Progress Report
of the Air Force Project

Covering research period 10/1/97 to 7/31/98

Agency No: DOD/F49620-98-1-0015
U of M No: 1613-189-6158
NRRI Technical Report No: NRRI/TR-98/14

Prediction of Health and Environmental Hazards of Chemical: A Hierarchical Approach Using QMSA and QSAR

August 13, 1998

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)720-9412
Email: sbasak@wyle.nrri.umn.edu

Table of Contents

II Objectives

During the past few years we have been involved in the development of new computational methods for quantifying similarity/dissimilarity of chemicals and applications of quantitative molecular similarity analysis (QMSA) techniques in analog selection and property estimation for use in the hazard assessment of chemicals. We have also explored the mathematical nature of the molecular similarity space in order to better understand the basis of analog selection by QMSA methods. The parameter spaces used for QMSA and analog selection were constructed from nonempirical parameters derived from computational chemical graph theory. Occasionally, graph invariants were supplemented with geometrical parameters and quantum chemical indices to study the relative effectiveness of graph invariants vis-à-vis geometrical and quantum chemical parameters in analog selection and property estimation. We carried out comparative studies of nonempirical descriptor spaces and physicochemical property spaces in selecting analogs. Molecular similarity methods were applied in predicting modes of toxic action (MOA) of chemicals. Our similarity/dissimilarity methods have also found successful applications in the discovery of new drug leads by US drug companies.

In this project, we will have four primary goals: 1) development of a hierarchical approach to molecular similarity, 2) formulation of quantitative structure-activity relationship (QSAR) models for predictive toxicology using a hierarchical approach, 3) applications of hierarchical QSAR and QMSA approaches in computational toxicology related to human health and ecological hazard assessment, and 4) the application of hierarchical QMSA and QSAR approaches in estimating potential toxicity of deicing agents.

The first goal of the project is the use of parameters of gradually increasing complexity, viz., topological, topochemical, geometrical, and quantum chemical indices, in the quantification of molecular similarity/dissimilarity of chemicals. We will take a two-tier approach in this area. First, similarity methods will be used in ordering sets of molecules and in selecting structural analogs of toxic chemicals which pose human health and ecological hazards. Secondly, we will use the properties of selected analogs in estimating toxicologically important properties for chemicals. Although different classes of parameters have been used in the characterization of molecular similarity, no systematic study has been carried out in the use of all four classes of parameters, mentioned above, in analog selection and property estimation. We will apply a hierarchical approach to the use of these four types of theoretical molecular descriptors in the quantification of molecular similarity/dissimilarity.

The second goal consists of the development of hierarchical QSAR models for predicting the toxic potential of chemicals using topological and quantum chemical indices. Initially, we will use parameters calculated by semi-empirical methods such as MOPAC and AMPAC. Parameters calculated by *ab initio* quantum chemical methods will be used in limited cases of QSAR model development, if they are considered necessary.

The third goal of the project will be the prediction of human health hazard and ecotoxicological effects of chemicals using QSAR and QMSA methods developed in the project. Attempts will be made to estimate endpoints, such as, carcinogenicity, mutagenicity, xenoestrogenicity, acute toxicity, transport of chemicals through the blood-brain barrier, biodegradation, and bioconcentration factor.

The fourth goal will involve the utilization of QMSA and QSAR methods developed as part of this project in predicting the potential toxicity of deicing agents.

III Status of Efforts

During the first year of the project the majority of effort was spent in the development of novel hierarchical QSAR methods, QMSA techniques and the applications of these methods in the prediction of toxicological, physicochemical as well as biomedicinal properties of different sets of chemicals. Our dissimilarity methods were used in the clustering of JP8 constituents into a small number of clusters which can be useful in the selection of surrogate mixtures for JP8 in toxicological studies by the Air Force. The clustering was done using algorithmically derived molecular descriptors calculated by our computer program POLLY. Such parameters can be calculated for any molecular structure, real or hypothetical. This makes the clustering methods independent of any experimentally determined property of the JP8 constituents.

IV Accomplishments/ New Findings

The following is the summary of accomplishments of the various tasks of the project during the reporting period:

Task 1: Development of Databases
Databases of toxicological endpoints and physicochemical properties have been developed from published literature. Such data have been used in the hierarchical QSAR and QMSA studies (vide infra).

Task 2: Development of a Comprehensive Computer Program for Calculating Topological Molecular Descriptors
POLLY can currently calculate more than one hundred topological indices (TIs). We are currently developing algorithms which can calculate other topological descriptors like local invariants. Such indices will be tested in hierarchical QSAR and QMSA research.

Task 3: Integration of Graph Theory and Quantum Chemistry for QSAR
Ongoing research in this area focused on the use of weighted graphs, pseudographs in the development of novel descriptors. This will lead to novel invariants which can encode information not quantified by existing molecular descriptors.

Task 6: Characterization of Structure Using Theoretical Structural Descriptors
We have used topological indices and principal components (PCs) derived from them in the characterization of a set of isospectral graphs which cannot be discriminated by the eigenvalues of the adjacence matrix of molecular graphs. This result has been published in the Journal of Chemical Information and Computer Sciences (see publication list below).

Task 7: Development of Hierarchical QMSA Models
Topostructural, topochemical, geometrical as well as quantum chemical parameters are being used in the development of QMSA methods. We carried out a dissimilarity based clustering of JP8 constituents into fourteen clusters. A mixture of compounds selected from these cluster can be used as surrogates for JP8 which is a very complex mixture. The results of the cluster analysis is given in Appendix 1.

Task 8: Development of Hierarchical Approach to QSAR
We have used quantum chemical parameters calculated by semiempirical methods in hierarchical

QSAR models for predicting toxicity and toxicologically relevant physicochemical properties. Several manuscripts have been published/ submitted for publication in peer-reviewed journals.

V Personnel supported

Subhash C. Basak, Principal Investigator
Greg Grunwald, Applications Programmer
Doug Dilla, Undergraduate student
Jassen Dagit, Undergraduate student

VI Publications

The following peer-reviewed papers, which are currently either published, in press, or submitted, report results of research carried out between August 1, 1997 and July 31, 1998.

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, 1998.

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

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*, submitted, 1998.

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*, JAI Press, submitted, 1998.

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.*, submitted, 1998.

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.*, in press, 1998.

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.*, in press, 1998.

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. Schuurmann, eds., SETAC Press, Pensacola, FL, 1998, Chapter 17, p 245-261.

Predicting acute toxicity (LC50) 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.

Characterization of isospectral graphs using graph invariants and derived orthogonal parameters, K. Balasubramanian and S. C. Basak, *J. Chem. Inf. Comput. Sci.*,38, 367, 1998.

Optimal molecular descriptors based on weighted path numbers, M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci., submitted.*

VII Interactions/ Transitions

A. Transitions

1. Applied computational methods in the design of 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 Dr. Alexander Gybin, The Chormaline Corporation, Duluth, MN in the computer-assisted design of photoactive chemicals

B. Meetings/ Seminars/ Invited Presentations

1. 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.

2. Dr. S.C. Basak gave several invited lectures at various national and international symposia during his stay in India from December 23, 1997 through January 31, 1998. These lectures included:

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.

3. Dr. S.C. Basak chaired a session at the DIMACS Workshop on Discrete Mathematical Chemistry, March 23-25, 1998, held at Rutgers 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.

4. 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.

5. 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

6. 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.

7. 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.

8. 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.

9. Dr. S.C. Basak attended the First Environmental Management Science Program Workshop organized jointly by the American Chemical Society and the Office of Environmental Management, Department of Energy, Chicago, IL, July 27-30, 1998.

C. Honors and Awards

1. 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.

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

VIII New Discoveries/ Inventions, Patent Disclosures

A. We fond that constituents of complex of mixtures like JP8 can be clustered into different structural groups using structure spaces derived from topological indices calculated by POLLY

B. 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.

c) For various sets of toxicological and physicochemical properties the topostructural and topochemical parameters explain most of the variance in the data; the addition of geometrical and quantum chemical parameters to the set of independent variables did small or no improvement in the predicting power of models.

IX Appendices

Appendix 1. Clustering of JP8 Constituents

Appendix 2. Publications

Appendix 2.1 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.*, in Press, 1998.

Appendix 2.2 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.*, in Press, 1998.

Appendix 2.3 Assessment of the Mutagenicity of Chemicals from Theoretical Structural Parameters: a Hierarchical Approach, S.C. Basak, B.D. Gute, and G.D. Grunwald, *Sar QSAR Environ. Res.*, Submitted, 1998.

Appendix 2.4 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, 1998.

Appendix 2.5 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*, Submitted, 1998.

Appendix 2.6 Characterizations of The Molecular Similarity of Chemicals Using Topological Invariants, S.C. Basak, B.D. Gute, and G.D. Grunwald, In: *Advance in Molecular Similarity – Volume 2,* R. Carbo-dorca and P.g. Mezey, Eds., Jai Press, Submitted, 1998.

Appendix 2.7 Characterization of Molecular Structures Using Topological Indices, S.C. Basak and B.D. Gute, *Sar. QSAR Environ. Res..*, 7, 1-21, 1997.

Appendix 2.8 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 2.9 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 2.10 Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters, K Balasubramanian and Sc Basak, *J. Chem. Inf. Comput. Sci.*,38, 367, 1998.

Appendix. 2.11 Optimal Molecular Descriptors Based on Weighted Path Numbers, M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci., submitted.*

Seq	Name	Avg_Conc	Sd_Conc	Cluster
1	ISTD (d10-anthracene)	0.000	0	1
24	9-methylanthracene ← centroid	.100	0	1
26	pyrene	.100	0	1
212	1,8-dimethylnaphthalene	2.200	3	1
220	1,4,6,7-tetramethylnaphthalene	3.900	5	1
37	iH-indene	13.700	27	1
218	fluorene	27.500	33	1
205	1-ethylnaphthalene	119.000	89	1
208	1,4-dimethylnaphthalene	191.000	140	1
209	1,5-dimethylnaphthalene	194.000	140	1
210	1,2-dimethylnaphthalene	204.000	160	1
91	cyclohexylbenzene	387.000	280	1
204	2-ethylnaphthalene	455.000	240	1
201	1,1'-biphenyl	542.000	330	1
178	naphthalene	2140.000	1400	1
89	2-methylnaphthalene	2980.000	1500	1
190	1-methylnaphthalene	3590.000	1800	1
93	1,3,5,5-tetramethyl-1,3-cyclohexadiene	1.300	2	2
21	4-methylcyclohexene	6.700	6	2
177	1-t-butyl-4-ethylbenzene	6.800	6	2
54	3,3,5-trimethylcyclohexene	9.000	8	2
82	2,3,3- or 3,4,4-trimethylcyclohexene	11.700	10	2
229	2,3,3- or 3,4,4-trimethylcyclohexene	11.700	10	2
94	t-1,1,3,5-tetramethylcyclohexane	15.800	21	2
176	1-t-butyl-3,5-dimethylbenzene	17.200	22	2
192	1-t-butyl-3,4,5-trimethylbenzene	20.200	28	2
54	1-t-butyl-3-methylbenzene	39.200	41	2
63	3,5,5-trimethylcyclohexene	65.800	95	2
120	1-isopropyl-4-methylcyclohexane	89.400	230	2
59	c,c,t-1,3,5-trimethylcyclohexane	91.800	170	2
83	c,c,t-1,3,5-trimethylcyclohexane	91.800	170	2
58	c,c,c-1,3,5-trimethylcyclohexane	113.000	180	2
90	c,c,c-1,2,3=trimethylcyclohexane	113.000	86	2
62	1,3,5-trimethylcyclohexane	115.000	180	2
97	1,1,6-trimethyltetralin	122.000	120	2
109	1,2,3,5-tetramethylcyclohexane	136.000	180	2
89	c,c,t-1,2,3-trimethylcyclohexane	175.000	130	2
43	t-1,3-dimethylcyclohexane	316.000	370	2
42	c-1,4-dimethylcyclohexane	327.000	380	2
55	c-1,2-dimethylcyclohexane	440.000	510	2
68	1-t-butyl-2-methylbenzene	467.000	320	2
92	1-ethyl-1-methylcyclohexane	840.000	760	2
67	1,1,4-trimethylcyclohexane	1120.000	1900	2

Seq	Name	Avg_Conc	Sd_Conc	Cluster
111	2,3-dimethyloctane	1220.000	700	2
35	c-1,3-dimethylcyclohexane	1630.000	980	2
31	1,2,4-trimethylcyclohexane	1860.000	2400	2
79	c-1,2,3-trimethylcyclohexane	2050.000	2900	2
183	2,2,3-trimethyldecane	2670.000	3500	2
98	isopropylcyclohexane ← centroid	2740.000	1200	2
84	1-ethyl-3-methylcyclohexane	2940.000	1600	2
102	n-propylcyclohexane	4030.000	1400	2
64	ethylcyclohexane	4150.000	2400	2
03	2,6-dimethyloctane	5960.000	2600	2
134	butylcyclohexane	6880.000	2900	2
07	3,4,5-trimethylheptane	3.600	5	3
91	3,3,5-trimethylheptane	4.900	4	3
7	3-ethylpentane	19.000	17	3
23	2,3,4-trimethylpentane	20.500	19	3
30	3,4-dimethylhexane	46.500	49	3
108	3-ethyl-3-methylheptane ← centroid	51.200	61	3
68	2,3,4-trimethylhexane	114.000	130	3
19	2,4-dimethylhexane	126.000	130	3
66	2,4-dimethyl-3-ethylpentane	147.000	190	3
78	2,4,6-trimethylheptane	173.000	460	3
04	3,4-diethylhexane	302.000	210	3
50	2,4-dimethylheptane	356.000	280	3
72	3,4-dimethylheptane	536.000	1600	3
33	3-ethylhexane	620.000	440	3
147	3-ethylnonane	806.000	300	3
112	4-ethyloctane	1180.000	670	3
17	3-ethyloctane	1270.000	1600	3
65	3,3- or 2,5-dimethylheptane	1570.000	1200	3
71	2,3-dimethylheptane	1570.000	910	3
170	1,2,3,4-tetramethylbenzene	1810.000	920	3
75	4-methyloctane	1980.000	1100	3
166	1,2,3,5-tetramethylbenzene	2040.000	880	3
99	3,5-dimethyloctane	2900.000	3500	3
06	3-ethyl-2-methylheptane	3260.000	1800	3
69	2,2,3,3-tetramethylpentane	12.600	22	4
5	benzene	270.000	650	5
13	2,2-dimethylhexane	11.700	18	6

Seq	Name	Avg_Conc	Sd_Conc	Cluster
123	2,2,4,6,6-pentamethylheptane	12.100	18	6
146	neopentylbenzene	14.100	12	6
96	2,2,6,6-tetramethylheptane	77.800	100	6
56	2,2,5,5-tetramethylhexane	474.000	430	6
158	t-pentylbenzene	483.000	360	6
95	2,2-dimethyloctane	542.000	470	6
222	hexaethylbenzene	.200	0	7
94	1,3,5-triisopropylbenzene	.400	1	7
174	1,4-diisopropylbenzene	14.400	11	7
187	(1,1-diethylpropyl-)-benzene	22.100	30	7
69	1,2-diisopropylbenzene	22.200	21	7
14	5,5-dibutylnonane	76.300	65	7
173	4,4-dipropylheptane	338.000	320	7
184	1,3,5-triethylbenzene	541.000	420	7
63	1,2,4,5-tetramethylbenzene	1950.000	900	7
18	2,5-dimethylhexane	6.600	6	8
81	(2-methylpentyl-)benzene	23.800	31	8
48	2-octene	27.500	24	8
36	2-ethyl-1-hexene	29.700	45	8
87	4-nonene	56.400	110	8
10	1-heptene	115.000	140	8
165	(2-methylbutyl-)-benzene	131.000	120	8
107	2,3-dimethylnaphthalene	145.000	110	8
167	(3-methylbutyl-)-benzene	242.000	150	8
161	sec-pentylbenzene	253.000	300	8
52	(1,2-dimethylpropyl-)-benzene	260.000	190	8
73	4-ethylheptane	261.000	420	8
136	1-isopropyl-2-methylbenzene	261.000	160	8
6	2-methylhexane	285.000	380	8
60	1-ethyl-3-isopropylbenzene	352.000	210	8
155	(1-ethylpropyl-)-benzene	399.000	210	8
127	isobutylbenzene	458.000	230	8
32	3-methylheptane	502.000	480	8
157	2-ethyl-1,3-dimethylbenzene	505.000	230	8
185	n-hexylbenzene	550.000	260	8
44	1,2-diethylbenzene	576.000	350	8
71	n-pentylbenzene	621.000	320	8
140	1,4-diethylbenzene	648.000	280	8
85	1-nonene	712.000	480	8
28	sec-butylbenzene	832.000	380	8
206	2,6-dimethylnaphthalene	923.000	610	8
28	4-methylheptane	1050.000	980	8

Seq	Name	Avg_Conc	Sd_Conc	Cluster
141	butylbenzene	1080.000	420	8
11	3-heptene	1170.000	960	8
39	1-propyl-4-methylbenzene	1200.000	410	8
30	1-isopropyl-3-methylbenzene	1250.000	490	8
74	ethylbenzene	1270.000	700	8
57	2,6-dimethylheptane	1330.000	840	8
62	1,2-dimethyl-3-ethylbenzene	1400.000	530	8
138	1,3-diethylbenzene	1430.000	500	8
100	isopropylbenzene ← centroid	1450.000	850	8
27	2-methylheptane	1490.000	820	8
124	t-butylbenzene	1520.000	610	8
227	3,3- or 2,5-dimethylheptane	1570.000	1200	8
12	n-heptane	1700.000	1400	8
32	1-ethyl-2,5-dimethylbenzene	1720.000	670	8
29	toluene	1750.000	1200	8
101	2,7-dimethyloctane	1840.000	940	8
13	5-methylnonane	1940.000	780	8
110	propylbenzene	1950.000	890	8
150	1-isopropyl-4-methylbenzene	2100.000	790	8
53	1-ethyl-3,4-dimethylbenzene	2140.000	790	8
48	1-propyl-2-methylbenzene	2570.000	1000	8
151	1-ethyl-2,4-dimethylbenzene	2680.000	1100	8
42	1-ethyl-3,5-dimethylbenzene	2730.000	1000	8
21	1-ethyl-2-methylbenzene	2750.000	1100	8
105	3,6-dimethyloctane	2910.000	1700	8
116	1-ethyl-4-methylbenzene	3150.000	1800	8
86	o-xylene	3450.000	1800	8
115	1-ethyl-3-methylbenzene	3590.000	1300	8
135	indane (2,3-dihydro-1H-indene)	4050.000	1500	8
77	3-methyloctane	4160.000	1900	8
18	1,3,5-trimethylbenzene	4270.000	1600	8
40	n-octane	5050.000	3000	8
131	1,2,3-trimethylbenzene	5660.000	1900	8
76	m & p-xylenes (co-eluting)	6050.000	3600	8
228	m- & p-xylenes (co-eluting)	6050.000	3600	8
125	1,2,4-trimethylbenzene	12300.000	4300	8
17	3,3,5,5-tetramethylcyclopentene	.700	1	9
26	1,1,3,3-tetramethylcyclopentane	3.800	5	9
37	2,2,4-trimethylhexane	5.500	4	9
16	2,4,4-trimethyl-2-pentene	6.900	6	9
9	iso-octane ← centroid	9.800	8	9
4	3,3-dimethylpentane	21.400	35	9
49	(c&t)-2,2,4-trimethyl-3-hexene	32.200	24	9
44	2,4,4-trimethylhexane	36.800	51	9

Seq	Name	Avg_Conc	Sd_Conc	Cluster
20	3,3-dimethylhexane	51.300	48	9
97	4,4-dimethyloctane	115.000	100	9
63	4,4-dimethylheptane	141.000	240	9
50	2,4,4-trimethyl-1-hexene	143.000	530	9
31	2,2,4,4-tetramethylpentane	353.000	310	9
30	3,3-diethylpentane	703.000	1200	9
193	cyclododecane	39.500	52	10
25	t-3,4,4-trimethyl-2-pentene	3.900	4	11
15	2,3,3-trimethyl-1,4-pentadiene	4.200	5	11
24	2,3,3-trimethylpentane ← centroid	9.200	8	11
2	2,2,3-trimethylbutane	12.400	14	11
52	2,3,3-trimethyl-1-hexene	18.200	52	11
51	2,2,3-trimethylhexane	42.500	70	11
3	2,3,3-trimethyl-1-butene	96.000	79	11
70	2,3,3,4-tetramethylpentane	1370.000	1300	11
47	1,2,4,4-tetramethylcyclopentene	1.600	3	12
45	1,2,3-trimethylcyclopentene	3.000	3	12
33	dicyclopentadiene	4.300	2	12
46	c-1,1,3,4-tetramethylcyclopentane	6.800	5	12
38	1-ethyl-1-methylcyclopentane ← centroid	20.600	17	12
34	t-1,1,3,4-tetramethylcyclopentane	35.300	46	12
8	t-1,3-dimethylcyclopentane	44.600	52	12
41	c,c,c-1,2,3-trimethylcyclopentane	50.700	170	12
39	c,t,c-1,2,3,4-tetramethylcyclopentane	104.000	110	12
61	propylcyclopentane	118.000	180	12
22	c,t,c-1,2,3-trimethylcyclopentane	192.000	190	12
14	1,1,3-trimethylcyclopentane	2210.000	1500	12
129	3,7,7-trimethylbicyclo(4.1.0)-3-heptene	10.700	8	13
225	1-phenyltridecane	1.000	1	14
219	n-decylbenzene	5.300	7	14
223	2,6,11,15-tetramethylhexadecane	8.800	12	14
216	1-hexadecene	25.000	18	14
215	n-nonylbenzene	25.700	29	14
202	1-tetradecene	36.700	25	14
211	n-octylbenzene	109.000	80	14
221	2,6,10,14-tetramethylpentadecane(pristan	180.000	160	14

Seq	Name	Avg_Conc	Sd_Conc	Cluster
122	1-decene	259.000	270	14
198	n-heptylbenzene	273.000	160	14
186	1-tridecene	303.000	290	14
196	3,7,11-trimethyl-1-dodecene	372.000	290	14
179	1-dodecene	473.000	750	14
156	1-undecene	490.000	450	14
195	heptylcyclohexane	1660.000	680	14
199	2,6,11-trimethyldodecane	1860.000	1100	14
217	n-hexadecane	2390.000	1100	14
164	2,6-dimethyldecane	4230.000	2500	14
200	2,6,10-trimethyldodecane	5020.000	2300	14
119	3-methylnonane	6050.000	2000	14
143	4-methyldecane	6100.000	2000	14
114	4-methylnonane	6410.000	2500	14
213	n-pentadecane	7170.000	2500	14
172	2-methylundecane ← centroid	8220.000	4400	14
149	3-methyldecane	8470.000	3000	14
145	2-methyldecane	9020.000	3900	14
182	2,6-dimethylundecane	10700.000	3900	14
203	n-tetradecane	16800.000	6500	14
88	n-nonane	18400.000	8200	14
188	n-tridecane	26000.000	11000	14
180	n-dodecane	27200.000	14000	14
159	n-undecane	28800.000	14000	14
126	n-decane	33700.000	14000	14

J. Chem. Inf. Comput. Sci., submitted

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

Subhash C. Basak , Brian D. Gute and Shibnath Ghatak[1]

Natural Resources Research Institute, The University of Minnesota
5013 Miller Trunk Highway, Duluth, Minnesota 55811, U.S.A.

[1] Department of Biology, Dana Laboratory, Tufts University
Medford, MA 02155, U.S.A.

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

Office: (218) 720-4230
E-mail: sbasak@wyle.nrri.umn.edu

ABSTRACT

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 step-wise fashion to build increasingly more complex models. The results show that topostructural indices alone, specifically ^{D}I, 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 3-dimensional

whether computable parameters such as TIs and geometric indices, can give 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-a-vis calculated geometric parameters in predicting the complement-inhibitory potencies of this set of benzamidines.

2. METHODS

2.1 Database. The 107 benzamidines used in this study are those presented in the work of Hansch and Yoshimoto.[47] This data was 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 percent 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.

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.[65] The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional.

(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:

$$1/\log C = -0.6428 + 0.0490(^{3D}W) \qquad \text{Eq.2}$$

(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.

[Insert Figure 2 here]

4. DISCUSSION

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.

ACKNOWLEDGEMENTS

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.

(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: Quant. Biol. Symp.* **1984**, *11*, 137-153.

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

(37) 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.

(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. Molec. 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 Walls' 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. and McConnell, I., Eds.; Blackwell Scientific Publications: Philadelphia, 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

(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., and 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, pp 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*

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*-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

* This SO_2F group should be *meta*- instead of *para*-.

Table 2. Side-chain structures and biological property data for 107 benzamidines.

		1 / Log C		
No.	X	Obsd.	Predict.[a]	Resid.
1	3,5-$(OCH_3)_2$	-0.452	-0.367[b]	-0.085
2	2-CH_3	-0.444	-0.405	-0.040
3	3,4-$(CH_3)_2$	-0.425	-0.389	-0.036
4	H	-0.418	-0.417	-0.002
5	3-OH	-0.415	-0.402	-0.012
6	3-$NHCO(CH_2)_2C_6H_5$	-0.412	-0.302[b]	-0.110
7	3-CF_3	-0.410	-0.369	-0.041
8	3-NO_2	-0.410	-0.378	-0.032
9	3-Br	-0.405	-0.401	-0.004
10	3-CH_3	-0.398	-0.402	0.004
11	3-OCH_3	-0.397	-0.389	-0.008
12	3-$CH_2C_6H_5$	-0.373	-0.339	-0.034
13	3,5-$(CH_3)_2$	-0.361	-0.389	0.028
14	3-OC_3H_7	-0.355	-0.362	0.007
15	3-*i*-C_5H_{11}	-0.355	-0.353	-0.002
16	3-OC_4H_9	-0.351	-0.349	-0.001
17	3-C_4H_9	-0.338	-0.362	0.024
18	3-$CH{=}CHC_6H_5$	-0.339	-0.325	-0.014
19	3-$OCH_2C_6H_5$	-0.331	-0.326	-0.005
20	3-$(CH_2)_2C_6H_5$	-0.330	-0.326	-0.004
21	3-OC_6H_{13}	-0.329	-0.327	-0.002
22	3-$O(CH_2)_4OC_6H_5$	-0.325	-0.288	-0.037
23	3-$O(CH_2)_2OC_6H_5$	-0.323	-0.306	-0.017
24	3-C_6H_5	-0.323	-0.347	0.025
25	3-$O(CH_2)_3OC_6H_4$-4-COOH	-0.321	-0.277	-0.044
26	3-OC_5H_{11}	-0.320	-0.338	0.017
27	3-O-*i*-C_5H_{11}	-0.318	-0.341	0.022
28	3-$O(CH_2)_2OC_{10}H_7$-α	-0.312	-0.283	-0.030
29	3-$O(CH_2)_4OC_6H_4$-4-NH_2	-0.306	-0.282	-0.024
30	3-$(CH_2)_4C_6H_5$	-0.302	-0.306	0.004
31	3-$O(CH_2)_3OC_6H_4$-4-NO_2	-0.301	-0.277	-0.024
32	3-$O(CH_2)_3OC_6H_4$-4-NH_2	-0.300	-0.290	-0.010
33	3-$(CH_2)_2$-4-C_5H_4N	-0.299	-0.326	0.026
34	3-$O(CH_2)_3OC_6H_5$	-0.299	-0.297	-0.003

35	3-O$(CH_2)_3C_6H_5$	-0.296	-0.306	0.010
36	3-$(CH_2)_2$-3-C_5H_4N	-0.294	-0.326	0.032
37	3-$(CH_2)_4C_6H_4$-4-NHAc	-0.294	-0.273	-0.021
38	3-$(CH_2)_2$-2-C_5H_4N	-0.291	-0.326	0.035
39	3-O$(CH_2)_3OC_6H_4$-2-NH_2	-0.283	-0.291	0.009
40	3-O$(CH_2)_3OC_6H_4$-4-NHAc	-0.278	-0.265	-0.012
41	3-$(CH_2)_4$-3-C_5H_4N	-0.276	-0.306	0.030
42	3-O$(CH_2)_4C_6H_5$	-0.276	-0.297	0.020
43	3-O$(CH_2)_3OC_6H_4$-3-NHAc	-0.270	-0.267	-0.003
44	3-O$(CH_2)_3OC_6H_3$-3,4-Cl_2	-0.265	-0.283	0.018
45	3-O$(CH_2)_3OC_6H_4$-3-NH_2	-0.265	-0.290	0.025
46	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_4-4-SO_2F	-0.265	-0.237	-0.028
47	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_5	-0.265	-0.253	-0.012
48	3-O$(CH_2)_3OC_6H_4$-4-OCH_3	-0.262	-0.283	0.022
49	3-O$(CH_2)_4OC_6H_4$-4-NHCONHC_6H_4-4-SO_2F	-0.260	-0.219	-0.040
50	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_3-2-OCH_3-5-SO_2F	-0.260	-0.233	-0.027
51	3-O$(CH_2)_3OC_6H_4$-4-Cl	-0.257	-0.290	0.033
52	3-O$(CH_2)_3OC_6H_4$-2-NO_2	-0.257	-0.281	0.024
53	3-O$(CH_2)_3OC_6H_4$-3-NO_2	-0.257	-0.278	0.021
54	3-O$(CH_2)_3OC_6H_4$-3-OCH_3	-0.256	-0.283	0.027
55	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_3-2-Cl-6-SO_2F	-0.255	-0.237	-0.018
56	3-O$(CH_2)_3OC_6H_4$-2-NHCONHC_6H_5	-0.255	-0.249	-0.006
57	3-O$(CH_2)_2OC_6H_4$-2-NHCONHC_6H_3-2-Cl-5-SO_2F	-0.250	-0.236	-0.014
58	3-O$(CH_2)_3OC_6H_4$-2-NHCONH$CH_2C_6H_4$-4-SO_2F	-0.250	-0.228	-0.022
59	3-O$(CH_2)_3OC_6H_4$-2-NHCONH-C_6H_2-2,4-$(CH_3)_2$-5-SO_2F	-0.248	-0.229	-0.019
60	3-O$(CH_2)_3OC_6H_4$-4-$COOCH_3$	-0.247	-0.271	0.025
61	3-O$(CH_2)_3OC_6H_3$-3-NO_2-4-CH_3	-0.245	-0.273	0.028
62	3-O$(CH_2)_3OC_6H_4$-3-CF_3	-0.245	-0.273	0.028
63	3-O$(CH_2)_3OC_6H_4$-2-NHCONHC_6H_4-4-CH_3-3-SO_2F	-0.245	-0.229	-0.015
64	3-O$(CH_2)_3OC_6H_4$-4-NHCOC_6H_5	-0.244	-0.246	0.002
65	3-O$(CH_2)_3OC_6H_4$-2-NHCO$CH_2OC_6H_4$-4-SO_2F	-0.244	-0.227	-0.017
66	3-O$(CH_2)_3OC_6H_4$-4-NHCOC_6H_4-4-OCH_3	-0.243	-0.236	-0.007
67	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_4-3-SO_2F	-0.243	-0.238	-0.005
68	3-O$(CH_2)_3OC_6H_4$-2-NHCO$CH_2C_6H_4$-4-SO_2F	-0.243	-0.233	-0.010
69	3-O$(CH_2)_3OC_6H_4$-3-$COOCH_3$	-0.242	-0.272	0.030
70	3-O$(CH_2)_3OC_6H_4$-2-NHCO$(CH_2)_2C_6H_4$-4-SO_2F	-0.242	-0.227	-0.014
71	3-O$(CH_2)_3OC_6H_4$-4-NHCOC_6H_4-4-NO_2	-0.239	-0.232	-0.007
72	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_4-4-NO_2	-0.239	-0.241	0.002

73	3-O$(CH_2)_3OC_6H_4$-4-NHCONHC_6H_5	-0.237	-0.241	0.004
74	3-O$(CH_2)_3OC_6H_4$-4-NHCOC_6H_4-3-NO_2	-0.237	-0.233	-0.005
75	3-O$(CH_2)_3OC_6H_4$-2-NHCO$(CH_2)_4C_6H_4$-4-SO_2F	-0.237	-0.217	-0.020
76	3-O$(CH_2)_3OC_6H_4$-2-NHCONHC_6H_4-4-SO_2F	-0.237	-0.233	-0.004
77	3-O$(CH_2)_3OC_6H_4$-3-NHCONHC_6H_4-4-SO_2F	-0.236	-0.225	-0.011
78	3-O$(CH_2)_3OC_6H_4$-2-NHCONH$(CH_2)_2C_6H_4$-4-SO_2F	-0.236	-0.223	-0.014
79	3-O$(CH_2)_4OC_6H_4$-3-NHCOC_6H_4-4-SO_2F	-0.236	-0.223	-0.013
80	3-O$(CH_2)_3OC_6H_4$-2-NHCONHC_6H_3-4-Cl-3-SO_2F	-0.235	-0.229	-0.006
81	3-O$(CH_2)_4OC_6H_4$-2-NHCOC_6H_3-4-CH_3-3-SO_2F	-0.235	-0.229	-0.006
82	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_2-2,4-$(CH_3)_2$-5-SO_2F	-0.234	-0.233	-0.001
83	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_2-2,4-Cl_2-5-SO_2F	-0.234	-0.233	-0.001
84	3-$(CH_2)_4C_6H_4$-2-NHCONHC_6H_4-3-SO_2F	-0.234	-0.239	0.005
85	3-O$(CH_2)_3OC_6H_4$-3-NHCOC_6H_4-4-OCH_3	-0.233	-0.237	0.004
86	3-$(CH_2)_4C_6H_4$-2-NHCONHC_6H_4-4-SO_2F	-0.233	-0.239	0.007
87	3-O$(CH_2)_3OC_6H_4$-4-NHCOC_6H_4-4-Cl	-0.232	-0.241	0.009
88	3-O$(CH_2)_3OC_6H_4$-2-NHCOC_6H_3-2-CH_3-5-SO_2F	-0.232	-0.236	0.004
89	3-O$(CH_2)_4OC_6H_4$-4-NHCONHC_6H_3-2-OCH_3-5-SO_2F	-0.232	-0.214	-0.018
90	3-O$(CH_2)_3OC_6H_4$-4-C_6H_5	-0.230	-0.261	0.031
91	3-O$(CH_2)_3OC_6H_4$-2-NHCONHC_6H_4-3-SO_2F	-0.230	-0.233	0.003
92	3-O$(CH_2)_3OC_6H_4$-3-NHCOC_6H_4-3-SO_2F	-0.230	-0.230	-0.000
93	3-O$(CH_2)_2OC_6H_4$-3-NHCOC_6H_4-3-SO_2F	-0.229	-0.236	0.007
94	3-O$(CH_2)_3OC_6H_4$-4-CH_3-3-NHCOC_6H_4-4-SO_2F	-0.229	-0.226	-0.003
95	3-O$(CH_2)_3OC_6H_4$-3-NHCONHC_6H_4-3-SO_2F	-0.222	-0.226	0.004
96	3-O$(CH_2)_3OC_6H_4$-3-NHCO$CH_2C_6H_4$-4-SO_2F	-0.220	-0.226	0.006
97	3-O$(CH_2)_3OC_6H_4$-3-NHCOC_6H_4-4-SO_2F	-0.219	-0.229	0.010
98	3-O$(CH_2)_3OC_6H_4$-2-NHCONHC_6H_3-2-Cl-5-SO_2F	-0.217	-0.230	0.013
99	3-O$(CH_2)_3OC_6H_4$-3-NHCO$CH_2OC_6H_4$-4-SO_2F	-0.217	-0.219	0.002
100	3-O$(CH_2)_2OC_6H_4$-3-NHCONHC_6H_4-4-SO_2F	-0.216	-0.231	0.015
101	3-O$(CH_2)_4OC_6H_4$-3-NHCONHC_6H_4-4-SO_2F	-0.215	-0.220	0.005
102	3-O$(CH_2)_3OC_6H_4$-3-NHCOC_6H_4-4-NO_2	-0.214	-0.233	0.019
103	3-O$(CH_2)_2OC_6H_4$-3-NHCOC_6H_4-4-SO_2F	-0.214	-0.235	0.021
104	3-O$(CH_2)_4OC_6H_4$-2-NHCONHC_6H_3-2-Cl-5-SO_2F	-0.207	-0.225	0.018
105	3-O$(CH_2)_3OC_6H_4$-3-NHCONHC_6H_4-4-NO_2	-0.204	-0.230	0.025
106	3-O$(CH_2)_3OC_6H_4$-4-CH_3-3-NHCONHC_6H_4-4-SO_2F	-0.204	-0.223	0.018
107	3-O$(CH_2)_3OC_6H_4$-3-NHCONH$(CH_2)_2C_6H_4$-4-SO_2F	-0.193	-0.215	0.022

[a]Predicted values based on equation 2.

[b]Values for compounds excluded from final modeling, provided to show lack of fit.

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 r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi$	Path connectivity index of order h = 0-6
${}^h\chi_C$	Cluster connectivity index of order h = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 6
${}^h\chi^b$	Bond path connectivity index of order h = 0-6
${}^h\chi_C^b$	Bond cluster connectivity index of order h = 3-6
${}^h\chi_{Ch}^b$	Bond chain connectivity index of order h = 6

$^{h}\chi^{b}_{PC}$	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 = 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

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		

Figure Captions:

Figure 1: Neutral base structure for the 107 benzamidines.

Figure 2: Scatterplot for observed 1/Log C versus predicted 1/Log C using equation 2 for the set of 107 benzamidines.

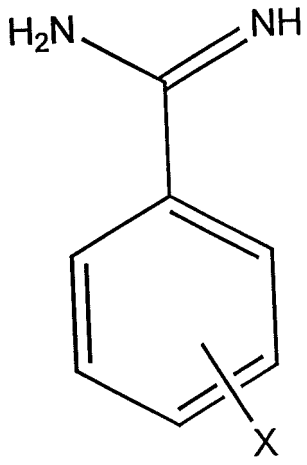


Fig. 1

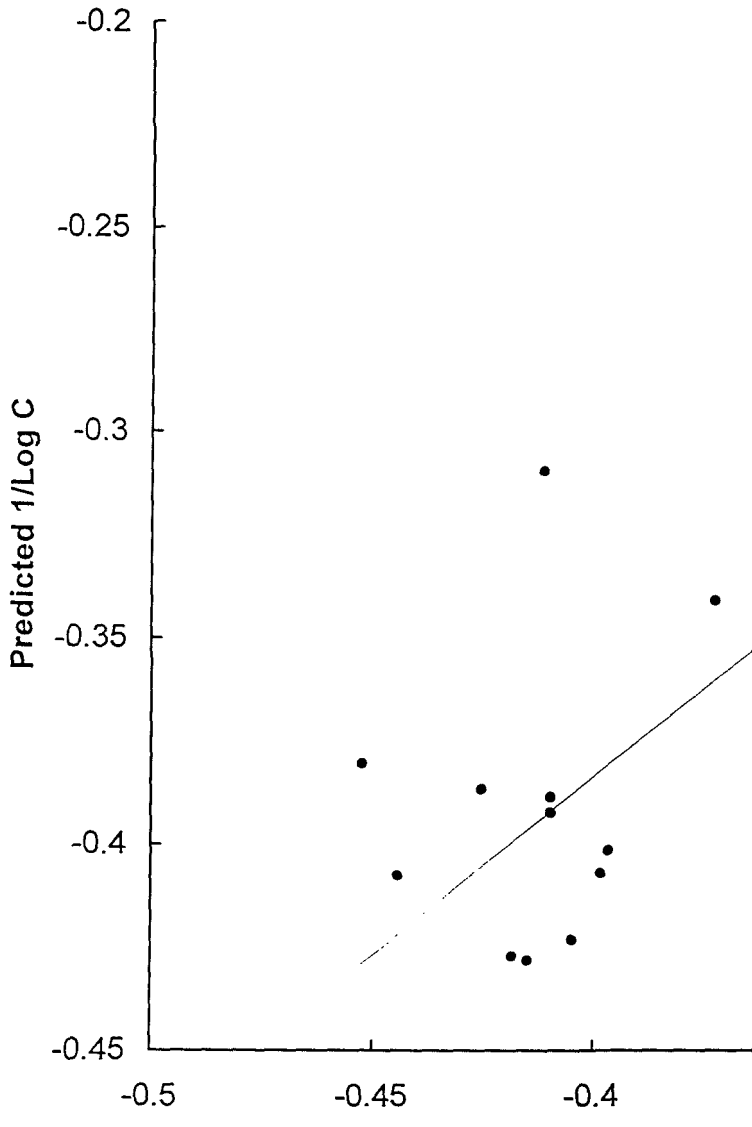
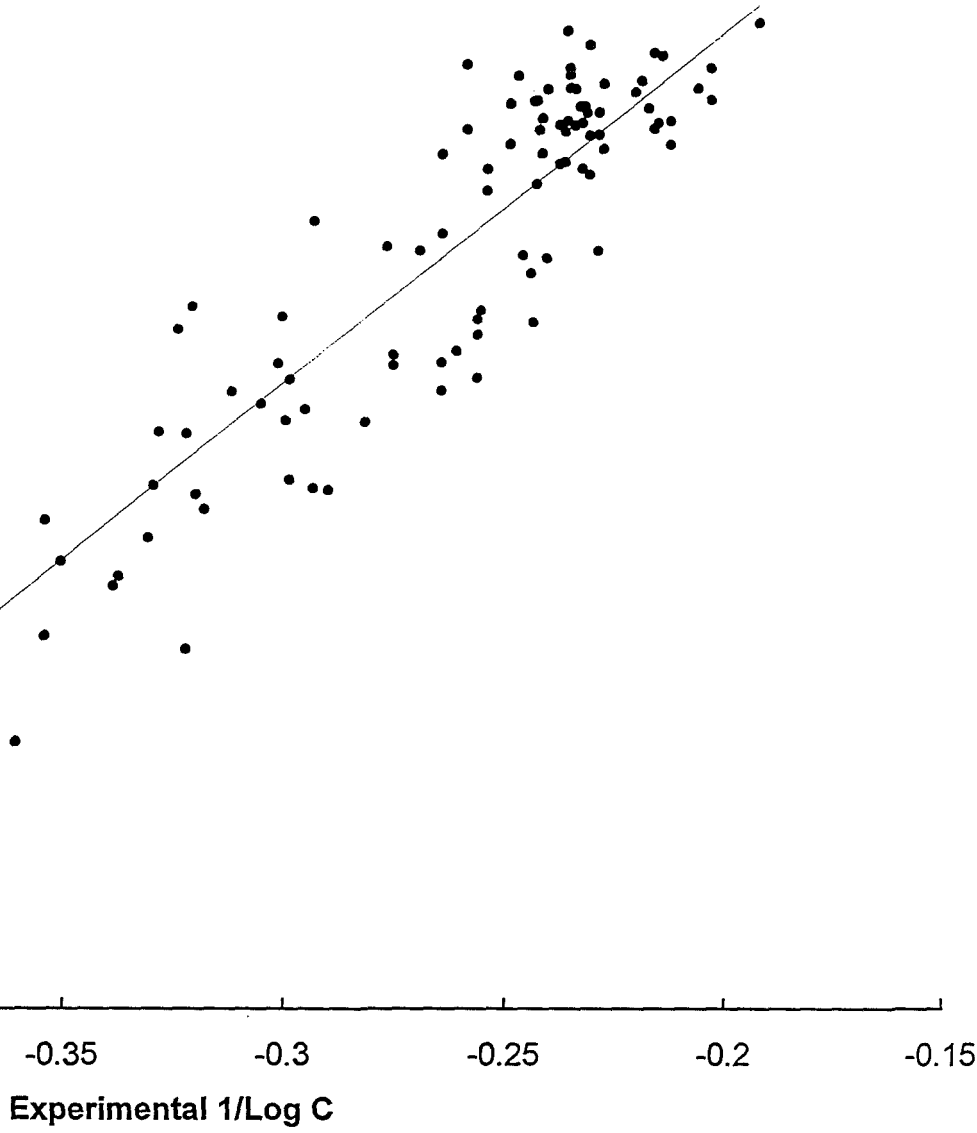


Fig. 2


-0.35
-0.3
-0.25
-0.2
-0.15
Experimental 1/Log C

SAR QSAR Environ. Res., in press

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

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

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

*Author to whom all correspondence should be addressed

ABSTRACT

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.

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 coefficent. 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-hour 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 Abstract Serivces (CAS) registry number (when available), and measured dermal

penetration; are presented in Table I.

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.

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 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 Table II.

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 lab [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 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 combined 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: IC, O, $^{3}\chi$-$^{5}\chi$, $^{6}\chi_{Ch}$, 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:

$$\%DP = 224.1 - 67.9\ P_0 \qquad \text{Eq. 1}$$

$$n = 60\ \ R^2 = 0.675\ \ s = 7.4\ \ F = 120.6$$

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

$$\%DP = 179.7 - 78.8\ ^{1}\chi^{b} \qquad \text{Eq. 2}$$

$$n = 60\ \ R^2 = 0.695\ \ s = 7.1\ \ F = 132.0$$

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, $^{1}\chi^{b}$ 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:

$$\%DP = 186.0 - 25.4\ {}^{3D}W \qquad \text{Eq. 3}$$

$$n = 60 \quad R^2 = 0.673 \quad s = 7.4 \quad F = 119.3$$

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:

$$\%DP = 90.6 - 0.3\ MW \qquad \text{Eq. 4}$$

$$n = 60 \quad R^2 = 0.674 \quad s = 7.4 \quad F = 120.0$$

The values for the parameters used in the final models (P_0, $^{1}\chi^{b}$, ^{3D}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 ($^1\chi^b$), geometrical descriptors (^{3D}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, $^1\chi^b$, 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 $^1\chi^b$ and figure 2 demonstrates the scatter of the residuals. Thus, there are no apparent co-variance problems within this model.

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, $^1\chi^b$, ^{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 characterization of structure for prediction of properties. *J. Math. Chem.* **4**, 185-205.

7. Basak, S.C., Niemi, G.J., and Veith, G.D. (1990). Recent developments in the characterization of chemical structure using graph-theoretic indices, in *Computational Chemical Graph Theory and Combinatorics* (D. H. Rouvray, Ed). Nova, New York, pp. 235-277.

8. Kier, L.B. and Hall, L.H. (1986). *Molecular Connectivity in Structure-Activity Analysis.* Research Studies Press, Letchworth, Hertfordshire, U.K.

9. Basak, S.C. and Grunwald, G.D. (1993). Use of graph invariants, volume and total surface area in predicting boiling point of alkanes. *Mathl. 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. (1997). The relative effectiveness of topological, geometrical and quantum chemical parameters in estimating mutagenicity of chemicals, in *Proceedings of the 7th International Workshop on QSAR in Environmental Sciences* (F. Chen, et al., Eds.). SETAC Press, in press.

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.*, in press.

18. Basak, S.C. and Gute, B.D. (1997). Characterization of molecular structures using topological indices. *SAR QSAR Environ. Res.*, in press.

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.*, in press.

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, UK.

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.

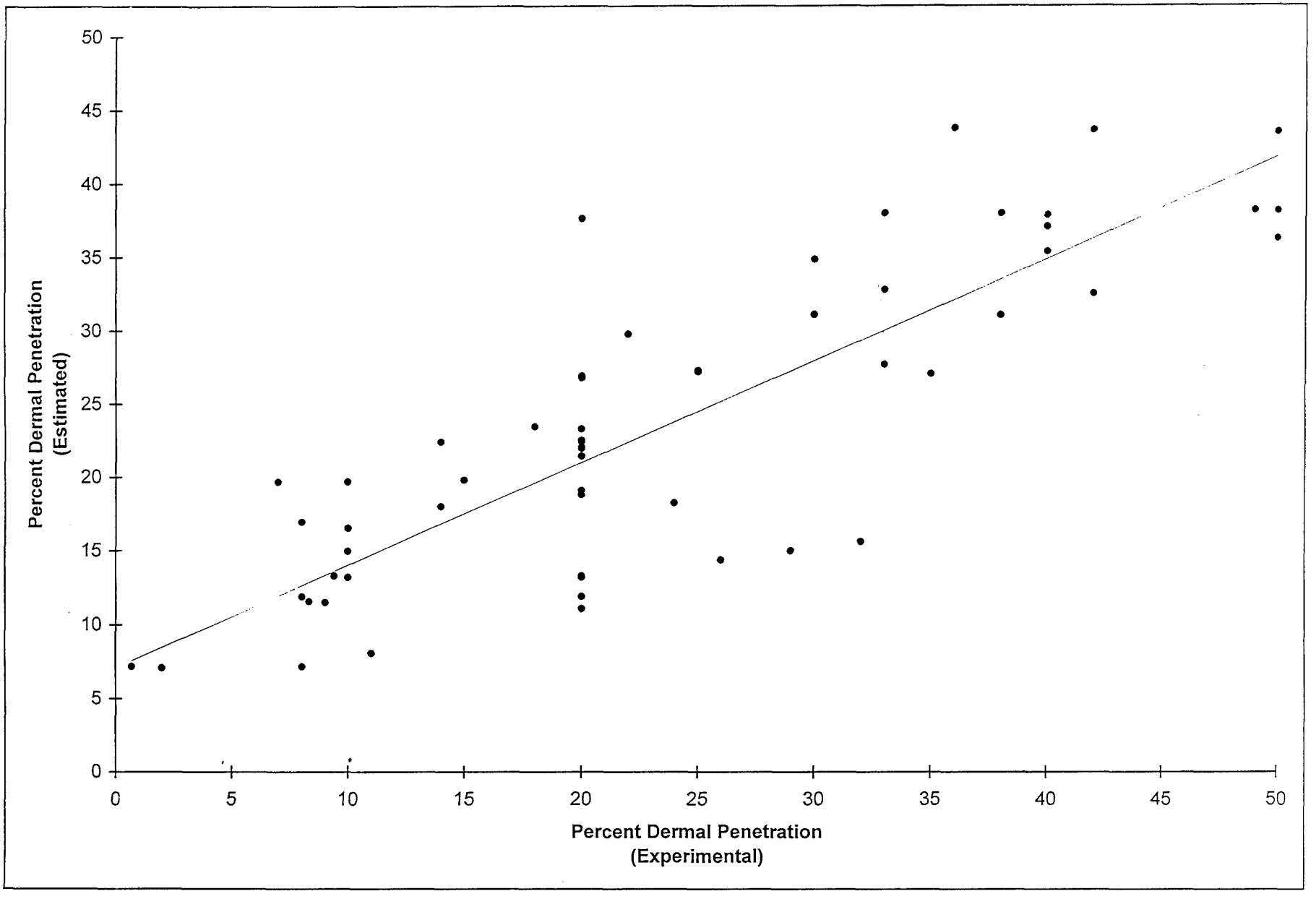
41. Topliss, J.G., and Edwards, R.P. (1979). Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **22**, 1238-1244.

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

FIGURE CAPTIONS

Figure 1. Scatterplot of experimentally determined percent dermal penetration (%DP) *vs.* estimated %DP using equation 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 equation 2.


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

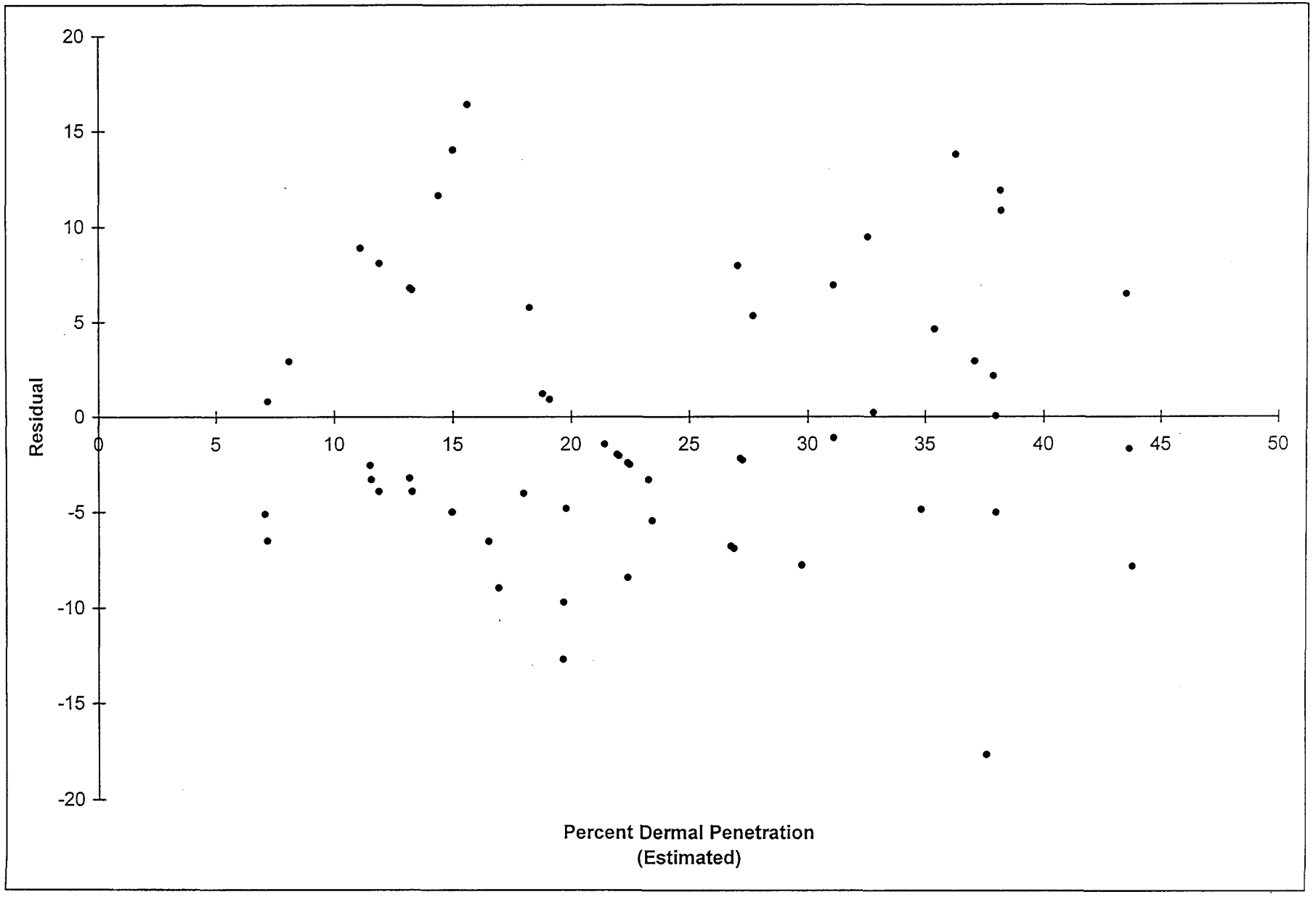

20
15
10
5
0
-5
-10
-15
-20
Residual
0
5
10
15
20
25
30
35
40
45
50
Percent Dermal Penetration
(Estimated)

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
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

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

Topostructural	Topochemical	Geometric	Quantum Chemical
I^W_D	I_{ORB}	V_W	E_{HOMO}
$\tilde{I}^W_D$	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	$^0X^b$-$^6X^b$		ΔHf
H^D	$^3X^b_c$ & $^5X^b_c$		μ
$\tilde{IC}$	$^5X^b_{Ch}$ & $^6X^b_{Ch}$		
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			

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
$\overline{I_D^W}$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^hX$	Path connectivity index of order h = 0-6
${}^hX_C$	Cluster connectivity index of order h = 3-5
${}^hX_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^hX_{Ch}$	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
$^{h}\chi^{b}_{C}$	Bond cluster connectivity index of order h = 3 & 5
$^{h}\chi^{b}_{Ch}$	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 & 5
$^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 5 & 6
$^{h}\chi^{v}_{PC}$	Valence path-cluster connectivity index of order h = 4-6
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

Table IV. Calculated values for molecular weight (MW), lipophilicity (LogP), P_0, $^1\chi^b$, ^{3D}W.

No.	MW	LogP	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

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

SAR QSAR Environ. Res., submitted

Assessment of the Mutagenicity of Chemicals from Theoretical Structural Parameters: A Hierarchical Approach

Subhash C. Basak*, Brian D. Gute and Gregory D. Grunwald
Natural Resources Research Institute,
5013 Miller Trunk Hwy., Duluth, MN, USA

Corresponding Author:
Subhash C. Basak, Natural Resources Research Institute, 5013 Miller Trunk Hwy., Duluth, MN 55811
Phone: (218) 720-4230
Fax: (218) 720-9412
Email: sbasak@wyle.nrri.umn.edu

Running Head: Hierarchical Assessment of Mutagenicity

ABSTRACT

A hierarchical approach has been used in this paper in predicting mutagenicity/ non-mutagenicity of a set of 127 chemical from their molecular descriptors. The set of descriptors consisted of topostructural and topochemical parameters, experimental properties like logP, 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 powers of the models.

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

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]. See Table I for a list of the compounds included in this study and their mutagenic classification based on experimentally determined mutagenic potency.

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 for modeling: I_D^W, $\overline{IC}$, P_3, IC_0, 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, $\in_{HOMO}$ and $\in_{LUMO}$, were combined with the set of six indices and all subsets regression was used again to select the best parameters for model construction. This procedure resulted in the selection of the following model: I_D^W, $\overline{IC}$, P_3, log P, $\in_{LUMO}$.

Discriminant function analysis, using the SAS procedure DISCRIM [34], was used to develop models for predicting mutagenicity/non-mutagenicity 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 III 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 III. 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, biochemical as well properties which showed that topostructural and topochemical indices explain most of the variance in the data [3, 6-10]

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., Lučić, 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.* Accepted.

(7) 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.

(8) Basak, S. C.; Gute, B. D.; 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 – Vol. VII* (F. Chen and G. Schüürman, Eds.). SETAC Press: Pensacola, FL, pp 245-261.

(9) Gute, B. D.; Grunwald, G. D.; Basak, S. C. (1997). 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.; 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.

(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.; 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.

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-aminofuoranthene	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

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
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.

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
$\overline{IC}$	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
${}^h\chi^b_C$	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

Table III. 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, $\in_{LUMO}$	33.3	95.3

Math. Model. Comput. Sci., vol. 8, in press

Quantitative Comparison of Five Molecular Structure Spaces in Selecting Analogs of Chemicals

Subhash C. Basak, Brian D. Gute and Gregory G. 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 76 compounds. These methods include an atom pair (AP) based similarity measure and principal component spaces derived from topostructural indices, topochemical indices, the combined set of all (topostructural and topochemical) indices, and 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 76 compounds. The degree of overlap of the selected analogs 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 structures (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 et al, 1990; Randić, 1992; Willet et al, 1986). Chemical structures can be represented by various types of models, *e.g.*, simple molecular graphs, multigraphs, pseudographs, 3-D representations, 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 et al, 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.

DIMACS Series in Discrete Mathematics and Theoretical Computer Science, submitted

Use of Graph Invariants in QMSA and Predictive Toxicology

S.C. Basak[1] and B.D. Gute
Natural Resources Research Institute,
5013 Miller Trunk Hwy., Duluth, MN, 55811 USA

[1]Corresponding author

Phone: (218) 720-4230
Fax: (218) 720-9412
Email: sbasak@wyle.nrri.umn.edu

1

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

2

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.

Methods

Databases

A set of 38 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 38 isospectral graphs are shown in Figure 1.

A subset of the set of 277 chemicals presented by Yamaguchi *et al.* [27] was used in the current study. This subset consisted of all the chemicals in the larger set 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.

3

[Insert Fig. 1 here]

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 labeled hydrogen-suppressed graph G_1 of acetic acid *(Fig.2)*:

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

W is calculated as:

$$W = \tfrac{1}{2}\sum_{ij} d_{ij} = \sum_{h} h \cdot g_h \qquad (1)$$

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. 2 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:

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

where v_i, v_j,..., 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; $\sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, \ldots\ldots, A_h$$

$$p_1, p_2, \ldots\ldots, p_h$$

5

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.

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 $u_o, u_1, \ldots, u_r$ of length r, there is a distinct path $v_o, v_1, \ldots, v_r$ of the same length such that the

7

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.

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, $\overline{I}_D^W$, 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* [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:

$$J = q(\mu + 1)^{-1} \sum_{i,j,edges} (s_i s_j)^{-½} \qquad (7)$$

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].

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:

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

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:

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

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.

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

12

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.

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.

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

13

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.

14

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 XXX 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

20

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

21

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.

22

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 r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi$	Path connectivity index of order h = 0-6
${}^h\chi_C$	Cluster connectivity index of order h = 3-6
${}^h\chi_{Ch}$	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

24

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^v (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%)

25

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

26

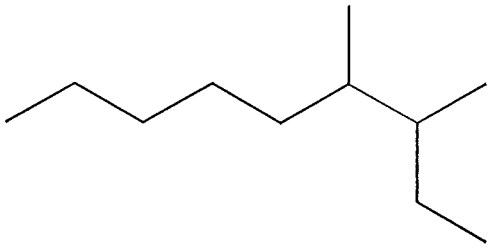
Figure Captions

Figure 1 – Structures of 38 isospectral graphs.

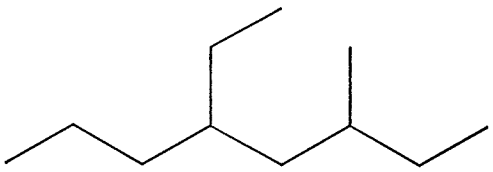
Figure 2 – Unlabeled, hydrogen-suppressed graph of acetic acid (G_1).

Figure 3 – Analogs selected for isospectral graph 10.1.1.

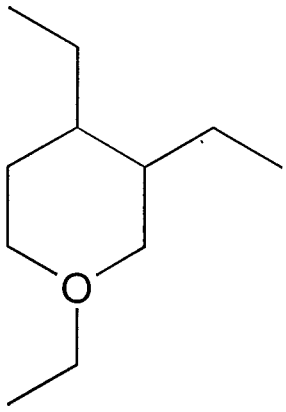
27



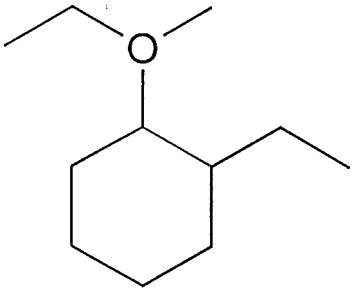
1.1



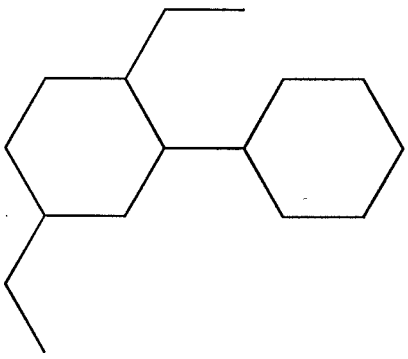
1.2



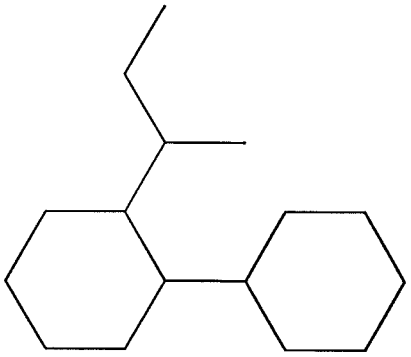
2.1



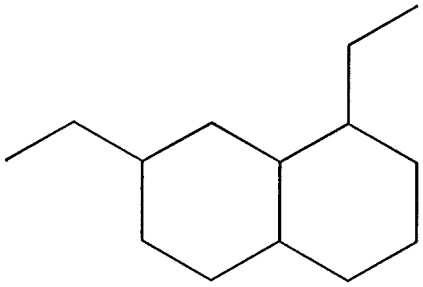
2.2



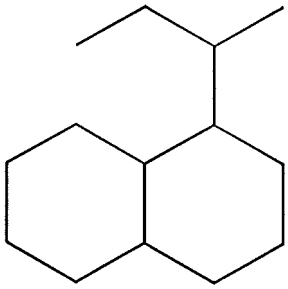
3.1



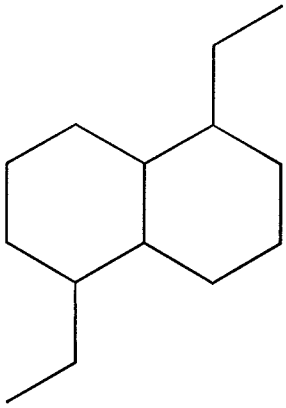
3.2



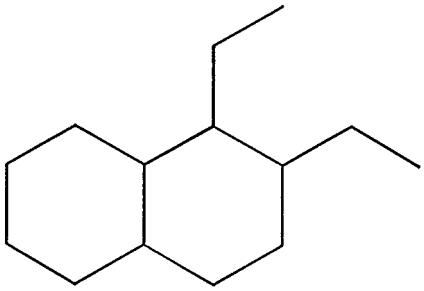
4.1.1



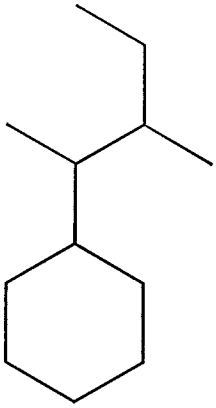
4.1.2



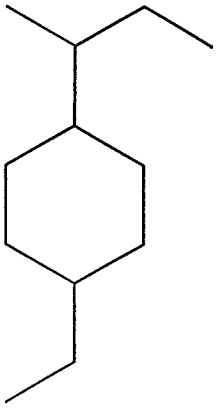
4.2.1



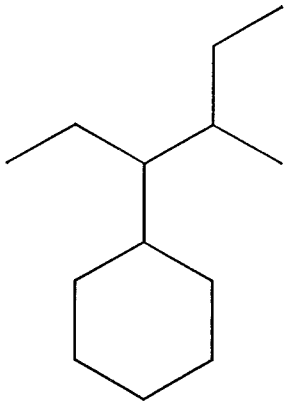
4.2.2



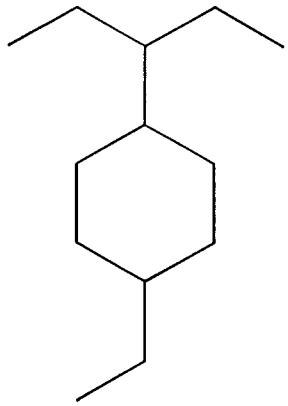
5.1



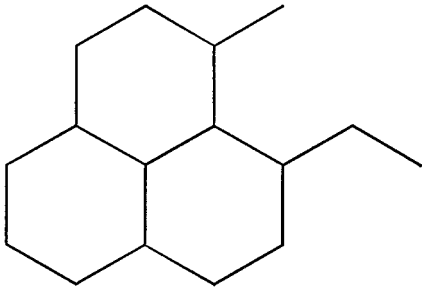
5.2



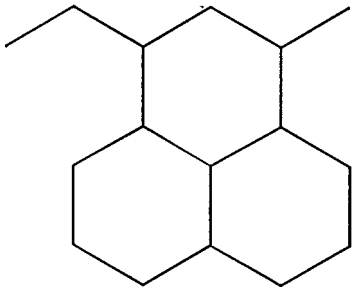
6.1



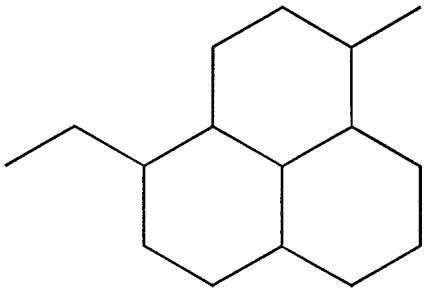
6.2



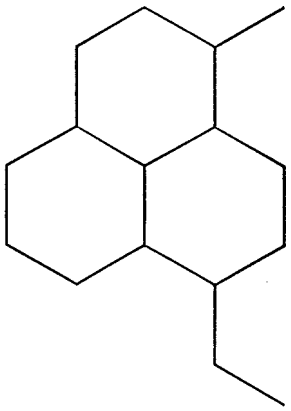
7.1.1



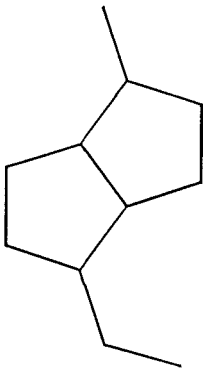
7.1.2



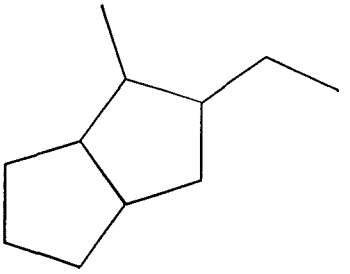
7.2.1



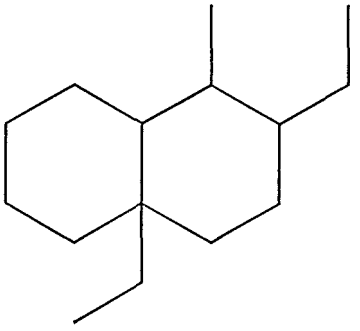
7.2.2



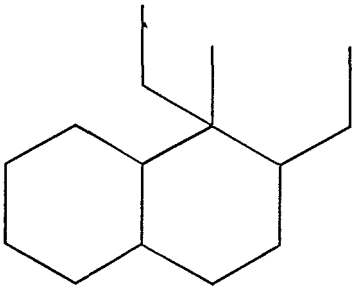
8.1



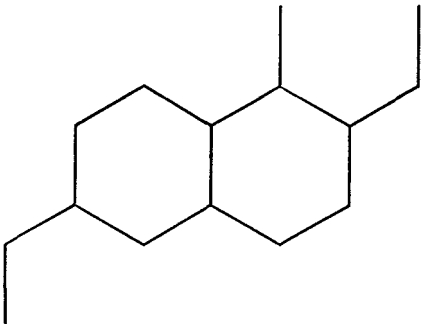
8.2



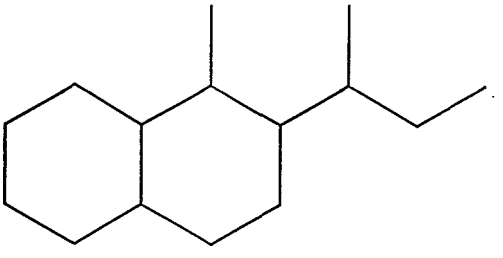
9.1.1



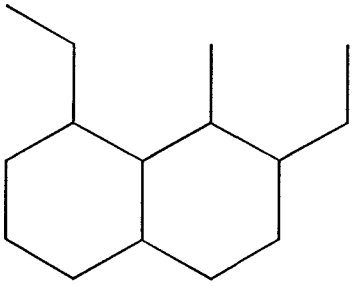
9.1.2



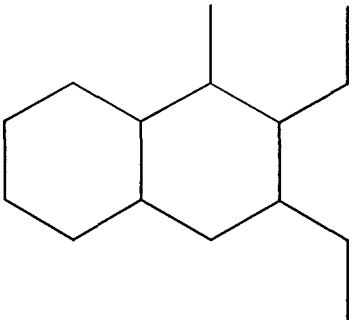
9.2.1



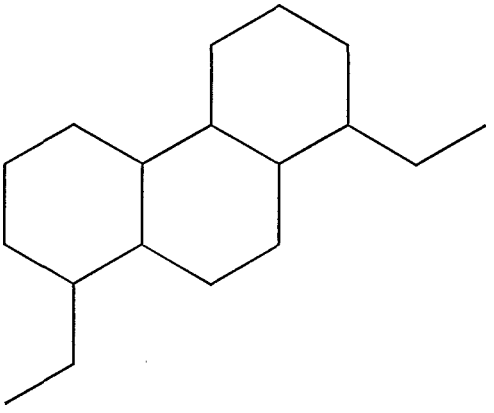
9.2.2



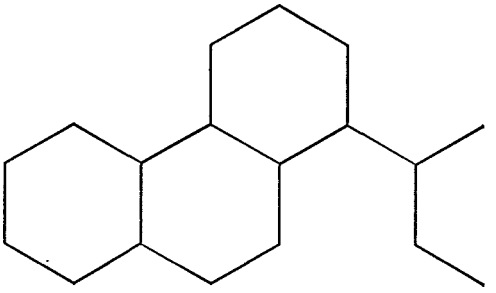
9.3.1



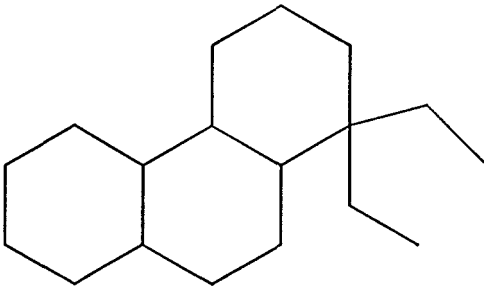
2



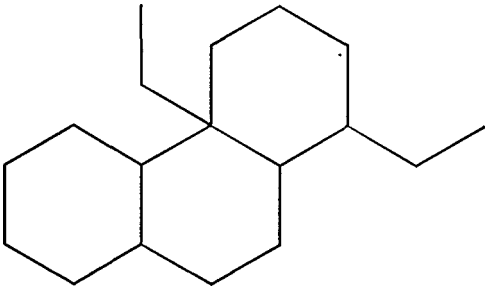
10.1.1



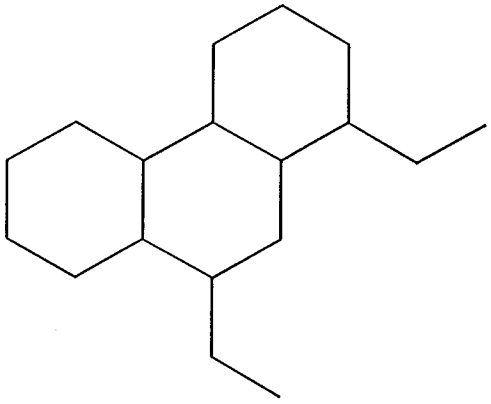
10.1.2



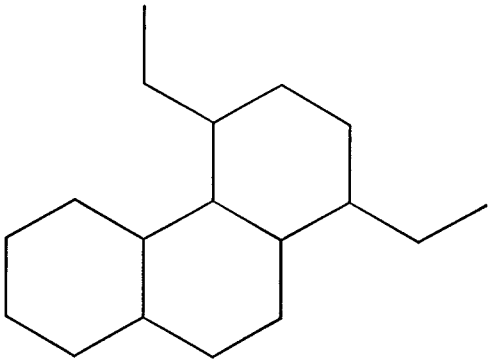
10.2.1



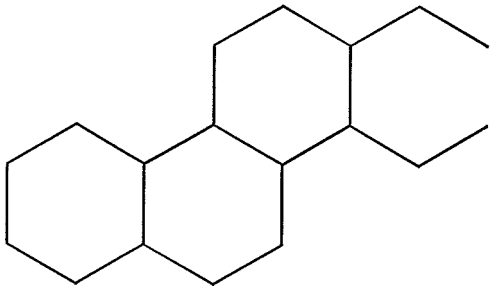
10.2.2



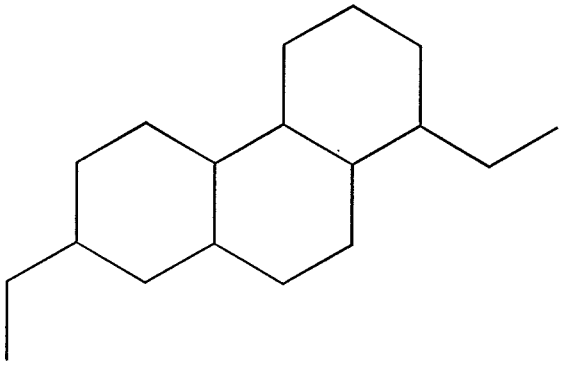
10.3.1



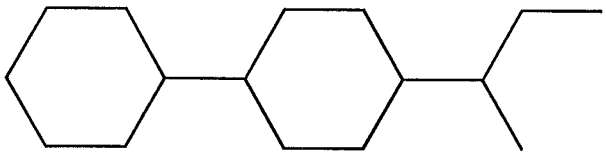
10.3.2



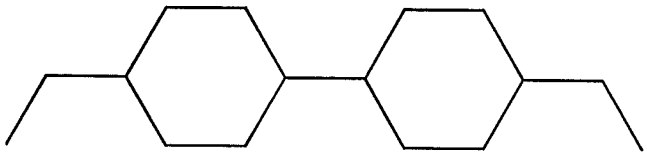
10.4.1



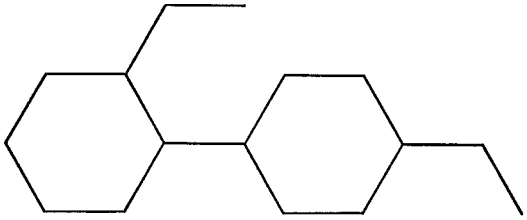
10.4.2



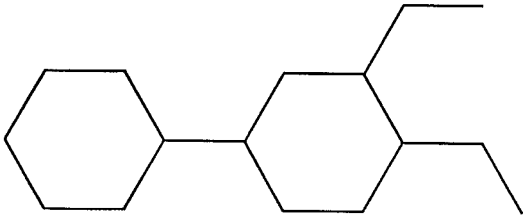
11.1.1



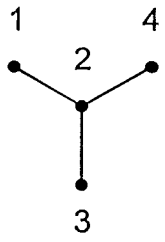
11.1.2



11.2.1



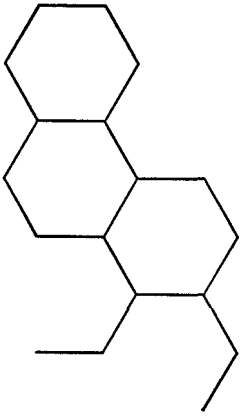
11.2.2

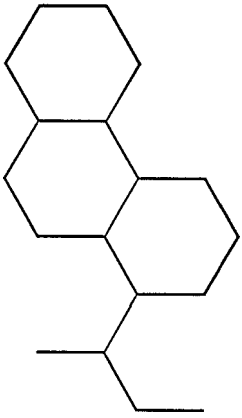


G_1

Probe

Atom Pair Method



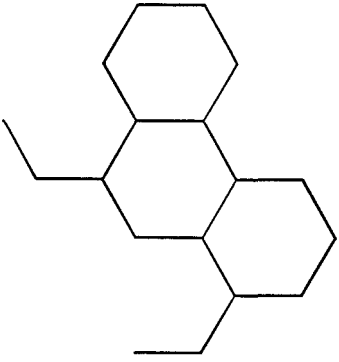


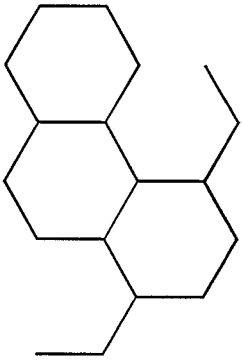
Similarity Score

S=0.95

S=0.93

Euclidean Distance Method

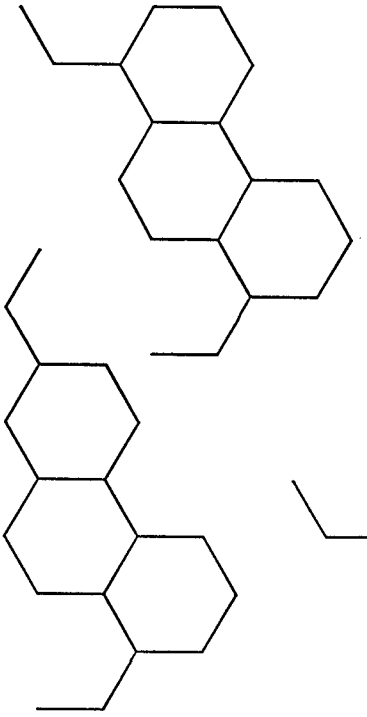




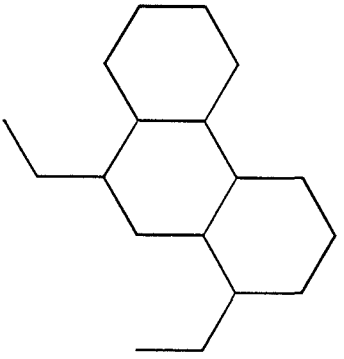
Euclidean Distance

ED=0.19

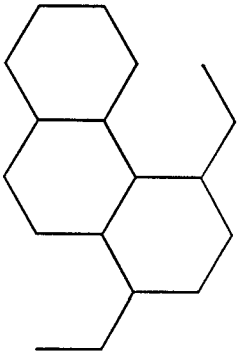
ED=0.20



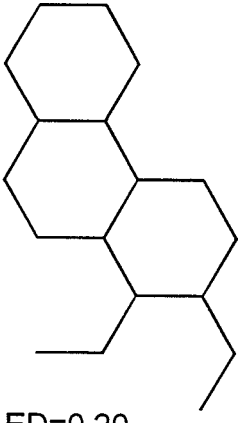
S=0.88



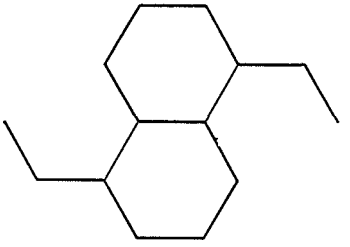
S=0.86



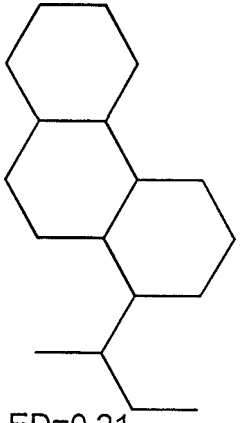
S=0.86



ED=0.20



ED=0.20



ED=0.21

Advances in Molecular Similarity, JAI Press, submitted

Characterization of the Molecular Similarity of Chemicals Using Topological Invariants

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

Center for Water and the Environment
Natural Resources Research Institute
University of Minnesota, Duluth
5013 Miller Trunk Highway
Duluth, MN 55811, USA

* To whom all correspondence should be addressed.

Abstract

Three similarity spaces were used in the selection of analogs 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 (PCs) derived from: 1) 40 topostructural indices, 2) 61 topochemical parameters and 3) the full set 101 topostructural and topochemical indices. The three methods selected sets of analogs with a substantial number of structurally analogous molecules. For the KNN method of property estimation, the similarity space which used the full set of indices was superior to either of the subsets (topostructural or topochemical). For all three methods, $K = 6\text{-}10$ gave the best estimated values for boiling point.

1. 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 called 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 analogs 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 analog selection and KNN based estimation of normal boiling point using : a) a set of 40 topostructural indices, b) a group of 61 topochemical indices, and c) the combined set of 101 indices.

2. Methods

2.1 DATABASE

The normal boiling point database consisted of 2926 compounds taken from the U.S. EPA ASTER [18] system. This 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.

2.2 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 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 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:

$$W = \tfrac{1}{2} \sum_{i,j} d_{ij} = \sum_{h} h \bullet g_h \qquad (1)$$

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 ten.

[Insert Fig. 1 here]

Randić's connectivity index [20], 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 [21]. The generalized form of the simple path connectivity index is as follows:

$$^{h}\chi = \sum_{paths} (v_i v_j ... v_{h+1})^{-½} \qquad (2)$$

where v_i, v_j,..., 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; $\sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, ..., A_h$$

$$p_1, p_2, ..., p_h$$

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

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

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

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*.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.* [32] 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, $S(v, 0) = \phi$, $S(v, r) = v$ 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 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 [26].

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 [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.

[Insert Fig. 2 here]

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]:

$$I_D^W = W \log_2 W - \sum_h g_h \bullet h \log_2 h \qquad (6)$$

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 upon 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 range of values are independent of molecular size. The general form of the J index calculation is as follows:

$$J = q(\mu + 1)^{-1} \sum_{i,j,\ edges} (s_i s_j)^{-½} \qquad (7)$$

where the cyclomatic number μ (or number of rings in the graph) is μ = 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].

2.3 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.

2.4 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 which 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 the molecules *i* and *j* is defined as:

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

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 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 analogs (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.

3. Results

3.1 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.

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.

3.2 ANALOG SELECTION

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

[Insert Fig. 3]

3.3 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, however, there was only a small difference between models.

[Insert Fig. 4]

4. 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 analogs and in the KNN based estimation of properties. The similarity spaces were created using a principal component analysis 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 which 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 analogs of a particular structure (Figure 3). Perusal of the three sets of structures show 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 analogs selected by similarity methods derived from experimental physical properties, atom pairs and topological indices select very similar sets of analogs [10].

In the case of KNN based estimation of boiling points of chemicals from their analogs, 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 analogs can be used for the estimation of boiling points of diverse chemicals if more accurate methods are not available.

5. Acknowledgments

This 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, a grant from Exxon Corporation and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

6. References

[1] M.A. Johnson and G.M. Maggiora, eds., *Concepts and applications of molecular similarity* (Wiley, 1990).

[2] R. Carbó, L. Leyda and M. Arnau, Int. J. Quantum Chem. 17(1980) 1185.

[3] P.E. Bowen-Jenkins, D.L. Cooper and G. Richards, J. Phys. Chem. 89 (1985) 2195.

[4] S.C. Basak, V.R. Magnuson, G.J. Niemi and R.R. Regal, Discrete Appl. Math. 19 (1988) 17.

[5] S.C. Basak, S. Bertelsen and G. Grunwald, J. Chem. Inf. Comput. Sci. 34 (1994) 270.

[6] G. Rum and W.C. Herndon, J. Am. Chem. Soc. 113 (1991) 9055.

[7] P. Willett and V. Winterman, Quant. Struct.-Act. Relat. 5 (1986) 18.

[8] C.L. Wilkins and M. Randić, Theoret. Chim. Acta (Berl.) 58 (1980) 45.

[9] N. Trinajstić, *Chemical Graph Theory Vols. I & II* (CRC Press, Boca Raton, Florida, 1983).

[10] S.C. Basak and G.D. Grunwald, Mathl. Modelling Sci. Comput., in press.

[11] S.C. Basak and G.D. Grunwald, SAR QSAR Environ. Res. 2 (1994) 289.

[12] S.C. Basak and G.D. Grunwald, New J. Chem. 19 (1995) 231.

[13] S.C. Basak and G.D. Grunwald, J. Chem. Inf. Comput. Sci. 35 (1995) 366.

[14] S.C. Basak and G.D. Grunwald, SAR QSAR Environ. Res. 3 (1995) 265.

[15] S.C. Basak and G.D. Grunwald, Chemosphere 31 (1995) 2529.

[16] S.C. Basak, B.D. Gute and G.D. Grunwald, Croat. Chim. Acta, 69 (1996) 1159.

[17] M.S. Lajiness, in: *Computational Chemical Graph Theory*, ed. D.H. Rouvray (Nova Science Publishers, New York, 1990) p. 300.

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

[19] H. Wiener, J. Am. Chem. Soc. 69 (1947) 17.

[20] M. Randić, J. Am. Chem. Soc. 97 (1975) 6609.

[21] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis* (Research Studies Press, Hertfordshire, U.K., 1986).

[22] D. Bonchev and N. Trinajstić, J. Chem. Phys. 67 (1977) 4517.

[23] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy and S.C. Basak, J. Comput. Chem. 5 (1984) 581.

[24] S.C. Basak, A.B. Roy and J.J. Ghosh, in: *Proceedings of the Second International Conference on Mathematical Modelling,* eds. X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler (University of Missouri - Rolla, 1980) p. 851.

[25] S.C. Basak and V.R. Magnuson, Arzneim.-Forsch. Drug Res. 33 (1983) 501.

[26] A.B. Roy, S.C. Basak, D.K. Harriss and V.R. Magnuson, in: *Mathematical Modelling in Science and Technology,* eds. X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin (Pergamon Press, New York, 1984) p. 745.

[27] A.T. Balaban, Chem. Phys. Lett. 89 (1982) 399.

[28] A.T. Balaban, Pure and Appl. Chem. 55 (1983) 199.

[29] A.T. Balaban, Math. Chem. (MATCH) 21 (1985) 115.

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

[31] C.E. Shannon, Bell Syst. Tech. J. 27 (1948) 379.

[32] R. Sarkar, A.B. Roy and R.K. Sarkar, Math. Biosci. 39 (1978) 299.

[33] V.R. Magnuson, D.K. Harriss and S.C. Basak, in: *Studies in Physical and Theoretical Chemistry,* ed. R.B. King (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] S.C. Basak, G.J. Niemi and G.D. Veith, J. Math. Chem., 7 (1991) 243.

[36] S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal and G.D. Veith, Mathl. Modelling 8 (1987) 300.

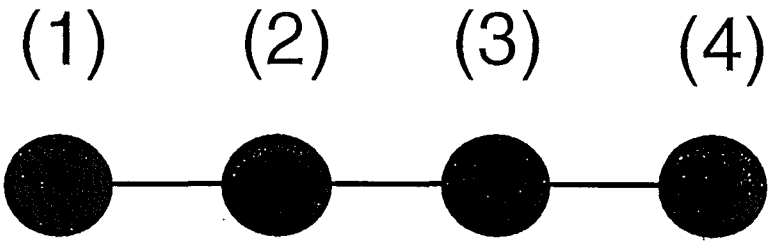
Figure Legend

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

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

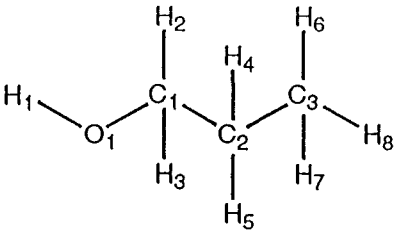
Figure 3 The five analogs selected for the probe 3-methyl-4-chlorophenol using three molecular similarity spaces: topostructural, topochemical, and all indices. The numbers under the structures indicate the ranking of the analogs and the Euclidean distance to the probe.

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



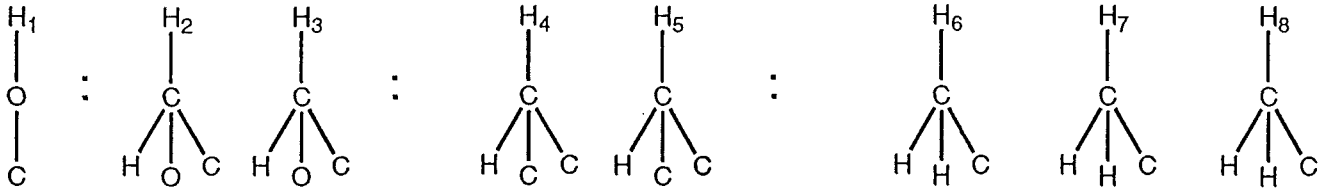
G_1

G_2: n-propanol

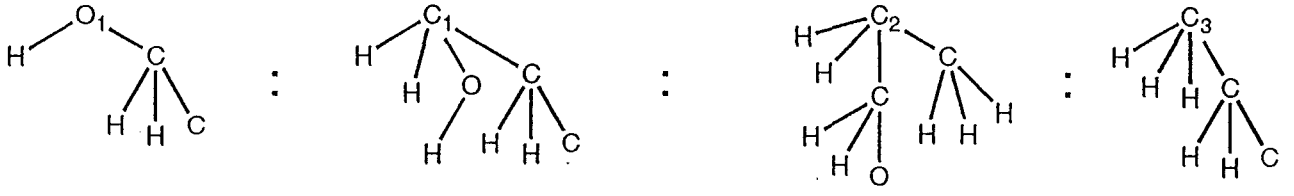


Second order neighbors:

I II III IV



V VI VII VIII



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)

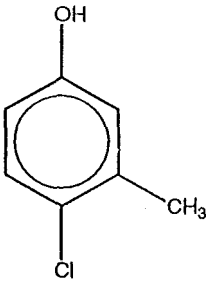
Probability:

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

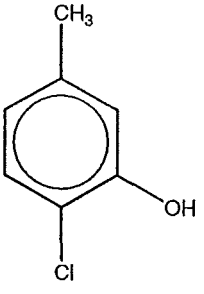
$SIC_2 = IC_1/\log_2 12 = 0.796$ bits

$CIC_2 = \log_2 12 - IC_2 = 0.730$ bits

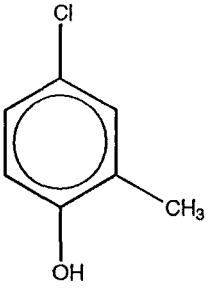


Probe: 3-methyl-4-chlorophenol

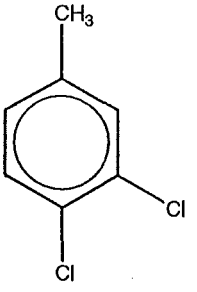
Structural:



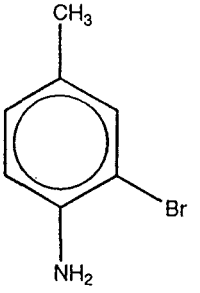
(1) 0.00



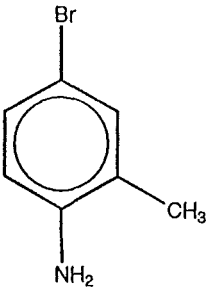
(2) 0.00



(3) 0.01

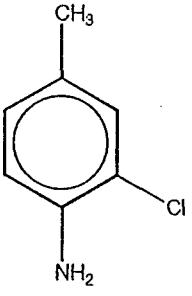


(4) 0.01

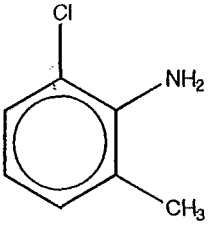


(5) 0.01

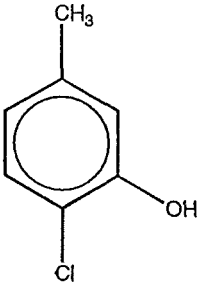
Chemical:



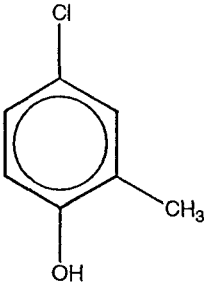
(1) 0.01



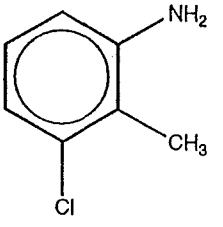
(2) 0.02



(3) 0.02

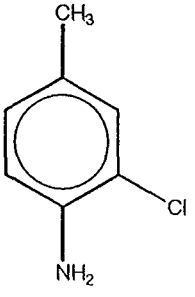


(4) 0.02

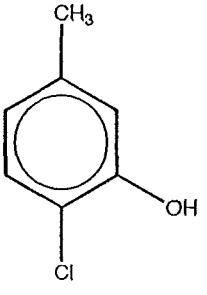


(5) 0.03

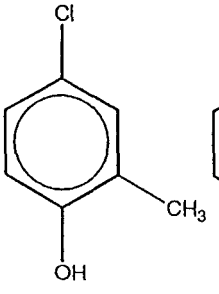
All:



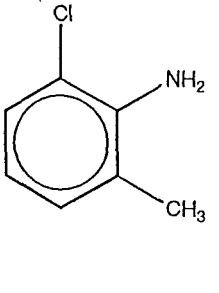
(1) 0.01



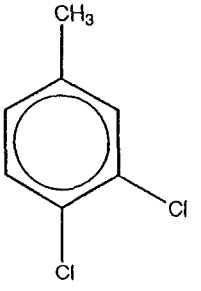
(2) 0.02



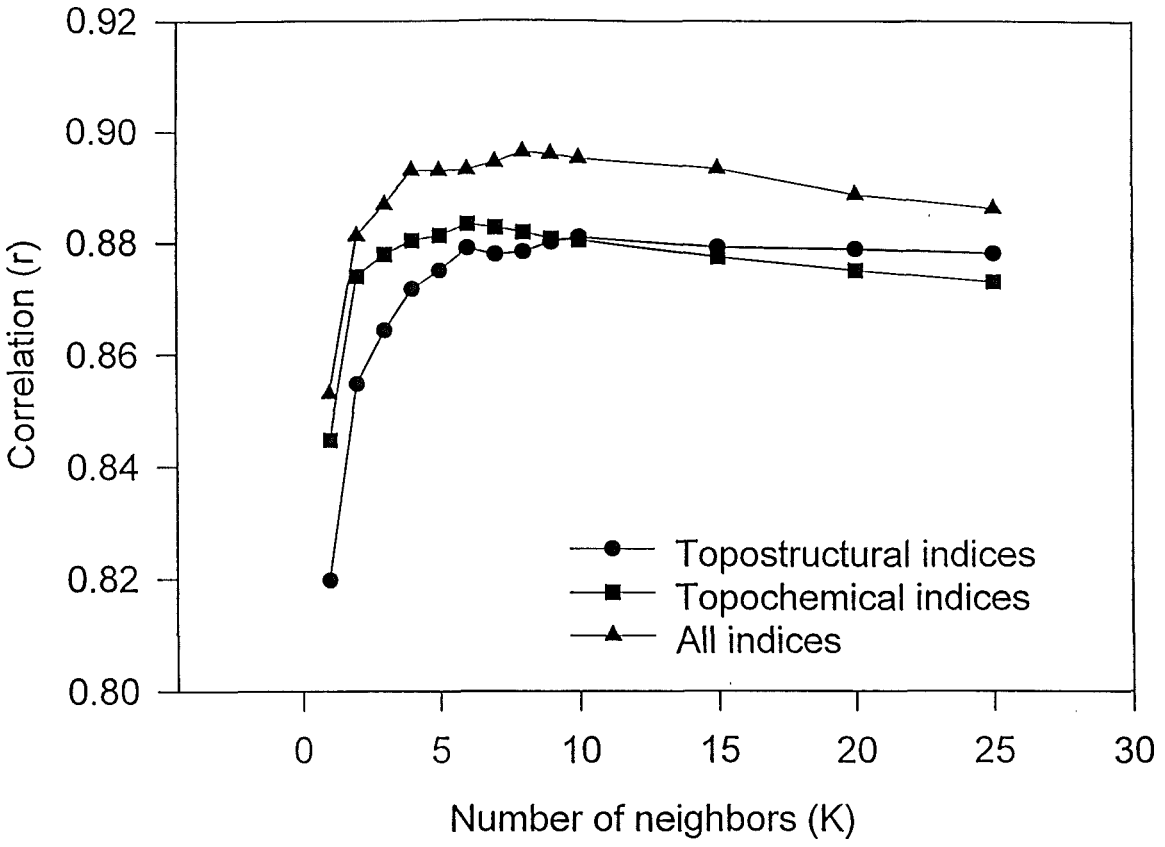
(3) 0.02



(4) 0.03



(5) 0.03


0.92
0.90
0.88
0.86
0.84
0.82
0.80
Correlation (r)
0
5
10
15
20
25
30
Number of neighbors (K)
Topostructural indices
Topochemical indices
All indices

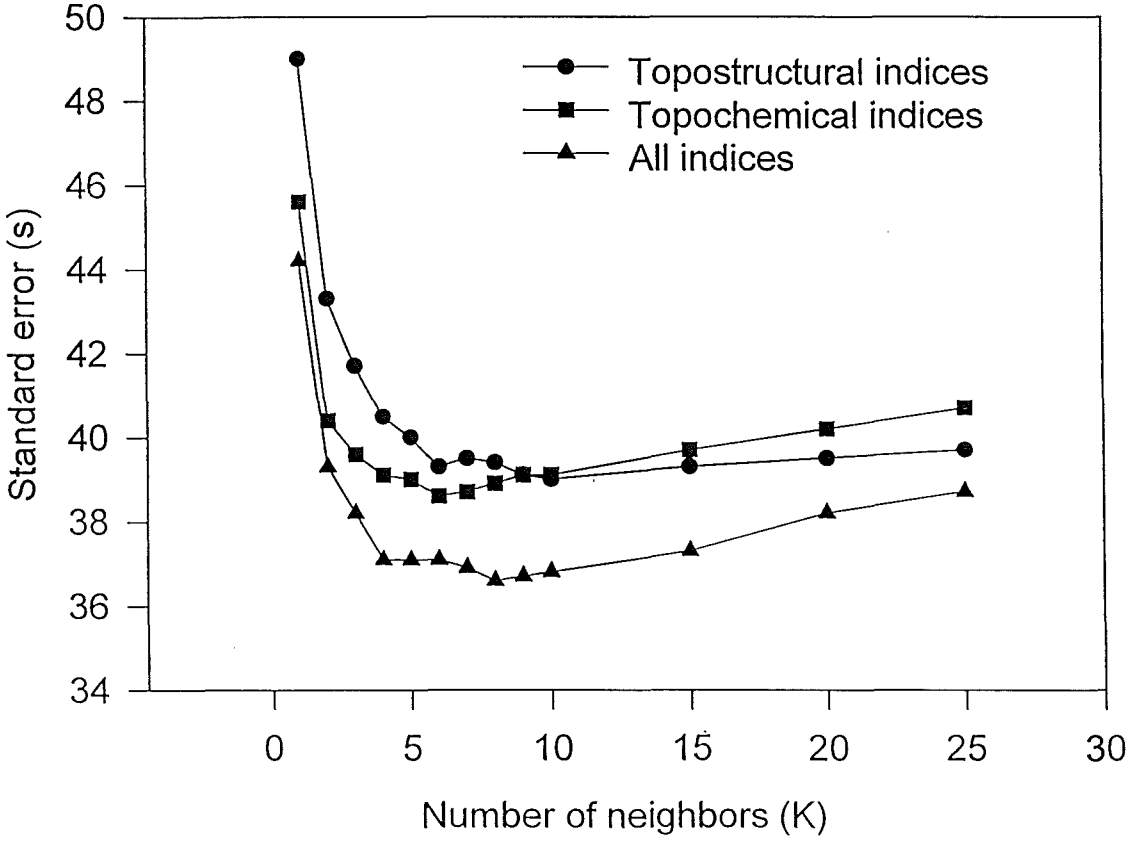

50
48
46
44
42
40
38
36
34
Standard error (s)
0
5
10
15
20
25
30
Number of neighbors (K)
Topostructural indices
Topochemical indices
All indices

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 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 = 3-6
$^{h}\chi^{b}_{Ch}$	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

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	$^{4}\chi^{b}_{C}$, $^{4}\chi^{v}_{C}$, $^{6}\chi^{v}_{PC}$

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):

$$D(G_1) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 2 & 3 & 3 & 4 & 5 \\ 1 & 0 & 1 & 1 & 2 & 2 & 3 & 4 \\ 2 & 1 & 0 & 2 & 3 & 3 & 4 & 5 \\ 2 & 1 & 2 & 0 & 1 & 1 & 2 & 3 \\ 3 & 2 & 3 & 1 & 0 & 2 & 3 & 4 \\ 3 & 2 & 3 & 1 & 2 & 0 & 1 & 2 \\ 4 & 3 & 4 & 2 & 3 & 1 & 0 & 1 \\ 5 & 4 & 5 & 3 & 4 & 2 & 1 & 0 \end{bmatrix} \end{array}$$

W is calculated as:

$$W = 1/2 \sum_{i,j} d_{ij} = \sum_{h} h \cdot g_h \tag{1}$$

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 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.

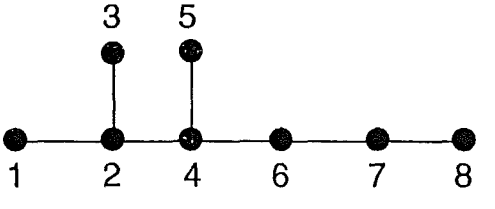


FIGURE 1 Hydrogen-suppressed graph of 2,3-dimethylhexane.

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:

$$A_1, A_2, \ldots, A_h$$
$$p_1, p_2, \ldots, p_h$$

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 [45]:

$$\mathrm{IC} = -\sum_{i=1}^{h} p_i \log_2 p_i \tag{2}$$

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 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 [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 r th 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 r th 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:

$$SIC_r = IC_r / \log_2 n \tag{3}$$

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

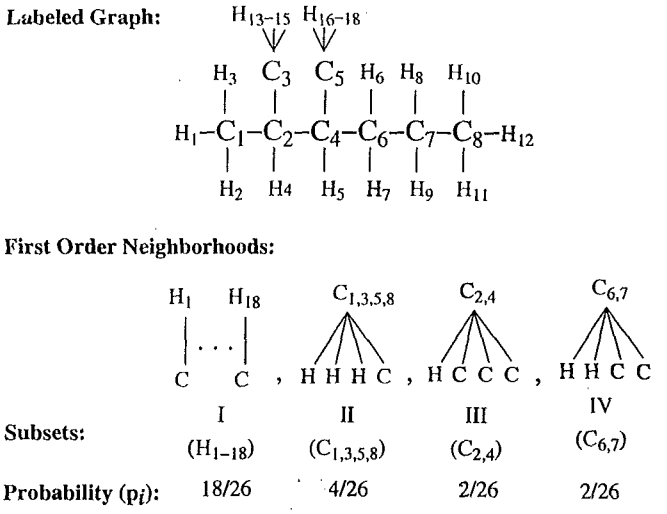
Another information-theoretic invariant, complementary information content (CIC_r) [50], is defined as:

$$CIC_r = \log_2 n - IC_r \tag{4}$$

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]:



$$IC_1 = -\Sigma\, p_i \cdot \log_2 p_i$$
$$= 2 \cdot 2/26 \cdot \log_2 26/2 + 4/26 \cdot \log_2 26 + 18/26 \cdot \log_2 26/18$$
$$= 1.150 \text{ bits}$$

$$SIC_1 = IC_1 / \log_2 26$$
$$= 0.353 \text{ bits}$$

$$CIC_1 = \log_2 26 - IC_1$$
$$= 2.108 \text{ bits}$$

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.

$$I_D^W = W \log_2 W - \sum g_{h_h} \cdot h \log_2 h \tag{5}$$

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 I 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 $h=0-6$
${}^hX^b_C$	Bond cluster connectivity index of order $h=3-6$
${}^hX^b_{Ch}$	Bond chain connectivity index of order $h=3-6$
${}^hX^b_{PC}$	Bond path-cluster connectivity index of order $h=4-6$
${}^hX^v$	Valence path connectivity index of order $h=0-6$
${}^hX^v_C$	Valence cluster connectivity index of order $h=3-6$
${}^hX^v_{Ch}$	Valence chain connectivity index of order $h=3-6$
${}^hX^v_{PC}$	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

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, ${}^4X$, ${}^4X_{Ch}$, ${}^6X^v_{Ch}$, ${}^3X^b_C$, ${}^5X^b_C$, ${}^3X^b_{PC}$, 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^v	SIC_0	SIC_1	SIC_4	^3X_C	^5X_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)	^1X(0.96)	^0X(0.96)	P_1(0.96)	^3X(0.95)	W(0.95)	$^1X^b$(0.95)	$^4X^b$(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)	$^5X^v_{Ch}$(0.63)	CIC_1(−0.63)	^5X_C(0.61)	$^5X^b_{Ch}$(0.61)	SIC_0(0.61)	^6X_C(0.60)	$^6X^v_{Ch}$(0.59)	CIC_3(−0.58)
4	J(0.83)	J^Y(0.73)	J^B(0.73)	J^X(0.62)	$^3X^b_C$(0.56)	^3X_C(0.55)	$^6X_{Ch}$(−0.44)	P_{10}(−0.42)	$^5X^b_{Ch}$(−0.41)	$^6X^b_{Ch}$(−0.41)
5	IC_0(−0.45)	SIC_0(−0.43)	J^X(0.36)	J(0.35)	$^4X^b_{Ch}$(0.35)	$^4X_{Ch}$(0.35)	CIC_0(0.35)	SIC_1(−0.34)	$^6X^v_{PC}$(−0.33)	$^5X_{Ch}$(0.33)
6	$^4X^b_C$(0.57)	^4X_C(0.57)	P_8(0.48)	P_9(0.46)	$^4X^v_C$(0.44)	P_{10}(0.42)	$^3X^b_C$(0.35)	^3X_C(0.33)	$^5X^v_C$(−0.32)	P_7(0.31)
7	^6X_C(−0.43)	$^6X^b_C$(−0.42)	$^5X^b_C$(−0.42)	$^3X_{Ch}$(0.40)	^5X_C(−0.39)	$^4X_{Ch}$(0.31)	$^4X^b_{Ch}$(0.29)	$^4X^b_{PC}$(−0.26)	$^5X_{Ch}$(0.26)	$^5X^v_{Ch}$(0.21)
8	$^4X^v_C$(0.49)	$^3X^v_C$(0.40)	$^2X^v$(0.29)	$^6X^b_C$(−0.27)	^6X_C(−0.26)	J^Y(−0.24)	$^1X^v$(0.23)	$^0X^v$(0.22)	J^b(−0.21)	P_9(−0.20)
9	$^3X_{Ch}$(0.73)	$^4X_{Ch}$(0.47)	$^4X^b_{Ch}$(0.43)	$^6X^v_{Ch}$(−0.21)	$^5X^v_{Ch}$(−0.21)	$^5X_{Ch}$(−0.19)	^6X_C(−0.16)	$^6X_{Ch}$(−0.16)	J^X(−0.16)	$^6X^b_{Ch}$(−0.15)
10	IC_0(0.35)	H^v(0.25)	J^X(−0.24)	$^1X^v$(0.24)	SIC_0(0.21)	I_{ORB}(−0.21)	$^6X_{PC}$(−0.21)	J^Y(−0.20)	J^B(−0.19)	$^1X^b$(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) ${}^3X_C$ and ${}^5X_C$ 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^w, P_8 and ${}^4X$ (general shape and size);
b) IC_4, SIC_3 and CIC_1 (complexity);

c) ${}^{4}X^{v}_{Ch}$ and ${}^{6}X^{v}_{Ch}$ (cyclicity); and
d) ${}^{3}X^{b}_{C}$, ${}^{5}X^{b}_{C}$, ${}^{3}X^{b}_{PC}$ 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). Physico-chemical 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.

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

Nr.	Compound	Exp. ln(R)	Est. 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 r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
$^h\chi$	Path connectivity index of order h = 0-6
$^h\chi_C$	Cluster connectivity index of order h = 3-5
$^h\chi_{PC}$	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
$^h\chi^b_C$	Bond cluster connectivity index of order h = 3, 5
$^h\chi^b_{Ch}$	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, 5
$^h\chi^v_{Ch}$	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	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
$\underline{H^D}$	$^3\chi^b_C$ and $^5\chi^b_C$		μ
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	$^3\chi^v_C$ and $^5\chi^v_C$		
$^0\chi$ - $^6\chi$	$^5\chi^b_{Ch}$ and $^6\chi^b_{Ch}$		
$^3\chi_C$ and $^5\chi_C$	$^4\chi^b_{PC}$ - $^6\chi^b_{PC}$		
$^5\chi_{Ch}$ and $^6\chi_{Ch}$	J^B		
$^4\chi_{PC}$ - $^6\chi_{PC}$	J^X		
P_0 - P_{10}	J^Y		
J			

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}, ΔHf	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}, ΔHf, μ	35.8	0.791	0.92
17-11	topostructural + topochemical + geometric + MNDO	$^4\chi_{PC}$, P_0, J, SIC_2, SIC_4, $^5\chi^b_C$, ΔHf	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, $^5\chi^b_C$, ΔHf	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 ($^5\chi_C^b$). 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.

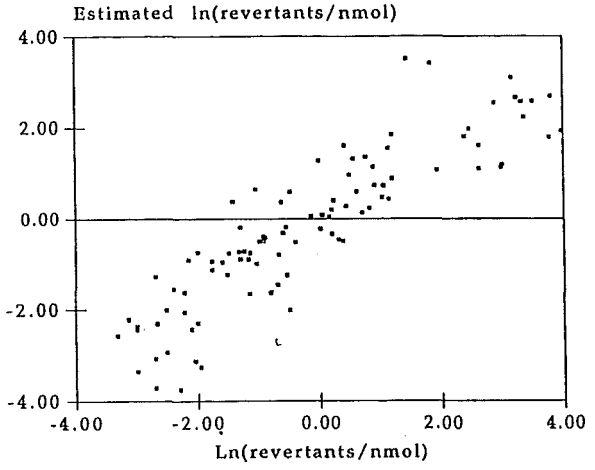


Figure 17–1 Scatterplot for observed ln(R) versus estimated ln(R) using Equation 17-10 for set of 95 aromatic and heteroaromatic amines

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}	ΔHf	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.

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 ninnow (*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
${}^h\chi_c^b$	Bond cluster connectivity index of order h = 3, 5
${}^h\chi_{ch}^b$	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, 5
${}^h\chi_{Pc}^v$	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	${}^0\chi^b - {}^6\chi^b$		ΔH_f
H^D	${}^3\chi_c^b$ and ${}^5\chi_c^b$		μ
$\overline{IC}$	${}^6\chi_{ch}^b$		
O	${}^4\chi_{pc}^b - {}^6\chi_{pc}^b$		
M_1	${}^0\chi^v - {}^6\chi^V$		
M_2	${}^3\chi_c^v$ and ${}^5\chi_c^v$		
${}^0\chi - {}^6\chi$	${}^4\chi_{pc}^v - {}^6\chi_{pc}^v$		
${}^3\chi_c$ 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: $SIC_0, SIC_1, SIC_4, CIC_0, {}^2\chi^b, {}^5\chi^bC, {}^5\chi^vC, {}^6\chi^v_{PC}$, 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_{HOMO}, $E_{HOMO\ 1}$, E_{LUMO}, $E_{LUMO\ 1}$, 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 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, ${}^2\chi^b$, 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_{LUMO1}	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.

Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters

Krishnan Balasubramanian and Subhash C. Basak

Department of Chemistry, Arizona State University, Tempe, Arizona 85287-1604, and Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Journal of
Chemical Information and Computer Sciences®

Reprinted from
Volume 38, Number 3, Pages 367–373

Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters

Krishnan Balasubramanian† and Subhash C. Basak*,‡

Department of Chemistry, Arizona State University, Tempe, Arizona 85287-1604, and
Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received July 1, 1997

Numerical graph theoretic invariants or topological indices (TIs) and principal components (PCs) derived from TIs have been used in discriminating a set of isospectral graphs. Results show that lower order connectivity and information theoretic TIs suffer from a high degree of redundancy, whereas higher order indices can characterize the graphs reasonably well. On the other hand, PCs derived from the TIs had no redundancy for the set of isospectral graphs studied.

1. INTRODUCTION

Graph theoretical and topological techniques have been harnessed in numerous practical applications in recent years. In particular, the use of graph theoretical techniques for the characterization of structures and for the exploration of structure−property relations have received considerable attention.[1−24] The intimate relation between the structure of a molecule and its activity has been the topic of exploration for many years. Several novel techniques based primarily on graph theory and topology have been proposed for predicting activities from the structure, and such techniques have been successfully applied to molecules of pharmacological relevance.

Since graph theoretical techniques are based on the topological connectivity of a molecule rather than its three-dimensional molecular structure, there is always a question as to the suitability of a graph theoretically based technique for the characterization or prediction of properties that may depend on more complex factors than simple connectivity. For this reason techniques based on the three-dimensional molecular geometry have been proposed.[21−23]

A recognized problem with graph-theoretically based technique is in dealing with graphs called isospectral graphs.[24−27] Isospectral graphs are graphs with the same characteristic polynomial which is simply the secular determinant of the adjacency matrix of a graph. Thus isospectral graphs would have the same graph eigenvalues or spectra, which could be visualized as the Huckel energy levels associated with the molecule corresponding to the graphs under consideration. The isospectral graphs have thus received much attention due to their "pathological" nature. Prior to the discovery of isospectral graphs it was surmised that the characteristic polynomials or spectra might uniquely characterize graphs, but examples of isospectral graphs revealed that there are pairs of nonisomorphic graphs which are topologically distinct and yet they have the same characteristic polynomials and spectra. As a result of this isospectral graphs pose several problems. As discussed in the work of Liu et *al.*,[24] some of the vertex partitioning algorithms fail for isospectral graphs. Likewise, the topologically based indices such as the Wiener index[3] become identical for isospectral graphs.

Basak et al.[28] used a combination of graph invariants to characterize a large collection of complex graphs. The principal component analysis (PCA) which is performed on the basis of these indices and the Euclidian distance method have provided a promising avenue for the characterization of structures and structure−activity relationships. Thus, it is interesting to explore if these techniques are satisfactory for isospectral graphs which are considered to be pathological in a graph theoretical sense. The objective of this study is to consider a series of isospectral graphs for the purpose of computing these indices and the PCA on those indices. We show that while lower-order indices often fail to discriminate isospectral graphs, the PCs derived from indices discriminate all isospectral graphs considered here.

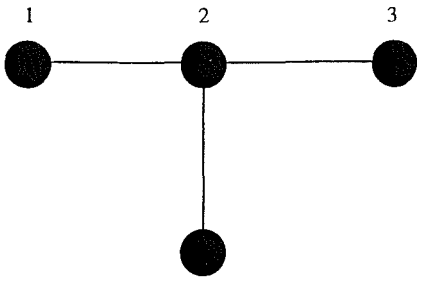
2. CALCULATION OF GRAPH THEORETICAL PARAMETERS

The calculation of the topological indices (TIs) used in this study has previously been described in detail.[1] The TIs for the isospectral pairs of graphs were calculated by POLLY.[2] The POLLY 2.3 version is capable of calculating 97 TIs from the SMILES line notation input of chemical structures. The TIs calculated by POLLY 2.3 include the Wiener index,[3] connectivity indices,[4,5] and information theoretic indices defined on distance matrices of graphs[6,7] as well as a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs.[8−11] We describe below the methods for the calculation of the TIs used in this paper.

The Wiener index W,[3] the first topological index reported in the chemical literature, may be calculated from the distance matrix $\mathbf{D}(\mathbf{G})$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $\mathbf{D}(\mathbf{G})$ of a nondirected graph G with n vertices is a real symmetric $n \times n$ matrix with

* Corresponding author.
† Arizona State University.
‡ University of Minnesota.

S0095-2338(97)00052-8 CCC: $15.00
Published on Web 04/21/1998



G_1

Figure 1. Hydrogen suppressed graph of isobutane.

elements d_{ij} 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 $\mathbf{D}(\mathbf{G}_1)$ of the unlabeled hydrogen-suppressed graph $\mathbf{G}_1$ of isobutane (Figure 1):

$$D(\mathbf{G}_1) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix} \end{array}$$

W is calculated as

$$W = {}^1/_2 \sum_{ij} d_{ij} = \sum_h h \cdot g_h \qquad (1)$$

where g_h is the number of unordered pairs of vertices whose distance is h.

Randićs[4] connectivity index as well as the higher-order path, cluster, and path-cluster types of simple and valence connectivity indices developed by Kier and Hall[5] were calculated by the computer program POLLY.[2] P_h parameters, the number of paths of length h ($h = 0-10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Information-theoretic topological indices are calculated by the application of information theory to 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 ($i = 1,2, \ldots h$; $\sum n_i = n$). A probability distribution is then assigned to the set of equivalence classes

$$A_1, A_2, \ldots\ldots, A_h$$

$$p_1, p_2, \ldots\ldots, p_h$$

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[12] relation

$$IC = -\sum_{i=1}^{h} p_i \log_2 p_i \qquad (2)$$

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 IC.

Rashevsky[13] was the first to calculate the information content of graphs where "topologically equivalent" vertices are placed in the same equivalence class. 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. Subsequently, Trucco[14] defined topological information of graphs on the basis of graph orbits. In this method, vertices which belong to the same orbit of the automorphism group are considered topologically equivalent. While Rashevsky[13] 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.*[15] 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 physicochemical 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 $d(v,vi) < r$. Then, $S(v,0) = \phi$, $S(v,r) = v$ 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 for $1 < r < 2$.

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 topological neighborhoods of vertices of up to rth order neighbors. Such an approach has already been initiated, and the information-theoretic indices calculated are called indices of neighborhood symmetry.[10]

In this method, chemical species 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 the rth order neighborhood if, and only if, corresponding to each path $u_o, u_1, \ldots, u_r$ of length r, there is a distinct path $v_o, v_1, \ldots, v_r$ 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 rth order bonded neighbors. The detailed equivalence relation is described in our earlier studies.

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated from eq 2. Basak, Roy, and Ghosh[9] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as

$$SIC_r = IC_r/\log_2 n \qquad (3)$$

Table 1. Topological Indexes: Symbols and Definitions

I^{w}_{D}	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}^{w}_{D}$	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
$\overline{IC}$	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 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 $h = 5-6$
$^{h}\chi_{PC}$	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

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),[11] is defined as

$$CIC_r = \log_2 n - IC_r \tag{4}$$

CIC_r represents the difference between the 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.

The information-theoretic index on graph distance, I^{W}_{D}, is calculated from the distance matrix **D(G)** of a chemical graph **G** by the method of Bonchev and Trinajstić:[7]

$$I^{W}_{D} = W \log_2 W - \sum_h g_h \cdot h \log_2 h \tag{5}$$

The mean information index, $\bar{I}^{W}_{D}$ is found by dividing the information index I^{W}_{D} by W. IC_r, SIC_r, CIC_r, I^{W}_{D}, and $\bar{I}^{W}_{D}$ were calculated by Polly.[6] The information theoretic parameters defined on the distance matrix, H^{D} and H^{v} were calculated by the method of Raychaudhury et *al.* Sixty TIs were calculated for each of the 38 molecular graphs in Figure 2.

3. STATISTICAL ANALYSIS

3.1. Data Reduction. The TIs used in this paper are shown in Table 1. Initially, all TIs were transformed by the natural logarithm of the value of the index plus one. This was done because the scale of some TIs may be several orders of magnitude greater than others.

3.2. Principal Components Analysis. The data for the isospectral graphs analyzed in this paper may be viewed as n (number of isospectral graphs) vectors in p (number of calculated parameters) dimensions. The data for each set can be represented by a matrix **X** which has n rows and p columns. For each of the graphs, the number of calculated parameters was 60 (TIs of Table 1). Each graph is therefore represented by a point in R^{60}, where R is the field of real numbers. If each graph s was represented in R^2, then one could plot and investigate the extent of relationship between individual parameters. In R^{60} such a simple analysis is not possible. However, since many of the TIs are highly intercorrelated, the points in R^{60} can likely be represented by a subspace of fewer dimensions. The method of PCA or the Karhunen-Loeve transformation is a standard method for reduction of dimensionality.[29] The first principal component (PC) is the line which comes closest to the points in the sense of minimizing the sum of the squared Euclidean distances from the points to the line. The second PC is given by projections onto the basis vector orthogonal to the first PC. For points in R^p, the first r PCs give the subspace which comes closest to approximating the n points. The first PC is the first axis of the points. Successive axes are major directions orthogonal to previous axes. The PCs are the closest approximating hyperplane, and because they are calculated from eigenvectors of a $p \times p$ matrix, the computations are relatively accessible. But there are important scaling choices, because PCs are scale dependent. To control this dependence, the most commonly used convention is to rescale the variables so that each variable has a mean of zero and a standard deviation of one. The covariance matrix for these rescaled variables is the correlation matrix. The PCA on the TIs for isospectral graphs has been carried out using SAS software.[30]

Table 2. Summary of Principal Components Analysis

	eigenvalue	% cumulative varnce explnd		eigenvalue	% cumulative varnce explnd
PC_1	34.1	56.8	PC_4	2.8	93.2
PC_2	14.4	80.8	PC_5	1.3	95.3
PC_3	4.6	88.5	PC_6	1.1	97. 1

4. RESULTS

The summary of PCA using 60 calculated TIs is shown in Table 2. The first three PCs explain nearly 90% of the variance in the data and the first six PCs with eigenvalue greater than 1.0 explain about 97% of the variance in the original data.

In Table 3 we give the values for PC_1-PC_6 for the 38 graphs analyzed in this paper. It is interesting to note that almost all PCs have distinct values for pairs (*e.g.*, 1.1 and 1.2; 2.1 and 2.2, etc.) of isospectral graphs.

Table 4 presents the values of connectivity indices $^{0}\chi-^{2}\chi$ and neighborhood complexity indices IC_0-IC_2 for the graphs.

Table 3. First Six PCs for the Set of 38 Isospectral Graphs (Figure 2)

graph	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6
1.1	−10.6828	−1.5214	0.0283	−2.3056	−0.2901	−0.7411
1.2	−11.2419	−0.7289	0.6454	−2.4077	−1.3287	−1.2562
2.1	−7.5623	−2.8765	0.4809	0.4976	−1.4914	1.6824
2.2	−7.6856	1.4163	−1.0238	−0.9141	−0.5908	−0.4823
3.1	1.6223	−1.7614	−4.5762	−0.4737	1.3809	0.1826
3.2	1.4956	−3.7201	−2.6087	0.5261	0.4641	2.0115
4.1.1	−2.1141	0.3656	−2.2068	0.2315	−0.1458	−0.3351
4.1.2	−2.5286	2.2386	−1.0309	−0.4577	−0.5908	−0.4608
4.2.1	−2.5555	−3.2923	3.9820	1.9017	−0.0220	0.5264
4.2.2	−2.4859	0.7047	−0.5478	0.1951	−1.4380	0.5363
5.1	−7.4612	−0.3102	−0.9097	−0.3816	0.8077	0.1601
5.2	−7.7603	0.9300	−0.9975	−1.7106	−0.3964	−1.3015
6.1	−5.8986	−0.5274	−0.4014	−1.1701	−0.3234	−0.3493
6.2	−5.8739	−5.7170	1.7090	0.1934	−0.2359	1.5281
7.1.1	4.1610	2.2536	0.1775	1.0976	−2.6734	0.1861
7.1.2	4.2882	4.4784	−1.1182	0.2768	0.0386	−2.4036
7.2.1	4.3117	3.0509	−0.2194	0.9898	0.1809	−1.9833
7.2.2	4.3284	3.2733	−0.9286	0.7415	−1.0757	−1.0430
8.1	−8.8239	5.4954	1.2720	4.7684	−0.1428	1.0801
8.2	−8.0694	4.3139	−1.5231	5.6667	3.1130	0.5582
9.1.1	0.6468	4.4113	3.3448	−2.6882	1.9146	−0.3797
9.1.2	1.2862	5.5270	1.7360	−3.5416	2.8117	3.1329
9.2.1	0.1561	0.2784	−0.9364	−0.8100	−0.5981	0.0892
9.2.2	−0.1287	0.9325	1.8555	−0.6934	0.5959	−1.1643
9.3.1	−0.3873	−0.1603	3.0373	−0.3006	−3.1157	0.9897
9.3.2	−0.2827	−0.6395	2.8592	−0.3025	−0.7054	−0.9675
10.1.1	7.5296	−3.0998	3.9813	1.5925	0.9975	−2.0763
10.1.2	7.6726	1.3574	−1.8310	−0.5161	−0.4564	−0.9028
10.2.1	8.3168	0.2849	4.3189	−0.5465	1.2660	0.6830
10.2.2	8.8218	3.3376	0.1809	−1.8163	0.8456	2.0753
10.3.1	7.9681	0.0713	−2.0128	0.5599	−1.5890	0.9229
10.3.2	7.5192	−2.1439	2.0070	1.3297	−1.9649	0.1591
10.4.1	7.9848	−1.1899	−1.6830	0.4285	−1.7291	1.5518
10.4.2	8.0537	−1.6182	−2.0384	0.4788	0.1030	−0.6392
11.1.1	1.2742	−2.3342	−2.4558	−1.0802	1.2188	−0.2514
11.1.2	1.2530	−7.5213	2.1144	0.9705	3.0859	−1.1629
11.2.1	1.5423	−3.1237	−3.2945	−0.4177	1.3260	0.1898
11.2.2	1.3098	−3.7138	−1.3866	0.0878	0.7538	−0.3457

Table 4. Selected Topological Indices for 38 Isospectral Graphs (Figure 2)

graph	$^0\chi$	$^1\chi$	$^2\chi$	IC_0	IC_1	IC_2
1.1	8.690	5.219	3.859	0.898	1.368	2.665
1.2	8.690	5.240	3.812	0.898	1.368	2.701
2.1	8.975	5.812	4.424	0.918	1.418	2.675
2.2	8.975	5.791	4.502	0.918	1.418	2.828
3.1	11.380	7.847	6.318	0.932	1.384	2.726
3.2	11.380	7.826	6.396	0.932	1.384	2.664
4.1.1	9.966	6.847	5.610	0.934	1.417	2.784
4.1.2	9.966	6.826	5.689	0.934	1.417	2.765
4.2.1	9.966	6.864	5.526	0.934	1.417	2.684
4.2.2	9.966	6.864	5.526	0.934	1.417	2.684
5.1	8.975	5.753	4.643	0.918	1.418	2.807
5.2	8.975	5.774	4.575	0.918	1.418	2.717
6.1	9.682	6.291	4.856	0.918	1.404	2.789
6.2	9.682	6.312	4.766	0.918	1.404	2.565
7.1.1	11.121	7.809	6.906	0.946	1.457	2.794
7.1.2	11.121	7.809	6.908	0.946	1.457	2.982
7.2.1	11.121	7.809	6.896	0.946	1.457	2.856
7.2.2	11.121	7.809	6.896	0.946	1.457	2.856
8.1	7.845	5.326	4.628	0.938	1.469	2.802
8.2	7.845	5.326	4.618	0.938	1.469	2.995
9.1.1	10.889	7.232	6.134	0.933	1.517	2.978
9.1.2	10.889	7.220	6.193	0.933	1.517	2.885
9.2.1	10.836	7.258	6.116	0.933	1.458	2.928
9.2.2	10.836	7.236	6.194	0.933	1.458	2.928
9.3.1	10.836	7.274	6.041	0.933	1.458	2.864
9.3.2	10.836	7.274	6.004	0.933	1.458	2.974
10.1.1	12.535	8.847	7.431	0.943	1.429	2.664
10.1.2	12.535	8.809	7.594	0.943	1.429	2.729
10.2.1	12.588	8.805	7.518	0.943	1.483	2.764
10.2.2	12.588	8.815	7.482	0.943	1.483	2.764
10.3.1	12.535	8.847	7.443	0.943	1.429	2.760
10.3.2	12.535	8.847	7.441	0.943	1.429	2.729
10.4.1	12.535	8.847	7.431	0.943	1.429	2.664
10.4.2	12.535	8.830	7.516	0.943	1.429	2.769
11.1.1	11.380	7.809	6.458	0.932	1.384	2.589
11.1.2	11.380	7.830	6.378	0.932	1.384	2.438
11.2.1	11.380	7.847	6.306	0.932	1.384	2.622
11.2.2	11.380	7.847	6.308	0.932	1.384	2.595

For most of the isospectral pairs, $^0\chi$, $^1\chi$, IC_0, and IC_1 could not discriminate between the isospectral pairs, whereas $^2\chi$ as well as complexity parameter IC_2 could discriminate the isospectral pairs reasonably well in most cases.

We retained the first six PCs with eigenvalues > 1.0. This is a substantial reduction in the number of parameters or the dimensionality of the parameter space as compared to the 60-dimensional space corresponding to the 60 TIs calculated originally. Our earlier work on PCA using large and diverse sets of molecular graphs show that a few first PCs explain a large fraction of the variance.[16-20]

In some of their earlier papers, Basak *et al.*[16-20] used the Euclidean distance (ED) in the *n*-dimensional PC-space in characterizing structural similarity/dissimilarity of molecules. In Table 5 we give the ED between 19 isospectral pairs of graphs. For all pairs of graphs considered in this paper, the value of ED was nonzero which shows the discriminating ability of the six-dimensional PC-space generated out of the calculated PCs.

Results and Discussion. We have considered a series of pairs of isospectral graphs shown in Figure 2. In this figure we have used the numbering convention *i.j.k*, where *i* is the same for two isospectral graphs. Based on the relation between the isospectral graphs, the index *j* will be kept the same if the two are closely related; in this case only the index *k* would differ. Thus we have isospectral graphs 9.1.1., 9.1.2, 9.2.1, 9.2.2, 9.3.1, and 9.3.2. As seen from Figure 2, 9.3.1 and 9.3.2 are more closely related compared to 9.1.1 and 9.3.1. Recall that the isospectral graphs have the same characteristic polynomials and spectra. Furthermore, many parameters computed based on the adjacency matrices of two isospectral graphs are identical. Commonly used topological indices such as the Wiener index, Randić's connectivity index, spectral index, indices based on path numbers, etc., become identical for such graphs. Consequently, many ordinary graph-theoretically based indices fail to discriminate isospectral graphs.

Table 5. Euclidean Distance in 7-Dimensional Principal Component Space for 19 Isospectral Graph Pairs

isospectral pairs		Euclidean distance	isospectral pairs		Euclidean distance
1.1	1.2	0.2142	9.1.1	9.1.2	0.6877
2.1	2.2	0.5781	9.2.1	9.2.2	0.4352
3.1	3.2	0.4627	9.3.1	9.3.2	0.5281
4.1.1	4.1.2	0.2230	10.1.1	10.1.2	0.8340
4.2.1	4.2.2	0.6627	10.2.1	10.2.2	0.5988
5.1	5.2	0.3705	10.3.1	10.3.2	0.5130
6.1	6.2	0.5929	10.4.1	10.4.2	0.4958
7.1.1	7.1.2	0.6831	11.1.1	11.1.2	0.7672
7.2.1	7.2.2	0.2773	11.2.1	11.2.2	0.2627
8.1	8.2	0.6324			

We have computed the connectivity indices $^0\chi$, $^1\chi$, and $^2\chi$ as well as the neighborhood complexity indices IC_0, IC_1, and IC_2 that are defined in the previous section for these

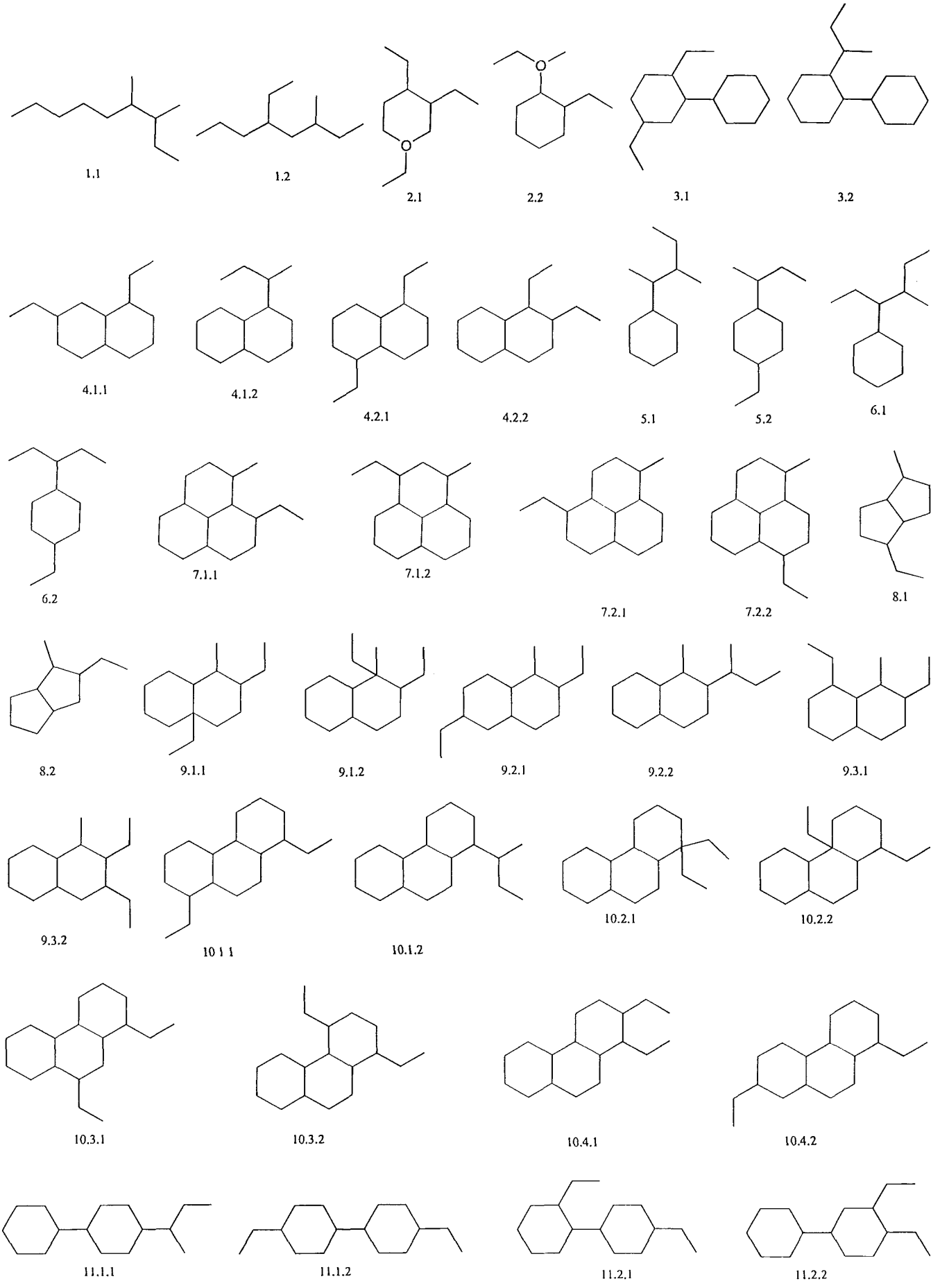


Figure 2. Structures of 38 isospectral graphs.

isospectral graphs shown in Figure 2. The isospectral graphs in Figure 2 are generated by attaching the same fragment at vertices called the isospectral vertices. As discussed before in the literature, certain vertices in some graphs are called isospectral vertices. For example, consider the graphs 2.1 and 2.2 in Figure 2. These two graphs are generated by attaching a fragment containing two vertices connected by a bond to either the para position of the six-membered ring, as in the graph 2.1 in Figure 2 (where the para position is defined as the fourth vertex in 2.1) or by attaching the same fragment to the other circled vertex of the pending fragment which results in the graph 2.2 in Figure 2. All of the isospectral graphs in Figure 2 are constructed in this manner by attaching an identical fragment to one of the isospectral vertices.

Table 4 shows the computed values for the indices ${}^0\chi$, ${}^1\chi$, and ${}^2\chi$ as well as the neighborhood complexity indices IC_0, IC_1, and IC_2. First let us discuss the discriminating powers of these indices before proceeding to the PCA. As seen from Table 4, the index ${}^0\chi$ is the least discriminating while ${}^2\chi$ is somewhat more discriminating. For all isospectral pairs of graphs the ${}^0\chi$ indices are identical as expected since the ${}^0\chi$ index is based on simple topological connectivity.

It is seen from Table 4 that although the ${}^2\chi$ index is relatively more discriminating compared to the ${}^0\chi$ index, the actual ${}^2\chi$ values are numerically too close for some of the isospectral graphs to consider these values to be truly discriminating. This is particularly exemplified by the graphs 11.2.1 and 11.2.2 whose ${}^2\chi$ values are 6.306 and 6.308, respectively (see Table 4). Likewise the ${}^2\chi$ values for the graphs 10.3.1 and 10.3.2 are 7.443 and 7.441, respectively. The ${}^2\chi$ values for the graphs 7.2.1 and 7.2.2. are identical (6.896). Likewise the ${}^2\chi$ values for the graphs 4.2.1 and 4.2.2 are the same (5.526). However, for other graphs considered here the ${}^2\chi$ values are more discriminating. Consequently, it is concluded that although the ${}^2\chi$ values are more discriminating than the zeroth-order index, these values are still not sufficiently discriminating for more complex isospectral graphs, although these indices work well for simpler isospectral graphs, as seen from Table 4.

As evidenced from Table 4, the neighborhood complexity indices IC_0, IC_1, and IC_2 have some similarity to the χ indices in that the higher-order indices are slightly more discriminating compared to the lower-order indices. Thus the IC_0 and the IC_1 indices do not discriminate isospectral graphs at all (see, Table 4). When ${}^2\chi$ is identical, IC_2 is as well. When ${}^2\chi$ is nearly identical, IC_2 is slightly more discriminating.

Since neither the ${}^n\chi$ indexes nor the IC_r indexes seem to be fully satisfactory in terms of discriminating complex isospectral graphs, it was decided to carry out the PCA on these graphs using the indices computed thus far. The philosophy behind the PCA technique and the algorithms derived from the technique have been illustrated in the previous section. The procedure uses an n-dimensional space of these indices and computes the Euclidian distances.

Table 3 shows the numerical values for the first six PCs which are labeled PC_1 through PC_6 in Table 4 for the isospectral graphs that are considered in this study. In this analysis we retained only the first six PCs with eigenvalues > 1.0 which leads to a substantial reduction in the number of parameters or the dimensionality of the parameter space as compared to the original 60-dimensional parameter space that we begin with. Earlier work on PCA using large and diverse sets of molecular graphs show that the first few PCs explain a large fraction of the variance.[17−20]

As seen from Table 3, the PC indices are far more powerful and discriminating compared to the simple topological indices considered in Table 4. Let us consider graphs 11.2.1 and 11.2.2 which are considered to be "pathological" from numerical and similarity standpoints in that the χ values and IC_r values are virtually the same. However, as seen from Table 3, the PC_1 and PC_2 values are very different (PC_1: 1.5423, 1.3098; PC_2: −3.1237, −3.7138). As a matter of fact all of the PC_1 through PC_6 values are sufficiently different to discriminate these isospectral graphs.

Let us consider graphs 7.2.1 and 7.2.2 that are not discriminated by their ${}^2\chi$ values. As seen from Table 3, while the PC_1 values for these two graphs are somewhat close (4.3117 and 4.3284) their PC_2 values are 3.0509 and 3.2733. Other higher order PC values differ even more thereby providing a sound and powerful basis for discriminating isospectral graphs.

Next we consider the pairs 4.2.1 and 4.2.2. These two graphs have identical ${}^2\chi$ values and IC_2 values. However, as seen from Table 4 these graphs have very different PC values for all n. Thus PCA seems to be a powerful technique to discriminate even isospectral graphs that are not so easily contrasted by topologically based techniques.

It should be pointed out that for a few isospectral graphs the first principal component value, PC_1 is not as discriminating as the higher-order PCs values. For example, the PC_1 values for the isospectral graphs 2.1 and 2.2 are −7.5623 and −7.6856, respectively. However, the PC_2 values are −2.8764 and 1.4163 for the same graphs. Likewise the graphs 7.2.1 and 7.2.2 have the PC_1 values of 4.3117 and 4.3284. However their PC_2 values are 3.0509 and 3.2733. We thus conclude that one needs more than the PC_1 value to discriminate complex isospectral graphs, but often the PC_2 values for those graphs are sufficiently different to contrast them.

ACKNOWLEDGMENT

The research effort of Subhash C. Basak reported in this paper was supported by cooperative agreement CR 819621 from the United States Environmental Protection Agency, Grant F49620-94-1-0401 from the United States Air Force and Exxon Biomedical, Inc., through the structure−activity relationship consortium (SARCON) of the Natural Resources Research Institute. This is contribution number 222 from the Center of Water and the Environment of the Natural Resources Research Institute.

REFERENCES AND NOTES

(1) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-activity Studies. *J. Chem. Ink Comput. Sci.* **1994**, *34*, 270−276.

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

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

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

(5) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis;* Research Studies Press: Letchworth, Hertfordshire, England, 1986.

(6) 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.
(7) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. J. *Chem. Phys.* 1977, *67*, 4517−4533.
(8) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* **1987**, *15*, 605−609.
(9) 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.,Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, Missouri, 1980; pp 851−856.
(10) 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., Rodin, E. Y., Eds.; Pergamon Press 1984; pp 745−750.
(11) 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./ Drug Res.* **1983**, *33*, 501−503.
(12) Shannon, C. E. A Mathematical Theory of Communication. *Bell Sys. Tech. J.* **1948**, *27*, 379−423.
(13) Rashevsky, N. Life, Information Theory and Topology. *Bull. Math. Biophys.* **1955**, *17*, 229−235.
(14) Trucco, E. A Note on Rashevsky's Theorem about Point Bases in Topological Biology. *Bull. Math. Biophys.* **1956**, *18*, 65−85.
(15) Sarkar, R.; Roy, A. B.; Sarkar, P. K. Topological Information Content of Genetic Molecules-I. *Math. Biosci.* **1978**, *39*, 299−312.
(16) 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.
(17) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathl. Model. Sci. Comput.* **1994**, in press.
(18) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity from Structural Similarity. *New* J. *Chem.* **1995**, *19*, 231−237.
(19) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR QSAR Environ. Res.* **1994**, *2*, 289−307.
(20) 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.
(21) Balasubramanian, K. *Chem. Phys Lett.* **1990**, *169*, 224; **1995**, *232*, 415.
(22) Balasubramanian, K. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 243.
(23) Balasubramanian, K. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 761.
(24) Liu, X.; Balasubramanian, K.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 263.
(25) Herndon, W. C. *Inorg. Chem.* **1983**, *22*, 654.
(26) Herndon, W. C. In *Chemical Applications of Topology & Graph Theory;* King, R. B., Ed.; Elsevier: Amsterdam, 1983.
(27) Razinger, M.; Balasubramanian, K.; Munk, M. C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 197.
(28) Balaban, A. T.; Liu, X.; Klein, O. J.; Babic, D.; Schmalz, T. G.; Seitz, W. A.; Ravic, M. Graph invariants for Fullerenes. *J. Chem. Inf. Comput. Sci,* **1995**, *35*, 396.
(29) Gnanadesikan, R. Methods for Statistical Analysis of Multivariate Observations; J Wiley: New York, 1977.
(30) SAS/STAT User's Guide. Version 6, 4th ed.; SAS Institute Inc.: Cary, NC, 1989; Vol. 2, p 846.

CI970052G

J. Chem. Inf. Comput. Sci., submitted

Optimal Molecular Descriptors Based on Weighted Path Numbers

Milan Randić and Subhash C. Basak[1]

Dept. of Mathematics and Comput. Sci.,
Drake University, Des Moines IA 50311

1Natural Resources Research Institute, Center for Water and the Environment,
University of Minnesota, Duluth MN 55811

ABSTRACT

We consider weighted path numbers as molecular descriptors for structure-property-activity studies. However, instead of using prescribed weights for paths we will optimize the weights so that the standard error in regression analysis is as small as possible. In particular we consider use of weighted paths to differentiate oxygen from carbon atoms in alcohols.

INTRODUCTION

Despite the large number of molecular descriptors available for use in multiple regression analysis (MRA) we still have no good descriptors for many molecular properties of interest. Most descriptors are size dependent and can produce apparently good correlations when used for samples of compounds that include molecules of different size. However, when the same descriptors are used to describe variations in properties among molecules of the same size, for example isomeric variations, then they show limitations. This is reflected in low coefficient of regression, low Fisher ratio, and generally the standard error which are comparable to the variations of the property considered. To illustrate the point consider correlation of molar refraction Rm of alkanes with the connectivity indices ${}^1\chi$ [1] and ${}^2\chi$ [2]. According to Kier and Hall [3] for a sample of alkanes that included molecules having from five to nine carbon atoms (the number of molecules in the sample was 46) a correlation with single connectivity index gives:

$$Rm = 10.440\ {}^1\chi + 1.369 \qquad \text{with } r = 0.9520 \text{ and } s = 1.81 \qquad (1)$$

where r is the coefficient of regression and s is the standard error. When the same data is regressed using two connectivity indices one obtains [ref. 3 p.108]:

$$Rm = 2.207\ {}^2\chi + 7.756\ {}^1\chi + 3.707 \qquad \text{with } r = 0.9997 \text{ and } s = 0.121. \qquad (2)$$

Superficially the eq. (1) appears fair if you recollect that for the set of compounds considered molar refraction varies from 25.267 (n-pentane) to 48.757 (2,2,5,5-tetramethylhexane). However, in fact ${}^1\chi$ only well correlates with the molecular size. If we look at the variation of the property (Rm)

within the isomers, for the 17 octane isomers included in the regression Rm varies from 38.719 (3-methyl-3-ethylpentane) to 39.264 (2,2,4-trimethylpentane). This gives the maximal difference for Rm of only 0.545, which is well within the standard error of the regression based on 1χ (s = 1.81).

Use of two descriptors substantially improves the correlation as is evident by the decrease in the standard error by an order of magnitude when both 1χ and 2χ are used. Apparently when we use two descriptors that individually correlate well with size (for octane isomer, hence molecues of the same size, the correlation between 1χ and 2χ is very high r = 0.9757), when combined it appears that they are able to account well for the isomeric variations of Rm, i.e., for the variation of Rm with the shapes among isomers. How is this to be understood?

One should recall that the connectivity index 1χ was designed so that it parallel isomeric variations in the boiling points of smaller alkanes. All the bond have first been classified in (m, n) bond types. For example, in 3-methylhexane (Fig. 1) we have: 2 (1, 2) + (1, 3) + (2, 2) + 2 (2, 3) while for n-heptane we have: 2 (1,2) + 4 (2,2). Because the boiling point of n-heptane is greater than that of 3-methylhexane we require that

$$2\,(1,2) + 4\,(2,2) > 2\,(1,2) + (1,3) + (2,2) + 2\,(2,3)$$

where bond types (1, 2), (2, 2), (1, 3), and (2, 3) are considered unknown variables yet to be determined so that the above inequality is satisfied. When similar inequalities based on other hexane and heptane isomers are constructed one find, as outlined in ref. [1], that the algorithm $(m\ n) = 1/\sqrt{mn}$ (but not for example: $(m\ n) = 1/mn$, or $(m\ n) = 1/(m + n)$, etc.) satisfy all the considered inequalities. It is not surprising therefore that the connectivity index can reasonably well correlates with the boiling points of

smaller alkanes. Kier and Hall [3] made comparison of the performance of several simple topological indices, and found that $^1\chi$ outperforms others -- well, it was designed to do that!

Equally, it should not be surprising that when larger alkanes are considered no single descriptor will perform admirably. The reason for this is that as the size of molecules increases *additional* structural elements play a role besides the connectivity and branching that simple indices consider. One such factor is the crowding of atoms (close methyl groups). Wiener [4] was very successful with his molecular descriptors W and P precisely so because his second descriptor P (which counted paths of length three in a molecule) well characterized the crowing of atoms. A pair of well selected descriptors have a better chance to account for a pair of molecular structural features: connectivity (i.e., branching) and crowding. One such pair is W, P, the other pair is $^1\chi$ and $^2\chi$. Hence, use of $^1\chi$ and $^2\chi$ for correlating the boiling points of alkanes is expected to produce good correlation, even if not surpassing the original correlation of the boiling points of alkanes as reported by Wiener's classic work.

The question we posed is why two descriptors, in particular $^1\chi$ and $^2\chi$, could correlate with molecular isomeric variations, while individually each fails to show any such trend. The failure of 1χ or 2χ to separately correlate with Rm obscures the fact that these descriptors are sensitive to molecular branching. If we correlated Rm against the boiling points, as illustrated in Fig. 2 for 17 octane isomers (Rm for 2,2,3,3-tetramethylbutane apparently was not available in Kier and Hall's study reviewed here) we see that there is no correlation. Clearly the two properties are dominated by *different* structural factors. One could say, perhaps at the risk of becoming somewhat simplistic, that the boiling points, the heats of atomization, and many other molecular

thermodynamic properties critically depend on molecular surface, while in the case of molar refraction, index of refraction, and molecular liquid density the dominant structural component is molecular volume, not the surface. Similarly the melting points critically depend on molecular shape and the packing, which calls for additional considerations, and not surprisingly melting points fail to be well described by simple topological indices.

If we want to examine closely how various descriptors describe molecular branching, i.e., the variations among isomers, we should pay attention to molecules of the same size and apply descriptors to various properties of such reduced samples, e.g., octanes, rather than considering alkanes in general. To illustrate the point consider again correlations of molar refraction Rm with the connectivity indices $^1\chi$ and $^2\chi$, but instead of alkanes view only octane isomers. Now the correlation with single connectivity index gives (Fig. 3):

$$Rm = 0.1150\ ^1\chi + 38.6065 \qquad \text{with } r = 0.0867 \text{ and } s = 0.187 \qquad (3)$$

where r is the coefficient of regression and s is the standard error. When the same data is regressed using two connectivity indices one obtains:

$$Rm = 1.3717\ ^2\chi + 4.6939\ ^1\chi + 17.5539 \qquad \text{with } r = 0.9708 \text{ and } s = 0.047. \qquad (4)$$

The eq. (1) clearly shows no correlation, even though in comparison with eq. (1) the standard errors has decreased tenfold! In the eq. (4) the standard error for the first time becomes respectable. However, if one uses Wiener index W and P one obtains:

$$Rm = -0.0050\ W - 0.1396\ P + 40.3148 \qquad \text{with } r = 0.9947 \text{ and } s = 0.020. \qquad (5)$$

which is even better, the standard error being reduced more than by factor of two.

In summary, the critical factor for quality regression analysis is the choice of molecular descriptors. This can be succinctly stated as: "If your experiment needs better statistics, you ought to have used better descriptors" [5]. This is a paraphrase of a hostile comment of Rutherford [6] on statistics at the time when statistic was not fully appreciated: "If your experiment needs statistics, you ought to have done better experiment."

OPTIMIZED MOLECULAR DESCRIPTORS

In Table 1 we illustrate a number of properties of octanes and the best single descriptor. As we see from several hundreds of the descriptors reported in the literature at most a dozen appear to emerge as the best single characterization of diverse physico-chemical properties of octanes. Are the descriptors shown in Table 1 the best, or could there be descriptors that are even better, but we have hitherto not been successful to find them?

Two questions we want to consider in here:

(1) How to search for the best descriptors, whether single or combined with others, that can best describe physicochemical properties of alkanes.

(2) How to extend such descriptors to molecules having double bonds, triple bonds, aromatic CC bond, and heteroatoms.

In contrast to the prevailing practice, which has its advantages and disadvantages, in which one selects few descriptors from a large pool of descriptors, we will consider the other extremal view: Use of as few descriptors as possible. In doing this nevertheless we will require that descriptors have a direct, even if not necessarily transparent, structural

meaning. Use of minimal number of descriptors has not only the obvious statistical advantage but it may allow simpler and easier interpretation of the resulting regression equation. How can one plan to have very good regression with few descriptors where others used many? The answer is in optimization of molecular descriptors for the particular application.

Search for optimized molecular descriptors has been outlined in QSAR [5, 7-9] but apparently have not yet received due attention. There are at least three distinctive routes to optimization of descriptors:

(1) One may try to optimize the functional form [7]. For example, instead of using the connectivity index 1χ, the Wiener index W, and the Hosoya index Z [10], as single descriptor one can consider various powers of such indices. It turns out in the case of the boiling points of smaller alkanes (from ethane to heptane isomers) the smallest standard deviations are found when one uses $\chi^{1/2}$ (s = 2.83), $W^{1/4}$ (s = 4.42) and $Z^{-1/3}$ (s = 3.54) [7].

(2) One may try to optimize the diagonal entries of an adjacency matrix so to differentiate between atoms of different time [8]. For example, when considering the boiling points of hexanols if one does not differentiate between carbons and oxygen one obtains for standard error s = 7.86 oC when the connectivity index is used as the descriptor. If however the connective index is constructed from the adjacency matrix in which entries on the main diagonal are viewed as two variables (one for carbon atoms and one for oxygen) so modified connectivity index leads to regression with the standard error of only 3.43 oC [8].

(3) One may try to optimize off-diagonal entries of adjacency matrix, i.e., one can introduce variable weights for bonds of different kind [9]. For example, let us consider molar refraction Rm for akenes. Kier and Hall [4] reported three parameter and six parameter equations using a combination of

valence connectivity indices and ordinary connectivity indices which produced standard error s = 0.233 and s = 0.147 respectively. When instead of the connectivity indices one use path numbers the standard error decreases from s = 0.171 when a single path number is used, to s = 0.139 when four path numbers are used [9]. Clearly the path numbers outperform the connectivity indices in this particular application even though they do not differentiate between C-C bonds and C=C bond.

A comparison was made using paths and optimally weighted paths by considering a set of 17 isomers of 1- heptene. Without weights the standard error for Rm is 0.105, but when weight are introduced weighted paths, by associating with C=C bond the optimal weight x=0.6 that minimizes the standard error s has further decrease to 0.079.

CONSTRUCTION OF DESCRIPTORS SUITABLE FOR OPTIMIZATION

The first task when considering optimization of molecular descriptors is to find a generalized form for the descriptor that allows introduction of variables to be optimized. Most molecular descriptors have been designed "rigidly," i.e., the algorithm for their construction is rigid so that once molecule is selected (including molecules having heteroatoms) the invariant of interest can be computed exactly. In Table 2 we list some better known and some perhaps less known molecular topological indices. They can be classified as integers [4, 11, 12] (the first generation indices according to Balaban [13]), as real numbers [1, 14] (the second order indices according to Balaban) and as molecular complexity indices applying information-theoretic formalism on chemical graphs [15], as indices derived from matrices either by matrix algebra [12, 14], (the third generation), or indices derived from novel

matrices considered for molecular graphs [15]. Finally some indices can be group in a natural way to that they form sequentially related family of molecular descriptors [2, 16]. To this list of diverse molecular descripotors we should add as the last class indices that have inherent flexibility involving variable part that can be optimized for different applications [8, 9].

In the case of invariants that are computed from the adjacency, the distance matrix, and other graph theoretic or structural matrices that have zero diagonal entries, one can arrive at invariants that will contain variable (to be optimized) by using the diagonal elements x_{ii} as variables. For example, this is an "algebraic" way which one can use to generalize the connectivity indices, the Wiener index, and Balaban's J index. However, Hosoya Z topological index, the path numbers, self-returning walks and other "geometrical" invariants, all which are of considerable interest, can not be in this way generalized.

An alternative way to obtain the connectivity index, the Wiener index, and even Hosoya Z index, that involve variable x to be optimized is to consider weighted paths. In the case of path numbers the generalization is based on partitioning of the Wiener index using path numbers [5]:

$$W = c_1 p_1 + c_2 p_2 + c_3 p_3 + \ldots + c_k p_k.$$

To obtain the Wiener number one has to use: $c_1 = 1$, $c_2 = 2$, $c_3 = 3$, $\ldots$ $c_k = k$. However, one can view the coefficients c_i as the weights and select the weights so to optimize a regression correlation.

An alternative generalization of the Wiener number suitable for discriminating two kinds of distinct bonds in a molecule, such as CC single and CC double bonds, or bonds between heteroatoms, has been outlined [9]. Here a distinct bond, such as CC double bond in alkenes, is given a weight x.

The count of paths, which leads to the Wiener number, now differentiates between CC single and CC double bond. Each time the CC double bond is involved in the count it involves the weight x. In Table 3 we illustrate the count of paths for 3-methyl-3-hexene. If we assume x=1 then the count represents paths in 3-methylhexane, if we assume x=3 then the count represent paths in 3-methyl-3-hexene. However, x can be viewed as a variable to be determined from regression analysis.

The same approach allows one to generalize Hosoya's Z topological index for alkenes [17]. Just instead of counting paths one counts disjoint edges. In Table 4 we give the result for the same alkene of Table 3. Finally, the weight x can be directly incorporated in the connectivity index $^1\chi$ and the higher connectivity indices as illustrated at the bottom of Table 4 [17]. The expression $^1\chi = 2/\sqrt{2} + 1/\sqrt{(2 + x)} + 1/\sqrt{[2 (2 + x)]} + 1/\sqrt{[2 (1 + x)]} + 1/\sqrt{[(1 + x) (2 + x)]}$ is obtained by viewing the end vertices of the edge given the weight x to have valences (2 + x) and (1 + x). When x = 1 the graph reduces to that of 3-methylhexane and $^1\chi$ becomes $2/\sqrt{2} + 1/\sqrt{3} + 1/2 + 2/\sqrt{6} = 3.30806$, and when x = 2 it becomes molecular graph of 3-methyl-3-hexene with $^1\chi = 2/\sqrt{2} + 1/\sqrt{8} + 1/2 + 1/\sqrt{6} + 1/\sqrt{12} = 2.96469$. In general x can take non-integer values.

WEIGHTED PATHS FOR ALCOHOLS

We will extend the application of weighted paths to molecules having a heteroatom (rather then a hetero-bond as was the case with CC double bond in alkenes). We will re-examine the boiling points of 58 alcohols already investigates in several structure-property studies [18-20]. In Table 5 we listed the weighted paths of length one to length four for the molecules considered.

We view the weight x as a variable to be freely adjusted and will seek the optimal weight for the C-O bond by minimizing the standard error in a stepwise multiple regression of the boiling points of alcohols.

In Table 6 we show the regression coefficient (r), the standard error (s) and the Fisher ratio (F) for various values of x when using from one to four weighted paths as descriptors. When p_1 is used as a single descriptor the outcome of the regression does not depends on x, because x is an additive constant for all paths of length one. Clearly p_1 is not adequate descriptor for boiling points in alcohols, as could have been expected, since p_1 only reflects the molecular size, having the same value for all isomers. With two descriptors we see an impressive improvement in the regression statistics. The standard error now depends on the value of x and has decreased already for the value x=1 to half (6.64 ^{o}C)of the previous case (13.28 ^{o}C). The case x=1 corresponds to treating carbon and oxygen atoms equally, hence represents a model in which we do not differentiate the presence of heteroatom. However, even this crude model has significantly smaller standard error than the regressions based on Wiener number W, the Shultz index MTI [13] or the valence connectivity index χ^{v}. These above indices were all considered in ref. [18], in regressions based on a single descriptor and produced the standard errors above 9 ^{o}C. In contrast we use two descriptors (p_1 and p_2), but on the other hand our descriptors do not differentiate heteroatoms while the descriptors used in ref. 18 were designed to differentiate oxygen and carbon.

Now we will consider x as a variable. From Table 6 we see that as x increases the standard error decreases and has minimum for x = 2.6. Fig. 4 shows the variation of the standard error with x in the interval x = 1 to x = 3 (fitted to quartic function).

The regression equation using first only p_1; then p_1, p_2 and finally p_1, p_2, p_3 are:

$$BP = 17.65758\ p_1 + 10.62758 \quad (6)$$

$$BP = 25.01704\ p_1 - 6.00248\ p_2 + 11.03094 \quad (7)$$

$$BP = 25.68984\ p_1 - 6.02031 p_2 - 0.42308\ p_3 + 8.72600 \quad (8)$$

which after the descriptors have been orthogonalized are respectively:

$$BP = 17.65758\ \Omega_1 + 10.62758 \quad (9)$$

$$BP = 17.65758\ \Omega_1 - 6.00248\ \Omega_2 + 10.62758 \quad (10)$$

$$BP = 17.65758\ \Omega_1 - 6.00248\ \Omega_2 - 0.44723\ \Omega_3 + 10.62758 \quad (11)$$

Here Ω_1 is p_1, Ω_2 is the part of p_2 not paralleling p_1, and Ω_3 is the part of p_3 not paralleling both p_1 and p_2 (more correctly not paralleling Ω_1 and Ω_2). The orthogonalization process has been described in ref. 21. Introduction of p_3 improves the correlation somewhat but apparently not considerably. Hence the equation (7) or its equivalent, the equation (10), can be taken as the best characterization of the correlation of the boiling points of aliphatic alcohols with weighted path numbers. Fig. 2 shows the regression of the calculated BP versus the experimental BP, and Table 7 lists the experimental and calculated BP.

CONCLUSION

The examples given clearly show the high quality results based on optimal molecular descriptors which on one hand use fewer descriptors and on the other hand give correlations with significantly if not dramatically reduced standard error. . When the multiple regression based on optimal

descriptors are combined with the orthogonalization of the descriptors [10] one can expect results that will not only give satisfactory structure-property-activity relationship but will also lead to meaningful interpretation of the results --- something that is currently missing in structure-property relationship studies. In order to further illuminate structure-property-activity relationship it seems appropriate not only to use orthogonalized molecular descriptors but to use whenever possible the same set of descriptors that serve as a basis for structure-property relationship valid for a wider pool of structures and wider range of properties.

LITERATURE

1 The connectivity index 1χ was first time introduced in: Randic, M.; On the characterization of molecular branching, J. Am. Chem. Soc. **1975**, *97*, 6609-6615.

2 The connectivity index 2χ and higher order connectivity indices $m\chi$ were first time introduced in: Kier, L. B.; Murray, W. J.; Randic, M.; Hall, L. H.; Molecular connectivity V: Connectivity series applied to density, J. Pharm. Sci. **1975**, *65*, 1226-1230.

3 Kier, L. B.; Hall, L. H.; Molecular Connectivity in Chemistry and Drug Research; Academic Press, New York, 1976.

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

5 Randić, M.; Linear combination of path numbers as molecular descriptors, New J. Chem., **1997**,*21*, 945.

6 See: D. Rogers, in: Genetic Algorithms in Molecular Modeling, (J. Devillers, ed.), Academic Press, London, 1996, p. 87.

7 Randić, M.; Hansen, P. J.; Jurs, P. C.; Search for useful graph theoretical invariants of molecular structure, J. Chem. Inf. Comput. Sci., **1988**, *28*, 60-68.

8 Randic, M.; Novel graph theoretical approach to heteroatom in quantitative structure-activity relationship, Chemometrics & Intel. Lab. Syst., **1991**, *12*, 970-980.

Randić, M.; On computation of optimal parameters for multivariate analysis of structure-property relationship, J. Comput. Chem., **1991**, *12*, 70-980.

Randic, M.; Dobrowolski, J. Cz.; Optimal molecular connectivity descriptors for nitrogen containing molcules, Int. J. Quant. Chem: Quant. Biol. Symp. (in press).

9 Randić, M.; Pompe, M.; On characterization of CC double bond in alkenes, New J. Chem. (submitted).

10 Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of graph theoretical and geometrical descriptors in structure-activity relationships, in: 3-D Molecular Structure and Chemical Graph Theory, Balaban, A. T., Ed., Plenum Publ. Corp., New York (1996).

Basak, S. C.; Niemi, G. J.; Veith, G. D. Optimal characterization of structure for prediction of properties, J. Math. Chem., **1990** *14*, 511.

Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. Correlation between structure and normal boiling points of haloalkanes C1-C4 using neural networks, J. Chem. Inf. Comput. Sci., **1994**,*34*, 118.

Basak, S. C.; Niemi, G. J.; Veith, G. D. A graph-theoretic approach to predicting molecular properties, Mathematical and Computer Modelling, **19904**, 185.

Basak, S. C.; Grunwald, G. D. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study, Chemosphere (submitted)

Randić, M. Topological indices, in: Encyclopedia of Computational Chemistry (P. v. R. Schleyer, Ed.), Wiley, Chichester, UK (in press)

11 Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons, Bull. Chem. Soc. Japan, **1971**, *44*, 2332-2339.

12 Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins, J. Chem. Phys., **1947**, *15*, 419-420.

Balaban, A. T. Theor. Chim. Acta, **1979**, *53*, 335.

Schults, H. P., J. Chem. Inf. Comput. Sci., **1989**,*29*, 277.

13 A. T. Balaban, Using real numbers as vertex invariants for third-generation topological indices, J. Chem. Inf. Comput. Sci., **1992**, *32*, 23-28.

14 Balaban, A. T. Highly discriminating distance-based topological index, Chem. Phys. Lett. **1982**, *89*, 399-404.

Randic, M. On molecular identification numbers, J. Chem. Inf. Comput. Sci., **1984**,*24*, 164-175.

Lovasz, L.; Pelikan, J. Period. Math. Hung., **1973**, *3*, 175.

Kirby, E. C. Sensitivity of topological indices to methyl group branching in octanes and azulenes, or what does a topological index index? J. Chem. Inf. Comput. Sci., **1994**, *34*, 1030-1035.

Randic, M. On structural ordering and branching of acyclic saturated hydrocarbons, J. Math. Chem. (in press)

Randić, M.; Guo, X.; Bobst, S. Use of path matrices for characterization of molecular structure, Discrete Appl. Math., (submitted)

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

Basac, S. C. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach, Med. Sci. Res., **1987**, *15*, 605.

16 Filip, P.; Balaban, T. S.; Balaban, A. T. A new approach for devising local graph invariants: Derived topological indices with low degeneracy and good correlation ability, J. Math. Chem., **1987**, *1*, 61-83.

Balaban, A. T.; Ciubotariu, D.; Medeleanu. M. Topological indices and real number vertx invariants based on graph eigenvalues or eignvectors, J. Chem. Inf. Comput. Sci., **199131**, 517-523.

Balaban, A. T. Lowering the intra- and intermolecular degenracy of topological invariants, Croat. Chem. Acta, **1993**, *66*, 447-458.

17 Tratch, S. S.; Stankevich, M. V.; Zefirov, N. S. Combinatorial models and algorithms in chemistry. The expanded Wiener number - a novel topological index, J. Comput. Chem., **1990**, *11*, 899-908.

Hall, L. H. Computational aspects of molecular connectivity and its role in structure-property modeling, in: Computational Graph Theory, (Rouvray D. H., Ed., Nova Publishers,: New York, 1990; pp. 202-233.

Randic, M.; Guo, X.; Oxley, T.; Krishnapriyan, H. Wiener matrix: Source of novel graph invariants, J. Chem. Inf. Comput. Sci., **1993**, *33*, 709-716.

Randić, M. Hosoya matrix - a source of novel molecular descriptors, Croat. Chem. Acta, **1994**, *67*, 415-429.

Amic, D.; Trinajstić, N. On the detour matrix, Croat. Chem. Acta, **1995**, *68*, 53-62.

Diudea, M. V.; Minailiuc, O.; Katona, G.; Gutman I. Szeged matrices and related numbers, MATCH, **1977**, *35*, 119-143.

Diudea, M. V. Cluj matrix, Cju: Source of various graph descriptors, MATCH, **1977**, *35*, 163-183.

Randić, M.; Plavšić, D.; Razinger, M. Double invariants, MATCH, **1997**, *35*, 243-259.

18 Kier, L. B.; Hall, L. H. Molecular connectivity VII: Specific treatement of heteroatoms, J. Pharm. Sci., **1976**, *65*, 1806-1809.

Kier, L. B. kappa

Hermann, A.; Zinn, P. List operations on chemical graphs. 6. Coparative study of combinatorial topological indexes of the Hosoya type, J. Chem. Inf. Comput. Sci., **1995**, *35*, 551.

Randić, M. ; Morales D. A.; Araujo, O. Higher order Fibonacci numbers, J. Math. Chem., **1996**, *20*, 79-94.

19 Randić, M.; work in progress

20 Nikolić, S.; Trinajstić, N.; Mihalić, Z. J. Math. Chem., **1993**, *12*, 251-264.

21 Seybold, P. G.; May, M.; Bagal, U. A. J. Chem. Educ., **1987**, *64*, 575

22 Amidon, G. L.; Yalkovsky, S. H.; Leung, S. J. Pharm. Sci., **1974**, *63*, 1858

23 Randić, M. Orthogonal molecular descriptors, New J. Chem., **1991**, *15*, 517-525.

Randić, M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors, J. Chem. Inf. Comput. Sci., **1991**, *31*, 311-320.

Randić, M. Correlation of enthalpy of octanes with orthogonal connectivity indices, J. Mol. Struct. (Theochem), **1991**, *233*, 45-59.

Randić, M. Fitting non-linear regressions by orthogonalized power series, J. Comput. Chem., **1993**, *14*, 363-370.

Randić, M. Curve fitting paradox, Int. J. Quant. Chem: Quant. Biol. Symp., **1994**, *21*, 215-225.

Table 1 The best single descriptors for various properties of octanes C_8H_{18}

Property	r	s	Descriptor
eccentric factor	0.992	0.0039	2χ
Density	0.979	0.0025	3χ
Molecular volume	0.978	0.554	3χ
Molar refraction	0.970	0.046	3χ
Surface tension	0.964	0.241	$2\chi - 3\chi$
Motor octane number	0.959	7.27	I_{WD}
Heat of vaporization	0.958	0.429	Z
Entropy	0.954	1.40	$m_{1/2}$
Heat of atomization	0.931	0.725	$1/2\chi$
Heat of formation	0.931	0.471	$1/2\chi$
C^{13} chemical shift sum	0.929	19.1	W/Z
Critical temperature	0.889	4.59	$1\chi - 2\chi$
Boiling points	0.888	2.90	Z
Critical volume	0.849	8.67	$\chi(V)$
Critical pressure	0.668	1.10	$1/2\chi$

Table 2 Classification of topological indices. For an extensive listing of use of topological descriptors in QSAR see ref. [10]:

Integers			Reference
	Wiener number	W	[4]
	Hosoya index	Z	[10]
	Path numbers	P1, P2, P_3, . . .	[11]
	Centric index	C	[11]
	Schultz index	MTI	[11]
Real numbers			
	Balaban,s index	J	[13]
	Identification number	ID	[13]
	The leading eigenvalue	$\lambda 1$	[13]
	Branching index	β	[13]
	Weighted ID number	WID	[13]
Information theoretic			
	Bonchev, Trinajstic		
	Basak et al.	IC	
	Basak et al.	SIC	
	Basak et al.	CIC	
Novel matrices			
	Expanded Wiener		[15]
	Total Topological State		[15]
	Wiener matrix index		[15]
	Hosoya matrix index		[15]
	Detour matrix index		[15]
	Szeged matrix index		[15]

	Cluj matrix index		[15]
Sequential descriptors			
	Path numbers	$^1P, ^2P, ^3P, \ldots$	[11]
	Connectivity indices	$^1\chi, ^2\chi, ^3\chi, \ldots$	[1, 2]
	Valence Connectivity	$^1\chi^m, ^2\chi^m, ^3\chi^m, \ldots$	[16]
	Kappa indices	$\kappa_1, \kappa_2, \kappa_3, \ldots$	[16]
	Hosoya type indices	$^1Z, ^2Z, ^3Z, \ldots$	[16]

Table 3 The count of paths and weighted paths in 3-methyl-3-hexene

Carbon atom	p_1	p_2	p_3	p_4	p_5
1	1	1	1 + x	x	x
2	2	1 + x	x	x	
3	2 + x	1 + x	x		
4	1 + x	1 + 2x	x		
5	2	x	2x	x	
6	1	1	x	2x	x
7	1	1 + x	1 + x	x	
Molecule	5 + x	3 + 3x	1 + 4x	3x	x

Wiener Number: $W = 9 + 12x$

Table 4 Enumeration of contributions to the Hosoya Z index and the higher order Hosoya-type indices based on enumeration of disjoint longer paths

^1Z		^2Z		^3Z	
1		1		1	
5 + x	isolated p_1	3 + 3x	isolated p_2	1 + 4x	isolated p_3
7 + 2x	two isolated p_1	2	two isolated p_2		
2 + x	three isolated p_1				
Molecule:					
15 + 4x		6 + 3x		2 + 4x	

Construction of "weighted" connectivity index $^1\chi$ and the higher order variable connectivity indices (for graph of Fig. 1):

$$^1\chi = 2/\sqrt{2} + 1/\sqrt{(2+x)} + 1/\sqrt{[2(2+x)]} + 1/\sqrt{[2(1+x)]} + 1/\sqrt{[(1+x)(2+x)]}$$

$$^2\chi = 1/\sqrt{[2(1+x)]} + 2/\sqrt{[2(2+x)]} + 1/\sqrt{[(1+x)(2+x)]} + 2/\sqrt{[2(1+x)(2+x)]}$$

$$^3\chi = 2/\sqrt{[2(1+x)(2+x)]} + 3/\{2\sqrt{[2(1+x)(2+x)]}\} + 1/\sqrt{[2(2+x)]}$$

Table 5 The weighted path numbers for aliphatic alcohols

	Compound	p_1	p_2	p_3	p_4
1	methanol	x	0	0	0
2	ethanol	1+x	x	0	0
3	1-propanol	2+x	1+x	x	0
4	2-propanol	2+x	1+2x	0	0
5	1-butanol	3+x	2+x	1+x	0
6	2-butanol	3+x	2+2x	1+x	0
7	2-methyl-1-propanol	3+x	3+x	2x	0
8	2-methyl-2-propanol	3+x	3+3x	0	0
9	1-pentanol	4+x	3+x	2+x	1+x
10	2-pentanol	4+x	3+2x	2+x	1+x
11	3-pentanol	4+x	3+2x	2+2x	1
12	2-methyl-1-butanol	4+x	4+x	2+2x	x
13	3-methyl-1-butanol	4+x	4+x	2+x	2x
14	2-methyl-2-butanol	4+x	4+3x	2+x	0
15	3-methyl-2-butanol	4+x	4+2x	2+2x	0
16	2,2-dimethyl-1-propanol	4+x	6+x	3x	0
17	1-hexanol	5+x	4+x	3+x	2+x
18	2-hexanol	5+x	4+2x	3+x	2+x
19	3-hexanol	5+x	4+2x	3+2x	2+x
20	2-methyl-1-pentanol	5+x	5+x	3+2x	2+x
21	3-methyl-1-pentanol	5+x	5+x	4+x	1+2x
22	4-methyl-1-pentanol	5+x	5+x	3+x	2+x
23	2-methyl-2-pentanol	5+x	5+3x	3+x	2+x
24	3-methyl-2-pentanol	5+x	5+2x	4+2x	1+x

25	4-methyl-2-pentanol	5+x	5+2x	3+x	2+2x
26	2-methyl-3-pentanol	5+x	5+2x	3 + 3x	2
27	3-methyl-3-pentanol	5+x	5+3x	4+2x	1
28	2-ethyl-1-butanol	5+x	5+x	4+2x	1 + 2x
29	2,2-dimethyl-1-butanol	5+x	7+x	3+3x	x
30	2,3-dimethyl-1-butanol	5+x	6+x	4+2x	2x
31	3,3-dimethyl-1-butanol	5+x	7+x	3+x	3x
32	2,3-dimethyl-2-butanol	5+x	6+3x	4+2x	0
33	3,3-dimethyl-2-butanol	5+x	7+2x	3+3x	0
34	1-heptanol	6+x	5+x	4+x	3+x
35	3-heptanol	6+x	5+2x	4+2x	3+x
36	4-heptanol	6+x	5+2x	4+2x	3+2x
37	2-methyl-2-hexanol	6+x	6+3x	4+x	3+x
38	3-methyl-3-hexanol	6+x	6+3x	5+2x	3+x
39	3-ethyl-3-pentanol	6+x	6+3x	6+3x	3
40	2,3-dimethyl-2-pentanol	6+x	7+3x	6+2x	2+x
41	3,3-dimethyl-2-pentanol	6+x	8+2x	6+3x	1+x
42	2,2-dimethyl-3-pentanol	6+x	8+2x	4+4x	3
43	2,3-dimethyl-3-pentanol	6+x	7+3x	6+3x	2
44	2,4-dimethyl-3-pentanol	6+x	7+2x	4+4x	4
45	1-octanol	7+x	6+x	5+x	4+x
46	2-octanol	7+x	6+2x	5+x	4+x
47	2-ethyl-1-hexanol	7+x	7+x	6+2x	4+2x
48	2,2,3-trimethyl-3-pentanol	7+x	10+3x	8+4x	3
49	1-nonanol	8+x	7+x	6+x	5+x
50	2-nonanol	8+x	7+2x	6+x	5+x

51	3-nonanol	8+x	7+2x	6+2x	5+x
52	4-nonanol	8+x	7+2x	6+2x	5+2x
53	5-nonanol	8+x	7+2x	6+2x	5+2x
54	7-methyl-1-octanol	8+x	8+x	6+x	5+x
55	2,6-dimethyl-4-heptanol	8+x	9+2x	6+2x	5+4x
56	3,5-dimethyl-4-heptanol	8+x	9+2x	8+4x	5+2x
57	3,5,5-trimethyl-1-hexanol	8+x	11+x	7+x	7+2x
58	1-decanol	9+x	8+x	7+x	6+x

Table 6 The regression coefficient (r), the standard error (s) and the Fisher ratio (F) for various values of x when weighted path are used as descriptors

Single descriptor		r	s	F
p_1		0.9294	13.277	355
Two descriptors				
p_1, p_2	x=1	0.96530	6.6424	794
	x=2	0.9931	4.269	1961
	x=2.2	0.9935	4.131	2096
	x=2.5	0.9938	4.045	2188
	x=2.6	0.9938	4.039	2193
	x=2.7	0.9938	4.044	2188
	x=2.8	0.9937	4.056	2175
	x=3	0.9936	4.098	2130
Three descriptors				
p_1, p_2, p_3	x=1	0.96647	6.5290	549
	x=2	0.9931	4.295	1292
	x=2.2	0.9936	4.128	1400
	x=2.5	0.9941	3.979	1508
	x=2.6	0.9942	3.949	1531
	x=2.7	0.9942	3.926	1549
	x=2.8	0.9943	3.910	1562
	x=3	0.9943	3.893	1576
	x=3.1	0.9943	3.891	1578
	x=3.5	0.9943	3.912	1561

Table 6 The experimental descriptors p1, p2

	Compound	BP exp	BP calc	Residual
1	methanol	64.7	65.24	-0.54
2	ethanol	78.3	77.69	0.61
3	1-propanol	97.2	96.42	0.77
4	2-propanol	82.3	84.11	-1.81
5	1-butanol	117.7	115.67	2.03
6	2-butanol	99.6	102.43	-2.83
7	2-methyl-1-propanol	107.9	109.15	-1.25
8	2-methyl-2-propanol	82.4	84.52	-2.12
9	1-pentanol	137.8	134.92	2.88
10	2-pentanol	119.0	121.68	-2.68
11	3-pentanol	115.3	120.75	-5.45
12	2-methyl-1-butanol	128.7	127.97	0.73
13	3-methyl-1-butanol	131.2	128.90	2.30
14	2-methyl-2-butanol	102.0	102.41	-0.41
15	3-methyl-2-butanol	111.5	114.72	-3.22
16	2,2-dimethyl-1-propanol	113.1	115.84	-2.74
17	1-hexanol	157.0	154.17	2.83
18	2-hexanol	139.9	140.92	-1.02
19	3-hexanol	135.4	139.99	-4.59
20	2-methyl-1-pentanol	148.0	147.22	0.78
21	3-methyl-1-pentanol	152.4	147.72	4.68

22	4-methyl-1-pentanol	151.8	148.15	3.65
23	2-methyl-2-pentanol	121.4	121.66	-0.25
24	3-methyl-2-pentanol	134.2	133.55	0.65
25	4-methyl-2-pentanol	131.7	134.90	-3.20
26	2-methyl-3-pentanol	126.5	134.31	-7.81
27	3-methyl-3-pentanol	122.4	120.30	2.10
28	2-ethyl-1-butanol	146.5	146.79	-0.29
29	2,2-dimethyl-1-butanol	136.8	134.37	2.43
30	2,3-dimethyl-1-butanol	149.0	140.77	8.23
31	3,3-dimethyl-1-butanol	143.0	136.11	6.89
32	2,3-dimethyl-2-butanol	118.6	114.28	4.32
33	3,3-dimethyl-2-butanol	120.0	121.00	-1.00
34	1-heptanol	176.3	173.41	2.87
35	3-heptanol	156.8	159.24	-2.44
36	4-heptanol	155.0	159.24	-4.24
37	2-methyl-2-hexanol	142.5	140.90	1.60
38	3-methyl-3-hexanol	142.4	139.55	2.85
39	3-ethyl-3-pentanol	142.5	138.37	4.13
40	2,3-dimethyl-2-pentanol	139.7	133.11	6.59
41	3,3-dimethyl-2-pentanol	133.0	139.67	-6.57
42	2,2-dimethyl-3-pentanol	136.0	139.32	-3.32
43	2,3-dimethyl-3-pentanol	139.0	132.18	6.82
44	2,4-dimethyl-3-pentanol	138.8	145.34	-6.54
45	1-octanol	195.2	192.58	2.62
46	2-octanol	179.8	179.33	0.47
47	2-ethyl-1-hexanol	184.6	185.29	-0.69

48	2,2,3-trimethyl-3-pentanol	152.2	152.78	-0.57
49	1-nonanol	213.1	211.91	1.19
50	2-nonanol	198.5	198.66	-0.16
51	3-nonanol	194.7	197.73	-3.03
52	4-nonanol	193.0	197.73	-4.73
53	5-nonanol	195.1	197.73	-2.63
54	7-methyl-1-octanol	206.0	205.46	0.54
55	2,6-dimethyl-4-heptanol	178.0	185.69	-7.69
56	3,5-dimethyl-4-heptanol	187.0	183.83	3.17
57	3,5,5-trimethyl-1-hexanol	193.0	186.98	6.02
58	1-decanol	230.2	231.15	-0.95

Figure captions

Fig. 1 Molecular graphs of 3-methylhexane, 3methyl-3-hexene, and weighted bond graphs of the same connectivity

Fig. 2 Correlation of molar refraction Rm against the boiling points BP for n-octane isomers

Fig. 3 and Plot of molar refraction Rm against the connectivity index 1χ plot of the boiling points BP against the connectivity index 1χ

Fig. 4 The variation of the standard error with the weight x (in the interval $1 \leq x \leq 3$) fitted to quartic polynomial

Fig. 5 The regression of calculated against the experimental BP for alcohols of Table 5

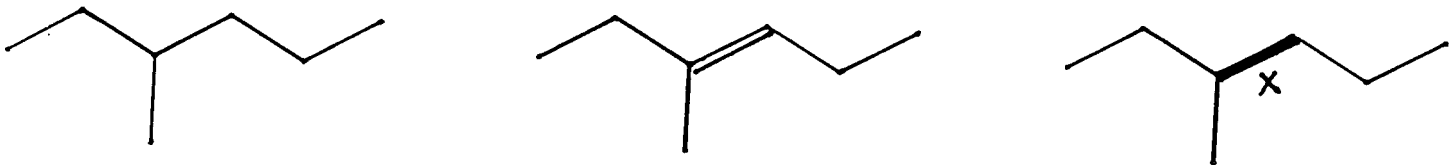


FIG. 1

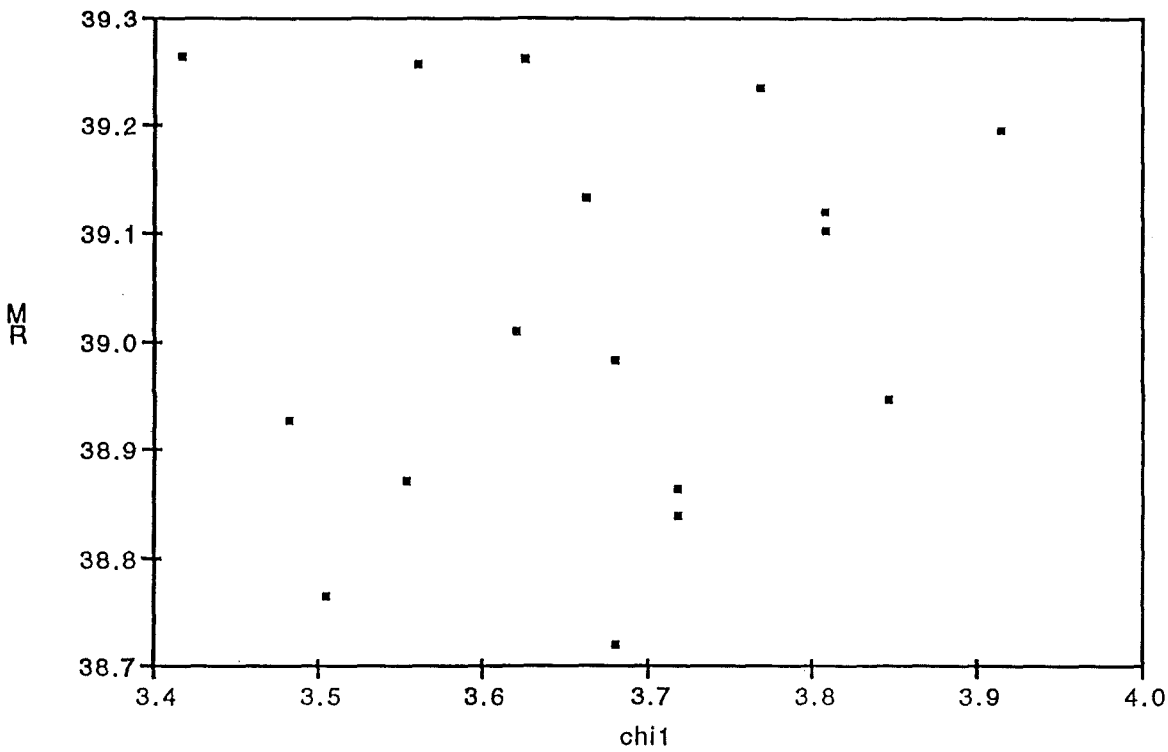


FIG. 3a

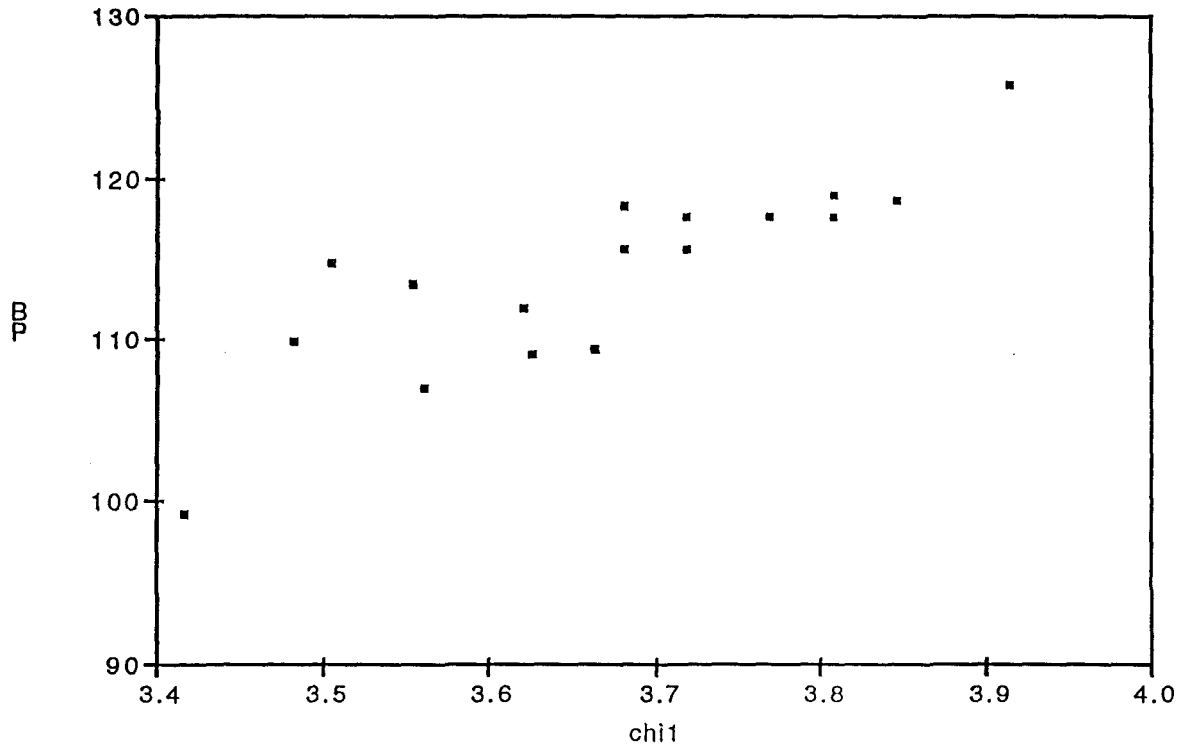


Fig. 36

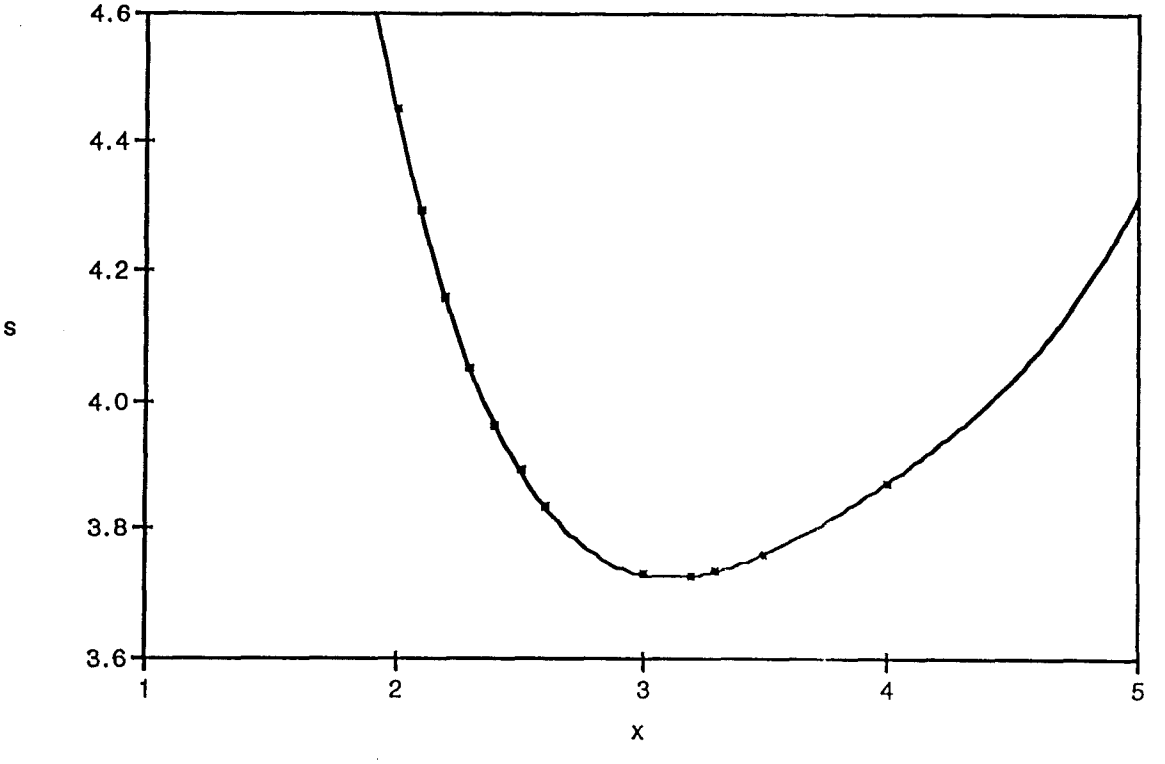


FIG. 4

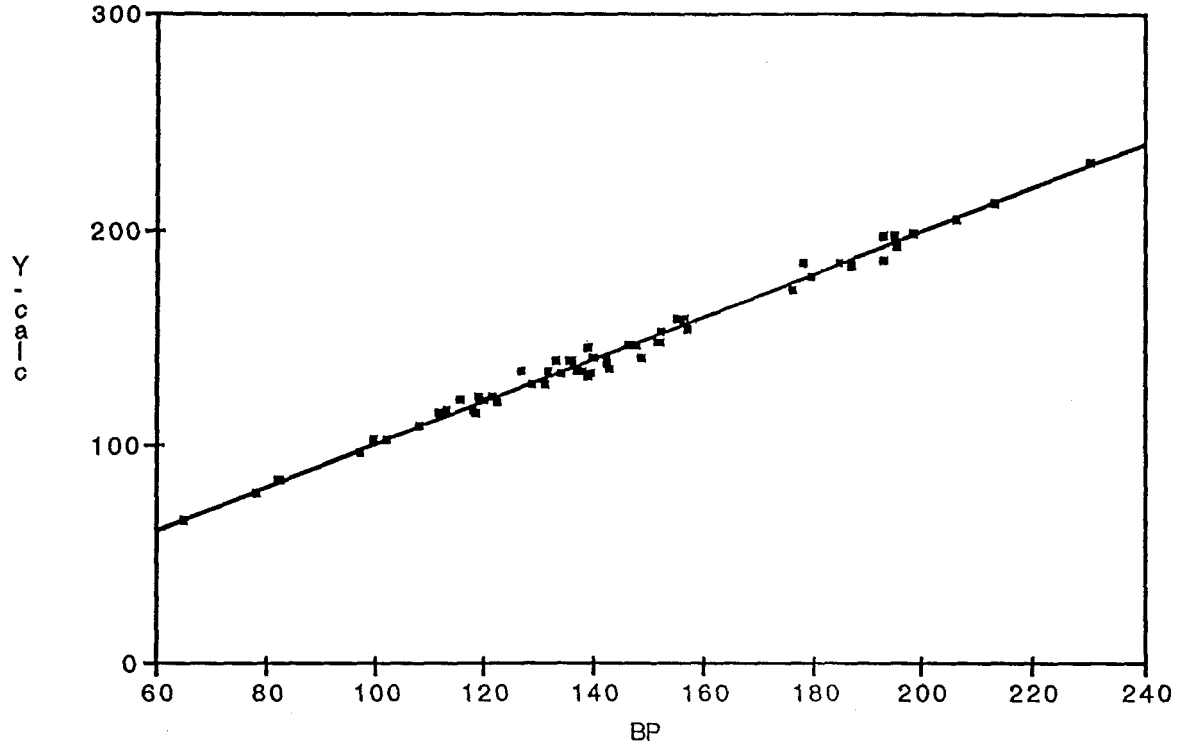


FIG. 5