OPTIMAL CHARACTERIZATION OF STRUCTURE FOR PREDICTION OF PROPERTIES

Subhash C. BASAK and Gerald J. NIEMI
Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway, Duluth, MN 55811, USA

Gilman D. VEITH
U.S. Environmental Protection Agency, Environmental Research Laboratory – Duluth, 6201 Congdon Boulevard, Duluth, MN 55804, USA

> "In order to describe an aspect of holistic reality we have to ignore certain factors such that the remainder separates into facts. Inevitably, such a description is true only within the adopted partition of the world, that is, within the chosen context".
>
> Hans Primas
> *Chemistry, Quantum Mechanics and Reductionism* [57]

Abstract

Different topological and physicochemical parameters have been used to predict hydrophobicity ($\log P$, octanol–water) of chemicals. We calculated a hydrogen bonding parameter (HB_1) and a large number of molecular connectivity and complexity indices for a diverse set of 382 molecules. It is known from earlier studies that topological indices (TIs) predict properties of congeneric sets reasonably well. Since HB_1 is an approximate quantifier of hydrogen bonding and has integral values, we used HB_1 to classify the diverse set into strongly and weakly hydrogen bonding subsets. In an attempt to examine the utility of TIs in predicting properties of relatively similar groups of molecules, we carried out a correlation of $\log P$ with TIs for a subset ($n = 139$) of the original diverse set ($n = 382$) with a weak hydrogen bonding ability ($HB_1 = 0$). Results show that TIs give a better predictive model for the more homogeneous subset as compared to the diverse set of molecules.

1. Introduction

A current trend in chemistry [1–11], pharmacology [12–24], toxicology [25–32], pharmaceutical drug design [33–36], and risk assessment of chemicals [37–40] is the prediction of behavior properties of molecules from their structure. The basic assumption underlying this field of research, called quantitative structure–activity relationships (QSAR), is that the structure of a molecule determines its behavior. This paradigm [22] can be expressed by the relationship:

$$P = f(S), \tag{1}$$

where P is any physical, biomedicinal, toxicological or environmental activity/endpoint of interest and S may represent either an empirical property of the total molecular structure, a relevant substructural fragment or a theoretical structural descriptor (or a set of descriptors) quantitating some aspects of molecular structure.

A review of QSAR studies of the past two decades shows that S in eq. (1) may frequently represent an empirical physical property or physicochemical substituent constants [41–43], quantum chemical structural parameter(s) calculated by *ab initio*/ semi-empirical methods [44], or substructural and topological parameters defined on chemical graphs of molecules [1–10,13–24,27–35].

A large number of QSARs have been published using physicochemical and quantum chemical parameters. Unfortunately, empirical parameters are not readily available for a large fraction of known chemical structures [40,45,46]. In drug design, one has to evaluate a large number (200,000 or more) of probable analogs derived from a lead to develop a new therapeutic agent [41]. Quantum chemical methods are not effective when considering a large number of molecules because computation time is excessively long. A similar situation exists in hazard assessment of chemicals. More than six million distinct chemical substances are known, and humans are exposed to about 66,000 of them [47]. This number is based on chemicals listed in the Toxic Substances Control Act (TSCA) inventory as well as those regulated as pesticides, drugs, food additives and cosmetics. Another sobering fact is that the number of new organic compounds synthesized worldwide is increasing by more than 400,000 per year.

Table 1
Important QSAR endpoints (properties)

Physicochemical	Biological
Molar volume	Receptor binding (K_D)
Boiling point	Michaelis constant (K_m)
Melting point	Inhibitor constant (K_i)
Vapor pressure	Biodegradation
Water solubility	Bioconcentration
Dissociation constant (pKa)	Alkylation profile (with DNA)
Partition coefficient	Metabolic profile
Octanol-water (log P)	Chronic toxicity
Air–water	Carcinogenicity
Sediment–water	Mutagenicity
Reactivity (electrophile)	Acute toxicity
	LD_{50}
	LC_{50}

Of these new chemicals, about 1000 are introduced yearly into societal use [48]. In drug research, toxicology and risk assessment of chemicals, there is need for reliable prediction of a large number of properties. Table 1 gives a sample of some of the more

frequently used properties. Although many of these properties can be determined empirically, because of cost and time limitations only a small fraction of the large number of candidate chemicals can be rigorously tested. Therefore, there is a need for the development of methods which could rapidly screen chemicals for their biomedicinal/toxicological properties to focus resources on chemicals with the greatest potential [48–50]. QSAR models which are based on parameters that are calculable for all chemical structures are gradually emerging as the method of choice in such cases.

2. Structure–activity relationship (SAR)

Structure–activity relationships (SARs) are models which attempt to relate certain structural aspects of molecules to their physicochemical/biological/toxicological properties [22]. High quality and reproducible data on the property of interest for an appropriate set of chemicals and "optimal characterization" of structure of the selected chemicals are the two pre-requisites for the development of SAR. Although physicochemical properties of molecules and data on the biological effects of chemicals at different levels of organization, viz., macromolecule (isolated receptor, protein, DNA or enzyme), membrane (transport through membranes and ion channels, interaction with membrane-bound enzymes), organelle, organ, whole organism and ecosystem are gradually becoming available [51–56], the factor S of eq. (1) has remained elusive to this day. A survey of SAR literature of the past two decades indicates that there is no unifying approach in the representation and optimal characterization of molecular structure [6,13,19,21,22,57]. By optimal characterization, we mean (a) delineation and quantitation of those aspects of molecular structure which determine a particular property, and (b) development of quantitative models which predict properties from structural variables. Part of the problem arises because we need to predict different properties of molecules which might not originate from analogous molecular or submolecular phenomena. However, at a more fundamental level, the principal hurdle has been the lack of uniformity in our definition and quantification of molecular structure [57,58].

The term *molecular structure* represents a set of nonequivalent and probably disjoint concepts [57]. 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, the different meanings we attach to the single term "molecular structure" originate from the same fundamental concept [59]. On the contrary, there is a theoretical and philosophical basis for the nonhomogeneity of concepts covered by the term molecular structure. In their famous paper, Einstein, Podolsky and Rosen [60] discussed correlations in spatially isolated quantal systems, and pointed out that such systems possess interference arising out of nonlocal interactions. This indicates that nature is entangled, holistic, and nonseparable. Experimental evidences indicate that EPR (Einstein–Podolsky–Rosen) correlations are genuine characteristics of nature [57]. On the other hand, the dominant

preconception of science is that modelling and analysis of nature in terms of approximately independent parts is in accordance with nature. This is popularly known as reductionism. EPR correlations, however, clearly contradict the reductionist view of science. This paradox and the plurality of concepts underlying the term molecular structure is explained in terms of abstractions from EPR correlations. We break the holistic symmetry of nature when we abstract deliberately from some EPR correlations. Each abstraction creates its own reality. To describe an aspect of reality, we have to ignore certain factors so that the remainder separates into distinct facts. Obviously, such a description is true only within the adopted partition of the world.

In the context of molecular science, the various concepts of molecular structure (e.g. classical valence bond representation, various chemical graph-theoretic representations, ball and spoke model of a molecule, representation of the molecule by minimum energy conformation, semisymbolic contour map of a molecule, or symbolic representation of chemical species by Hamiltonian operators) are model objects [61] derived through different abstractions of the same chemical reality or molecule [57,58]. In each instance, the equivalence class (concept or model of molecular structure) is generated by selecting certain aspects while ignoring some unique properties of those actual events. This explains the plurality of the concept 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 [57].

3. Characterization of molecular structure

Any concept of molecular structure is a hypothetical sketch of the organization of molecules. Such a *model object* is a general theory and remains empirically untestable. A model object has to be grafted onto a specific theory to generate a *theoretical model* [61]. A theoretical model of an object can be empirically tested. For example, when it was suggested by Sylvester [62] in 1878 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 objects), 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 nonhydrogen atoms present in C_4H_{10} [63]. 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 property. This is because of the fact that any empirical property P maps a set of chemical structures into the set $\mathbb{R}$ 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 [6,34]. 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) [2–9,10–23].

The predictive potential of a theoretical model depends both on: (1) efficacy of representation of the relevant aspects of reality by the model object, and (2) optimal treatment of the model object by appropriate mathematical theories. Effective prediction also depends on the quality of available data and the level of complexity (i.e. lack of our understanding) of the property of interest. The more complex a property, the less is the chance of optimal characterization of structural determinants (and prediction of the property) by a particular theoretical model. This is the case with prediction of carcinogenicity of chemicals, where experts recommend to supplement structural criteria with various functional criteria (pharmacological/toxicological effects of chemicals at different levels of biological organization) in order to attain an acceptable level of prediction [48]. At this level of SAR (level I), the central question becomes: Whether a particular activity is possible for a given chemical. No attempt is made to predict the property quantitatively. At the next level of SAR (level II), predictions are within an order of magnitude. Acute toxicity, e.g. LD_{50} in rodents, LC_{50} in fish, fall in this category. The situation improves in level III SAR when we attempt to predict

Table 2

Different levels of SAR

Level	Qualitative/ quantitative	Central question	Extent of accuracy	Example
I	Qualitative	Is the activity possible?	None	Carcinogenicity hydrolysis
II	Qualitative	Is a rough estimate of the potency possible?	Order of magnitude	Chronic toxicity, mutagenicity, bioconcentration factors
III	Semiquantitative	Is an estimate of the activity possible?	Factor of 2	Log P, toxicity of specific classes of chemicals
IV	Quantitative	What is the predicted numerical value?	Within 20% of measured data	Boiling point and other chemical properties

bioactivity of specific groups of chemicals with a well-defined mode of action, e.g. narcotics, polar narcotics, uncouplers of oxidative phosphorylation, etc. At this stage, our predictive capability comes within a factor of two. Finally, at the level IV SAR or quantitative SAR (QSAR), we have less complex properties of molecules (e.g. boiling point) which can be predicted from structure within 20% of the experimental value [2,3,6,9,12,42,43,49,51,53,64–66]. Table 2 briefly summarizes various levels of SAR – from purely qualitative to the quantitative.

4. Nonempirical approach to SAR

The ideal goal of SAR research is to predict behavior of chemical species from a minimal set of input data. In principle, the set could consist of: (1) empirical properties or parameters [41], (2) a combination of empirical and nonempirical parameters [10,26,29,38], and (3) purely nonempirical parameters [1–8,11–25]. An alternative classification was outlined in ref. [24]. However, two important considerations suggest the use of as many nonempirical parameters as possible. Firstly, in drug design one can easily envisage thousands of structures derivable from a particular pharmacological "lead" and many of these might not even be synthesized at the time of evaluation of their property [67]. Empirically-based SARs are not useful for predicting properties of such chemicals. Secondly, from a more practical point of view, even simple properties such as boiling point, melting point, or vapor pressure are not available for a very large fraction of known chemicals [46]. Consequently, in recent years, nonempirical graph-theoretic parameters have been used in SAR studies for predicting chemical behavior [1–9,11–25]. These are graph invariants [68], usually a single number or a vector, which can be used to characterize and order molecules, and predict properties. It is evident from published literature that SAR models work well for reasonably homogeneous sets of chemicals. This reflects the age-old wisdom of biomedicinal chemistry: similar structures usually have similar properties. In the case of bioactive molecules, structures recognized by a particular enzyme or receptor are usually reasonably similar, and may be looked upon as derived from a "core structure" (pharmacophore or toxophore). No such structural homogeneity is evident in non-specific (narcotic) interaction [69]. Good prediction of bioactivity of a diverse group of narcotic chemicals can be achieved either through classification of the original set into chemically (based on some arbitrary concept of structural similarity) or biochemically (on the basis of some well-defined biochemical mechanism of action, e.g. narcosis, polar narcosis, etc.) homogeneous subsets or by accounting for structural heterogeneity in terms of multiple physicochemical factors, e.g. molecular size, dipolarity, hydrogen bonding, etc. [70]. For a set of molecules with limited structural diversity, presence or absence of certain functional group(s) or selected substructure(s) may act as classifiers. However, this method fails for a very diverse group of chemicals. At the topological level, when paths of length two (P_2) and paths of length three (P_3) are taken as coordinates of chemical structure on a coordinate grid, useful ordering of isomers can be achieved [71].

In an earlier paper, we found that lipophilicity (log P), octanol–water) of a large ($n = 382$) and structurally diverse group of chemicals could be predicted reasonably well with a combination of molecular connectivity indices, molecular complexity indices, and a hydrogen bonding parameter HB_1 [6]. This result is in line with the notion that hydrophobicity of a chemical is primarily determined by its size, polarity, and hydrogen bonding properties [41]. The HB_1 parameter is algorithmically defined and can be calculated for all molecular structures [72,73]. Since it has been found that SAR models work more efficiently for homogeneous groups of molecules as opposed to diverse data

sets, it was of interest to see whether a combination of connectivity and complexity indices can provide a viable model for the prediction of log P of solutes without strong hydrogen bonding ability. Therefore, in this paper we attempted to predict log P for a subset (n = 139) of nonhydrogen bonding ($HB_1 = 0$) chemicals derived from the original diverse set (n = 382) of molecules analyzed in our previous study [6].

5. Theoretical foundation, definition, and computation of parameters

In this paper, we have used three types of parameters: (a) molecular connectivity indices, (b) molecular complexity indices, and (c) hydrogen bonding parameter HB_1.

A graph G is defined as an ordered pair consisting of two sets V and R,

$$G = [V, R],$$

where V is a finite nonempty set and R is a binary relation defined on V. The elements of V are called vertices and the elements of R, sometimes symbolized by $E(G)$ or E, are called edges. Such an abstract graph may be visualized by representing elements of V as points and by connecting a pair $x = (v_i, v_j)$ of elements of V with a line if and only if $(v_i, v_j) \in R$. Two vertices of G are called adjacent if they are connected by a line. A walk of the form $v_0, x_1, v_1, x_2, \ldots, v_n$ joins vertices v_0 and v_n in G. The length of a walk is the number of occurrences of lines in it. A walk is closed if $v_0 = v_n$. A path is an open walk in which all vertices are distinct. A graph G is connected if every pair of its vertices is connected by a path. A graph G is a multigraph if it contains more than one edge between at least one pair of adjacent vertices, otherwise G is a linear graph. The distance $d(v_i, v_j)$ between vertices v_i and v_j in G is the length of any shortest path connecting v_i and v_j. The degree δ_i of the vertex v_i in G is equal to the number of lines incident with v_i. The radius ρ of a graph is given by $\rho = \min \max_{u,v \in V} d(u, v)$. For a vertex $v \in V$, the first-order neighborhood $\Gamma^1(v)$ is a subset of V such that $\Gamma^1(v) = \{u \in V \mid d(u, v) = 1\}$. The first-order closed neighborhood $N^1(v)$ is a subest of V such that $N^1(v) = (v) \cup \Gamma^1(v) = \Gamma^0(v) \cup \Gamma^1(v)$, where (v) is the one-point set consisting of v only and may be looked upon as $\Gamma^0(v)$. If ρ is the radius of a graph G, we can construct $\Gamma^i(v)$ and $N^i(v)$, $i = 1, 2, \ldots, \rho$, for each vertex v in G. Detailed definitions of terms used in this paper may be found in books by Harary [68] and Trinajstić [74].

A graph $G = [V, E]$ becomes a *model object* in chemistry when elements of V represent a prescribed set of atoms in a molecule and the edge set E depicts the bonding relationship among them. Any pair of atoms in a molecule is involved in a binary relation: either the pair is bonded or not bonded. This pattern of connectedness of atoms in a molecule is adequately represented by graphs. Figure 1 gives the chemical structure, hydrogen-suppressed multigraph (G_1) and simple hydrogen-suppressed linear graph (G_2) of acetamide. While in G_1 all hydrogen atoms present in the molecule are ignored, G_2 does not take care of atom types or bond multiplicity. Molecular graphs like G_1 and G_2 are of little use in comparing chemical structures and predicting their properties unless they lead to a more precise mathematical description or a *theoretical model* [61].

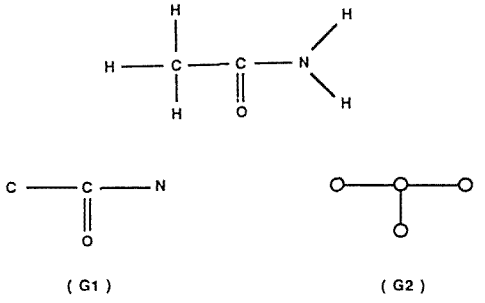

Fig. 1. Structural formula, hydrogen-suppressed multigraph and simple linear graph of acetamide.

In the realm of chemical graph theory, this has been accomplished by defining specific graph invariants [4,12,16,22,28,35,36,63,74]. A graph invariant may be a polynomial, a vector (sequence), or a single number. The Wiener index [75], connectivity indices [4,12], and molecular complexity indices [22,74] are examples of numerical graph invariants. From the simple linear graphs of a molecule, the zero-order connectivity index ${}^0\chi$ is calculated as:

$$ {}^0\chi = \sum_i (\delta_i)^{-1/2}. \tag{2} $$

Randić's connectivity index ${}^1\chi$ is calculated as [4]:

$$ {}^1\chi = \sum_{\text{all edges}} (\delta_i \, \delta_j)^{-1/2}. \tag{3} $$

A generalized connectivity index ${}^h\chi$ considering paths of the type $v_0, v_1, \ldots, v_h$ of length h is defined as [12]:

$$ {}^h\chi = \sum_{\text{all paths}} (\delta_0 \, \delta_1 \ldots \delta_h)^{-1/2}. \tag{4} $$

Cluster, path–cluster, and cycle types of connectivity indices are calculated by the method of Kier and Hall [76].

Simple connectivity indices are computed from the linear graph model of a molecule, where the weight of a vertex v_i in G is equal to its degree δ_i or the topological valency of the ith atom. This picture over-simplifies the chemical reality of a molecule, neglecting features such as bond angle, bond length, chirality, nature of individual atoms, etc. Improvements over simple linear graph models of molecules have been done by representing molecules using weighted graphs [12,35,76–78]. Valence connectivity

indices result from one such weighting scheme, where the degree δ_i^v of the vertex v_i in the weighted graph is given by [12]:

$$\delta_i^v = (Z_i - h_i)/(Z - Z_i - 1), \tag{5}$$

where Z is the atomic number of the ith atom, Z_i is the number of valence electrons, and h_i is the number of hydrogen atoms attached to the ith atom. Valence connectivity indices are calculated by substituting δ_i^v for δ_i in the above relevant equations for calculation of simple connectivity indices.

Molecular complexity indices constitute another way of deriving numerical descriptors from molecular graphs [9,11,16,22,77,79–82]. The science of information theory has grown mainly out of the pioneering studies of Shannon [83], Wiener [84], Ashby [85], and Kolmogorov [86]. There is more than one version of information theory [81]. In Shannon's [83] statistical information theory, information is measured as reduced uncertainty of the system. In the algorithmic theory of Kolmogorov [86], the quantity of information is defined as the minimal length of a program which allows a one-to-one transformation of an object (set) into another. In applying information-theoretic formalism on chemical graphs, one looks upon the information content (or complexity) of a graph as a measure of its degree of variety or heterogeneity, as suggested by Ashby [85]. An appropriate set A of n elements is derived from a molecular graph G depending on certain preselected criteria. On the basis of an equivalence relation defined on A, the set A is partitioned into equivalence classes A_i of order n_i ($i = 1, 2, \ldots, h$, $\sum_i n_i = n$). A probability scheme is then assigned to the set of equivalence classes:

$$\begin{pmatrix} A_1, A_2 & , \ldots, & A_h \\ p_1, p_2 & , \ldots, & p_h \end{pmatrix},$$

where $p_i = n_i/n$, n_i and n being the cardinalities of A_i and A, respectively. The mean information content (or complexity) of an element A is defined by Shannon's [83] relation:

$$\mathrm{IC} = -\sum_i p_i \log_2 p_i \,. \tag{6}$$

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

It is to be noted that the complexity of a real object or a model object is not uniquely defined. While there could be more than one way of defining a model object [57,61] corresponding to the same piece of reality, complexity of the same model object may vary depending on the nature of the equivalence relation. In science, we deal with equivalence classes of events generated by grouping actual events and ignoring, at the same time, some unique properties of those events [57]. For example, when A represents

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 transitive sets or orbits of the automorphism group of G, whereafter vertices are considered equivalent if they belong to the same orbit [87–90]. Excellent reviews are available on measures of complexity and computation of complexity parameters [22,81,87].

Rashevsky [91] symbolized molecules by simple linear graphs and calculated molecular complexity. In this approach, two vertices u and v of a graph G 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, various authors have computed complexity of molecules where linear graphs [11,80,87–90] or multigraphs [82] with indistinguishable vertices were used to symbolize the chemical species. On the other hand, to account for the unique nature of atoms and their bonding pattern in a molecule, Sarkar et al. [93], Roy et al. [94], Basak et al. [28,79], Ray et al. [77] calculated complexity of graphs on the basis of equivalence relations where both the nature of the atom (vertex) and the number and chemical nature of bonded neighbors of all atoms are taken into account. This was accomplished by defining open spheres for all vertices of the molecular graph [95]. If r is any nonnegative real number and v is a vertex of the graph G, then the open r-sphere $S(v, r)$ is defined as the subset $V(G)$ consisting of all vertices v_i such that $d(v, v_i) < r$. Obviously, $S(v, 0) = \varnothing$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r) = (v) \cup \Gamma^1(v) = N^1(v)$ for $0 < r < 2$. One can construct open r-spheres of each vertex of G for all integral values of r, $0 \leq r \leq \rho$. For a particular value of r, the collection of all such open spheres $S(v, r)$, where v runs over the entire vertex set V, forms a neighborhood system of the vertices of G. A suitably defined equivalence relation can then partition V into disjoint subsets based on the equivalence of nature, connectedness, and bonding pattern of neighbors up to rth-order neighborhoods [94]. It is noteworthy that this approach incorporates the effects of distant neighbors (i.e. neighbors of immediately bonded neighbors) on an atom or a reaction center. After partitioning of the vertices for a particular order (r) of neighborhood, IC_r is calculated by eq. (6). Subsequently, Basak, Roy and Ghosh [79] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r/\log_2 n, \tag{7}$$

where IC_r is calculated by eq. (6) and n is the total number of vertices of the graph. It is noted that SIC_r is related to Brillouin's [96] measure of redundancy of a system. Another information-theoretic invariant, complementary information content (CIC), was defined as [28]:

$$CIC_r = \log_2 n - IC_r. \tag{8}$$

The Wiener index W [75], and the information-theoretic indices I_D^W and $\bar{I}_D^W$ are calculated from the distance matrix of chemical graphs [81]. The set of topological indices used in this paper are shown in table 3. Topological parameters were calculated

Table 3

Definition and symbols for topological indices

W	Half-sum of the off-diagonal elements of the distance matrix of a graph.
I_D^W	Information index for the magnitudes of the distances between all possible pairs of vertices of a graph.
$\bar{I}_D^W$	Mean information index for the magnitude of the distance.
IC_r	Mean information content or complexity of a graph based on the rth ($r = 0, 1, \ldots, 6$) order neighborhood of vertices in a graph.
SIC_r	Structural information content of a graph based on rth ($r = 0, 1, \ldots, 6$) order neighborhood of vertices.
CIC_r	Complementary information content of a graph G calculated from the rth ($r = 0, 1, \ldots, 6$) neighborhood of vertices.
${}^h\chi$	Path terms of hth order ($h = 0, 1, \ldots, 6$).
${}^h\chi_C$	Cluster terms of hth order ($h = 3, \ldots, 6$).
${}^h\chi_{PC}$	Path–cluster terms of hth order ($h = 4, \ldots, 6$).
${}^h\chi_{CH}$	Chain or cycle terms of different orders ($h = 3, \ldots, 6$).
${}^h\chi^v$	Valence connectivity type path terms of hth order ($h = 0, 1, \ldots, 6$).
${}^h\chi^v_C$	Valence connectivity type cluster terms of hth order ($h = 3, \ldots, 6$).
${}^h\chi^v_{PC}$	Valence connectivity type path–cluster terms of hth order ($h = 4, \ldots, 6$).
${}^h\chi^v_{CH}$	Valence connectivity type chain or cycle terms of hth order ($h = 3, \ldots, 6$).
P_h	Number of paths of length h ($h = 0, 1, \ldots, 10$) in the hydrogen deleted graph.

by the computer program POLLY [97], where SMILES line notation [98] is the input. Log P values of the 139 chemicals analyzed in this paper were calculated by CLOGP3 (version 3.53) of MedChem software [99] at the U.S. EPA's Environmental Research Laboratory in Duluth. The hydrogen bonding parameter HB_1 was calculated by using a computer program developed by Basak [72], based on the ideas of Lien and co-workers [73]. HB_1 includes both hydrogen bond donor and hydrogen bond acceptor properties.

6. Statistical analyses

The present study is the continuation of our ongoing QSAR research to develop models for the prediction of properties of molecules using parameters which can be

calculated directly from structure. In a recent study, using a set of 70 topological parameters and HB_1 as independent variables, we found that a combination of molecular connectivity indices, molecular complexity indices, and the hydrogen bonding parameter HB_1 could predict (R^2 = 0.91) reasonably well the octanol–water partition coefficient (log P) of a diverse group of 382 chemicals [6]. The following parameters appeared in linear regression equations containing up to ten independent variables:

$$P_6,\ IC_0,\ CIC_1,\ {}^5\chi,\ {}^5\chi_C,\ {}^5\chi_{PC},\ {}^6\chi_{CH},\ {}^0\chi^v,\ {}^3\chi^v,\ {}^4\chi^v,\ {}^4\chi^v_{PC}.$$

We used the maximum R^2 method to identify the prediction model for log P [100]. This method finds the "best" one-variable model, the "best" two-variable model, and so forth, for the prediction of the dependent variable. Of the independent variables, HB_1 quantitates the hydrogen bonding ability of a molecule approximately and has integral values [73]. Therefore, we decided to use HB_1 to classify a diverse set into more homogeneous subsets instead of using HB_1 as an independent variable. As a first exploratory analysis, we decided to take those molecules (n = 139) of the set of 382 which lack any hydrogen-bonding potential with respect to the scale of Ou et al. [73], and investigated to what extent topological indices could predict log P values for the homogeneous non-hydrogen bonding group.

Multiple regression analysis showed that there was an improvement in the prediction of log P up to step 6 (table 4). In the six-variable model, there was a significant regression of log P with ${}^0\chi^v$, ${}^4\chi^v$, ${}^4\chi^v_{PC}$, ${}^6\chi_{CH}$, ${}^5\chi_C$ and IC_0 parameters. This

Table 4

Summary of multiple regression analysis for prediction of log P from topological indices

Step	Variables	F	R^2	Standard error of estimate
1	${}^0\chi^v$	1180	0.90	0.46
2	CIC_1, ${}^0\chi^v$	928	0.93	0.37
3	CIC_1, ${}^6\chi_{CH}$, ${}^0\chi^v$	677	0.94	0.36
4	IC_0, ${}^6\chi_{CH}$, ${}^0\chi^v$, ${}^4\chi^v$	593	0.95	0.33
5	IC_0, ${}^5\chi$, ${}^6\chi_{CH}$, ${}^0\chi^v$, ${}^4\chi^v$	507	0.95	0.32
6	IC_0, ${}^5\chi_C$, ${}^6\chi_{CH}$, ${}^0\chi^v$, ${}^4\chi^v$, ${}^4\chi^v_{PC}$	446	0.95	0.31

model was developed using the set of 139 compounds. However, the model had two influential outliers (compounds 14 and 101 in table 5) as determined by Cook's D statistic [101]. Deletion of these two compounds resulted in the following highly significant 6-variable model:

Table 5

Log P, estimated log P, and six topological indices for 139 compounds

Sequence number	Chemical name	Log P	Predicted log P (eq. (9))	IC_0	$^5\chi_C$	$^6\chi_{CH}$	$^0\chi^v$	$^4\chi^v$	$^4\chi^v_{PC}$
1	1,1,1-trichloroethane	2.481	2.797	0.940	0.000	0.000	1.775	0.000	0.000
2	1,1,2,2-tetrachloroethane	2.644	2.872	0.916	0.287	0.000	1.901	0.000	1.080
3	1,2,3,4-tetrachlorobenzene	4.994	4.875	0.900	0.324	0.054	2.162	0.808	0.983
4	1,2,3,4-tetramethylbenzene	4.738	4.707	0.869	0.324	0.054	2.099	0.738	0.862
5	1,2,3,5-tetracholorobenzene	4.994	4.912	0.900	0.241	0.054	2.162	0.896	0.864
6	1,2,3,5-tetramethylbenzene	4.738	4.712	0.869	0.241	0.054	2.099	0.812	0.760
7	1,2,3-trichlorobenzene	4.281	4.165	0.916	0.241	0.066	2.033	0.694	0.793
8	1,2,3-trimethylbenzene	4.089	4.067	0.867	0.241	0.066	1.979	0.641	0.690
9	1,2,4,5-tetrachlorobenzene	4.994	4.896	0.900	0.287	0.054	2.162	0.818	0.895
10	1,2,4,5-tetramethylbenzene	4.738	4.711	0.869	0.287	0.054	2.099	0.744	0.788
11	1,2,4-trichlorobenzene	4.281	4.113	0.916	0.154	0.066	2.033	0.692	0.663
12	1,2,4-trimethylbenzene	4.089	3.991	0.867	0.154	0.066	1.979	0.637	0.582
13	1,2-dibromobenzene	3.588	4.063	0.900	0.154	0.080	2.109	0.698	1.018
14	1,2-dibromoethane	1.738		0.916	0.000	0.000	1.847	0.000	0.000
15	1,2-dichlorobenzene	3.568	3.414	0.900	0.154	0.080	1.884	0.537	0.559
16	1,2-dicholoroethane	1.458	1.862	0.916	0.000	0.000	1.544	0.000	0.000
17	1,2-diphenylethane	4.888	4.931	0.831	0.000	0.186	2.218	0.947	0.364
18	1,3,5-tricholorobenzene	4.281	4.195	0.916	0.000	0.066	2.033	0.873	0.449
19	1,3,5-trimethylbenzene	4.089	4.005	0.867	0.000	0.066	1.979	0.789	0.405
20	1,3-dicholorobenzene	3.568	3.417	0.900	0.000	0.080	1.884	0.640	0.341
21	1,3-dimethylnaphthalene	4.614	4.657	0.965	0.152	0.130	2.136	0.989	0.595
22	1,4,5-trimethylnaphthalene	5.263	5.259	0.972	0.262	0.117	2.239	1.079	0.782
23	1,4-dibromobenzene	3.868	4.172	0.900	0.000	0.080	2.109	0.672	0.563
24	1,4-dicholorobenzene	3.568	3.269	0.900	0.000	0.080	1.884	0.520	0.362
25	1,4-dimethylnaphthalene	4.614	4.698	0.965	0.222	0.130	2.136	0.956	0.665
26	1,5-dimethylnaphthalene	4.614	4.716	0.965	0.222	0.128	2.136	0.965	0.665
27	1-butene	2.266	1.627	0.651	0.000	0.000	1.384	0.000	0.000
28	1-chlorobutane	2.523	2.917	0.788	0.000	0.000	1.659	0.337	0.000
29	1-chloroheptane	4.110	4.719	0.752	0.000	0.000	1.998	0.635	0.000
30	1-chlorohexane	3.581	4.175	0.761	0.000	0.000	1.897	0.536	0.000
31	1-chloronaphthalene	4.029	4.166	0.969	0.152	0.141	2.038	0.888	0.534
32	1-chloropentane	3.052	3.568	0.773	0.000	0.000	1.785	0.427	0.000
33	1-chloropropane	1.994	1.920	0.807	0.000	0.000	1.515	0.000	0.000
34	1-ethylnaphthalene	4.494	4.513	0.965	0.128	0.141	2.110	0.956	0.516
35	1-hexene	3.324	3.220	0.651	0.000	0.000	1.688	0.299	0.000
36	1-isopropyl-4-methylbenzene	4.368	4.323	0.869	0.154	0.080	2.065	0.668	0.604
37	1-methylbenz(a)anthracene	6.313	6.178	0.977	0.286	0.233	2.472	1.384	0.870
38	1-methylfluorene	4.874	4.946	0.839	0.243	0.310	2.225	1.199	0.777
39	1-methylnaphthalene	3.965	4.136	0.945	0.152	0.141	2.020	0.870	0.509
40	1-pentene	2.795	2.510	0.651	0.000	0.000	1.547	0.186	0.000
41	12-methylbenz(a)anthracene	6.313	6.237	0.977	0.322	0.237	2.472	1.410	0.901
42	2,2',4,5-tetrachlorobiphenyl	6.882	6.462	0.890	0.347	0.130	2.485	1.192	1.005
43	2,2',4-trichlorobiphenyl	6.169	5.964	0.873	0.222	0.141	2.393	1.143	0.797
44	2,2,4-trimethylpentane	4.536	4.701	0.637	0.000	0.000	2.052	0.800	0.597

Table 5 (continued)

Sequence number	Chemical name	Log P	Predicted log P (eq. (9))	IC_0	$^5\chi_C$	$^6\chi_{CH}$	$^0\chi^v$	$^4\chi^v$	$^4\chi^v_{PC}$
45	2,3,4,5-tetrachlorobiphenyl	6.882	6.451	0.890	0.409	0.137	2.485	1.198	1.129
46	2,3-dimethylnaphthalene	4.614	4.632	0.965	0.223	0.130	2.136	0.903	0.679
47	2,4'-dichlorobiphenyl	5.456	5.385	0.842	0.152	0.154	2.291	0.992	0.654
48	2,4,5-trichlorobiphenyl	6.169	5.957	0.873	0.286	0.147	2.393	1.104	0.880
49	2,4,6-trichlorobiphenyl	6.169	6.024	0.873	0.193	0.147	2.393	1.241	0.764
50	2,5-dichlorobiphenyl	5.456	5.404	0.842	0.152	0.157	2.291	1.006	0.643
51	2,6-dichlorobiphenyl	5.456	5.546	0.842	0.193	0.157	2.291	1.097	0.684
52	2,6-dimethylnaphthalene	4.614	4.476	0.965	0.080	0.128	2.136	0.910	0.549
53	2-chlorobiphenyl	4.743	5.016	0.792	0.152	0.170	2.178	0.939	0.534
54	2-chloronaphthalene	4.029	4.049	0.969	0.080	0.141	2.038	0.851	0.459
55	2-chlorophenanthrene	5.203	5.221	0.985	0.176	0.188	2.285	1.152	0.681
56	2-chlorotoluene	3.504	3.283	0.941	0.154	0.080	1.863	0.523	0.521
57	2-methylanthracene	5.139	5.140	0.969	0.154	0.188	2.272	1.118	0.651
58	2-methylbutane	3.209	2.496	0.628	0.000	0.000	1.665	0.000	0.342
59	2-methylhexane	4.267	4.080	0.635	0.000	0.000	1.902	0.477	0.254
60	2-methylnaphthalene	3.965	4.015	0.945	0.080	0.141	2.020	0.837	0.443
61	2-methylpentane	3.738	3.592	0.632	0.000	0.000	1.790	0.456	0.254
62	2-methylphenanthrene	5.139	5.192	0.969	0.176	0.188	2.272	1.142	0.668
63	3-chlorotoluene	3.504	3.254	0.941	0.000	0.080	1.863	0.615	0.324
64	4-chlorotoluene	3.504	3.118	0.941	0.000	0.080	1.863	0.506	0.344
65	5,6-dimethylchrysene	6.962	6.657	0.997	0.421	0.230	2.547	1.466	1.052
66	5-methylchrysene	6.313	6.231	0.977	0.308	0.237	2.472	1.410	0.878
67	6-methylbenzo(e)pyrene	6.773	6.687	0.990	0.400	0.264	2.553	1.575	1.031
68	6-methylchrysene	6.313	6.244	0.977	0.326	0.237	2.472	1.397	0.890
69	7-ethylbenz(a)anthracene	6.842	6.466	0.997	0.306	0.237	2.530	1.457	0.899
70	7-methylbenz(a)anthracene	6.313	6.249	0.977	0.339	0.237	2.472	1.397	0.912
71	9,10-dimethylanthracene	5.788	5.832	0.989	0.354	0.186	2.363	1.271	0.896
72	9-methylanthracene	5.139	5.316	0.969	0.257	0.192	2.272	1.176	0.740
73	acenapthene	4.070	4.026	0.985	0.154	0.284	2.064	1.159	0.631
74	adamantane	3.982	4.376	0.673	0.000	0.241	2.022	1.606	0.881
75	anthracene	4.490	4.815	0.909	0.154	0.200	2.172	1.059	0.556
76	benz(a)anthracene	5.664	5.826	0.925	0.243	0.245	2.391	1.307	0.759
77	benz(b)anthracene	5.664	5.783	0.925	0.223	0.245	2.391	1.290	0.743
78	benzene	2.142	2.031	0.693	0.000	0.118	1.496	0.326	0.000
79	benzo(a)fluorene	5.399	5.298	0.961	0.265	0.360	2.350	1.363	0.820
80	benzo(a)pyrene	6.124	6.299	0.940	0.317	0.268	2.479	1.502	0.922
81	benzo(b)fluoranthene	6.124	5.984	0.940	0.355	0.405	2.479	1.516	0.935
82	benzo(b)fluorene	5.399	5.232	0.961	0.243	0.367	2.350	1.342	0.799
83	benzo(e)pyrene	6.124	6.346	0.940	0.341	0.272	2.479	1.517	0.933
84	benzo(ghi)perylene	6.584	6.732	0.948	0.392	0.293	2.559	1.656	1.065
85	benzo(j)fluoranthene	6.124	5.986	0.940	0.341	0.395	2.479	1.517	0.933
86	benzo(k)fluoranthene	6.124	5.942	0.940	0.334	0.401	2.479	1.499	0.926
87	biphenyl	4.030	4.529	0.690	0.080	0.186	2.051	0.816	0.326
88	bromobenze	3.005	3.230	0.844	0.000	0.097	1.849	0.542	0.321
89	carbon tetrachloride	2.875	3.544	0.543	0.000	0.000	1.798	0.000	0.000

Table 5 (continued)

Sequence number	Chemical name	Log P	Predicted log P (eq. (9))	IC_0	$^5\chi_C$	$^6\chi_{CH}$	$^0\chi^v$	$^4\chi^v$	$^4\chi^v_{PC}$
90	chloanthrene	6.418	6.065	1.008	0.330	0.358	2.500	1.577	0.981
91	chlorobenzene	2.855	2.663	0.844	0.000	0.097	1.709	0.445	0.197
92	chrysene	5.664	5.877	0.925	0.265	0.245	2.391	1.327	0.776
93	cycloheptane	3.913	4.160	0.651	0.000	0.000	1.783	0.805	0.000
94	cyclohexane	3.354	3.198	0.651	0.000	0.118	1.657	0.723	0.000
95	cyclohexene	2.810	2.786	0.670	0.000	0.118	1.606	0.565	0.000
96	cyclooctane	4.472	4.713	0.651	0.000	0.000	1.896	0.881	0.000
97	cyclopentane	2.795	2.836	0.651	0.000	0.000	1.512	0.633	0.000
98	cyclopentene	2.251	2.371	0.673	0.000	0.000	1.453	0.463	0.000
99	dibenz(ah)anthracene	6.838	6.659	0.933	0.324	0.287	2.571	1.506	0.929
100	dibenz(aj)anthracene	6.838	6.659	0.933	0.324	0.287	2.571	1.507	0.929
101	diethyl sulfide	1.900		0.770	0.000	0.000	1.730	0.477	0.000
102	dimethyl sulfide	0.842	1.622	0.799	0.000	0.000	1.441	0.000	0.000
103	ethyl chloride	1.465	1.167	0.832	0.000	0.000	1.346	0.000	0.000
104	ethylbenzene	3.320	3.129	0.855	0.000	0.097	1.807	0.539	0.226
105	fluoranthene	4.950	5.058	0.936	0.272	0.363	2.280	1.320	0.772
106	fluorene	4.225	4.461	0.793	0.176	0.329	2.119	1.125	0.623
107	fluorobenzene	2.285	2.069	0.844	0.000	0.097	1.561	0.347	0.070
108	fluorotrichloromethane	2.435	2.455	0.863	0.000	0.000	1.664	0.000	0.000
109	hexachlorobenzene	6.420	6.534	0.693	0.511	0.036	2.380	1.099	1.329
110	hexamethylbenzene	6.036	5.966	0.863	0.511	0.036	2.303	0.989	1.179
111	iodobenzene	3.265	3.579	0.844	0.000	0.097	1.935	0.604	0.397
112	isopropylbenzene	3.719	3.802	0.867	0.154	0.097	1.941	0.609	0.493
113	naphthalene	3.316	3.582	0.872	0.080	0.154	1.890	0.758	0.326
114	pentachlorobenzene	5.707	5.667	0.844	0.401	0.044	2.277	0.976	1.135
115	pentachloroethane	3.627	3.844	0.832	0.624	0.000	2.047	0.000	1.566
116	pentamethylbenzene	5.387	5.365	0.867	0.401	0.044	2.206	0.881	1.001
117	perylene	6.124	6.458	0.878	0.339	0.269	2.479	1.522	0.932
118	phenanthrene	4.490	4.866	0.909	0.176	0.200	2.172	1.083	0.575
119	pyrene	4.950	5.379	0.936	0.230	0.221	2.280	1.300	0.757
120	tetrachloroethylene	3.020	3.394	0.651	0.287	0.000	1.877	0.000	0.899
121	toluene	2.791	2.601	0.821	0.000	0.097	1.684	0.428	0.176
122	tricholoethylene	2.267	2.193	0.900	0.000	0.000	1.701	0.000	0.351
123	triphenylene	5.664	5.924	0.925	0.287	0.252	2.391	1.350	0.781
124	m-xylene	3.440	3.300	0.855	0.000	0.080	1.842	0.592	0.307
125	n-decane	5.984	5.743	0.640	0.000	0.000	2.158	0.792	0.000
126	n-nonane	5.455	5.297	0.639	0.000	0.000	2.073	0.708	0.000
127	n-octane	4.926	4.813	0.637	0.000	0.000	1.980	0.617	0.000
128	n-undecane	6.513	6.153	0.641	0.000	0.000	2.237	0.869	0.000
129	n-heptane	4.397	4.277	0.635	0.000	0.000	1.877	0.517	0.000
130	n-butane	2.810	2.098	0.622	0.000	0.000	1.485	0.000	0.000
131	n-butylbenzene	4.378	4.191	0.869	0.000	0.097	2.016	0.706	0.198
132	n-hexane	3.868	3.681	0.632	0.000	0.000	1.763	0.405	0.000
133	n-pentane	3.339	3.035	0.628	0.000	0.000	1.633	0.303	0.000
134	n-propylbenzene	3.849	3.729	0.867	0.000	0.097	1.917	0.659	0.198

Table 5 (continued)

Sequence number	Chemical name	Log P	Predicted log P (eq. (9))	IC_0	$^5\chi_C$	$^6\chi_{CH}$	$^0\chi^v$	$^4\chi^v$	$^4\chi^v_{PC}$
135	o-xylene	3.440	3.359	0.855	0.154	0.080	1.842	0.509	0.484
136	p-xylene	3.440	3.176	0.855	0.000	0.080	1.842	0.493	0.326
137	tert-amylbenzene	4.647	4.609	0.867	0.299	0.097	2.151	0.739	1.003
138	tert-butylbenzene	4.118	4.506	0.869	0.360	0.097	2.065	0.662	0.794
139	trans-1,2-dichloroethylene	1.514	1.567	0.950	0.000	0.000	1.487	0.000	0.000

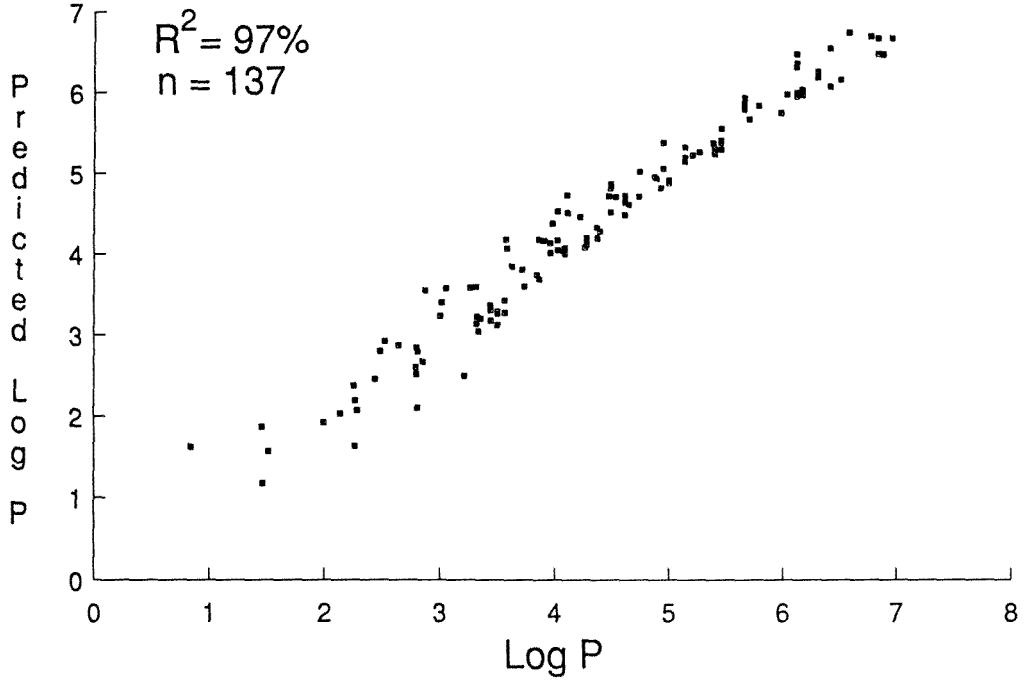

Fig. 2. The plot of predicted log P (using the six-parameter model of eq. (9)) versus log P for 137 compounds (table 5).

$$\log P = -3.127 - 1.644(IC_0) + 2.120(^5\chi_C) - 2.914(^6\chi_{CH}) + 4.208(^0\chi^v + 1.060(^4\chi^v) - 1.020(^4\chi^v_{PC}), \quad (n = 137,\ R^2 = 0.97,\ se = 0.26). \quad (9)$$

The distribution of residuals based on eq. (9) was normal (Wilks–Shapiro, $p = 0.082$) and no outliers were determined through Cook's D statistic [102,101]. For the group of 139 compounds, the values of the six algorithmically defined predictors which appeared in eq. (9) are given in table 5. A plot of predicted log P (using eq. (9)) versus log P

is shown in fig. 2. Results show that there is an improvement of the predictability of log P (R^2 = 0.97) for the set of 137 chemicals as compared to the total set of 382 compounds (R^2 = 0.91) reported earlier [6].

7. Discussion

The primary purpose of this paper was to develop a scheme for classification of diverse sets of molecules and the prediction of their properties using algorithmically defined structural variables. The success of the approach is evident from results presented by eq. (9), which shows that a combination of molecular connectivity indices and molecular complexity indices can efficiently predict (R^2 = 0.965) log P values of a relatively homogeneous group of weakly hydrogen-bonding (HB_1 = 0) chemicals. In an earlier study [6], we reported that log P (octanol–water) of a large and diverse set of 382 chemicals could be predicted reasonably well (R^2 = 0.91) with a combination of molecular connectivity indices, molecular complexity indices, and the hydrogen bonding parameter HB_1. This is in agreement with the finding that lipophilicity of a molecule is related to its size, polarizability, and ability to form hydrogen bonds [41]. It is known that many of the connectivity indices, the Wiener index W, I_D^W and $\bar{I}_D^W$ are highly correlated with molecular size [6,18,35,40,46,103]. Although each of these indices may quantitate different proportions of bulk and shape factors, size seems to be the principal molecular factor encoded by these indices [6,35,46,103]. Many of these invariants are based on simple linear graph models of molecules. A linear graph grossly oversimplifies the complex reality of a molecule by depicting only its primary structure (i.e. connectivity of atoms) and neglecting other structural features, e.g. bond length, bond angle, stereochemistry, chirality, etc. [74]. Yet, the success of graph-theoretic invariants derived from linear graphs in predicting physicochemical/biological properties of congeners is well known [1–14,74,75]. This indicates that for reasonably homogeneous groups of structures or for molecules with a specific biochemical mode of action, a property is primarily governed by the pattern of connectedness of atoms as opposed to specific properties of certain atoms, functional groups or substructures. On the other hand, molecular complexity indices are defined on weighted multigraphs which account for the heterogeneity of atomic envirnoment in the molecule [16,22,27,28,77,79,93]. Our earlier study on a set of 3692 structurally diverse chemicals showed that complexity indices contain information not encoded by simple connectivity, valence connectivity, W, I_D^W and $\bar{I}_D^W$ parameters [35,46].

Although molecular size and heterogeneity are accounted for, in terms of connectivity and complexity parameters, respectively, they are not able to predict log P very efficiently [6]. This is because these parameters are incapable of quantifying hydrogen bonding, a proximity effect of substituents. A substituent may modify properties of the parent structure (or reaction center) through a variety of interactions: (1) *proximity effect*, e.g. hydrogen bonding, neighboring group involvement, steric retardation or acceleration, (2) *resonance effects* via delocalization, (3) *direct electrostatic interactions* arising from substituent poles or dipoles, and (4) *indirect actions* by means of

polar effects [104]. The prediction of lipophilicity of molecules should be done using parameters which *optimally characterize* major determinants of that property, viz., molecular size, polarity and hydrogen bonding. For the set of 382 chemicals, this could be achieved at an acceptable level (R^2 = 0.91) through a combination of molecular connectivity indices, molecular complexity indices, and HB_1 [6].

The quantitation of hydrogen bonding ability of a molecule is a complicated process. Different empirical and theoretical methods have been used to quantitate hydrogen bonding capacity of molecules [43,73,104–107]. HB_1 is a convenient quantifier of hydrogen bonding and can be calculated directly from structure. Also, it is an approximate parameter and has integral values.

In our earlier studies, we used HB_1 as an independent variable [6]. An alternative use of HB_1 could be in the classification of diverse sets of molecules into relatively homogeneous subsets based on the degree of hydrogen bonding. The simplest classification could partition a set of chemicals into strongly hydrogen bonding ($HB_1 \geq 1$) and weakly hydrogen bonding ($HB_1 = 0$). Of the 382 chemicals analyzed in our previous study, 139 have $HB_1 = 0$. Results of regression analysis (eq. (9)) show that a preselected set of graph-theoretic invariants can effectively predict (R^2 = 0.97) log P values of the more homogeneous set of weakly hydrogen bonding chemicals (fig. 2). The parameters used for this study are a subset of 70 graph invariants (table 3) which appeared in linear regression models of log P containing up to 10 independent variables for the entire set of 382 chemicals [6].

In this paper, we have used graph invariants defined on linear graphs and multigraphs. While the list of descriptors chosen for this study is not exhaustive, the set of indices used in this paper appear to optimally characterize aspects of molecular structure pertinent to the prediction of log P (octanol–water). Further studies with other properties are needed to evaluate the utility of this approach in the characterization of molecular structure and the prediction of properties.

Acknowledgements

The authors are thankful to Greg Grunwald, Ann Lima and Anda Bellamy for their assistance in the project. Contribution number 64 from the Center for Water and the Environment.

References

[1] D.E. Needham, I.C. Wei and P.G. Seybold, J. Amer. Chem. Soc. 110(1988)4186.
[2] D.H. Rouvray and W. Tatong, Z. Naturforsch. 41a(1986)1238.
[3] Y. Gao and H. Hosoya, Bull. Chem. Soc. Japan 61(1988)3093.
[4] M. Randić, J. Amer. Chem. Soc. 97(1975)6609.
[5] O. Mekenyan, S. Dimitrov and D. Bonchev, Eur. Polymer J. 19(1983)1185.
[6] S.C. Basak, G.J. Niemi and G.D. Veith, in: *Computational Chemical Graph Theory*, ed. D.H. Rouvray (Nova, New York, 1989), in press.

[7] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Ind. J. Chem. 20B(1981)894.
[8] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Arzneim. Forsch. 32(1982)322.
[9] V.R. Magnuson, D.K. Harriss and S.C. Basak, in: *Chemical Applications of Topology and Graph Theory*, ed. R.B. King (Elsevier, Amsterdam, 1983), p. 178.
[10] R.J. Baker, W.E. Acree and C.-C. Tsai, Quant. Struct.-Act. Relat. 3(1984)10.
[11] O. Mekenyan, D. Bonchev and N. Trinajstić, Int. J. Quant. Chem. 18(1980)369.
[12] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure–Activity Analysis* (Research Studies Press, Letchworth, Hertfordshire, UK, 1986).
[13] M. Randić, Int. J. Quant. Chem. Quant. Biol. Symp. 11(1984)137.
[14] A. Sabljić and N. Trinajstić, Acta Pharm. Yugosl. 31(1981)189.
[15] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Ind. J. Pharmacol. 13(1982)301.
[16] C. Raychaudhury, S.C. Basak, A.B. Roy and J.J. Ghosh, Indian Drugs 18(1980)97.
[17] S.C. Basak, S.K. Ray, C. Raychaudhury, A.B. Roy and J.J. Ghosh, IRCS Med. Sci. 10(1982)145.
[18] S.C. Basak, D.P. Gieschen, V.R. Magnuson and D.K. Harriss, IRCS Med. Sci. 10(1982)619.
[19] S.C. Basak, D.K. Harriss and V.R. Magnuson, J. Pharm. Sci. 73(1984)429.
[20] S.C. Basak, D.P. Gieschen, D.K. Harriss and V.R. Magnuson, J. Pharm. Sci. 72(1983)934.
[21] S.C. Basak, L.J. Monsrud, M.E. Rosen, C.M. Frane and V.R. Magnuson, Acta Pharm. Yugosl. 36(1986)81.
[22] S.C. Basak, Med. Sci. Res. 15(1987)605.
[23] S.C. Basak, Med. Sci. Res. 16(1988)281.
[24] N. Trinajstić, M. Randić and D.J. Klein, Acta Pharm. Yugosl. 36(1986)267.
[25] M. Randić, S.C. Grossman, B. Jerman-Blažič, D.H. Rouvray and S. El-Basil, Math. Comput. Modelling 11(1988)837.
[26] M. Yuan and P.C. Jurs, Toxicol. Appl. Pharmacol. 52(1980)294.
[27] G. Klopman and C. Raychaudhury, J. Comput. Chem. 9(1988)232.
[28] S.C. Basak and V.R. Magnuson, Arzneim. Forsch. 33(1983)501.
[29] S.C. Basak, D.P. Gieschen and V.R. Magnuson, Environ. Toxicol. Chem. 3(1984)191.
[30] S.C. Basak, Med. Sci. Res. 16(1988)281.
[31] S.C. Basak, C.M. Frane, M.E. Rosen and V.R. Magnuson, IRCS Med. Sci. 14(1986)848.
[32] P.G. Seybold, Int. J. Quant. Chem. Quant. Biol. Symp. 10(1983)103.
[33] M. Johnson, in: *Graph Theory and its Applications to Algorithms and Computer Science*, ed. Y. Alavi, G. Chartrand, L. Lesniak, D.R. Lick and C.E. Wall (Wiley, New York, 1985), p. 457.
[34] M. Johnson, S.C. Basak and G. Maggiora, Math. Comput. Modelling 11(1988)630.
[35] S.C. Basak, V.R. Magnuson, G.J. Niemi and R.R. Regal, Discr. Appl. Math. 19(1988)17.
[36] M. Randić, in: *Computer Based Methods of Molecular Similarity*, ed. G.M. Maggiora (Wiley, New York, 1989), in press.
[37] K. Enslein, H.H. Borgstedt, M.E. Tomb, B.W. Blake and J.B. Hart, Toxicol. Ind. Health 3(1987)267.
[38] M. Vighi and D. Calamari, Chemosphere 16(1987)1043.
[39] G.J. Niemi, G.D. Veith and R.R. Regal, Environ. Toxicol. Chem. 6(1987)515.
[40] G.J. Niemi, R.R. Regal and G.D. Veith, in: *Environmental Applications of Chemometrics*, ed. J.J. Breen and P.E. Robinson (American Chemical Society, Washington, DC, 1984), p. 148.
[41] C. Hansch, in: *Correlation Analysis in Chemistry*, ed. N.B. Chapman and J. Shorter (Plenum, New York, 1978), p. 397.
[42] R.W. Taft, T. Gramstad and M.J. Kamlet, J. Org. Chem. 47(1982)4557.
[43] M.J. Kamlet, R.M. Doherty, R.W. Taft, M.H. Abraham, G.D. Veith and D.J. Abraham, Environ. Sci. Tech. 21(1987)149.
[44] W.G. Richards, *Quantum Pharmacology* (Butterworths, London, 1977).
[45] A.J. Stuper, W.E. Brugger and P.C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function* (Wiley, New York, 1979).
[46] S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal and G.D. Veith, Math. Modelling 8(1987)300.

[47] C.T. Helmes, C.C. Sigman and P.A. Papa, CHEMTECH 15(1985)48.
[48] J.C. Arcos, Environ. Sci. Tech. 21(1987)743.
[49] G.D. Veith, D.J. Call and L.T. Brokke, Can. J. Fish Aquat. Sci. 40(1983)743.
[50] G.M. Maggiora, M.A. Johnson, M.S. Lajiness, A.B. Miller and T.R. Hagadone, Math. Comput. Modelling 11(1988)626.
[51] W.J. Lyman, W.F. Reehl and D.H. Rosenblatt, *Handbook of Chemical Property Estimation Methods* (McGraw-Hill, New York, 1982).
[52] T. Nogrady, *Medicinal Chemistry: A Biochemical Approach* (Oxford University Press, New York, 1985).
[53] S.H. Yalkowsky, A.A. Sinkula and S.C. Valvani, *Physical Chemical Properties of Drugs* (Marcel Dekker, New York, 1980).
[54] H.G.S. van Raalte, Ecotoxicol. Environ. Safety 4(1980)466.
[55] F. Moriarty, *Ecotoxicology: The Study of Pollutants in Ecosystems* (Academic Press, London, 1988).
[56] C.E. Searle, *Chemical Carcinogens*, Vols. 1, 2 (American Chemical Society, Washington, DC, 1984).
[57] H. Primas, *Chemistry, Quantum Mechanics and Reductionism* (Springer-Verlag, Berlin, 1981).
[58] S.J. Weininger, J. Chem. Educ. 61(1984)939.
[59] R.G.J. Woolley, J. Amer. Chem. Soc. 100(1978)1073.
[60] A. Einstein, B. Podolsky and N. Rosen, Phys. Rev. 47(1935)777.
[61] M. Bunge, *Method, Model and Matter* (Reidel, Dordrecht–Boston, 1973).
[62] J.J. Sylvester, Amer. J. Math. 1(1878)64.
[63] A.T. Balaban, J. Chem. Inf. Comput. Sci. 25(1985)334.
[64] A.K. Ghose, A. Pritchett and G.M. Crippen, J. Comput. Chem. 9(1988)80.
[65] A. Leo, P. Jow, C. Silipo and C. Hansch, J. Med. Chem. 18(1975)865.
[66] R. Rekker, *The Hydrophobic Fragmental Constant* (Elsevier, Amsterdam, 1977).
[67] C. Hansch, in: *Advances in Pharmacology and Chemotherapy*, ed. S. Garattini, A. Goldin, F. Hawking and I.J. Kopin (Acadmic Press, New York, 1975), p. 45.
[68] F. Harary, *Graph Theory* (Addison-Wesley, Reading, MA, 1969).
[69] A. Albert, *Selective Toxicity* (Chapman and Hill, London, 1973).
[70] M.J. Kamlet, R.M. Doherty, G.D. Veith, R.W. Taft and M.H. Abraham, Environ. Sci. Tech. 20(1986)690.
[71] M. Randić and C.L. Wilkins, J. Phys. Chem. 83(1979)1525.
[72] S.C. Basak, H-BOND: A program for calculating hydrogen bonding parameters, University of Minnesota–Duluth (1988).
[73] X.-C. Ou, Y. Ouyang and E.J. Lien, J. Mol. Sci. (China) 4(1986)89.
[74] N. Trinajstić, *Chemical Graph Theory* (CRC Press, Boca Raton, FL, 1983).
[75] H. Wiener, J. Amer. Chem. Soc. 69(1947)17.
[76] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
[77] S.K. Ray, S. Gupta, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Ind. J. Chem. 24B(1985)1149.
[78] E.J. Kupchik, Quant. Struct.-Act. Relat. 7(1988)57.
[79] S.C. Basak, A.B. Roy and J.J. Ghosh, in: *Proc. 2nd Int. Conf. on Mathematical Modelling*, Vol. 2, ed. X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler (University of Missouri, Rolla, MO, 1980), p.851.
[80] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy and S.C. Basak, J. Comput. Chem. 5(1984)581.
[81] D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures* (Research Studies Press, Chichester, 1983);
D. Bonchev and N. Trinajstić, J. Chem. Phys. 67(1977)4517.
[82] S.H. Bertz, Bull. Math. Biol. 45(1983)849.
[83] C.E. Shannon, Bell Syst. Tech. J. 27(1948)379.
[84] N. Wiener, *Cybernetics* (Wiley, New York, 1948).

[85] W. Ashby, *An Introduction to Cybernetics* (Wiley, New York, 1956).
[86] A.N. Kolmogorov, Probl. Peredachi. Inf. 5(1969)3.
[87] C.W. Marshall, *Applied Graph Theory* (Wiley–Interscience, New York, 1971).
[88] E. Trucco, Bull. Math. Biophys. 18(1956)129
[89] A. Mowshowitz, Bull. Math. Biophys. 30(1968)175.
[90] A. Mowshowtiz, Bull. Math. Biophys. 30(1968)225.
[91] N. Rashevsky, Bull. Math. Biophys. 17(1955)229.
[92] L.B. Kier, J. Pharm. Sci. 69(1980)807.
[93] R. Sarkar, A.B. Roy and P.K. Sarkar, Math. Biosci. 39(1978)299.
[94] A.B. Roy, S.C. Basak, D.K. Harriss and V.R. Magnuson, in: *Mathematical Modelling in Science and Technology*, ed. X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin (Pergamon, New York, 1984), p. 745.
[95] P.E. Long, *An Introduction to General Topology* (Merrill, Columbus, Ohio, 1971).
[96] L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1956).
[97] S.C. Basak, D.K. Harriss and V.R. Magnuson, POLLY: Copyright of the University of Minnesota (1988).
[98] E. Anderson, G.D. Veith and D. Weininger, Report No. EPA/600/M-87/021, Environmental Research Laboratory–Duluth (1987).
[99] MedChem Software, version 3.53, Daylight Chemical Information Systems Inc., Claremont, CA (1988).
[100] SAS/STAT User Guide, release 6.03 edition, SAS Institute Inc., Cary, NC (1988), p. 1028.
[101] R.D. Cook, Technometrics 19(1977)15.
[102] S.S. Shapiro and M.B. Wilk, Biometrika 52(1965)591.
[103] I. Motoc, A.T. Balaban, O. Mekenyan and D. Bonchev, MATCH 13(1982)369.
[104] P.R. Wells, *Linear Free Energy Relationships* (Academic Press, London, 1968).
[105] M.H. Abraham, P.P. Duce, P.L. Grellier, D.V. Prior, J.J. Morris and P.J. Taylor, Tetrahedron Lett. 29(1988)1587.
[106] V.A. Terent'ev, Russ. J. Phys. Chem. 46(1972)1103.
[107] B.L. Karger, L.R. Snyder and C. Eon, J. Chromatogr. 125(1976)71.