease in the explained variance. As part of model construction, it became necessary to eliminate P_0 from the set of indices when adding the hydrogen-suppressed 3D Wiener number because of resulting problems with variance inflation between the 2 parameters. As a result, the model that retained the geometric parameter had slightly lower R^2 and s values than the model using topostructural and topochemical only (Equation 17-9). This being the case, we chose to use the parameters from Equation 17-8 in the following modeling with the quantum chemical parameters. Thus, the last 4 models were all constructed with the 6 parameters from Equation 17-8 and all 6 quantum chemical parameters for the particular Hamiltonian methodology available for modeling.

As can be seen from Table 17-4, the AM1 parameters made the most significant contribution to our hierarchical modeling procedure ($R^2 = 79.1\%$, $s = 0.92$). The other 3 methods showed only minimal improvement over the topostructural + topochemical model.

Finally, individual models using only topochemical, only geometrical, and only quantum chemical parameters were constructed to further our understanding of the individual contribution of these different types of parameters. The topochemical model was the strongest of the 3, with the geometrical and quantum chemical models showing little effectiveness. The topochemical model included 6 parameters and did show a slight increase in explained variance and standard error over the topostructural model.

The goal of this chapter is to investigate the relative effectiveness of theoretical structural parameters — namely topostructural, topochemical, geometrical, and quantum chemical parameters — in predicting the mutagenicity of a set of aromatic and heteroaromatic amines. To this end, we used a hierarchical approach in the development of QSARs using 4 classes of molecular descriptors.

The results show that the topostructural parameters explain a large fraction of the variance (R^2) in the mutagenic potency of the amines. The best model in this area explained about 72% of variance in mutagenicity using O, ${}^4\chi_{PC}$, P_0, J. These indices do not contain any explicit chemical information about the molecules. The large explained variance probably indicates that general structural features like size, shape, symmetry, and branching play a major role in determining mutagenic potency. The addition of topochemical variables made some improvement in the explained variance. The best model using topostructural and topochemical indices explained about 75% of variance in mutagenicity. The addition of geometrical parameters, however, did not make any improvement in estimation. Finally, the addition of quantum chemical parameters was attempted. Indices from AM1, PM3, MNDO, and MINDO3 were used separately in developing the QSAR models. While addition of the heat of formation, dipole moment, and E_{HOMO1} parameters calculated by the AM1 method provided some improvement in the estimation of ln(R), parameters calculated by PM3, MINDO3, and MNDO did not make any significant improvement in the estimation of mutagenic potency. The calculated values for the parameters used in the hierarchical model that included the AM1 parameters (Equation 17-10) are presented in Table 17-5. These values represent the original, nontransformed values for all indices used in Equation 17-10. Additionally, Figure 17-1 presents a scatterplot of observed versus estimated mutagenic potency based on Equation 17-10.

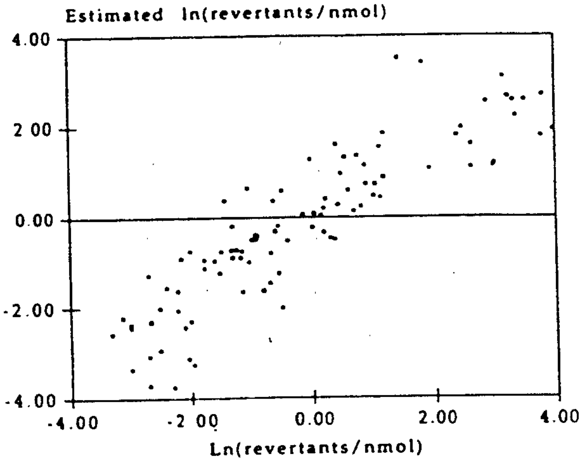


Figure 17–1 Scatterplot for observed ln(R) versus estimated ln(R) using Equation 17-10 for set of 95 aromatic and heteroaromatic amines

Table 17-5 Calculated values for topostructural, topochemical, and AM1 quantum chemical parameters used in Equation 17-1

Nr.	${}^4\chi_{PC}$	P_0	J	SIC_2	SIC_4	${}^5\chi^b_C$	E_{HOMO1}	ΔHf	μ
1	2.482	15	1.722	0.780	0.966	0.080	−9.510998	57.462489	3.246
2	1.409	10	2.356	0.824	0.875	0.059	−9.198889	−24.061979	1.613
3	1.440	11	1.993	0.831	0.975	0.058	−9.528133	51.959364	2.993
4	0.841	10	2.132	0.775	0.818	0.000	−9.761040	−22.045505	1.782
5	1.440	11	1.993	0.639	0.931	0.058	−9.342732	40.325881	1.549
6	2.209	14	1.800	0.697	0.931	0.109	−9.019172	53.561923	1.377
7	2.148	15	1.673	0.613	0.885	0.049	−8.752501	61.467301	1.686
8	3.051	17	1.694	0.616	0.890	0.119	−8.883560	90.631004	1.061
9	1.440	11	1.993	0.807	0.975	0.058	−9.497513	49.496038	1.140
10	2.650	16	1.701	0.703	0.967	0.083	−8.759018	93.256750	2.202
11	1.292	11	1.932	0.648	0.907	0.025	−8.981140	39.152911	1.625
12	3.058	17	1.692	0.593	0.890	0.112	−9.017251	86.180524	1.025
13	2.289	16	1.879	0.722	0.951	0.065	−9.635184	49.692122	5.732
14	2.154	10	2.462	0.622	0.786	0.167	−9.195396	−1.116909	1.386
15	2.136	14	1.751	0.704	0.948	0.080	−8.880375	53.383623	1.407
16	3.115	16	1.884	0.677	0.755	0.194	−9.010987	29.747467	1.402
17	1.478	9	2.346	0.719	0.867	0.083	−9.402700	5.680026	1.423
18	2.482	15	1.722	0.692	0.766	0.080	−9.008264	51.483002	0.749
19	3.131	17	1.679	0.592	0.890	0.128	−8.745169	113.597721	1.348
20	2.132	14	1.739	0.704	0.948	0.080	−9.316509	53.266008	1:795
21	2.481	16	1.832	0.699	0.902	0.103	−10.009252	50.464895	5.573
22	1.351	13	1.789	0.570	0.836	0.028	−9.611345	45.922022	1.682
23	1.418	10	2.376	0.824	0.875	0.059	−9.233259	−23.899670	2.229
24	2.132	14	1.739	0.715	0.981	0.057	−8.382162	66.295627	1.688
25	2.126	11	2.396	0.874	0.942	0.121	−10.236383	−21.118276	6.030
26	1.945	14	1.963	0.591	0.755	0.104	−8.411351	45.503434	0.270
27	2.482	15	1.722	0.791	0.967	0.080	−9.366850	8.492721	1.867
28	2.332	15	1.763	0.600	0.951	0.091	−8.782735	57.726120	1.543
29	1.478	9	2.346	0.696	0.867	0.083	−9.229828	5.699677	1.431
30	2.293	16	1.944	0.699	0.902	0.075	−9.850974	54.711440	5.793
31	1.478	9	2.346	0.847	0.910	0.083	−9.261839	−30.703134	1.260
32	2.148	15	1.673	0.651	0.891	0.049	−9.205497	91.251439	1.882
33	1.221	14	1.685	0.593	0.845	0.000	−9.510446	52.769884	1.912
34	2.499	13	2.526	0.777	0.920	0.107	−11.360524	25.435777	7.257
35	1.838	11	2.437	0.722	0.815	0.131	−8.792416	3.913795	2.561
36	1.478	9	2.346	0.836	0.962	0.083	−10.029053	−69.256743	2.575
37	1.630	15	1.681	0.603	0.659	0.000	−8.406652	39.288132	1.394
38	3.115	16	1.884	0.656	0.716	0.194	−8.782407	29.805987	2.494
39	2.913	17	1.674	0.604	0.905	0.093	−8.844299	113.962366	0.866
40	2.437	16	1.921	0.716	0.967	0.103	−9.940798	79.401262	6.265
41	3.058	17	1.700	0.616	0.920	0.119	−8.657007	101.911673	1.867
42	1.683	16	1.601	0.606	0.660	0.000	−8.707849	57.273517	2.562
43	0.816	8	2.192	0.737	0.812	0.000	−9.948850	13.095294	2.631
44	2.176	15	1.722	0.606	0.951	0.057	−8.807318	59.927756	1.359

Table 17-5 *continued*

Nr.	$^4\chi_{PC}$	P_0	J	SIC_2	SIC_4	$^5\chi^b_C$	E_{HOMO1}	ΔH*f*	μ
45	0.816	8	2.192	0.737	0.812	0.000	−10.025071	−24.569648	2.776
46	2.280	15	1.787	0.603	0.885	0.091	−8.826091	57.985510	1.608
47	1.641	14	1.861	0.624	0.755	0.028	−9.637290	52.825739	0.355
48	2.888	17	1.654	0.569	0.807	0.077	−8.537199	81.775262	1.644
49	2.006	10	2.487	0.719	0.812	0.144	−9.653936	6.122184	0.948
50	2.727	18	1.612	0.786	0.920	0.080	−9.409869	19.708295	4.954
51	2.497	16	1.667	0.644	0.771	0.049	−9.614724	124.753819	2.050
52	1.292	11	1.932	0.831	0.975	0.025	−9.345759	50.639120	2.728
53	1.574	10	2.330	0.824	0.875	0.083	−9.524426	−23.745777	1.831
54	2.234	16	1.984	0.716	0.967	0.075	−9.701876	55.625683	6.167
55	1.848	14	1.867	0.628	0.902	0.066	−8.529041	45.389658	1.889
56	2.802	16	1.739	0.677	0.755	0.117	−8.724272	87.859343	1.995
57	1.683	16	1.601	0.584	0.643	0.000	−8.694071	52.783142	3.652
58	3.074	14	2.661	0.813	0.920	0.174	−11.175279	33.261219	6.162
59	1.360	12	2.246	0.740	0.890	0.059	−8.803533	−7.047410	2.543
60	1.630	15	1.681	0.579	0.642	0.000	−8.589188	21.521611	2.589
61	1.292	13	1.833	0.588	0.884	0.028	−9.075139	46.291223	1.526
62	2.802	16	1.744	0.677	0.771	0.117	−8.760423	87.878976	2.958
63	2.293	14	1.786	0.697	0.931	0.127	−8.809819	52.914796	1.658
64	2.972	17	1.656	0.613	0.896	0.093	−8.672342	86.560420	1.569
65	1.138	8	2.279	0.775	0.962	0.083	−9.647217	13.148070	1.773
66	2.214	11	2.461	0.788	0.903	0.250	−10.328717	−135.798912	4.070
67	2.274	14	2.092	0.732	0.939	0.093	−9.498965	42.132738	5.212
68	2.332	16	1.793	0.699	0.902	0.065	−9.707684	49.439690	6.645
69	0.816	8	2.192	0.737	0.812	0.000	−9.958995	24.673699	2.834
70	1.478	9	2.346	0.885	0.966	0.083	−9.512320	−30.257131	1.873
71	2.994	18	1.913	0.670	0.725	0.146	−8.597273	−29.701343	0.593
72	1.351	13	1.789	0.633	0.783	0.048	−9.618662	−11.036978	1.453
73	1.221	14	1.685	0.593	0.845	0.000	−9.519593	24.038959	3.243
74	2.855	19	1.809	0.670	0.738	0.118	−8.322206	14.345758	1.347
75	3.130	17	1.674	0.786	0.953	0.117	−9.907587	57.088597	7.715
76	1.759	14	1.780	0.558	0.624	0.028	−8.898246	44.312986	2.417
77	2.390	14	2.079	0.760	0.939	0.103	−9.995923	44.945430	7.318
78	2.348	16	1.843	0.699	0.902	0.065	−10.065351	48.997787	5.907
79	2.391	16	1.760	0.656	0.836	0.065	−10.153390	48.597189	7.636
80	2.300	15	1.714	0.655	0.884	0.083	−9.466774	90.375028	1.894
81	2.975	17	1.775	0.705	0.773	0.167	−8.668864	−51.583170	2.233
82	1.851	11	2.471	0.863	0.938	0.070	−10.795945	14.958329	5.163
83	1.292	11	1.932	0.807	0.975	0.025	−9.250508	61.289442	2.564
84	2.136	14	1.751	0.715	0.981	0.057	−8.650669	70.561209	2.432
85	1.478	9	2.346	0.738	0.875	0.083	−9.338439	12.337686	1.935
86	2.180	15	1.741	0.606	0.935	0.057	−8.832492	56.103853	1.663
87	1.700	14	1.820	0.611	0.869	0.028	−8.581538	44.585899	2.808
88	2.300	15	1.714	0.617	0.896	0.083	−9.168383	66.520403	1.216

Table 17-5 *continued*

Nr.	${}^4\chi_{PC}$	P_0	J	SIC_2	SIC_4	${}^5\chi^b_C$	E_{HOMO1}	ΔHf	μ
89	2.293	14	1.786	0.708	0.962	0.091	–8.617125	69.956608	1.276
90	2.357	15	1.760	0.587	0.787	0.103	–9.179235	64.230081	1.689
91	2.209	14	1.800	0.708	0.962	0.082	–8.497152	66.236222	1.211
92	3.175	19	1.575	0.553	0.913	0.124	–8.830777	100.875189	1.130
93	3.110	17	1.677	0.577	0.890	0.112	–8.958369	70.826740	1.287
94	3.721	21	1.867	0.638	0.674	0.263	–8.315255	10.633206	1.225
95	2.497	16	1.664	0.644	0.755	0.049	–9.634497	124.742897	0.004

Using the same set of aromatic amines Debnath et al. (1992) developed various QSAR models using hydrophobicity (log P, octanol/water), E_{HOMO}, and E_{LUMO} calculated by the AM1 Hamiltonian and some indicator variables. For the largest subset (n = 88), they derived the following model:

$$\ln(R) = 7.20 + 1.08(\log P) + 1.28(E_{HOMO}) - 0.73(E_{LUMO}) + 1.46(I_L) \qquad (17\text{-}14)$$

$$s = 0.860,\ F = 12.6,\ R^2 = 0.806$$

The model in Equation 17-10 is comparable to the model developed by Debnath et al. (1992) and uses all the 95 aromatic amines as compared to a smaller subset (n=88) used in their study.

Acknowledgments

This is contribution number 197 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, a grant from Exxon Corporation, and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

References

Unless otherwise noted, sources are U.S.

Arcos JC. 1987. Structure-activity relationships: criteria for predicting the carcinogenic activity of chemical compounds. *Environ Sci Tech* 21:743–745.

Auer CM, Nabholz JV, Baetcke KP. 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.

Balaban AT, Basak SC, Colburn T, Grunwald G. 1994. Correlation between structure and normal boiling points of haloalkanes C1-C4 using neural networks. *J Chem Inf Comput Sci* 34:1118–1121.

Balaban AT. 1982. Highly discriminating distance-based topological index. *Chem Phys Lett* 89:399–404.

Balaban AT. 1983. Topological indices based on topological distances in molecular graphs. *Pure and Appl Chem* 55:199–206.

Balaban AT. 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.

Basak SC. 1987. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med Sci Res* 15:605–609.

Basak SC. 1990. A nonempirical approach to predicting molecular properties using graph- theoretic invariants. In: Karcher W, Devillers J, editors. Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology. Dordrecht/Boston/London: Kluwer Academic. p 83–103.

Basak SC, Bertelsen S, Grunwald G. 1994. Application of graph theoretical parameters in quantifying molecular similarity and structure-activity studies. *J Chem Inf Comput Sci* 34:270–276.

Basak SC, Bertelsen S, Grunwald GD. 1995. Use of graph theoretic parameters in risk assessment of chemicals. *Toxicology Letters* 79:239–250.

Basak SC, Grunwald GD. 1994a. In press. Use of topological space and property space in selecting structural analogs. *Mathematical Modeling and Scientific Computing.*

Basak SC, Grunwald GD. 1994b. Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. *SAR and QSAR in Environmental Research* 2:289–307.

Basak SC, Grunwald GD. 1995a. Estimation of lipophilicity from molecular structural similarity. *New Journal of Chemistry* 19:231–237.

Basak SC, Grunwald GD. 1995b. Predicting genotoxicity of chemicals using nonempirical parameters. In: Rao RS, Deo MG, Sanghui LD, editors. Proceeding of the XVI International Cancer Congress. Bologna, Italy: Monduzzi. p 413–416.

Basak SC, Grunwald GD. 1995c. Molecular similarity and estimation of molecular properties. *J Chem Inf Comput Sci* 35:366–372.

Basak SC, Grunwald GD. 1995d. Tolerance space and molecular similarity. *SAR and QSAR in Environmental Research* 3:265–277.

Basak SC, Grunwald GD. 1995e. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. *Chemosphere* 31:2529–2546.

Basak SC, Grunwald GD. In preparation 1996. Characterization of relative proximity of molecules in structure space: development of a molecular ruler using octane isomers. *Mathl Modeling Sci Computing.*

Basak SC, Grunwald GD, Niemi GJ. 1997. Use of graph theoretic and geometrical molecular descriptors in structure-activity relationships. In: Balaban AT, editor. From chemical topology to three dimensional molecular geometry. New York: Plenum Pr. p 73-116.

Basak SC, Gute BD. 1997. Use of graph theoretic parameters in predicting inhibition of microsomal r-hydroxylation of anilines by alcohols: a molecular similarity approach. In: Johnson BL, Xintaras C, Andrews Jr JS, editors. Impacts on human and ecological health. New Jersey NJ: Princeton Scientific. p 492-504.

Basak SC, Gute BD, Drewes LR. 1996. Predicting blood-brain transport of drugs: a computational approach. *Pharm Res* 13:775–778.

Basak SC, Gute BD, Grunwald GD. 1995. Development and applications of molecular similarity methods using nonempirical parameters. *Mathl Modeling Sci Computing.* In press.

Basak SC, Gute BD, Grunwald GD. 1996a. Estimation of normal boiling points of haloalkanes using molecular similarity. *Croat Chim Acta* 69:1159-1173.

Basak SC, Gute BD, Grunwald GD. 1996b. 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.

Basak SC, Gute BD, Grunwald GD. 1997. Use of topostructural, topochemical and geometrical parameters in the prediction of vapor pressure: a hierarchical QSAR approach. *J Chem Inf Comput Sci* 37:651–655.

Basak SC, Harriss DK, Magnuson VR. 1988. POLLY 2.3: [computer software]. Copyright of the University of Minnesota.

Basak SC, Magnuson VR. 1983. Molecular topology and narcosis: a quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim Forsch* 33:501–503.

Basak SC, Magnuson VR, Niemi GJ, Regal RR, Veith GD. 1987. Topological indices: their nature, mutual relatedness, and applications. *Mathematical Modeling* 8:300–305.

Basak SC, Roy AB, Ghosh JJ. 1980. Study of the structure-function relationship of pharmacological and toxicological agents using information theory. In: Avula XJR, Bellman R, Luke YL, Rigler AK, editors. Proceedings of the Second International Conference on Mathematical Modeling. Rolla MO: Univ Missouri - Rolla Pr. p 851–856.

Bogdanov B, Nikolic S, Trinajstic N. 1989. On the three-dimensional Wiener number. *J Math Chem* 3:299–309.

Bonchev D, Trinajstic N. 1977. Information theory, distance matrix, and molecular branching. *J Chem Phys* 67:4517–4533.

Bondi A. 1964. van der Waal's volumes and radii. *J Phys Chem* 68:441–451.

Debnath AK, Debnath G, Shusterman AJ, Hansch C. 1992. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ Mol Mutagen* 19:37–52.

[GAO] General Accounting Office. 1993. EPA toxic substances program: long-standing information planning problems must be addressed. Washington DC: U.S. General Accounting Office (USGAO), Accounting and Information Management Division. GAO/AIMD-94-25.

Kier LB, Hall LH. 1986. Molecular connectivity in structure-activity analysis. Letchworth, Hertfordshire UK: Research Studies Pr. 262 p.

Moriguchi I, Kanada Y. 1977. Use of van der Waal's volume in structure-activity studies. *Chem Pharm Bull* 25:926–935.

Moriguchi I, Kanada Y, Komatsu K. 1976. van der Waal's volume and the related parameters for hydrophobicity in structure-activity studies. *Chem Pharm Bull* 24:1799–1806.

[NRC] National Research Council. 1984. Toxicity testing: strategies to determine needs and priorities. Washington DC: National Academy Pr. 84 p.

Randic M. 1975. On characterization of molecular branching. *J Am Chem Soc* 97:6609–6615.

Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC. 1984. Discrimination of isomeric structures using information theoretic topological indices. *J Comput Chem* 5:581–588.

Roy AB, Basak SC, Harris DK, Magnuson VR. 1984. Neighborhood complexities and symmetry of chemical graphs and their biological applications. In: Avula XJR, Kalman RE, Liapis AI, Rodin EY, editors. Mathematical modeling in science and technology. New York: Pergamon Pr. p 745-750.

SAS Institute Inc. 1988. In: SAS/STAT User's Guide, Release 6.03 Edition. Cary NC: SAS Institute Inc. Chapters 28 and 34; p 773–875, 949–965.

Stewart JJP. 1990. MOPAC Version 6.00. QCPE #455. U.S. Air Force Academy CO: Frank J Seiler Research Laboratory.

Topliss JG, Edwards RP. 1979. Chance factor in studies of quantitative structure-activity relationships. *J Med Chem* 22:1238–1244.

Tripos Associates, Inc. 1993. CONCORD [computer software]. Version 3.2.1. St. Louis MO: Tripos Associates Inc.

Tripos Associates, Inc. 1994. SYBYL [computer software]. Version 6.2. St. Louis MO: Tripos Associates Inc.

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

APPENDIX 1.14 Characterization of molecular structures using topological indices

SAR and QSAR in Environmental Research, 1997, Vol. 7, pp. 1-21
Reprints available directly from the publisher
Photocopying permitted by license only

CHARACTERIZATION OF MOLECULAR STRUCTURES USING TOPOLOGICAL INDICES

S. C. BASAK* and B. D. GUTE

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

(Received 3 April 1997; In final form 15 June 1997)

The characterization of molecular structure using structural invariants has increased greatly over the last ten years. Specifically, topological indices have become more widely used in the quantification of molecular structure for use in quantitative structure-activity relationship studies, chemical documentation, and molecular similarity studies. The basis, calculation, and utility of topological indices has been examined, with an eye to the specific advantages and problems in their use. In addition, variable clustering and principal component analysis are examined as two potential solutions to the problem of index intercorrelation.

Keywords: Topological indices; molecular structure; graph theory; graph invariants; variable clustering; principal component analysis

INTRODUCTION

An important area of research in computational and mathematical chemistry is the characterization of molecular structure using structural invariants [1 - 14]. The impetus for this research trend comes from various directions. Researchers in chemical documentation have searched for a set of invariants which will be more convenient than the adjacency matrix (or connection table) for the storage and comparison of chemical structures [15]. Invariants have been used to order sets of molecules [3 - 5, 8, 16]. With the substantial increase in available databases of chemical structures and properties, attempts have been made to develop structure-activity relation-

*Author to whom all correspondence should be addressed.

ships (SARs) whereby existing molecules can be compared with other molecules (real or hypothetical) on the basis of these structural invariants. The properties of the molecules of interest can then be predicted based on molecular structure without the need for experimental data.

In this age of combinatorial chemistry thousands of molecules of known structure can be produced rapidly. However, at the same time resources for determining even the simplest properties of all these molecules in the laboratory are unavailable. In the USA, the Toxic Substances Control Act (TSCA) Inventory includes nearly 74,000 chemicals and the list is growing at a rate of more than 2,000 new submissions to the United States Environmental Protection Agency (USEPA) for the Premanufacture Notification (PMN) process per year [17–20]. At present, risk assessment of the PMN chemicals is carried out using limited test data. For example, approximately 15% of PMN submissions have empirical mutagenicity data. Under such circumstances, structural descriptors will play a pivotal role in comparing molecules with one another and in predicting their properties.

MOLECULAR STRUCTURE – BEAUTY IN THE EYE OF THE BEHOLDER OR CONUNDRUM?

The main hurdle to the characterization of molecular structure is the lack of uniformity in its definition and quantification. The term *molecular structure* represents a set of nonequivalent and probably disjoint concepts [21]. For example, the term "molecule" means different things when it represents an assembly of identifiable atoms held together by fairly rigid bonds as compared to a collection of delocalized nuclei and electrons in which all identical particles are indistinguishable [21]. There is no reason to believe that when we discuss diverse topics (e.g., chemical synthesis, reaction rates, spectroscopic transitions, reaction mechanisms, and *ab initio* calculations) using the notion of *molecular structure*, that the different meanings we attach to this term originate from the same fundamental concept [21, 22]. This fundamental problem has been described succinctly by Woolley [22]:

> "... there is no reason to suppose that the same basic idea can provide a basis for the discussion of all molecular experiments. This is understandable if one recognizes that every physical and chemical concept is only defined with respect to a certain class of experiments, so that it is perfectly reasonable for different sets of concepts, although mutually incompatible, to be applicable to different experiments."

In the context of molecular science, the various concepts of molecular structure (e.g., classical valence bond representation, various chemical graph-theoretic representations, the ball-and-stick model, representation by minimum energy conformation, semi-symbolic contour maps, or symbolic representation by Hamiltonian operators) are distinct molecular models derived through different means of abstraction from the same chemical reality or molecule [23]. In each instance, the equivalence class (concept or model of molecular structure) is generated by selecting certain aspects while ignoring other unique properties of those actual events. This explains the plurality of the concepts of molecular structure and their autonomous nature, the word autonomous being used in the sense that one concept is not logically derived from the other.

GRAPHS AND MOLECULAR STRUCTURE

At the most fundamental level, the structural model of an assembled entity (e.g., a molecule consisting of atoms) may be defined as the pattern of relationship among its parts as distinct from the values associated with them [24]. Constitutional formulae of molecules are graphs where vertices represent the set of atoms and edges represent chemical bonds [25]. The pattern of connectedness of atoms in a molecule is preserved by constitutional graphs. A graph (more correctly a non-directed graph) $G = [V, E]$ consists of a finite nonempty set V of points together with a prescribed set E of unordered pairs of distinct points of V [26]. A *structural model* assigns to the points of G a realization in some applied field and each element of E indicates a pair of entities (elements of the structural model) which are in the finite nonempty irreflexive symmetric binary relation described by G. For example, when elements of the set V symbolize atomic cores without valence electrons and the elements of E represent covalent two-electron bonds, G is the molecular graph or constitutional graph of a covalent chemical species. Such a graph can represent structural formulae of a large number of organic compounds. Since more than 90% of chemical compounds described so far are either organic or contain organic ligands, such a graph has been found to be useful in chemistry [13]. The edge set need not always represent a covalent bond. In fact, elements of E may symbolize almost any type of bond (e.g., ionic, coordinate, hydrogen, or weak bonds representing transition states of an SN_2 reaction, etc.) [27–29]. If the interaction between a pair of atoms is asymmetric (e.g., in case of sufficiently polar covalent bonds, hydrogen bond donor acidity, hydrogen bond acceptor basicity, or charge transfer complex formation) the bonding pattern can be represented by a binary relation which is anti-reflexive and asymmetric [6]. Further refinement could be achieved through the assignment of weights to the vertices or edges [3], and use of multiple edges between a pair of atoms held together both by *sigma* and *pi* bonds. The weighted pseudograph appears to be the most general model capable of symbolizing the bonding pattern of a large number of organic and inorganic chemicals.

For a long time, chemists have relied on visual perception to relate various aspects of constitutional graphs to observable phenomena. The power of graph-theoretic formalism in chemistry is evident from its successful applications in chemical documentation, isomer discrimination and characterization of molecular branching, enumeration of constitutional isomers associated with a particular empirical formula, calculation of quantum chemical parameters, structure-physicochemical property correlations, and chemical structure-biological activity relationships [30–37].

GRAPHS AS MOLECULAR MODELS

Any concept of molecular structure is a hypothetical sketch of the organization of atoms within the molecule. Such a *model object* is a general theory and remains empirically untestable. A model object has to be grafted to a specific theory to generate a *theoretical model* which can be empirically tested [38]. For example, when it was suggested by Sylvester in 1878 [39] that the structural formula of a molecule is a special kind of graph, it was an innovative general theory without any predictive potential. When the idea of combinatorics was applied on chemical graphs (model object), it could be predicted that "there should be exactly two isomers of butane (C_4H_{10})" because "there are exactly two tree graphs with four vertices" when one considers only the non-hydrogen atoms present in C_4H_{10} [13]. This is a theoretical model of limited predictive potential. Although it predicts the existence of chemical species, given a set of molecules (e.g., isomers of hexane [C_6H_{14}]) the model is incapable of predicting any properties for these molecules. This is due to the fact that any empirical property *P* maps a set of chemical structures into the set *R* of real numbers and thereby orders the set empirically. Therefore, to predict the property from structure, we need a nonempirical (structural) ordering scheme which closely resembles the empirical ordering of structures as determined by *P*. This is a more specific theoretical model based on the same model object (chemical graph) and can be accomplished by using specific graph invariant(s).

CHARACTERIZATION OF MOLECULAR GRAPHS

Molecular graphs can be characterized by graph invariants. A graph invariant is a graph-theoretic property which is preserved by isomorphism [26]. A graph invariant could be a polynomial, a sequence of numbers, or a single number. The characteristic polynomial of a graph and the spectra of graphs are graph invariants. Numerical graph invariants derived from molecular graphs are called graph-theoretic indices or topological indices [25]. Topological indices quantitatively describe molecular topology and are sensitive to such structural attributes as size, shape, patterns of branching, bonding types, and cyclicity of molecules.

Topological indices (TIs) can sometimes be derived conveniently from different matrices such as the adjacency matrix and the distance matrix. The origins of such TIs illuminate the fundamental structural features that they quantify. On the other hand, some indices are derived to quantify a key structural feature which is qualitative and only understood intuitively. In deriving his original connectivity index (${}^1X$), Randić asked the question: which of the two heptane isomers, *viz.*, 3-methylhexane and 3-ethylpentane, is more branched [9]. Until that time, branching was understood only intuitively; Randić derived a quantitative description of branching based on the graph-theoretic treatment of the structures. In addition, information theoretic indices of chemical structures have been derived to answer the question: which of a collection of structures is more complex or heterogeneous? Different measures of molecular complexity attempt to answer this question from different points of view [40]. In the following section we discuss the structural basis and method of calculation for some of the major topological indices.

CALCULATION OF TOPOLOGICAL INDICES

The Wiener index (W) [41], the first topological index reported in the chemical literature, 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 $n \times n$ 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 2,3-dimethylhexane (Fig. 1):

$$D(G_1) = \begin{array}{c|cccccccc|} & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) \\ 1 & 0 & 1 & 2 & 2 & 3 & 3 & 4 & 5 \\ 2 & 1 & 0 & 1 & 1 & 2 & 2 & 3 & 4 \\ 3 & 2 & 1 & 0 & 2 & 3 & 3 & 4 & 5 \\ 4 & 2 & 1 & 2 & 0 & 1 & 1 & 2 & 3 \\ 5 & 3 & 2 & 3 & 1 & 0 & 2 & 3 & 4 \\ 6 & 3 & 2 & 3 & 1 & 2 & 0 & 1 & 2 \\ 7 & 4 & 3 & 4 & 2 & 3 & 1 & 0 & 1 \\ 8 & 5 & 4 & 5 & 3 & 4 & 2 & 1 & 0 \end{array}$$

W is calculated as:

$$W = 1/2 \sum_{i,j} d_{ij} = \sum_{h} h \cdot g_h \qquad (1)$$

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 seventy.

Randić's connectivity index [9], and higher-order connectivity path, cluster, path-cluster and chain types of simple, bond and valence connectivity parameters were calculated using the method of Kier and Hall [10]. P_h parameters, number of paths of length $h (h = 0, 1, \ldots, 10)$ in the hydrogen-suppressed graph, are calculated using standard algorithms.

Balaban defined a series of indices based upon distance sums within the distance matrix for a chemical graph which he designated as J indices [42–44]. These indices are highly discriminating with low degeneracy. Unlike W, the J indices range of values are independent of molecular size.

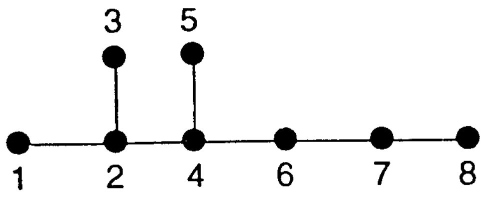


FIGURE 1 Hydrogen-suppressed graph of 2,3-dimethylhexane.

Information-theoretic topological indices are calculated by the application of information theory on chemical graphs. An appropriate set of A of n elements is derived from a molecule graph G depending upon 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 $n_i(i = 1, 2, \ldots, h; \sum_i n_i = n)$. A probability distribution is then assigned to the set of equivalence classes:

$$\begin{matrix} A_1, A_2, \ldots, A_h \\ p_1, p_2, \ldots, p_h \end{matrix}$$

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

The mean information content of an element of A is defined by Shannon's relation [45]:

$$\mathrm{IC} = -\sum_{i=1}^{h} p_i \log_2 n_i \tag{2}$$

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

It is to be noted that the information content of a graph G is not uniquely defined. It depends on how the set A is derived from G as well as on the equivalence relation which partitions A into disjoint subsets A_i. For example, when A constitutes the vertex set of a chemical graph G, two methods of partitioning have been widely used:

a) Chromatic-number coloring of G where two vertices of the same color are considered equivalent, and
b) Determination of the orbits of the automorphism group of G thereafter vertices belonging to the same orbit are considered equivalent.

Rashevsky was the first to calculate the information content of graphs where "topologically equivalent" vertices were placed in the same equivalence class [46]. In Rashevsky's approach, two vertices u and v of a graph are said to be topologically equivalent if and only if for each neighboring vertex $u_i (i = 1, 2, \ldots, k)$ of the vertex u, there is a distinct neighboring vertex v_i of the same degree for the vertex v. While Rashevsky used simple linear graphs with indistinguishable vertices to symbolize molecular structure, weighted linear graphs or multigraphs are better models for conjugated or aromatic molecules because they more properly reflect the actual bonding patterns, i.e., electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.* [47] calculated 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 stereo-electronic characteristics of distant neighbors, i.e., neighbors of neighbors, it was deemed essential to extent 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 non-negative 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 value 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 which are topologically equivalent for rth order neighborhood. Such an approach has been developed and the information-theoretic indices calculated based on this idea are called indices of neighborhood symmetry [40].

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 rth order neighborhood if any only if corresponding to each path u_0, $u_1, \ldots, u_r$ of length r, there is a distinct path v_0, $v_1, \ldots, v_r$ 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 rth order bonded neighbors. The detailed equivalence relation has been described in earlier studies [40, 48].

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by Eq. 2. Basak *et al.* [49] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$\mathrm{SIC}_r = \mathrm{IC}_r / \log_2 n \tag{3}$$

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

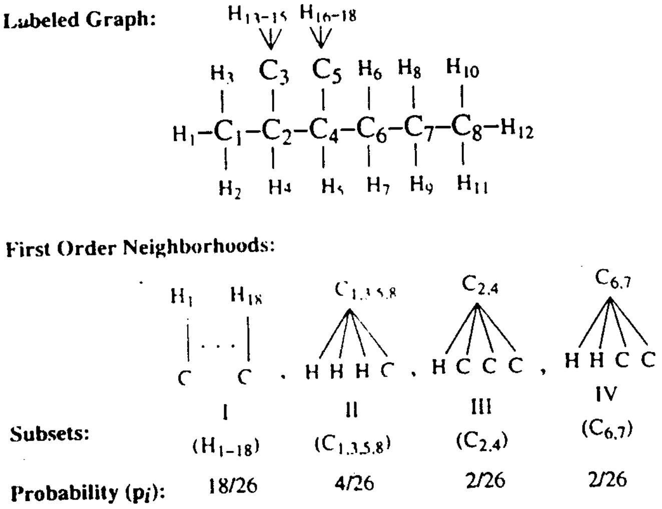
Another information-theoretic invariant, complementary information content (CIC_r) [50], is defined as:

$$CIC_r = \log_2 n - IC_r \qquad (4)$$

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.

In Figure 2, the calculation of IC_1, SIC_1 and CIC_1 is demonstrated for the hydrogen-filled graph of 2,3-dimethylhexane.

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 [11]:



$IC_1 = - \Sigma p_i \cdot \log_2 p_i$

$= 2 \cdot 2/26 \cdot \log_2 26/2 + 4/26 \cdot \log_2 26 + 18/26 \cdot \log_2 26/18$

$= 1.150$ bits

$SIC_1 = IC_1 / \log_2 26$

$= 0.353$ bits

$CIC_1 = \log_2 26 - IC_1$

$= 2.108$ bits

FIGURE 2 The calculation of IC_1, SIC_1 and CIC_1 based on the first order neighborhoods for the labeled graph of 2,3-dimethylhexane

$$I_D^W = W \log_2 W - \sum g_{h_h} \cdot h \log_2 h \qquad (5)$$

The mean information index, $\bar{I}_D^W$, 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.* [12].

THEORETICAL METHODS

Databases and Calculations

Two data sets were used for this study: the first consists of the seventy-four alkanes (C_2-C_9) and the second, more heterogeneous set was taken from the STARLIST group of chemicals [51]. The STARLIST subset includes 219 chemicals for which HB_1 was equal to zero and calculated log *P* values fell in the range of -2 to 5.5. HB_1 is a measure of the hydrogen bonding potential of a chemical. Chemical structures for these compounds were encoded using the SMILES line notation for chemical structures and entered into the computer program POLLY version 2.3 for the calculation of indices [52]. Table 1 provides a comprehensive list and brief descriptions for these indices.

STATISTICAL METHODS

Initially all TIs were tranformed by the natural logarithm of the index plus one. This is routinely done to scale the indices since there may be a difference of several orders of magnitude between indices and some may equal zero.

From the original sets of 102 indices calculated for both data sets, it was necessary to remove some indices. Some of the indices for the set of alkanes (e.g., the simple, valence and bond connectivity indices) were completely redundant. Other indices were removed because they had values of zero for all compounds. This "cleaning" of the sets of TIs left fifty-three indices for the alkanes and ninety-eight indices for the STARLIST set.

Variable clustering and principal component analysis were used on the remaining indices to minimize problems of intercorrelation amongst the indices. The variable clustering was conducted using the SAS procedure VARCLUS which divides the indices into disjoint clusters which are

TABLE I Symbols and definitions of topological indices

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
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
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
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
${}^hX$	Path connectivity index of order $h=0-6$
${}^hX_C$	Cluster connectivity index of order $h=3-6$
${}^hX_{Ch}$	Chain connectivity index of order $h=3-6$
${}^hX_{PC}$	Path-cluster connectivity index of order $h=4-6$
${}^hX^b$	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 $h=4-6$
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^Y	Balaban's J index based on relative covalent radii

essentially unidimensional based on the correlation matrix [53]. From each cluster, the index which was most correlated with the cluster was selected as the best representative of that cluster. In this way, individual indices are retained while minimizing intercorrelations. This procedure resulted in the retention of eight TIs for the alkanes; H^V, SIC_0, SIC_1, SIC_4, ${}^3X_C$, ${}^5X_C$, P_4, P_8; and twelve TIs for the STARLIST data; I_D^W, IC_4, SIC_3, CIC_1, ${}^4X$, ${}^4X_{Ch}$, ${}^6X_{Ch}^v$, ${}^3X_C^b$, ${}^5X_C^b$, ${}^3X_{PC}^b$, P_8, J^B. TI values for a subset of the alkanes, the eighteen octane isomers, are presented in Table II.

The principal component analysis (PCA) was accomplished using the SAS procedure PRINCOMP [54]. 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 previous PCs, eliminating the redundancy which can occur with TIs. The maximum number of PCs generated is equal to the number of individual 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.* [3]. The seven PCs with eigenvalues greater than one and the ten PCs with eigenvalues greater than one were retained for the alkanes and STAR-LIST set respectively. Table III presents the PCs for the octane isomers, a subset of the seventy-four alkanes.

DISCRIMINATION OF ISOMERS USING TOPOLOGICAL INDICES AND PRINCIPAL COMPONENTS DERIVED FROM THEM

Topological aspects of chemicals have been used in chemical documentation. One line of research in this area has been the development of

TABLE II TIs selected by variable clustering of the alkanes (octane isomers listed)

Isomer Name	H'	SIC_0	SIC_1	SIC_4	$^3\chi_C$	$^5\chi_C$	P_4	P_8
Octane	1.288	0.173	0.218	0.477	0.000	0.000	2	0
2-methylheptane	1.233	0.173	0.248	0.561	0.342	0.000	2	0
3-methylheptane	1.228	0.173	0.248	0.598	0.254	0.000	2	0
4-methylheptane	1.215	0.173	0.248	0.503	0.254	0.000	2	0
3-ethylhexane	1.177	0.173	0.248	0.532	0.186	0.000	2	0
2,2-dimethylhexane	1.157	0.173	0.248	0.495	0.940	0.000	2	0
2,3-dimethylhexane	1.170	0.173	0.253	0.557	0.450	0.212	2	0
2,4-dimethylhexane	1.171	0.173	0.253	0.557	0.529	0.000	2	0
2,5-dimethylhexane	1.183	0.173	0.253	0.384	0.597	0.000	2	0
3,3-dimethylhexane	1.137	0.173	0.248	0.548	0.792	0.000	2	0
3,4-dimethylhexane	1.157	0.173	0.253	0.469	0.386	0.154	2	0
3-ethyl-2-methylpentane	1.096	0.173	0.253	0.490	0.405	0.154	2	0
3-ethyl-3-methylpentane	1.073	0.173	0.248	0.421	0.656	0.000	1	0
2,2,3-trimethylpentane	1.075	0.173	0.255	0.490	0.944	0.477	1	0
2,2,4-trimethylpentane	1.083	0.173	0.255	0.450	1.088	0.000	2	0
2,3,3-trimethylpentane	1.065	0.173	0.255	0.506	0.850	0.529	1	0
2,3,4-trimethylpentane	1.097	0.173	0.225	0.413	0.620	0.326	2	0
2,2,3,3-tetramethylbutane	0.997	0.173	0.218	0.218	1.253	1.179	0	0

TABLE III Values of the first seven PCs for the eighteen octane isomers

Isomer Name	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
Octane	0.328	−1.744	5.807	0.602	−0.320	−0.473	−0.433
2-methylheptane	2.181	−4.236	1.097	0.386	1.100	0.300	−0.935
3-methylheptane	2.817	−4.857	−0.307	0.921	0.368	0.366	−0.513
4-methylheptane	1.338	−2.211	0.848	−0.821	0.005	−0.541	−0.904
3-ethylhexane	1.553	−2.077	−0.348	−0.494	−0.817	−0.651	−0.290
2,2-dimethylhexane	1.163	0.007	−0.436	−0.878	1.367	1.383	0.638
2,3,-dimethylhexane	2.122	−2.060	−1.546	0.502	−0.308	−0.253	−0.105
2,4-dimethylhexane	2.089	−2.306	−1.372	−0.289	−0.205	0.004	0.291
2,5-dimethylhexane	−0.769	1.340	1.473	−2.659	0.612	−0.387	−1.443
3,3-dimethylhexane	2.044	−0.573	−1.726	0.303	0.173	0.582	1.163
3,4-dimethylhexane	0.807	0.228	−0.825	−0.696	−0.730	−1.223	−0.545
3-ethyl-2-methylpentane	0.991	−0.035	−1.596	−0.672	−1.076	−1.438	0.110
3-ethyl-3-methylpentane	−0.035	2.870	−0.614	−0.909	−0.497	−1.178	0.271
2,2,3-trimethylpentane	1.136	2.191	−2.383	1.277	0.465	−0.075	0.548
2,2,4-trimethylpentane	0.377	2.377	−1.284	−1.846	0.726	0.461	1.676
2,3,3-trimethylpentane	1.318	1.825	−2.717	1.990	0.318	−0.400	0.251
2,3,4-trimethylpentane	−0.548	4.168	1.329	0.020	−1.745	−1.140	−0.039
2,2,3,3-tetramethylbutane	−4.473	12.522	2.681	4.256	1.345	−0.129	−2.627

topological indices which are more discriminatory. For example, the *J* index developed by Balaban is one of the most discriminatory indices. Randić developed the concept of molecular identification number (I. D. number) by combining a few topological aspects of structures. Other authors have used more than one index for this purpose. One example is the topological superindex proposed by Bonchev *et al.* [55] where they use a collection of indices as the superindex. Two structures are said to be distinct if the magnitudes of any one of the component indices differ for them.

In view of the intercorrelation of indices and the fact that a large number of TIs have been defined in the literature, we have been interested in deriving orthogonal parameters from TIs. We have employed two statistical methods: variable clustering and principal components analysis (PCA). In the former method, we begin with the TIs calculated by POLLY and derive a small set of original variables which are minimally intercorrelated. In the case of the seventy-four alkanes the method retained eight indices. In the PCA, seven principal components (PCs) are derived from original variables and these PCs are linear combinations of all the TIs. For the STARLIST set, twelve TIs were retained by variable clustering, while ten PCs were derived.

We are interested to see the discriminatory power of the TIs selected by variable clustering *vis-a-vis* the PCs. Values of the TIs selected by the variable clustering technique and the first seven PCs with eigenvalue greater than 1.0 for the set of eighteen octane isomers are presented in Tables II and III respectively. It is clear from the data that some individual TIs are not sufficiently discriminatory for the eighteen octane isomers. On the other hand, each PC is unique for any given structure, making them more discriminatory than any individual TI. In the interest of space, the values of the TIs and PCs for all of the alkanes and for the STARLIST set were not included in the tables, however, this information is available upon request from the authors.

TOPOLOGICAL INDEX SPACE VIS-A-VIS PC SPACE: WHAT DO THEY MEAN?

Each TI quantifies certain aspects of molecular structure. Distinct indices selected by the variable clustering procedure encode different information regarding molecular structure (model object). For example, indices like the connectivity index or Wiener index quantify adjacency information of the simple planar graph model of molecules. On the other hand, information theoretic graph invariants quantify the degree of complexity of the molecular graph. Intuitively, these are distinct aspects of molecular structure and this notion is borne out by the result of variable clustering analysis on the set of TIs calculated by POLLY. It is tempting to speculate that each index retained by variable clustering represents one distinct aspect of molecular architecture and that, collectively, the TIs form the structure space of the set of chemicals. Such a space can be used for the discrimination of structures and structure-property correlation. The magnitudes of eight TIs for the eighteen octane isomers show that the TIs selected by variable clustering have reasonable power for discriminating isomeric structures.

At the level of PCs, we have derived a certain number of orthogonal variables using PCA of the indices. For the alkanes we had seven PCs with eigenvalues greater than 1.0 (Tab. III) whereas for the structurally diverse set of 219 compounds we had ten PCs with eigenvalues greater than 1.0. This result indicates that the structure space for the set of 219 molecules is more complex than that for the set of seventy-four alkanes. This is in agreement with our intuitive notion that molecules with heteroatoms and many functional groups are more complex than molecules devoid of any heteroatom. Finally, the pattern of correlation of the individual PCs with the TIs can help us in understanding the nature of the axes derived by PCA (Tabs. IV and V).

TABLE IV PC loading for the seven principal components with eigenvalues greater than 1.0 for the 74 alkanes

PC	*Ten Most Correlated Indices*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	I^D(0.98)	W(−0.97)	^1X(0.97)	I_D^W(0.97)	P_1(0.97)	CIC$_0$(0.97)	P_0(0.97)	SIC$_0$(−0.97)	^0X(0.94)	IC$_0$(0.94)
2	CIC$_2$(0.89)	CIC$_3$(0.79)	CIC$_4$(0.77)	CIC$_5$(0.77)	^3X_C(0.76)	SIC$_2$(−0.74)	^4X_C(0.69)	$^4X_{PC}$(0.69)	^5X_C(0.65)	SIC$_3$(−0.64)
3	SIC$_1$(−0.76)	^6X(0.68)	P_6(0.67)	P_7(0.65)	P_8(0.55)	$^4X_{PC}$(−0.41)	IC$_1$(−0.40)	^5X(0.39)	$^5X_{PC}$(−0.39)	CIC$_2$(0.38)
4	^5X_C(0.64)	^6X_C(0.63)	^4X(−0.40)	P_4(−0.34)	$^4X_{PC}$(0.31)	I_{ORB}(0.29)	P_7(0.28)	CIC$_5$(−0.27)	CIC$_4$(−0.27)	SIC$_1$(−0.24)
5	^4X(−0.39)	^4X_C(0.38)	$^6X_{PC}$(−0.36)	^3X_C(0.35)	P_4(0.34)	$^5X_{PC}$(−0.31)	^3X(−0.29)	^2X(0.26)	P_3(−0.26)	SIC$_1$(0.25)
6	^4X_C(0.40)	P_5(0.39)	^5X(0.37)	^3X_C(0.35)	$^6X_{PC}$(0.34)	$\bar{I}_D^W$(−0.23)	H^D(−0.22)^{ns}	P_4(0.19)^{ns}	IC$_0$(−0.19)^{ns}	$\overline{IC}$(−0.18)^{ns}
7	P_8(0.59)	P_7(0.38)	^5X_C(0.30)	^6X_C(−0.23)	P_5(−0.20)^{ns}	^5X_C(−0.19)^{ns}	^5X(−0.19)^{ns}	^3X_C(0.17)^{ns}	^6X(−0.17)^{ns}	O(0.16)^{ns}

^{ns}Not significant at the $p \leq 0.05$ level.

TABLE V PC loading for the 10 principal components with eigenvalues greater than 1.0 for the 219 STARLIST chemicals

PC	*Ten Most Correlated Indices*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	P_0(0.97)	^{1}X (0.96)	^{0}X (0.96)	P_1(0.96)	^{1}X (0.95)	W (0.95)	$^{1}X^{b}$(0.95)	$^{4}X^{b}$(0.95)	M_2(0.95)	M_1(0.94)
2	SIC_4(−0.86)	SIC_3(−0.86)	SIC_5(−0.86)	SIC_6(−0.86)	CIC_5(0.80)	CIC_6(0.80)	CIC_4(0.80)	SIC_2(−0.78)	CIC_3(0.76)	IC_2(−0.74)
3	CIC_2(−0.67)	SIC_1(0.65)	$^{5}X^{v}_{Ch}$(0.63)	CIC_1(−0.63)	$^{5}X_{C}$(0.61)	$^{5}X^{b}_{Ch}$(0.61)	SIC_0(0.61)	$^{6}X_{C}$(0.60)	$^{6}X^{v}_{Ch}$(0.59)	CIC_3(−0.58)
4	J(0.83)	J^{Y}(0.73)	J^{B}(0.73)	J^{X}(0.62)	$^{3}X^{b}_{C}$(0.56)	$^{3}X_{C}$(0.55)	$^{6}X_{Ch}$(−0.44)	P_{10}(−0.42)	$^{5}X^{b}_{Ch}$(−0.41)	$^{6}X^{b}_{Ch}$(−0.41)
5	IC_0(−0.45)	SIC_0(−0.43)	J^{X}(0.36)	J (0.35)	$^{4}X^{b}_{Ch}$(0.35)	$^{4}X_{Ch}$(0.35)	CIC_0(0.35)	SIC_1(−0.34)	$^{6}X^{v}_{PC}$(−0.33)	$^{5}X_{Ch}$(0.33)
6	$^{4}X^{b}_{C}$(0.57)	$^{4}X_{C}$(0.57)	P_8(0.48)	P_9(0.46)	$^{4}X^{v}_{C}$(0.44)	P_{10}(0.42)	$^{3}X^{b}_{C}$(0.35)	$^{3}X_{C}$(0.33)	$^{5}X^{v}_{C}$(−0.32)	P_7(0.31)
7	$^{6}X_{C}$(−0.43)	$^{6}X^{b}_{C}$(−0.42)	$^{5}X^{b}_{C}$(−0.42)	$^{3}X_{Ch}$(0.40)	$^{5}X_{C}$(−0.39)	$^{4}X_{Ch}$(0.31)	$^{4}X^{b}_{Ch}$(0.29)	$^{4}X^{b}_{PC}$(−0.26)	$^{5}X_{Ch}$(0.26)	$^{5}X^{v}_{Ch}$(0.21)
8	$^{4}X^{v}_{C}$(0.49)	$^{3}X^{v}_{C}$(0.40)	$^{2}X^{v}$(0.29)	$^{6}X^{b}_{C}$(−0.27)	$^{6}X_{C}$(−0.26)	J^{Y}(−0.24)	$^{1}X^{v}$(0.23)	$^{0}X^{v}$(0.22)	J^{b}(−0.21)	P_9(−0.20)
9	$^{3}X_{Ch}$(0.73)	$^{4}X_{Ch}$(0.47)	$^{4}X^{b}_{Ch}$(0.43)	$^{6}X^{v}_{Ch}$(−0.21)	$^{5}X^{v}_{Ch}$(−0.21)	$^{5}X_{Ch}$(−0.19)	$^{6}X_{C}$(−0.16)	$^{6}X_{Ch}$(−0.16)	J^{X}(−0.16)	$^{6}X^{b}_{Ch}$(−0.15)
10	IC_0(0.35)	H^{v}(0.25)	J^{X}(−0.24)	$^{1}X^{v}$(0.24)	SIC_0(0.21)	I_{ORB}(−0.21)	$^{6}X_{PC}$(−0.21)	J^{Y}(−0.20)	J^{B}(−0.19)	$^{1}X^{b}$(0.19)

DISCUSSION

The major objectives of this paper were:

a) To illuminate the fundamental nature of mathematical invariants of molecular structure,

b) To study the utility of graph invariants in the characterization of molecular structure, and

c) To study the intercorrelation of indices and extraction of orthogonal variables from TIs.

It is clear from the results presented in this paper that the various classes of mathematical invariants quantify different aspects of molecular architecture. They depend principally on the structural model (model object) used for the calculation of the invariant as well as the intuitive aspect of molecular structure they are used to quantify. For example, connectivity indices and neighbor complexity indices were designed to quantify distinct aspects of molecular structure. The results of variable clustering of the congeneric set of alkanes and the diverse set of 219 chemicals show that these indices encode largely independent structural information about these molecules.

Many structural schemes have been developed for the derivation of numbers or sets of numbers which can discriminate closely related structures so that they can be useful in chemical documentation. The results presented in this paper show that both the collection of indices selected by variable clustering as well as the PCs can discriminate among the eighteen octane isomers (Tabs. II - V). It is also clear from the data that the PCs are more discriminatory than the individual indices. For example, each PC has distinct values for all eighteen octane isomers. PCs derived from TIs have also been used in the discrimination of isospectral molecular graphs where individual indices show a high degree of degeneracy [56].

Variable clustering of TIs for the set of seventy-four alkanes retained eight parameters which can be classified into three subsets:

a) H^V, P_4 and P_8 which represent generalized size and shape;

b) SIC_{10}, SIC_1, and SIC_4 which quantify molecular complexity; and

c) ${}^3\chi_C$ and ${}^5\chi_C$ which encode information about molecular branching.

In the case of the more diverse set of 219 chemicals, the indices retained after variable clustering fall into four subclasses:

a) I''_D, P_8 and ${}^4\chi$ (general shape and size);

b) IC_4, SIC_3 and CIC_1 (complexity);

c) ${}^{4}X^{v}_{\mathrm{Ch}}$ and ${}^{6}X^{v}_{\mathrm{Ch}}$ (cyclicity); and
d) ${}^{3}X^{h}_{\mathrm{C}}$, ${}^{5}X^{h}_{\mathrm{C}}$, ${}^{3}X^{h}_{\mathrm{PC}}$ and J^{B} (branching).

A perusal of results from both the sets indicate that distinct indices quantify different intuitive aspects of molecular structure.

A similar picture emerges from the principal component analysis of both sets of molecules. The first PC is strongly correlated with variables which quantify shape and size. The next important factor is molecular complexity which is encoded by the second PC (Tabs. IV and V). The higher order PCs (3 - 5) are strongly correlated with invariants which quantify such subtle structural factors as branching, cyclicity, etc. It may be mentioned that such a result emerged from our earlier studies on a large, diverse set of 3,692 chemicals [3, 57].

In conclusion, mathematical invariants derived from chemical topology quantify different aspects of molecular architecture which are intuitively understood by the chemist. One can create a structure space from these invariants taking uncorrelated structural information (indices or PCs). Such orthogonal factors can be useful in the discrimination of closely related structures like isomers and in the creation of structure spaces. Metrics defined on such spaces have been useful in the quantification of molecular similarity [3 - 5, 58 - 63]. Orthogonal variables derived by PCA or variable clustering can also be used in QSAR studies pertaining to pharmacology and toxicology [1, 2, 6, 33 - 36, 40, 48 - 50, 64 - 68].

Acknowledgement

This is contribution number 214 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, a grant from Exxon Corporation and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to extend their thanks to Greg Grunwald for technical support.

References

[1] 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

[2] 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.

[3] Basak, S. C., Magnuson, V. R., Niemi, G. J. and Regal, R. R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.*, **19**, 17–44.

[4] Basak, S. C., Bertelsen, S. and Grunwald, G. D. (1994). Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. *J. Chem. Inf. Comput. Sci.*, **34**, 270–276.

[5] Basak, S. C. and Grunwald, G. D. (1994). Use of topological space and property space in selecting structural analogs. *Mathl. Modelling and Sci. Comput.*, in press.

[6] 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.

[7] Fisanick, W., Cross, K. P. and Rusinko, III, A. (1992). Similarity searching on CAS registry substances. 1. Molecular property and generic atom triangle geometric searching. *J. Chem. Inf. Comput. Sci.*, **32**, 664–674.

[8] Carhart, R. E., Smith, D. H. and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, **25**, 64–73.

[9] Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609–6615.

[10] Kier, L. B. and Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis.* Research Studies Press, Letchworth, Hertfordshire, U.K.

[11] Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **67**, 4517–4533.

[12] Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B. and Basak, S. C. (1994). Discrimination of isomeric structures using information-theoretic topological indices. *J. Comput. Chem.*, **5**, 581–588.

[13] Balaban, A. T. (1985). Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.*, **25**, 334–343. ·

[14] Basak, S. C. and Grunwald, G. D. (1993). Use of graph invariants, volume and total surface area in predicting boiling point of alkanes. *Mathl. Modelling and Sci. Comput. Modelling*, **2**, 735–740.

[15] Randić, M. (1984). On molecular identification numbers. *J. Chem. Inf. Comput. Sci.*, **24**, 164–175.

[16] Wilkins, C. L. and Randić, M. (1980). A graph theoretical approach to structure-property and structure-activity correlations. *Theoretica Chimica Acta*, **58**, 45–68.

[17] 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.

[18] National Research Council (NRC). (1984). *Toxicity Testing Strategies to Determine Needs and Priorities.* National Academy Press, Washington, D. C.

[19] Arcos, J. C. (1987). Structure-activity relationships: criteria for predicting carcinogenic activity of chemical compounds. *Environ. Sci. Technol.*, **21**, 743–745.

[20] Toxic Substances Control Act (TSCA). Public Law 94-469, 90 Stat. 2003, October 11, 1976

[21] Weininger, S. J. (1984). The molecular structure conundrum: Can classical chemistry be reduced to quantum chemistry? *J. Chem. Educ.*, **61**, 939–944.

[22] Woolley, R. G. (1978). Must a molecule have a shape. *J. Am. Chem. Soc.*, **100**, 1073–1078.

[23] Primas, H. (1981). *Chemistry, Quantum Mechanics and Reductionism.* Springer-Verlag, Berlin

[24] Whyte, L. L. (1965). Atomism, structure and form: a report on the natural philosophy of form, in *Structure in Art and Science* (G. Keeps, Ed.). George Braziler, Inc., New York, pp. 20–28.

[25] Trinajstić, N. (1983). *Chemical Graph Theory*, **I and II**. CRC Press, Boca Raton, Florida.

[26] Harary, F. (1969). *Graph Theory*. Addison Wesley Publishing Co., Reading, Massachusetts

[27] Spialter, L. (1964). The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP): a new computer-oriented chemical nomenclature. *J. Am. Chem. Soc.*, **85**, 2012–2013.

[28] Spialter, L. (1964). The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP). *J. Chem. Doc.*, **4**, 261–269.

[29] Spialter, L. (1964). The atom connectivity matrix characteristic polynomial (ACMCP) and its physico-geometric (topological) significance. *J. Chem. Doc.*, **4**, 269–274.

[30] Kennedy, J. W. and Quintas, L. V. (1988). *Applications of Graphs in Chemistry and Physics*, North-Holland, Amsterdam.

[31] Randić, M. (1984). Nonempirical approach to structure-activity studies. *Int. J. Quantum Chem. Quantum Biol. Symp.*, **11**, 137–153.

[32] Sabljić, A. and Trinajstić, N. (1981). Quantitative structure-activity relationships: the role of topological indices. *Acta Pharm. Yugosl.*, **31**, 189–214.

[33] Basak, S. C., Gieschen, D. P., Harriss, D. K. and Magnuson, V. R. (1983). Physicochemical and topological correlates of the enzymatic acetyltransfer reaction. *J. Pharm. Sci.*, **72**, 934–937.

[34] Basak, S. C., Monsrud, L. J., Rosen, M. E., Frane, C. M. and Magnuson, V. R. (1986). A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharm. Yugosl.*, **36**, 81–95.

[35] Basak, S. C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR appraoch. *Med. Sci. Res.*, **15**, 605–609.

[36] Basak, S. C. (1988) Binding of barbiturates to cytochrome P_{450}: a QSAR study using log P and topological indices. *Med. Sci. Res.*, **16**, 281–282.

[37] Trinajstić, N., Randić, M. and Klein, D. J. (1986). On the quantitative structure-activity relationship in drug research. *Acta Pharm. Yugosl.*, **36**, 267–279.

[38] Bunge, M. (1973). *Method, Model and Matter.* Reidel, D. Publishing Co., Dordrecht-Holland/Boston.

[39] Sylvester, J. J. (1878) On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics. *Amer. J. Math.*, **1**, 64–83.

[40] 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. Liapis and E. Y Rodin, Eds.). Pergamon Press, Elmsford, New York, pp. 745–750.

[41] Wiener, H. (1947) Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.

[42] Balaban, A. T. (1982) Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399–404

[43] Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.*, **55**, 199–206.

[44] Balaban, A. T. (1985). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)*, **21**, 115–122.

[45] Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

[46] Rashevsky, N. (1955) Life, information theory and topology. *Bull. Math. Biophys.*, **17**, 229–235.

[47] Sarkar, R., Roy, A B and Sarkar, R. K. (1978). Topological information content of genetic molecules – I *Math. Biosci.*, **39**, 299–312.

[48] Magnuson, V. R, Harriss, D. K. and Basak, S. C. (1983). Topological indices based on neighborhood symmetry chemical and biological applications, in *Studies in Physical and Theoretical Chemistry* (R B King, Ed.). Elsevier, Amsterdam, pp. 178–191.

[49] 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

[50] Basak, S. C. and Magnuson, V. R. (1983). Molecular topology and narcosis. *Arzneim-Forsch. Drug Research*, 33, 501 - 503.

[51] Leo, A. and Weininger, D. (1984). *CLOGP Version 3.2 User Reference Manual*, Medicinal Chemistry Project, Pomona College, Claremont, CA.

[52] Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1988). POLLY 2.3: Copyright of the University of Minnesota.

[53] SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition*, SAS Institute Inc., Cary, NC, Chapter 34, pp. 949 - 965.

[54] SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition*, SAS Institute Inc., Cary, NC, Chapter 34, pp. 751 - 771.

[55] Bonchev, D., Mekenyan, O. and Trinajstić, N. (1981). Isomer discrimination by topological information approach. *J. Comput. Chem.*, 2, 127 - 148.

[56] Balasubramanian, K. and Basak, S. C. (1997). Characterization of isospectral graphs using graph invariants and derived orthogonal parameters. *J. Chem. Inf. Comput. Sci.*, in preparation.

[57] Basak, S. C., Magnuson, V. R., Niemi, G. J., Regal, R. R. and Veith, G. D. (1987). Topological indices: their nature, mutual relatedness and applications, in *Mathematical Modelling in Science and Technology*., (X. J. R. Avula, G. Leitmann, C. D. Mote, Jr. and E. Y. Rodin, Eds.). Pergamon Press: Oxford, pp. 300 - 305.

[58] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1996). Estimation of normal boiling points of haloalkanes using molecular similarity. *Croat. Chem. Acta*, 69, 1159 - 1173.

[59] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Development and application of molecular similarity methods: using nonempirical parameters. *Mathl. Model and Sci. Comput.*, in press.

[60] Basak, S. C. and Gute, B. D. (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal p-hydroxylation of aniline by alcohols: a molecular similarity approach, in *Proceedings of the 2nd International Congress on Hazardous Waste: Impact on Human and Ecological Health*. (B. L. Johnson, C. Xintaras and J. S. Andrews, Jr., Eds.). Princeton Scientific Publishing Co., Inc., New Jersey, pp. 492 - 504.

[61] Basak, S. C. and Grunwald, G. D. (1994). Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. *SAR QSAR Environ Res.*, 2, 289 - 307.

[62] Basak, S. C. and Grunwald, G. D. (1995). Molecular similarity and estimation of molecular properties. *J. Chem Inf Comput Sci.*, 35, 366 - 372.

[63] Basak, S. C. and Grunwald, G. D. (1995). Tolerance space and molecular similarity. *SAR QSAR Environ. Res.*, 3, 265 - 277.

[64] Basak, S. C., Gute, B. D. and Ghatak, S. (1997). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, submitted.

[65] 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.

[66] 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 QSAR approach. *J. Chem. Inf. Comput Sci.*, 37, 651–655.

[67] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, in *Quantitative Structure-Activity Relationships in Environmental Sciences-VII* (F. Chen, *et al.*, Eds.). SETAC Press, in press.

[68] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Development of a quantitative structure-activity relationship (QSAR) for estimating bioconcentration factor in *Pimephales promelas*: a hierarchical approach. In progress.

APPENDIX 1.15 Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular...

SAR and QSAR in Environmental Research, 1997, Vol. 7, pp. 117 - 131
Reprints available directly from the publisher
Photocopying permitted by license only

PREDICTING ACUTE TOXICITY (LC_{50}) OF BENZENE DERIVATIVES USING THEORETICAL MOLECULAR DESCRIPTORS: A HIERARCHICAL QSAR APPROACH

B. D. GUTE and S. C. BASAK*

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

(Received 3 March 1997; In final form 10 June 1997)

Four classes of theoretical structural parameters, viz., topostructural, topochemical, geometrical and quantum chemical descriptors, have been used in the development of quantitative structure-activity relationship (QSAR) models for a set of sixty-nine benzene derivatives. None of the individual classes of parameters was very effective in predicting toxicity. A hierarchical approach was followed in using a combination of the four classes of indices in QSAR model development. The reslults show that the hierarchical QSAR approach using the algorithmically derived molecular descriptors can estimate the LC_{50} values of the benzene derivatives reasonably well.

Keywords Hierarchical QSAR; topological indices; geometrical indices; quantum chemical parameters; aquatic toxicity; benzene derivatives

INTRODUCTION

Today's toxicologist is faced with a myriad of unknowns. In 1996 approximately 1.26 million new chemicals were registered with the Chemical Abstract Service (CAS), bringing the total number of registered chemicals to around 15.8 million [1]. With such a large number of chemicals being registered yearly, it is impossible to test all of them exhaustively for their

*Author to whom all correspondence should be addressed effects on the environment and human health. Chemicals can only be evaluated as they are called into question, and for many of these compounds there will be little or no test data available. Therefore, when the issue of hazard assessment comes up, it becomes difficult at best to provide any useful suggestions or analyses for many of the registered chemicals, including some which are in commerce today. To complete the battery of tests necessary for the proper hazard assessment of a single compound is an extremely costly procedure and there is simply not enough time or money to complete these test batteries for all compounds which are registered today [2]. As a result, when we need to evaluate the human health or ecological hazards posed by a chemical it becomes ever more important that we have accurate methods for estimating the physicochemical and biological properties of molecules.

Quantitative structure-activity relationships (QSARs) have come into widespread use for the prediction of various molecular properties and biological responses. Traditional QSARs use empirical properties; e.g., boiling point, melting point, octanol-water partition coefficient; or empirically derived parameters; e.g., linear free energy related (LFER) and linear solvation energy related (LSER) parameters; for the prediction of other endpoints [3 - 8]. However, due to the scarcity of available data for the majority of chemicals that need to be evaluated for ecotoxicological risk assessment, these physicochemical properties necessary for traditional QSAR model development may not be known. When this is the case, it is imperative that we have methods that make use of nonempirical parameters. One of the fundamental principles of biochemistry is that activity is dictated by structure [9]. Following this principle, one can use theoretical molecular descriptors which quantify structural aspects of the molecular structure [10 - 27]. These theoretical descriptors can be generated directly from the molecular structure alone, without any input of experimental data.

Topological indices (TIs) are numerical graph invariants that quantify certain aspects of molecular structure. TIs are sensitive to such structural features as size, shape, bond order, branching, and neighborhood patterns of atoms in molecules. They can be derived from simple linear graphs, multigraphs, weighted graphs, and weighted pseudographs. TIs derived from these different classes of graphs will encode different types of information about molecular architecture. The different classes of TIs provide us with nonempirical, quantitative descriptors that can be used in place of experimentally derived descriptors in QSARs for the prediction of properties.

Our recent studies have focused on the role of different classes of theoretical descriptors of increasing levels of complexity and their utility in QSAR [28 31] This takes the form of a hierarchical approach which examines the relative contributions of parameters of gradually increasing complexity; e.g., structural, chemical, shape and quantum chemical descriptors; in estimating physicochemical and biological properties.

In this paper we have reported the utility of this hierarchical approach in modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of sixty-nine benzene derivatives.

THEORETICAL METHODS

Database

Acute aquatic toxicity [$-\log(LC_{50})$] in fathead minnow (*Pimephales promelas*) data was taken from the work of Hall, Kier and Phipps [32]. 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. The complete set of fathead minnow data included 69 benzene derivatives. According to the authors, the set of benzene derivatives were tested using methodologies which were comparable to their 96-hour fathead minnow toxicity test system. The derivatives chosen for this study have seven different substituent groups that are all present in at least six of the molecules. These groups consist of chloro, bromo, nitro, methyl, methoxyl, hydroxyl, and amino substituents (Tab. I).

Computation of Indices

Four distinct sets of theoretical descriptors have been used in this study. These sets include topostructural, topochemical, geometric, and quantum chemical indices. The topostructural and topochemical indices fall into the category normally grouped together as topological indices. The geometrical indices are three-dimensional Wiener number for hydrogen-filled molecular structure, hydrogen-suppressed molecular structure, and van der Waals volume.

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

TABLE I Sixty-nine benzene derivatives and their fathead minnow toxicities, expressed as $-\log(LC_{50})$

No.	*Compound*	$-log(LC_{50})$ *(obs.)*	$-log(LC_{50})$ *(est. Eq. 4)*	*Residual*
1	Benzene	3.40	3.42	−0.02
2	Bromobenzene	3.89	3.77	0.12
3	Chlorobenzene	3.77	3.75	0.02
4	Phenol	3.51	3.38	0.13
5	Toluene	3.32	3.66	−0.34
6	1, 2 -dichlorobenzene	4.40	4.29	0.11
7	1, 3-dichlorobenzene	4.30	4.37	−0.07
8	1, 4-dichlorobenzene	4.62	4.51	0.11
9	2-chlorophenol	4.02	3.79	0.23
10	3-chlorotoluene	3.84	3.88	−0.04
11	4-chlorotoluene	4.33	3.87	0.46
12	1, 3-dihydroxybenzene	3.04	3.43	−0.39
13	3-hydroxyanisole	3.21	3.33	−0.12
14	2-methylphenol	3.77	3.64	0.13
15	3-methylphenol	3.29	3.60	−0.31
16	4-methylphenol	3.58	3.53	0.05
17	4-nitrophenol	3.36	3.61	−0.25
18	1, 4-dimethoxybenzene	3.07	3.28	−0.21
19	1, 2-dimethylbenzene	3.48	3.93	−0.45
20	1, 4-dimethylbenzene	4.21	3.87	0.34
21	2-nitrotoluene	3.57	3.66	−0.09
22	3-nitrotoluene	3.63	3.53	0.10
23	4-nitrotoluene	3.76	3.49	0.27
24	1, 2-dinitrobenzene	5.45	5.24	0.21
25	1, 3-dinitrobenzene	4.38	4.18	0.20
26	1, 4-dinitrobenzene	5.22	4.94	0.28
27	2-methyl-3-nitroaniline	3.48	3.79	−0.31
28	2-methyl-4-nitroaniline	3.24	3.51	−0.27
29	2-methyl-5-nitroaniline	3.35	3.68	−0.33
30	2-methyl-6-nitroaniline	3.80	3.84	−0.04
31	3-methyl-6-nitroaniline	3.80	3.78	0.02
32	4-methyl-2-nitroaniline	3.79	3.80	−0.01
33	4-hydroxy-3-nitroaniline	3.65	3.61	0.04
34	4-methyl-3-nitroaniline	3.77	3.73	0.04
35	1, 2, 3-trichlorobenzene	4.89	4.89	−0.00
36	1, 2, 4-trichlorobenzene	5.00	5.04	−0.04
37	1, 3, 5-trichlorobenzene	4.74	5.11	−0.37
38	2, 4-dichlorophenol	4.30	4.33	−0.03
39	3, 4-dichlorotoluene	4.74	4.26	0.48
40	2, 4-dichlorotoluene	4.54	4.36	0.18
41	4-chloro-3-methylphenol	4.27	3.87	0.40
42	2, 4-dimethylphenol	3.86	3.76	0.10
43	2, 6-dimethylphenol	3.75	3.80	−0.05
44	3, 4-dimethylphenol	3.90	3.80	0.10
45	2, 4-dinitrophenol	4.04	4.14	−0.10
46	1, 2, 4-trimethylbenzene	4.21	4.09	0.12
47	2, 3-dinitrotoluene	5.01	5.20	−0.19
48	2, 4-dinitrotoluene	3.75	4.10	−0.35
49	2, 5-dinitrotoluene	5.15	4.84	0.31
50	2, 6-dinitrotoluene	3.99	4.41	−0.42
51	3, 4-dinitrotoluene	5.08	5.11	−0.03

TABLE I (Continued)

52	3, 5-dinitrotoluene	3.91	4.05	−0.14
53	1, 3, 5,-trinitrobenzene	5.29	5.37	−0.08
54	2-methyl-3, 5-dinitroaniline	4.12	4.13	−0.01
55	2-methyl-3, 6-dinitroaniline	5.34	4.80	0.54
56	3-methyl-2, 4-dinitroaniline	4.26	4.28	−0.02
57	5-methyl-2, 4-dinitroaniline	4.92	4.14	0.78
58	4-methyl-2, 6-dinitroaniline	4.21	4.67	−0.46
59	5-methyl-2, 6-dinitroaniline	4.18	4.80	−0.62
60	4-methyl-3, 5-dinitroaniline	4.46	4.34	0.12
61	2, 4, 6-tribromophenol	4.70	4.89	−0.19
62	1, 2, 3, 4-tetrachlorobenzene	5.43	5.62	−0.19
63	1, 2, 4, 5-tetrachlorobenzene	5.85	5.80	0.05
64	2,4, 6-trichlorophenol	4.33	4.79	−0.46
65	2-methyl-4, 6-dinitrophenol	5.00	4.21	0.79
66	2, 3, 6-trinitrotoluene	6.37	6.36	0.01
67	2, 4, 6-trinitrotoluene	4.88	5.16	−0.28
68	2, 3, 4, 5-tetrachlorophenol	5.72	5.36	0.36
69	2, 3, 4, 5, 6-pentachlorophenol	6.06	6.03	0.03

(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. Brief definitions of the topological indices are shown in Table II.

Topological Indices

The 102 topological indices used in this study, both the topostructural and the topochemical, have been calculated by POLLY 2.3 [33] and software developed by the authors. These indices include Wiener index [34], connectivity indices developed by Randić [35] and higher order connectivity indices formulated by Kier and Hall [36], bonding connectivity indices defined by Basak *et al.* [37], a set of information theoretic indices defined on the distance matrices of simple molecular graphs [38, 39] and neighborhood complexity indices of hydrogen-filled molecular graphs [40, 41], and Balaban's *J* indices [42 - 44]. Table III provides the list of the topostructural, topochemical, geometrical and quantum chemical indices included in this study

Geometrical Indices

Van der Waals volume, V_w [45-47], was calculated using *Sybyl* 6.1 from Tripos Associates, Inc [48]. The 3-*D* Wiener numbers were calculated by *Sybyl* using an SPL (Sybyl Programming Language) program developed in

TABLE II 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
$\overline{IC}$	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 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
IC_r	Mean information content or complexity of a graph based on the r^{th}(r = 0–5) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th}(r = 0–5) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th}(r = 0–5) 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, 5
${}^h\chi_{Ch}$	Chain connectivity index of order h = 6
${}^h\chi_{PC}$	Path-Cluster connectivity index of order h = 4–6
${}^h\chi^b$	Bond path connectivity index of order h = 0–6
${}^h\chi^b_C$	Bond cluster connectivity index of order h = 3, 5
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order h = 6
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order h = 4–6
${}^h\chi^v$	Valence path connectivity index of order h = 0–6
${}^h\chi^v_C$	Valence cluster connectivity index of order h = 3, 5
${}^h\chi^v_{PC}$	Valence path- cluster connectivity index of order h = 4–6
P_h	Number of paths of length h = 1–9
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^Y	Balaban's J index based on relative covalent radii
V_W	van der Waals 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

TABLE III Classification of parameters used in developing models for acute aquatic toxicity (LC_{50}) in *Pimephales promelas*

Topological	*Topochemical*	*Geometric*	*Quantum Chemical AM1*
I_D^W	I_{ORB}	V_w	E_{HOMO}
$\overline{I_D^W}$	IC_0-IC_5	^{3D}W	E_{HOMO1}
W	SIC_0-SIC_5	$^{3D}W_H$	E_{LUMO}
I^D	CIC_0-CIC_5		E_{LUMO1}
H^V	$^0\chi^h - {}^6\chi^h$		ΔH_f
H^D	$^3\chi_c^h$ and $^5\chi_c^h$		μ
$\overline{IC}$	$^6\chi_{ch}^h$		
O	$^4\chi_{pc}^h - {}^6\chi_{pc}^h$		
M_1	$^0\chi^v - {}^6\chi^V$		
M_2	$^3\chi_c^v$ and $^5\chi_c^v$		
$^0\chi - {}^6\chi$	$^4\chi_{pc}^v - {}^6\chi_{pc}^v$		
$^3\chi_c$ and $^5\chi_c$	J^B		
$^6\chi_{ch}$	J^X		
$^4\chi_{pc} - {}^6\chi_{pc}$	J^Y		
P_1 - P_9			
J			

our lab [49]. 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 [50]. 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

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 [51].

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 transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of eighty-one topological indices was then partitioned into two distinct sets, the topostructural indices (thirty-three) and the topochemical indices (forty-seven). To further reduce the number of independent variables for model construction, the sets to topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [52]. 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 acute aquatic toxicity of benzene derivatives in fathead minnow. 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.

Reducing the number of independent variables is critical when attempting to model small datasets. The smaller the dataset is, the greater the chance of spurious error when using a large number of independent variables (descriptors). Topliss and Edwards have studied this issue of chance correlations [53]. For a set with about seventy dependent variables (observations), to keep the probability of chance correlations less than 0.01, we can use at most forty independent variables. This number is dependent on the actual correlation achieved in the modeling process, with a high correlation we have a better chance of using more variables with the same limited probability of chance correlations. In this study we are well below the cut-off of forty. In fact, the total number of descriptors which will be used for model construction and estimation is twenty-three, well within the bounds of the Topliss and Edwards criteria [53].

Statistical Analysis and Hierarchical QSAR

Regression modeling was accomplished using the SAS procedure REG on seven distinct sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest parameters, the TSIs. After using the TSIs to model the activity, the next level of complexity is added. To the indices included in the best TSI model, we add all of the TCIs and proceed to model the activity using these parameters. Likewise, the indices included in the best model from this procedure are combined with the indices from the next level, the geometrical indices and modeling is conducted once again. Finally, the best model utilizing TSIs, TCIs and geometrical indices is combined with the quantum chemical parameters. The regression analysis results in the final selection of indices for each of the models. The remaining three models which use TCIs, geometric, and quantum chemical parameters independently serve as a means of validating the utility of the hierarchical approach and the need for varying types of theoretical descriptors.

RESULTS

The variable clustering of the topostructural indices resulted in the retention of five indices: $M_1, \overline{IC}, O, P_8, P_9$. All-possible subsets regression resulted in the selection of a four-parameter model to estimate $-\log(LC_{50})$ with an explained variance (R^2) of 45.3% and a standard error (s) of 0.58. While this is an unsatisfactory model, the indices will still be retained and combined with the topochemical indices in the second step of model development. Table IV lists the indices used in each of the models.

The second step of the hierarchical method combined the four indices used in the first tier model with the nine topochemical indices selected in the variable clustering procedure: $SIC_0, SIC_1, SIC_4, CIC_0, {}^2\chi^b, {}^5\chi^b C, {}^5\chi^v C, {}^6\chi^v_{PC}, J_x$. Again all-possible subsets regression was conducted resulting in a four-parameter model with an explained variance (R^2) of 78.3% and a standard error(s) of 0.36. While this model retained two parameters from the topostructural model, it is evident that the addition of two topochemical indices made a significant contribution to the effectiveness of our model.

The four indices from the second tier model were then combined with the three geometric parameters: ${}^{3D}W_H, {}^{3D}W, V_W$. The resulting model from this procedure retained four indices, replacing the topochemical index CIC_0 with the geometric parameter ${}^{3D}W_H$. This model had an explained variance (R^2) of 79.2% and a standard error (s) of 0.36.

The final step in the hierarchical method combined the four parameters from the third tier model with the quantum chemical (AM1) parameters: $E_{\mathrm{HOMO}}, E_{\mathrm{HOMO\ 1}}, E_{\mathrm{LUMO}}, E_{\mathrm{LUMO\ 1}}, \Delta H_f, \mu$. This set of ten indices led to a seven-parameter model with an explained variance (R^2) of 86.3% and a standard error(s) of 0.30. This model retained all of the indices from the third model and added three quantum chemical parameters.

Three other models were constructed for the purpose of comparison. These include a five-parameter topochemical model, a three parameter geometric model, and a four-parameter quantum chemical model. The indices used in these models and the results of the models can be found in Table IV.

DISCUSSION

The goal of this paper was to investigate the utility of hierarchical QSAR using algorithmically derived molecular descriptors in predicting LC_{50} values for a set of sixty-nine benzene derives. To this end, we used four classes of parameters, viz., topostructural descriptors, topochemical indices, geometrical descriptors and semiempirical quantum chemical indices.

It is clear from the results described in Table IV that none of the individual classes of parameters correlate well with acute aquatic toxicity. The TSIs, the simplest of the four classes of parameters, explained about 45% of the variance in toxicity. The inclusion of topochemical indices in the set of independent variables made substantial improvement in the predictive capacity of the QSAR models. This is understandable since the benzene derivatives analyzed in this paper comprise a fairly congeneric set, and while the number and size of substituents may be important, the chemical nature of the substituents also plays an important role in determining the overall toxicity of the molecule. This is shown by the dramatic increase in predictive power between Eqs. 1 and 2. Equation 2 replaces two TSI descriptors with two TCI indices that are sensitive to the atom types in all zero-order neighborhoods. The addition of this basic chemical information results in an

TABLE IV Summary of the regression results for all models for the full set of sixty-nine benzene derivatives

Eq	*Parameter class*	*Variables Included*	*F*	R^2	*S*
1	TSI	M_1, $\overline{IC}$, P_8, P_9	13.3	0.453	0.58
2	TSI + TCI	M_1, P_9, SIC_0, CIC_0	57.9	0.783	0.36
3	TSI + TCI + Geometric	M_1, P_9, SIC_0, ${}^{3D}W'_H$	61.1	0.792	0.36
4	TSI + TCI + Geometric + Quantum Chemical	M_1, P_9, SIC_0, ${}^{3D}W'_H$ E_{LUMO1}, ΔH_f, μ	55.0	0.863	0.30
5	TCI	SIC_0, SIC_1, CIC_0, ${}^2\chi^h$, J^X	34.3	0.731	0.41
6	Geometric	${}^{3D}W_H$, ${}^{3D}W$, V_W	34.8	0.616	0.48
7	Quantum Chemical	E_{HOMO1}, E_{LUMO}, E_{LUMO1}, μ	23.8	0.598	0.50

TABLE V Calculated values for the topostructural, topochemical, geometric and quantum chemical parameters used in Eq. 4 (Tab. IV)

No.	M_1	P_9	SIC_0	$^{3D}W_H$	E_{LUMOI}	ΔH_f	μ
1	3	0	0.246	5.21	0.5540	22.0240	0.005
2	3	0	0.315	5.25	0.2447	26.7581	1.449
3	3	0	0.315	5.25	0.2632	14.8214	1.299
4	3	0	0.304	5.43	0.5095	−22.2334	1.233
5	3	0	0.227	5.79	0.5745	16.5004	0.279
6	4	0	0.341	5.28	−0.0203	9.2203	1.974
7	4	0	0.341	5.28	−0.0462	8.2544	1.218
8	4	0	0.341	5.28	−0.0988	10.4661	0.000
9	4	0	0.362	5.46	0.2406	−28.6621	0.934
10	4	0	0.284	5.81	0.2785	7.1915	1.478
11	4	0	0.284	5.82	0.3208	7.1066	1.623
12	4	0	0.323	5.64	0.3778	−66.4516	2.433
13	4	0	0.295	6.16	0.4618	−59.9961	2.338
14	4	0	0.276	5.95	0.5331	−28.9297	0.960
15	4	0	0.276	5.97	0.5610	−29.6368	1.079
16	4	0	0.276	5.97	0.4880	−29.7869	1.333
17	4	0	0.376	5.84	−0.4095	−19.5199	5.261
18	4	0	0.274	6.59	0.5766	−52.9350	2.424
19	4	0	0.213	6.22	0.6180	7.5221	0.465
20	4	0	0.213	6.28	0.6450	6.8236	0.003
21	4	0	0.341	6.11	−0.2692	19.0823	5.015
22	4	0	0.341	6.14	−0.2921	17.6145	5.443
23	4	0	0.341	6.15	−0.2334	17.2948	5.728
24	4	2	0.389	5.99	−1.2793	38.6210	7.804
25	4	0	0.389	6.01	−1.5339	33.1466	4.845
26	4	0	0.389	6.02	−1.0875	33.2941	0.013
27	4	0	0.344	6.38	−0.1596	20.4489	5.727
28	4	0	0.344	6.41	−0.0919	14.3213	7.434
29	4	0	0.344	6.41	−0.1084	19.7541	6.185
30	4	0	0.344	6.39	−0.0006	13.8471	5.374
31	4	0	0.344	6.42	0.1022	12.9086	5.649
32	4	0	0.344	6.42	0.0314	13.3128	5.280
33	4	0	0.376	6.15	−0.2384	−15.9560	6.801
34	4	0	0.344	6.41	−0.1379	18.0141	5.596
35	4	0	0.349	5.31	−0.3391	4.2313	2.070
36	4	0	0.349	5.31	−0.2761	2.9490	1.033
37	4	0	0.349	5.31	−0.3927	2.2158	0.020
38	4	0	0.385	5.49	−0.1034	−35.1296	0.395
39	4	0	0.312	5.84	0.0251	1.5862	2.296
40	4	0	0.312	5.84	0.0006	1.2199	1.464
41	4	0	0.326	5.99	0.2063	−36.1532	1.059
42	4	0	0.255	6.40	0.5006	−36.4200	1.052
43	4	0	0.255	6.38	0.5503	−35.5810	1.199
44	4	0	0.255	6.38	0.5387	−36.6403	1.229
45	4	0	0.383	6.17	−1.5210	−8.7887	6.201
46	4	0	0.202	6.64	0.6477	−0.1093	0.274
47	4	2	0.365	6.40	−1.2262	31.8226	7.909
48	4	0	0.365	6.43	−1.4332	26.3804	5.390
49	4	0	0.365	6.42	−1.0421	26.9397	0.797
50	4	0	0.365	6.39	−1.4076	30.3487	3.639
51	4	2	0.365	6.43	−1.1564	32.0703	8.256
52	4	0	0.365	6.44	−1.4923	25.3294	5.321

TABLE V (Continued)

No.	M_1	P_9	SIC_0	$^{3D}W_H$	E_{LUMO1}	ΔH_f	μ
53	4	0	0.378	6.33	−2.5221	44.8961	0.032
54	4	0	0.362	6.66	−1.2453	27.9172	6.590
55	4	0	0.362	6.65	−0.6994	25.1359	3.166
56	4	0	0.362	6.65	−1.1532	23.8377	5.797
57	4	0	0.362	6.67	−1.3084	51.2351	7.196
58	4	0	0.362	6.68	−1.0204	18.0757	2.366
59	4	0	0.362	6.66	−1.0160	54.7718	3.199
60	4	0	0.362	6.66	−1.2172	29.5227	5.090
61	4	0	0.392	5.54	−0.4993	2.2014	1.096
62	4	0	0.341	5.34	−0.5585	−0.5979	1.616
63	4	0	0.341	5.34	−0.6587	3.2072	0.000
64	4	0	0.392	5.52	−0.3777	−38.2930	1.083
65	4	0	0.362	6.56	−1.5102	−19.8380	4.669
66	4	2	0.365	6.66	−1.9189	46.0695	3.518
67	4	0	0.365	6.67	−2.3240	41.4239	1.418
68	4	0	0.385	5.54	−0.5526	−43.2613	1.231

improvement in the model. A similar conclusion is borne out from the QSAR analysis of the same set of benzene derivatives reported by Hall *et al.* where they found that the chemical nature of the substituent is important in determining toxicity [32].

In the next tier, Eq. 3 replaces one of the information content indices with the three-dimensional Wiener number, a descriptor that characterizes the three-dimensional aspects of molecular shape and size. This leads to refinement of the model developed in Eq. 2. Finally, the addition of the quantum chemical parameters; energy of the second lowest unoccupied molecular orbital, heat of formation, and dipole moment; leads to a marked improvement in the predictive power of the model (Eq. 4).

As can be seen from Eqs. 1 and 5–7 (Tab. IV), none of the four classes of indices do very well individually. The hierarchical QSAR approach using four classes of parameters resulted in acceptable predictive models (Eq. 4). We may conclude from the results presented in this paper that each of the four classes of theoretical descriptors that were used are necessary for the development of good QSARs for the acute aquatic toxicity of benzene derivatives in fathead minnow.

Acknowledgments

This is contribution number 213 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, a grant from

Exxon Corporation and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to extend their thanks to Greg Grunwald for technical support.

References

[1] Personal communication with W. Fisanick, 20, 1997.

[2] Menzel, D. B. (1995). *Extrapolating the future; research trends in modeling. Toxicol. Lett.*, 79, 299 - 303.

[3] Hansch, C. and Leo, A. (1995). *Exploring QSAR; Fundamental and Applications in Chemistry and Biology*. American Chemical Society, Washington, D.C., p. 557.

[4] Dearden, J. C. (1990). Physico-chemical descriptors. In, *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (W. Karcher and J. Devillers, Eds.). Kluwer Academic Publishers, Dordrecht, pp 25 - 59.

[5] Lipnick, R. L. (1990). Narcosis: Fundamental and baseline toxicity mechanism for nonelectrolyte organic chemicals. In, *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (W. Karcher and J. Devillers, Eds.). Kluwer Academic Publishers, Dordrecht, pp. 281 - 293.

[6] Van de Waterbeemd, H. (1995). Discriminant analysis for activity prediction. In *Chemometric Methods in Molecular Design* (H. Van de Waterbeemd, Ed.), VCH Publishers, Inc., New York, pp. 283 - 294.

[7] Kamlet, M. J., Abboud, J.-L. M. and Taft, R W. (1977). Solvatochromic comparison method 6. π^* scale of solvent polarities. *J. Am. Chem. Soc.*, **99**, 6027 - 6038.

[8] Kamlet, M J., Abboud, J.-L. M., Abraham, M. H. and Taft, R. W. (1983). Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, π^*, α and β, and some methods for simplifying the generalized solvatochromic equation. *J. Org. Chem.*, **48**, 2877 - 2887.

[9] Hansch, C (1976). On the structure of medicinal chemistry. *J. Med. Chem.*, **19**, 1 - 6.

[10] Randic, M. (1980). A graph theoretical approach to structure-property and structure-activity correlations. *Theoret. Chim. Acta* (*Berl.*), **58**, 45 - 68.

[11] Randic, M (1984). Nonempirical approach to structure-activity studies. *Int. J. Quant. Chem.*, **11**, 137 - 153.

[12] Randic, M (1995). Molecular topographic indices. *J. Chem. Inf. Comput. Sci.*, 35, 140 - 147

[13] Sabljic, A. and Trinajstic, N. (1981). Quantitative structure-activity relationships: the role of topological indices. *Acta Pharm. Jugosl.*, **31**, 189 - 214.

[14] Basak, S C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.*, **15**, 605 - 609.

[15] Balaban, A. T., Bertelsen, S. and Basak, S. C. (1994). New centric topological indexes for acyclic molecules (trees) and substituents (rooted trees), and coding of rooted trees. *MATCH*, **30**, 55 - 72.

[16] Basak, S C and Grunwald, G. D. (1995). Estimation of lipophilicity from molecular structural similarity. *New J. Chem.*, **19**, 231 - 237

[17] Diudea, M V., Horvath, D. and Graovac, A. (1995). Molecular topology. 15. 3*D* distance matrices and related topological indices *J Chem. Inf. Comput. Sci.*, **35**, 129 - 135.

[18] Estrada, E (1995). Three-dimensional molecular descriptors based on electron charge density, weighted graphs. *J. Chem. Inf. Comput Sci.*, **35**, 708 - 713.

[19] Voelkel, A (1994). Structural descriptors in organic chemistry – new topological parameter based on electrotopological state of graph vertices. *Computers Chem.*, **18**, 1 - 4.

[20] Basak, S C., Gute, B. D. and Grunwald, G D (1996). Estimation of normal boiling points of haloalkanes using molecular similarity *Croat. Chem. Acta*, **69**, 1159 - 1173.

[21] Basak, S. C., Gute, B. D. and Drewes, L. R. (1996). Predicting blood-brain transport of drugs: a computational approach. *Pharm. Res.*, **13**, 775–778.

[22] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Development and application of molecular similarity methods: using nonempirical parameters. *Mathl. Model. And Sci. Comput.*, in press.

[23] Basak, S. C. and Gute, B. D. (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal *p*-hydroxylation of aniline by alcohols: a molecular similarity approach. In *Proceedings of the 2nd international Congress on Hazardous Waste: Impact on Human and Ecological Health* (B. L. Johnson, C. Xintaras, and Jr. J. S. Andrews, Eds.). Princeton Scientific Publishing Co., Inc., New Jersey, pp. 492–504.

[24] Basak, S. C., Gute, B. D. and Ghatak, S. (1997). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, submitted.

[25] Famini, G. R., Penski, C. A. and Wilson, L. Y. (1992). Using theoretical descriptors in quantitative structure activity relationships: some physicochemical properties. *J. Phys. Org. Chem.*, **5**, 395–408.

[26] Cramer, C. J., Famini, G. R. and Lowrey, A. H. (1993). Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationships. *Acc. Chem. Res.*, **26**, 599–605.

[27] Famini, G. R., Wilson, L. Y. and DeVito, S. C. (1994). Modeling cytochrome P-450 mediated acute nitrile toxicity using theoretical linear solvation energy relationships. In, *Biomarkers of Human Exposures to Pesticides* (M. A. Saleh, J. N. Blancato and C. H. Nauman, Eds.) American Chemical Society, pp. 22–36.

[28] 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.

[29] 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 QSAR approach. *J. Chem. Inf. Comput. Sci.*, **37**, 651–655.

[30] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals. In, *Quantitative Structure-Activity Relationships in Environmental Sciences* (F. Chen and G. Schüürman, Eds.). SETAC Press, in press.

[31] Basak, S. C., Gute, B. D. and Grunwald, G. D. (1997). Development of a quantitative structure-activity relationship (QSAR) for estimating bioconcentration factor in *Pimephales promelas*: a hierarchical approach. In progress.

[32] Hall, L. H., Kier, L. B. and Phipps, G. (1984). Structure-activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem.*, **3**, 355–365.

[33] Basak, S. C., Harriss, D. K. and Magnuson, V. R. (1988). POLLY 2.3: Copyright of the University of Minnesota

[34] Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.

[35] Randic, M. (1975) On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609–6615.

[36] Kier, L. B. and Hall, L. H. (1986) *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, Hertfordshire, UK.

[37] Basak, S. C. and Magnuson, V. R. (1988) Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.*, **19**, 17–44.

[38] 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.

[39] Bonchev, D. and Trinajstic, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **67**, 4517–4533.

[40] 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.

[41] 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.

[42] Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399-404.

[43] Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.*, **55**, 199-206.

[44] Balaban, A. T. (1986). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into accout periodicities of element properties. *Math. Chem. (MATCH)*, **21**, 115-122.

[45] Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441-451.

[46] Moriguchi, I. and Kanada, Y. (1977) Use of van der Waals volume in structure-activity studies. *Chem. Pharm. Bull.*, **25**, 926-935.

[47] 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.

[48] *SYBYL Version* 6.1. (1994). Tripos Associates, Inc: St. Louis, MO.

[49] Mekenyan, O., Peitchev, D., Bonchev, D., Trinajstic, 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.

[50] *CONCORD Version* 3.0.1. (1993). Tripos Associates, Inc.: St. Louis, MO.

[51] Stewart, J. J. P. (1990). *MOPAC Version* 6.00. QCPE # 455. Frank J Seiler Research Laboratory: US Air Force Academy, CO

[52] SAS Institute Inc. (1998). In *SAS/STAT User's Guide, Release 6.03 Edition*, SAS Institute Inc.: Cary, NC.

[53] Topliss, J. G. and Edwards, R P (1979) Chance factors in studies of quantitative structure-activity relationships *J Med Chem*, **22**, 1238-1244.

APPENDIX 1.16 Use of topostructural, topochemical and geometric parameters in the prediction...

Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach

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

Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received December 31, 1996[⊗]

Numerous quantitative structure−activity relationships (QSARs) have been developed using topostructural, topochemical, and geometrical molecular descriptors. However, few systematic studies have been carried out on the relative effectiveness of these three classes of parameters in predicting properties. We have carried out a systematic analysis of the relative utility of the three types of structural descriptors in developing QSAR models for predicting vapor pressure at STP for a set of 476 diverse chemicals. The hierarchical technique has proven to be useful in illuminating the relationships of different types of molecular description information to physicochemical property and is a useful tool for limiting the number of independent variables in linear regression modeling to avoid the problems of chance correlations.

1. INTRODUCTION

A large number of quantitative structure−activity relationship (QSAR) studies have been reported in recent literature using theoretical molecular descriptors in predicting physicochemical, pharmacological, and toxicological properties of molecules.[1−15] Such descriptors comprise graph invariants, geometrical or 3-D parameters, and quantum chemical indices. One of the reasons for the current upsurge of interest is the fact that such descriptors can be derived algorithmically, i.e., can be computed for any molecule, real or hypothetical, using standard software. Both in pharmaceutical drug design and in risk assessment of chemicals, one has to evaluate potential biological effects of chemicals. Evaluation schemes based on property−property correlation paradigms are not very useful in practical situations, because, for most of the candidate structures, the experimental data necessary for proper evaluation are not available. This is especially true for the thousands of chemicals rapidly produced by methods of combinatoric chemistry[16] as well as for the large number of chemicals present in the Toxic Substances Control Act (TSCA) Inventory.[17]

A large number of physicochemical and biological endpoints are necessary for estimating the ecotoxicological fate, transport, and effects of environmental pollutants.[17−19] The vapor pressure of chemicals is important in determining the partitioning of chemicals among different phases once they are released in the environment. Many QSARs have been reported for predicting normal vapor pressure of chemicals. Such studies are usually carried out on small sets of congeneric chemicals. Also, many QSARs use experimental data as inputs in the model. Therefore, it becomes necessary to develop QSARs based on nonempirical parameters which can predict the vapor pressure for a heterogeneous collection of chemicals so that such models are generally applicable. With this end in mind, in the current paper we have carried out a QSAR study of 476 diverse chemicals using three types of nonempirical molecular descriptors.

2. MATERIALS AND METHODS

2.1. Normal Vapor Pressure Database. Measured values for a subset of the Toxic Substances Control Act (TSCA) Inventory[17] were obtained from the ASTER (Assessment Tools for the Evaluation of Risk) database.[20] This subset consisted of a diverse set of chemicals where vapor pressure (p_{vap}) was measured at 25 °C and over a pressure range of approximately 3−10 000 mmHg. Due to the size of the dataset being used in this study, data for these chemicals will not be listed in this paper. An electronic copy of the data may be obtained by contacting the authors.

2.2. Computation of Topological Indices. The majority of the topological indices (TIs) used in this study have been calculated by the computer program POLLY 2.3.[21] These indices include Wiener index,[22] the molecular connectivity indices developed by Randić and Kier and Hall,[1,23] information theoretic indices defined on distance matrices of graphs,[24,25] and a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs.[2,26−28] Balaban's *J* indices[29−31] were calculated using software developed by the authors.

van der Waal's volume (V_w)[32−34] was calculated using Sybyl 6.2.[35] The 3-D Wiener numbers[36] were calculated by Sybyl using an SPL (Sybyl Programming Language) program developed by the authors. Calculation of 3-D Wiener numbers consists of the summation of the 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 3.2.1.[37] Two variants of the 3-D Wiener number were calculated, $^{3D}W_H$ and ^{3D}W, where hydrogen atoms are included and excluded from the computations, respectively.

Table 1 provides a complete listing of all of the topological and geometrical parameters which have been used in this study. The listing includes the symbols used to represent the parameters and brief definitions for each of the parameters.

Two additional parameters were used in modeling normal vapor pressure, HB_1, and dipole moment (μ). HB_1 is a simple hydrogen bonding parameter calculated using a program developed by Basak,[38] which is based on the ideas

* All correspondence should be addressed to Dr. Subhash C. Basak, Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811

[⊗] Abstract published in *Advance ACS Abstracts*, June 1, 1997.

S0095-2338(96)00176-X CCC: $14.00 © 1997 American Chemical Society

Table 1. Symbols and Definitions of Topological and Geometrical Parameters

I^{W}_{D}	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I^{W}_{D}}$	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
$\overline{\mathrm{IC}}$	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 IC_r reaches it 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
IC_r	mean information content or complexity of a graph based on the r^{th}(r = 0−5) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth (r = 0−5) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complmentary information content for rth (r = 0−5) 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 = 5, 6
${}^{h}\chi^{b}$	bond path connectivity index of order h = 0−6
${}^{h}\chi^{b}_{C}$	bond cluster connectivity index of order h = 3−6
${}^{h}\chi^{b}_{Ch}$	bond chain connectivity index of order h = 5, 6
${}^{h}\chi^{b}_{PC}$	bond path-cluster connectivity index of order h = 4−6
${}^{h}\chi^{v}$	valence path connectivity index of order h = 0−6
${}^{h}\chi^{v}_{C}$	valence cluster connectivity index of order h = 3−6
${}^{h}\chi^{v}_{Ch}$	valence chain connectivity index of order h = 5, 6
${}^{h}\chi^{v}_{PC}$	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^{Y}	Balabən'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

of Ou *et al.*[39] Dipole moment was calculated using Sybyl 6.2.[35]

2.3. Data Reduction. The set of 92 TIs was partitioned into two distinct subsets: topostructural indices and topochemical indices. The distinction was made as follows: topostructural indices 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 like hybridization states of atoms and number of core/valence electrons in individual atoms, while topochemical indices quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. 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 2.

Table 2. Classification of Parameters used in Modeling Normal Vapor Pressure [$\log_{10}(p_{vap})$]

topological	topochemical	geometric	other parameters
I_D^{W}	I_{ORB}	V_W	HB_1
$\overline{I_D^{W}}$	IC_0-IC_5	${}^{3D}W$	μ
W	SIC_0-SIC_5	${}^{3D}W_H$	
I^{D}	CIC_0-CIC_5		
H^{V}	${}^{0}\chi^{b}-{}^{6}\chi^{b}$		
H^{D}	${}^{3}\chi^{b}_{C}-{}^{6}\chi^{b}_{C}$		
$\overline{IC}$	${}^{5}\chi^{b}_{Ch}$ and ${}^{6}\chi^{b}_{Ch}$		
O	${}^{4}\chi^{b}_{PC}-{}^{6}\chi^{b}_{PC}$		
M_1	${}^{0}\chi^{v}-{}^{6}\chi^{v}$		
M_2	${}^{3}\chi^{v}_{C}-{}^{6}\chi^{v}_{C}$		
${}^{0}\chi-{}^{6}\chi$	${}^{5}\chi^{b}_{Ch}$ and ${}^{6}\chi^{b}_{Ch}$		
${}^{3}\chi_{C}-{}^{6}\chi_{C}$	${}^{4}\chi^{b}_{PC}-{}^{6}\chi^{b}_{PC}$		
${}^{5}\chi_{Ch}$ and ${}^{6}\chi_{Ch}$	J^{B}		
${}^{4}\chi_{PC}-{}^{6}\chi_{PC}$	J^{X}		
P_0-P_{10}	J^{Y}		
J			

The partitioning of the indices left 38 topostructural indices and 54 topochemical indices. At this point no further data reduction is called for, since the ratio of the number of observations in the training set (342) to the total number of variables (92 maximum) falls well within the condition limits suggested by Topliss and Edwards[40] for reducing the probability of spurious correlations even at the more conservative $R^2 \geq 0.7$ level.

2.4. Statistical Analysis and Hierarchical QSAR. 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. The geometric parameters were transformed by the natural logarithm of the parameter.

Two regression procedures were used in developing the linear models. When the number of independent variables was high, typically greater than 25, a stepwise regression procedure was used to maximize the improvement of the explained variance (R^2). When the number of independent variables was smaller, all possible subsets regression was used. Models were then optimized to reduce problems of variance inflation and collinearity. Regression modeling was conducted using the REG procedure of the statistical package SAS.[41]

The vapor pressure data (p_{vap}) was split into a training set (342 compounds) and a test set (134 compounds), an approximately 75/25 split. Models were developed using the training set of chemicals and then used to predict the p_{vap} values of the test chemicals. Final models were then developed using the combined training and test set of chemicals.

Five sets of indices were used in model development. These sets were constructed as part of a hierarchical approach to QSAR modeling. The hierarchy begins with the simplest indices, the topostructural. After developing our initial model utilizing the topostructural indices, we increase the level of complexity. To the indices included in the best topostructural model, we add all of the topochemical indices and proceed to model p_{vap} using these parameters. Likewise, the indices included in the best model from this procedure are combined with the geometrical indices and modeling is conducted once again. In addition to this hierarchical approach, models were also constructed using the topochemical indices alone and the geometrical indices alone for purposes of comparison.

3. RESULTS

Stepwise regression analyses for $\log_{10}(p_{vap})$ of the training set of chemicals is summarized in Table 3. As shown in

Table 3. Summary of the Regression Results for the Training Set and the Prediction Results for the Test Set for the Hierarchical Analysis of $\log_{10}(p_{vap})$

	training set ($N = 342$)				test set ($N = 134$)	
parameter class	variables included	F	R^2	s	R^2	s
topostructural	$^1\chi$, $^6\chi_C$, P_9	104.6	48.1	0.56	57.9	0.46
topochemical	SIC_0, SIC_2, SIC_3, CIC_0, CIC_1, $^3\chi^b{}_C$, $^1\chi^v$, $^5\chi^v$, $^3\chi^v{}_C$, J^Y	126.3	79.2	0.36	85.8	0.27
geometrical	^{3D}W, $^{3D}W_H$, V_W	168.9	51.8	0.53	62.2	0.44
topostructural + topochemical	$^1\chi$, P_9, IC_1, SIC_2, CIC_1, $^3\chi^b{}_C$, $^1\chi^v$, $^3\chi^v$, $^6\chi^v$, $^3\chi^v{}_C$, $^5\chi^v{}_{Ch}$	112.5	80.4	0.35	84.7	0.28
all indices	H^V, SIC_1, SIC_2, CIC_0, CIC_3, $^6\chi_C$, $^1\chi^v$, $^3\chi^v$, $^6\chi^v{}_C$, P_6, P_{10}	117.4	79.6	0.35	84.2	0.28
ttg + HB_1 + μ	$^1\chi$, P_3, P_9, IC_0, $^1\chi^b$, $^3\chi^b{}_C$, $^1\chi^v$, $^3\chi^v$, $^3\chi^v{}_C$, HB_1	160.8	82.9	0.32	83.1	0.29

Table 3, the topostructural model using three parameters resulted in an explained variance (R^2) of 48.1% and a standard error (s) of 0.56. Addition of the topochemical parameters to the three topostructural parameters led to a significant increase in the effectiveness of the model. The resulting model used 12 parameters, two topostructural and ten topochemical. This model had an R^2 of 80.4% and s of 0.35. All subsets regression of the two topostructural and ten topochemical indices retained thus far and the three geometrical indices resulted in the selection of the same 12 parameter model, thus the geometrical indices did not contribute significantly to model development. Several other models were constructed for comparative purposes. Using topochemical indices only, a ten parameter model was developed which had an R^2 of 79.2% and s of 0.36. A geometrical model was developed which utilized all three geometrical indices and resulted in an R^2 of 51.8% and s of 0.53. Finally, two additional stepwise models were developed. One model simply used all indices for a comparison between a simple stepwise analysis of the data and the results of the hierarchical procedure. This resulted in an 11 parameter model with R^2 of 79.6% and s of 0.35. The second model added two new parameters, HB_1 and μ. We thought that it might be possible to improve our modeling by adding in some other nonempirical parameters which could be important to the determination of normal vapor pressure. We selected the parameters HB_1 and μ, since they would be important in intermolecular interactions which could have a dramatic effect on vapor pressure. To look at the addition of these parameters, we conducted a stepwise regression analysis using all topostructural, topochemical, and geometric indices so that we would be able to optimize our model, just as we had done with the previous models. The addition of these parameters led to the selection of a ten parameter model which included three topostructural indices, nine topochemical indices, and HB_1. This was the best model yet, with an R^2 of 82.9% and s of 0.32.

Application of these six models to the test set of chemicals resulted in comparable R^2 and s; actually all models improved slightly on their predictions of the test set, and these values are also listed in Table 3. Based on these results, we decided that it was pointless to develop further models using only geometrical parameters. Also, based on the findings that the geometrical indices did not contribute significantly to any of the training models, they were dropped from the development of final models for the full set of 476 chemicals. However, even though the topostructural indices did not perform well in modeling vapor pressure by themselves, they will be used in model development since they did contribute significantly to most of the models.

Regression analyses of the combined set of 476 chemicals showed similar results for estimating $\log_{10}(p_{vap})$ as analysis of the training set. Using only the topostructural indices, stepwise regression analysis resulted in a five parameter model to estimate vapor pressure:

$$\log_{10}(p_{vap}) = 4.88 + 0.20(\mathrm{O}) - 2.56(^1\chi) + 0.49(^4\chi_C) + 0.79(^6\chi_C) + 0.98(P_{10}) \quad (1)$$

$$n = 476,\quad R^2 = 51.5\%,\quad s = 0.53,\quad F = 99.7$$

Stepwise regression using the five topostructural parameters and all topochemical parameters resulted in the selection of the following seven parameter model:

$$\log_{10}(p_{vap}) = 8.44 - 1.77(^1\chi) + 1.25(P_{10}) - 5.69(IC_1) + 3.91(IC_2) - 1.24(IC_5) + 1.41(^3\chi^b{}_C) - 1.70(^1\chi^v) \quad (2)$$

$$n = 476,\quad R^2 = 79.3\%,\quad s = 0.34,\quad F = 224.0$$

Only two of the topostructural indices used in eq 1 were retained by the stepwise regression procedure used to produce eq 2: $^1\chi$ and P_{10}. The improvement in R^2 was significant, increasing from 51.5% for eq 1 to 79.3% for eq 2. Also, the model error decreased significantly, dropping by 0.19 logarithmic units. Since we have dropped the geometrical indices, this becomes our final hierarchical model.

The stepwise regression analysis of only topochemical parameters resulted in a 12 parameter model:

$$\log_{10}(p_{vap}) = 6.65 - 3.44(IC_0) - 1.33(IC_5) + 3.47(SIC_2) + 0.87(CIC_1) - 0.48(^4\chi^b) + 1.44(^3\chi^b{}_C) - 1.00(^1\chi^v) - 0.41(^3\chi^v) - 0.70(^5\chi^v) - 1.08(^3\chi^v{}_C) + 1.42(^6\chi^v{}_{Ch}) - 1.23(J^Y) \quad (3)$$

$$n = 476,\quad R^2 = 75.8\%,\quad s = 0.38,\quad F = 120.5$$

This model which is inferior to the topostructural + topochemical model (eq 2), because its variance explained is lower and, more importantly, it requires more independent variables (parameters) to achieve this explanation of variance.

Stepwise regression of all indices resulted in the selection of an 11 parameter model. This approach selected three topostructural indices and eight topochemical indices to arrive at the following model:

$$\log_{10}(p_{vap}) = 7.85 - 2.56(H^V) + 1.17(^6\chi_C) - 5.01(IC_1) + 3.65(IC_2) - 0.99(IC_5) + 0.51(CIC_1) - 1.54(^1\chi^v) - 0.36(^3\chi^v) - 0.36(^4\chi^v) - 1.40(^6\chi^v{}_C) \quad (4)$$

$$n = 476,\quad R^2 = 80.4\%,\quad s = 0.33,\quad F = 173.4$$

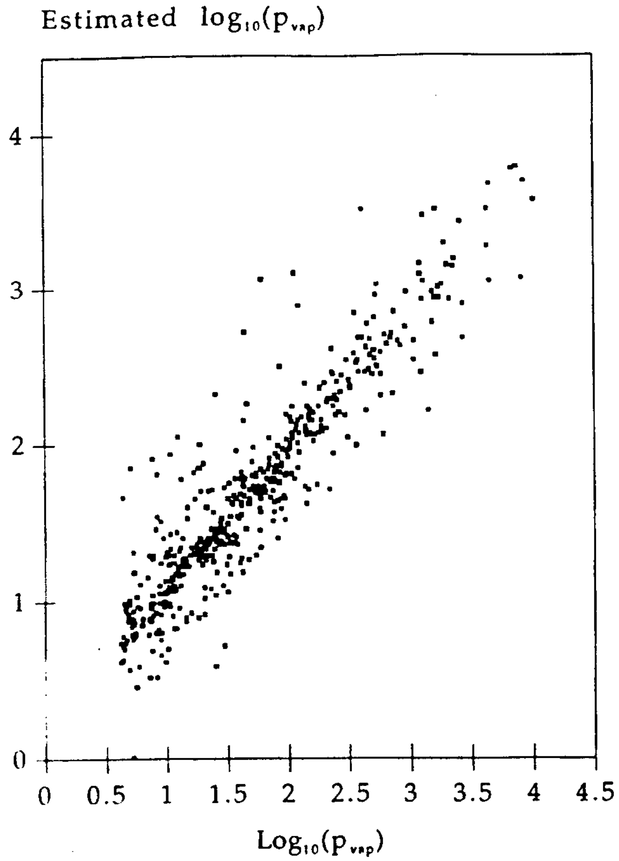


Figure 1. Scatterplot of observed $\log_{10}(p_{vap})$ *vs* estimated $\log_{10}(p_{vap})$ using eq 5 for 476 diverse compounds.

While eq 4 shows some slight improvements over eq 2, the hierarchical model, eq 2 is preferred since it is a simpler model using seven indices instead of 11 and based on a comparison of *F* values it is a more robust model than that in eq 4

Finally, we conducted the stepwise regression modeling using all topostructural and topochemical indices with HB_1 and *u* for the complete set of 476 chemicals. The resulting ten parameter model used three topostructural indices, six topochemical indices, and HB_1:

$$\log_{10}(p_{vap}) = 9.67 - 3.66(^{1}\chi) + 0.35(P_3) + 0.74(P_9) - 1.78(IC_0) - 3.33(SIC_1) - 0.81(CIC_2) + 2.05(^{2}\chi^{b}) - 1.73(^{2}\chi^{v}) - 0.79(^{3}\chi^{v}) - 0.29(HB_1) \quad (5)$$

$$n = 476, \quad R^2 = 84.3\%, \quad s = 0.29, \quad F = 249.5$$

Equation 5 shows marked improvement over eq 2, justifying the addition of indices to the model. Also, it meets the criteria on which eq 4 was judged to be lacking. Overall, there is an improvement in variance explained of 5%, with a comparable decrease in standard deviation. A scatter plot of observed $\log_{10}(p_{vap})$ versus estimated $\log_{10}(p_{vap})$ using eq 5 is presented in Figure 1.

4. DISCUSSION

The purpose of this paper was 2-fold: (a) to study the utility of algorithmically-derived molecular descriptors in developing QSAR models for predicting the vapor pressure of chemicals from structure and b) to investigate the relative roles of topostructural, topochemical, and geometrical indices in the estimation of standard vapor pressure.

Table 4. Summary of the Chemical Class Composition of the Normal Vapor Pressure Dataset

compd classification	no. of compds	pure	substituted
total normal vapor pressure dataset	476		
hydrocarbons	253		
non-hydrocarbons[a]	223		
nitro compounds	4	3	1
amines	20	17	3
nitriles	7	6	1
ketones	7	7	0
halogens	100	95	5
anhydrides	1	1	0
esters	18	16	2
carboxylic acids	2	2	0
alcohols	10	6	4
sulfides	39	38	1
thiols	4	4	0
imines	2	2	0
epoxides	1	1	0
aromatic compounds[b]	15	10	4
fused-ring compounds[c]	1	1	0

[a] The non-hydrocarbons are further broken down into the following groups. [b] The 15 aromatic compounds are a mixture of 11 aromatic hydrocarbons and four aromatic halides. [c] The only fused-ring compound was a polycyclic aromatic hydrocarbon.

Results described in this paper (eqs 1−5) show that nonempirical parameters derived predominantly from graph theoretic models of molecules can estimate normal vapor pressure of diverse chemicals reasonably well. The explained variance of data ($R^2 = 84.3\%$) is excellent in view of the fact that the database of chemicals analyzed in this paper is very diverse (see Table 4). It should be mentioned that most published QSAR models for the estimation of vapor pressure have dealt with much smaller data sets with limited structural variety.[42,43]

The relative effectiveness of topostructural, topochemical, and geometrical indices in predicting normal vapor pressure of chemicals is evident from the result presented above. Equation 1 explains over 51% of variance in the data. All parameters used to derive eq 1 are topostructural, *i.e.*, they are parameters which encode information about the adjacency and distance of vertices in skeletal molecular graphs without quantifying any explicit information about such chemical aspects like bond order, electronic character of atoms, etc. Yet, the high explained variance of the property indicates that adjacency and distance in chemical graphs, being general descriptors of molecular size, shape, and branching, are important in predicting properties. This may explain the success of parameters like simple connectivity indices in estimating many diverse properties.[1]

Equation 3 is derived only from topochemical indices. The explained variance of vapor pressure (75.8%) shows that topochemical parameters, as a class, explain a larger fraction of the variance as compared to models derived from only topostructural indices (eq 1). Geometrical parameters were dropped from the set of descriptors after their limited success in prediction for the training and test sets. This is in line with our earlier studies with normal boiling point and hydrophobicity, where it was reported that the addition of geometrical indices could not significantly improve the predictive power of QSAR models derived from a combined set of topostructural and topochemical parameters.[15] It would be interesting to see whether this pattern holds good for other properties as well. Finally, the addition of the simple nonempirical parameter, HB_1, which contains information relevant to intermolecular interactions further improves our ability to estimate normal vapor pressure resulting in an explained variance of 84.3% (eq 5).

ACKNOWLEDGMENT

This is contribution number 209 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, a grant from Exxon Corporation and the Structure−Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to dedicate this paper to Professor Milan Randić in appreciation of his contributions in chemical information, quantitative structure−activity relationships, and chemical graph theory.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986

(2) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res* **1987**, *15*, 605−609.

(3) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. Correlation Between Structure and Normal Boiling Points of Haloalkanes C_1-C_4 Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1118−1121

(4) Basak, S. C. A Nonempirical Approach to Predicting Molecular Properties Using Graph-Theoretic Invariants. In *Practical Applications of Quantitative Structure−Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht/Boston/London, 1990. pp 83−103.

(5) Basak, S. C.; Bertelsen, S.; Grunwald, G. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270−276

(6) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett* **1995**, *79*, 239−250

(7) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathematical Modelling and Scientific Computing*. In press.

(8) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR and QSAR in Environ. Res.* **1994**, 2, 289−307.

(9) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity From Molecular Structural Similarity. *New J. Chem.* **1995**, *19*, 231−237

(10) 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 Geometry*, Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73−116

(11) Basak, S. C.; Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: 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

(12) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood-Brain Transport of Drugs: A Computational Approach. *Pharm Res* **1996**, *13*, 775−778.

(13) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math Modelling Sci. Computing* In press.

(14) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Estimation of Normal Boiling Points of Haloalkanes Using Molecular Similarity. *Croat Chem Acta* **1996**, *69*, 1159−1173.

(15) 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.

(16) Martin, Y. C. Opportunities for Computational Chemists Afforded by the New Strategies in Drug Discovery: An Opinion. *Network Science* **1996**.

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

(18) NRC. *Toxicity Testing: Strategies to Determine Needs and Priorities*; National Academy Press: Washington, DC, 1984; p 84.

(19) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* **1991**, *7*, 243−272.

(20) Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. ASTER: An Integration of the AQUIRE Data Base and the QSAR System for Use in Ecological Risk Assessments. *Sci. Total Environ.* **1991**, *109/110*, 667−670.

(21) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY Version 2.3*; Copyright of the University of Minnesota, 1988.

(22) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(23) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(24) 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.

(25) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.

(26) 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.

(27) 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: Rolla, MO, 1980; p 745.

(28) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon: New York, 1984; p 745.

(29) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(30) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199−206.

(31) 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.

(32) Bondi, A. van der Waal's Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(33) Moriguchi, I.; Kanada, Y. Use of van der Waal's Volume in Structure−Activity Studies. *Chem. Pharm. Bull.* **1977**, *25*, 926−935.

(34) Moriguchi, I.; Kanada, Y.; Komatsu, K. van der Waal's Volume and the Related Parameters for Hydrophobicity in Structure−Activity Studies. *Chem. Pharm. Bull.* **1976**, *24*, 1799−1806.

(35) Tripos Associates, Inc. *SYBYL Version 6.2*; Tripos Associates, Inc.: St. Louis, MO, 1994.

(36) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the Three-Dimensional Wiener Number. *J. Math. Chem.* **1989**, *3*, 299−309.

(37) Tripos Associates, Inc. *CONCORD Version 3.2.7*; Tripos Associates, Inc.: St. Louis, MO, 1995.

(38) Basak, S. C. *H-Bond*; Copyright of the University of Minnesota, 1988.

(39) Ou, Y. C.; Ouyang, Y.; Lien, E. J. *J. Mol. Sci.* **1986**, *4*, 89.

(40) Topliss, J. G.; Edwards, R. P. Chance Factor in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(41) 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.

(42) Drefahl, A.; Reinhard, M. *Handbook for Estimating Physico-Chemical Properties of Organic Compounds*; Stanford University Bookstore, Stanford, CA, 1995.

(43) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*; McGraw-Hill Book Company: New York, 1982.

CI960176D