CROATICA CHEMICA ACTA CCACAA **64** (1) 43–54 (1991)

CCA-1978

YU ISSN 0011-1643
UDC 541
Original Scientific Paper

Search for Optimal Molecular Descriptors

Milan Randić

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, USA

Received July 10, 1990.

We illustrate, on a set of heptane isomers and van der Waals molecular areas as a property, a search for optimal descriptors to be used in multivariate analyses. We employed molecular connectivities as the basis descriptors and examined their various combinations as alternative descriptors that could lead to a better regression. The critical step in the search appears to be the selection of the first descriptor to which, subsequently, other descriptors are made orthogonal. A well constructed descriptor can yield a regression equation superior to regressions using several standard descriptors. The approach is illustrated by considering, among others, the molecular *ID* numbers, the Wiener numbers and Balaban's J index, as well as novel graphical bond orders based on the above indices. A by-product of the systematic search for better structure-property relationship is a novel property-property correlation (between octane number and molecular cavity areas).

INTRODUCTION

The multiple regression analysis is a useful tool in structure-property and structure-activity studies. It is one of the oldest, multidisciplinary, and the most widely used statistical methods. The quality of reported regression in structure-property and particularly in structure-activity studies varies considerably. The quality of a multivariate regression is indicated qualitatively by the ratio of the number of variables used and the size of the sample. Qualitatively, the quality can be conveniently measured by the correlation coefficient R and standard error S. In this paper, we will focus attention on improving the quality of the multivariable regression method primarily by outlining a way of reducing the number of descriptors employed while at the same time maintaining or improving the statistical parameters R and S. Our interest will be directed towards locating the best combinations of descriptors that have been selected for testing in a given problem. In particular, we consider variations in the molecular van der Waals areas of heptane isomers and will start with an examination of the connectivity

index[1] and higher connectivity indices.[2] The approach, however, equally applies to other molecules, other properties and other molecular descriptors, including also the properties used as descriptors. The main tool in our approach is a recently outlined procedure for orthogonalization of molecular descriptors.[3,4]

VAN DER WAALS AREAS OF HEPTANE ISOMERS

The effective molecular area as a property is of considerable interest in structure-property and structure-activity studies. Its dependence on the size of a molecule is quite apparent. However, its variation between molecules of a same size but different shape is more subtle. A number of theoretical approaches to molecular area and molecular shape have been reported in the literature.[5-10] One of the earlier studies is an ambitious approach of Hermann[5] who considered idealized areas of the molecular cavity surface that contained spherical representations for water molecules in the first layer around solute. Computationally, the approach is of considerable complexity since typical molecular cavities are nonspherical and depend on details of molecular geometry. It was, therefore, significant to find that a simply computed connectivity index correlated surprisingly well with so elaborately calculated theoretical quantities.[1,6] A variety of approaches concerning molecular shape, an ambiguous and elusive concept, focused attention at specific aspects of the concept. Rohrbaugh and Jurs[7] tried to characterize 3-dimensional molecular shape by half a dozen descriptors that encode the information from three orthogonal projections of a molecular model. Mezey[8] concentrated on the topological aspects of shape, in which boundaries between parts of a surface of diverse concavity (so called catchment regions) play an important role. Kler[9] introduced a scale that assigns a shape index κ to a graph, based on shapes of extreme acyclic graphs, a path and a »star«; while Motoc[10] used yet another approach in which shapes can be associated with codes (suitable for computer processing). Clearly, factors that are important in specifying molecular shape and which often discriminate between isomers is the number of »exposed« atoms and atomic groups, such as primary and secondary carbon atoms, as opposed to »buried« atoms and groups, such as tertiary and quaternary carbon atoms. The former, illustrated by methyl and methylene groups in hydrocarbons, are expected to make a larger contribution to the overall molecular area; the latter, illustrated by methine group, are expected to make smaller contributions. With this hindsight, it is not surprising to find that the connectivity index,[1] in which contributions to a molecular property are bond additive and bonds of different type are weighted appropriately, so well described numerous isomeric variations of molecular properties. However, most of the past structure-property regression analyses involved molecules of widely differing sizes. As size in the apparent dominant component of many physicochemical properties, such correlations do not reveal in sufficient detail how a property depends on molecular shape. Only recently it was shown that apparent interrelatedness of numerous molecular properties does not necessarily extend to isomers when attention is restricted to molecules of the same size. This has been illustrated by considering correlations involving alkane isomers rather then using data on all alkanes.[11] We have therefore decided to focus our attention on the role of the molecular shape as reflected in the skeletal diversity of heptane isomers and to study variations in their properties as a function of selected molecular descriptors.

OUTLINE OF THE SEARCH STRATEGY

Table I lists heptane isomers, their connectivity indices, and their van der Waals areas. The connectivity indices can be found in the appendix to Kier and Hall's book,[12] while the molecular areas have been taken from a recent study of Labanowski, Motoc, and Dammkoehler,[13] based on a method of Pearlman[14] and corresponding to solvent spheres of radius 1.5 Å. Calculation of the molecular areas was based on standard molecular geometries, effective van der Waals atomic radii,[15] and molecular conformations that avoid overlapping of envelopes representing non-bonded atoms.

TABLE I

The connectivity indices and molecular van der Waals cavity areas for heptane isomers

Isomer	1χ	2χ	3χ	A^2
n-heptane	3.4142	2.0607	1.2071	334.36
3-E	3.3461	2.0908	1.7321	303.15
3-M	3.3081	2.3021	1.4784	316.67
2-M	3.2701	2.5361	1.1350	322.91
2,3-MM	3.1807	2.6295	1.7820	303.11
2,4-MM	3.1259	3.0234	0.9428	309.97
3,3-MM	3.1213	2.8713	1.9142	297.20
2,2-MM	3.0607	3.3107	1.0000	306.53
2,2,3-MM	2.9432	3.5207	1.7321	292.39

The first step in a systematic search for optimal correlation is to find regressions of the property (here molecular surface area) against all combinations of the adopted basis (descriptors). In Table II, we show the coefficients of multiple regression (R) and standard errors (S) obtained using the connectivity indices 1χ–3χ as descriptors. The size of the test sample (nine heptanes) at most justifies the use of two descriptors. We adopted three descriptors to facilitate outlining the methodology of the search, rather than to suggest that the use of such an extensive basis would be statistically sound. We will use primarily S as the indicator of the quality of regression, although we also report the correlation coefficients.

From Table II we see that isomeric variations of the molecular area are not well accounted for by any single connectivity index. The connectivity index 1χ, which

TABLE II

The statistical parameters (R = regression coefficient, S = standard error) for regression of molecular surface areas against the connectivity indices and their various combinations

Descriptors	R	S
1χ	0.7533	9.1881
2χ	0.6279	10.8744
3χ	0.5992	11.1858
$1\chi,2\chi$	0.9466	4.8670
$1\chi,3\chi$	0.9186	5.9656
$2\chi,3\chi$	0.9060	6.3887
$1\chi,2\chi,3\chi$	0.9475	5.2849

shows a marginally better performance than the other connectivity indices with $R = 0.753$ and $S = 9.188$, gives a regression equation that is still far from satisfactory. There are physicochemical properties of alkanes that can be satisfactorily described by a single connectivity index. For example heats of formation correlate well with $^1\chi$, while indices of refraction correlated well with $^2\chi$.[1,2] The present situation is somewhat »aggravated« by the fact that *single* descriptors were reported that offer satisfactory regression of surface areas of heptanes. Labanowski and coworkers[13] reported several information-based indices with S less than 3.00 and R above 0.97. In view of the less satisfactory results obtained with connectivity indices, the question is whether our basis is deficient and lacks some important structural ingredients that could improve regressions or we have not yet found the functional form $F(^1\chi,^2\chi,^3\chi)$ that would better characterize structural variations among heptane isomers.

When two connectivity indices are combined, we see a substantially increased R and a considerably reduced S. This suggested that combinations of connectivity indices may contain essential structural ingredients of interest when considering molecular surface areas in alkanes. Interestingly, combination of all three connectivity indices does not improve standard error, and hence would not be of interest on these grounds alone, regardless of a danger that it may represent a statistically not significant result. The question to consider is: Can we arrive at regressions of acceptable R and S, but based on fewer descriptors?

When a single descriptor dominates a regression, *i.e.*, accounts for the major part of a correlation, the first step in searching for an improved regression is to construct descriptors orthogonal to the dominant descriptor and examine how additional orthogonalized descriptors correlate with the residual of the single-variable regression.[3,4] Here, however, we have no such single dominant variable (yet) that would allow us to take advantage of the above indicated approach. Hence, we have to consider the problem of how to improve a regression which already involves several descriptors and how to reduce the number of descriptors while at the same time maintaining acceptable R and S.

To answer the above questions, we will evaluate individual descriptors in a regression independently of one another. A tool for such a task is provided by the recently introduced orthogonalization process for molecular descriptors.[3,4] For a given regression, one calculates the property $<P>$ based on selected basis descriptors B_1, B_2, B_3,..., which will be assumed to have been made mutually orthogonal. The difference between the property P and the calculated property $<P>$, *i.e.* the *residual* of the regression of P against the already adopted basis descriptors B_1, B_2, B_3, ..., represents the part of the property that the regression cannot »explain«. Additional descriptor, B_k, made orthogonal to all previously used descriptors and representing an enlargement of the basis, is then selected according to how well it correlates with the preceding residual. If the coefficient of such a correlation points to a statistically significant relationship, the new descriptor B_k is incorporated into the basis, otherwise it is discarded.

The use of residuals in an orthogonalization process is not novel, being incorporated in the principal component method of Hotelling.[16] What is novel in comparison with the principal component method is the direct applications of the orthogonalization process, in a sequential manner, to individual descriptors, one at a time. In contrast, orthogonalization implied in the principal component method is simultaneously performed on all descriptors, some of which are significant but many

of which are statistically not significant and irrelevant. Thus, for the first time, it is possible to select *dominant* components in a multivariate regression and, hence, the label Dominant Component Analysis (DCA) appears a suitable and an adequate name for the approach that complements the well established Principal Component Analysis (PCA) method. Not only that in DCA a researcher can at will include or discard a descriptor but the orthogonal basis does not involve irrelevant variables and consequently, makes interpretation of the regression possible. As it is well known, interpretations of the resulting orthogonal linear combinations of initial variables in PCA are often qualitative, somewhat arbitrary, ambiguous and most often impossible. A part of the difficulty is the nonorthogonality of the initial descriptors, many of which are highly intercorrelated. Our orthogonalization procedure, in fact, effectively separates variables into mutually uncorrelated descriptors, and the process is fully controlled by the researcher, who can make decision depending on the available prior steps in the analysis.

Orthogonalization procedure was fully outlined elsewhere.[3,4] Briefly, a descriptor B_j is made orthogonal to descriptor B_i in a process analogous to the Schmidt orthogonalization process for vectors in linear algebra, by making it orthogonal successively one descriptor at a time. To obtain B_j orthogonal to B_i one correlates B_j against B_i and view the residual, the difference $B_j - <B_j>$, where $<B_j>$ represents the calculated B_j, as the new orthogonal Ω_j descriptor. The process of orthogonalization continues by using the derived orthogonal descriptor Ω_j together with all preceding orthogonalized descriptors in a correlation with the next descriptor. When the last descriptor is processed in this way, the procedure ends and an orthogonal basis is obtained.

LINEAR COMBINATIONS OF DESCRIPTORS

Table III lists regression equations using orthogonal combinations of the connectivity indices as descriptors and the molecular surface area as a property. Note that the correlation coefficient and standard error are the same whether a pair of orthogonal or nonorthogonal descriptors are used. Also note that the coefficients of the same variable (connectivity index) are the same in different regression equations; minor variations are due to numerical rounding errors affecting the last digits. Hence, the regression equations associated with orthogonalized descriptors are stable as compared to those based on nonorthogonal descriptors. Since the coefficients associated with regression variables are now constant, they allow a meaningful interpretation of

TABLE III

The regression equations and accompanying statistical parameters for regressions using orthogonalized connectivity indices

Regression equations
$A = -15.828\ {}^{2}\chi + 317.260\ {}^{1}\Omega^{2} + 352.459$
$A = -15.842\ {}^{2}\chi - 23.077\ {}^{3}\Omega^{2} + 352.508$
$A = -21.101\ {}^{3}\chi - 17.188\ {}^{2}\Omega^{3} + 339.943$
$A = 65.449\ {}^{1}\chi + 74.481\ {}^{2}\Omega^{1} + 100.422$

regression equations. Moreover, as already outlined elsewhere[3,4] the coefficients of orthogonalized regression equations occur as »diagonal« entries in nonorthogonal equations and, therefore, can be extracted without actually constructing orthogonal descriptors!

TABLE IV

Evaluation of the contributions of individual orthogonal components

Descriptors	Residual	R	S
$^{2}\Omega^{1}$	R(1)	0.871	4.506
$^{1}\Omega^{2}$	R(2)	0.910	4.510
$^{3}\Omega^{2}$	R(2)	0.849	5.917
$^{2}\Omega^{3}$	R(3)	0.849	5.916

$^{i}\Omega^{j}$ represent $^{i}\chi$ made orthogonal to $^{j}\chi$

We are now in a position to evaluate the roles of individual descriptors and judge if additional descriptors really significantly improve a correlation. In Table IV we give R and S for orthogonalized descriptors of Table III as regressed against the residuals obtained after applying the first descriptor. The first four possibilities clearly indicate a significant improvement of the regressions, measured by R and S; hence, one is fully justified to incorporate the second descriptor in these regressions. The relatively high regression coefficients in Table IV are in part due to the fact that in this case the first descriptors showed a somewhat limited performance. However, this does not generally guarantee that the next descriptor is likely to be successful. In the case of the hexanols recently examined,[17] using weighted path numbers as molecular descriptors, no simple combinations of the connectivity indices significantly improved relatively unsatisfactory regressions describing isomeric variations of solubilities. The here found regression coefficient $R = 0.753$ for $^{1}\chi$ points to the limitations of the connectivity index in this application. A two variable regression with overall $R = 0.947$ appears quite satisfactory, but are the two descriptors inherently essential or do they merely reflect our inability to find a more suitable characterization of the property?

REDUCTION OF NUMBER OF DESCRIPTORS

How can we reduce the number of descriptors to be used in a regression *without* dramatically reducing R or increasing S?

One way to achieve this goal, which translates into an improved significance of the regression variables and the statistics of the regression, is to consider a fixed combination of descriptors as a novel descriptor. We will illustrate this by examining the regularities of various isomeric properties of alkanes. Recollect that many properties of alkanes can be represented on a grid in which p_2 and p_3 (paths of length two and length three, respectively) are the coordinates for individual isomers.[18] By extending such considerations, novel theoretical molecular properties can be constructed. A sum of carbon-13 chemical shifts was found to show a regularity with isomeric variation.[19] Moreover, a closer look at the numerical values of the carbon-13 chemical shift sums shows that the sum critically depends not on p_2 and p_3 but on their difference p_2, p_3.[20] Thus, instead of using two descriptors, p_2, p_3, a single descriptor p_2–p_3 can account well for the dominant behavior of carbon-13 chemical shifts in octanes. This effectively reduces the number of critical descriptors that characterize the property. Miyashita and coworkers[21] have subsequently extended such studies and found that a simple

linear combination: $p_o+p_1+p_2-p_3$ is the critical parameter for correlations of carbon-13 chemical shift sums in alkanes. It is important to recognize the difference between a set of descriptors, such as p_o, p_1, p_2, p_3, as independent variables and fixed combinations such as p_2-p_3 and $p_o+p_1+p_2-p_3$, which represent a single descriptor even if expressed by several other variables. The significance of these results is that they demonstrate that descriptors can be combined into novel indices that can account for selected molecular properties. Hall[22] extended such considerations to the variety of topological indices and, in particular, illustrated the advantages of various combinations of connectivity indices.

We will examine a single linear combination of connectivity indices as a novel descriptor for heptane surface area. Consider $(2\chi+3\chi)/2$ as a novel descriptors, the average of two connectivity indices (factor 1/2 is not essential). We obtain a regression equation:

$$A=-38.207\ (2\chi+3\chi)/2 + 388.751 \quad \text{with } R=0.895 \text{ and } S=6.23$$

which is better than any of the single variable regressions of Table II, approaching in quality the regressions of Table II based on two variables. While the above definitely represents an improvement, even if we have not dramatically increased R and decreased S, we have arrived at a regression that is comparable in quality to those based on two descriptors, thus effectively reducing the number of variables, which should be viewed as a significant accomplishment. An alternative approach to the reduction of the number of descriptors, in which structural information from two and more descriptors is »compacted« into a single variable using orthogonalization methodology, has been recently discussed at length elsewhere.[23]

MOLECULAR ID NUMBERS AS DESCRIPTOR

An exhaustive search for simple combinations of descriptors is likely to lead to an even better single variable regression equation for correlating the heptane molecular surface areas. While such efforts may have educational merits, they also increase the likelihood of a chance correlation.[24] Hence, rather than continuing to explore such an avenue, we will examine molecular *ID* numbers as descriptors.[25] *ID* numbers are defined as a sum of all weighted paths in a molecule and can be derived from the ALLPATH program,[26] suitably modified to yield weighted paths. Molecular *ID* numbers have been previously used in few structure-property-activity studies[27] and, therefore, suggest themselves as a viable descriptor, closely related to connectivity indices, to be tested.

In Table V we have listed regression equations and the accompanying statistical data for the molecular *ID* numbers and the connectivity indices made orthogonal to *ID* as variables. First, note the numerical constancy of the constant term and *ID* coefficient in the regression equations, which again illustrate the already mentioned stability of such equations in contrast to a chaotic behaviour of the coefficients of regression equations using nonorthogonal descriptors. The regression based on the molecular *ID* as a single descriptor is better than regression based on individual connectivity indices and is comparable in quality (R and S values) to that based on the average of 2χ and 3χ. Hence, it seems that *ID* qualifies as a useful *first* descriptor that can initiate further search for optimal descriptors.

By identifying an acceptable first descriptor, which absorbs most of the correlation satisfactorily, we convert our problem of locating best descriptors to that already con-

TABLE V

Regression equations and the statistical parameters using molecular ID numbers and orthogonalized connectivity indices descriptors

Descriptor	R	S
${}^{1}\Omega^{ID}$	0.7796	4.2301
${}^{2}\Omega^{ID}$	0.8861	3.1703
${}^{3}\Omega^{ID}$	0.7422	4.5835
Descriptors		
ID ${}^{1}\Omega^{ID}$	0.9821	2.8401
ID ${}^{2}\Omega^{ID}$	0.9739	3.4232
ID ${}^{3}\Omega^{ID}$	0.9447	4.9491
Regression equations		
$A = 69.835\ ID - 180.928\ {}^{1}\Omega^{ID} - 567.465$		
$A = 69.834\ ID + 27.498\ {}^{2}\Omega^{ID} - 567.460$		
$A = 69.834\ ID - 13.358\ {}^{3}\Omega^{ID} - 567.459$		

sidered: How to improve upon a regression based on an acceptable single-descriptor.[3,4] The upper part of Table VI shows how connectivity indices, previously orthogonal to *ID* numbers, regress against the residual of the regression using *ID* numbers. We see that inclusion of 1χ–3χ substantially improves the initial regression, and in particular 1χ combined with *ID* results in the standard error, which is less than 1 % of the computed molecular areas and is probably within the accuracy that the particular theoretical model offers. We have thus arrived at the best 2-variable characterization of the heptane molecular surface areas so far, and one of the reasons for the success is that this time we started with a better *first* descriptor. Hence, a message to be taken is that it pays to strive to improve the first descriptor in order to improve a multivariate regression.

TABLE VI

Novel x/X descriptors based on the concept of graph bond orders using the molecular ID numbers, Hosoya's topological index Z, the Wiener number and Balaban's J index as a source invariants

	x/χ	id/ID	z/Z	w/W	j/J
n-hexane	5.6569	9.7256	2.9524	2.5000	7.5038
3-ethyl	5.5787	9.1994	3.3000	3.2500	5.9603
3-methyl	5.5939	9.1801	3.3158	3.0400	6.3913
2-methyl	5.6095	9.2153	3.3889	2.8462	6.9057
2,3-dimethyl	5.5419	8.6942	3.6471	3.3913	5.8111
2,4-dimethyl	5.5577	8.7218	3.8667	3.2083	6.2378
3,3-dimethyl	5.5353	8.5659	3.7500	3.5455	5.6206
2,2-dimethyl	5.5599	8.6025	4.0000	3.3478	6.0332
2,2,3-trimethyl	5.5025	8.1427	4.2308	3.7143	5.4429

ALTERNATIVE FIRST DESCRIPTORS

So far, the best single descriptors were based on the sum of 2χ and 3χ connectivity indices and the use of *ID* numbers. Both examples involved *summation* of components as the underlying mathematical operation. We want to explore different mathematical operations and different functional forms using connectivity indices and a few

topological indices as the basic input for characterization of a molecular structure. Alternative weighting algorithms, such as the use of cube roots instead of square roots in the construction of bond weights, were discussed elsewhere.[28] Here we will consider the generalized graph bond order, a recently introduced local molecular descriptor.[29] The graph theoretical bond order x/X is derived for any bond in a structure by considering a graph invariant X for a graph G and the same invariant x for graph g obtained from G by erasing the edge considered. The x values for g if disjoint are obtained by adding the corresponding values for its components. If x values for all edges are summed up we obtain a molecular invariant, designated as x/X, which is by definition bond additive. Since many molecular properties are also bond additive, the new constructions derived in this way may be of interest in structure-property studies. The concept of graph bond orders x/X can be viewed as a novel *algorithm* to generate graph invariants analogous to the use of bond weights $1/\sqrt{(m\,n)}$ in the definition of connectivity indices, weighted paths, and Balaban's J index.[30] They both illustrate the power of algorithmic routes to structural invariants and have the advantage of not proliferating *ad hoc* invariants, but rather enriching the existing pool of descriptors with structurally related additional variables. The so derived novel descriptors are, nevertheless, subject to the same stringent requirements suggested for curbing proliferation of unwarranted topological indices,[31] the foremost being a demonstrated utility in structure-property-activity studies.

TABLE VII

Regression equations and the accompanying statistics for selected novel descriptors of Table VI

Descriptors		R	S
Z/z	$A = -25.429\ z/Z + 401.33$	0.781	8.730
id/ID	$A = -216.57\ id/ID + 1406.77$	0.924	5.348
χ/χ	$A = 275.26\ \chi/\chi - 1223/74$	0.958	4.011
w/W	$A = -35.070\ w/W + 422.037$	0.988	2.119
j/J	$A = 19.956\ j/J + 185.677$	0.9925	1.703

Table VI lists several novel x/X descriptors based on selected well known topological indices. Table VII gives the corresponding regression equations and statistical parameters R and S when novel descriptors are used in regression against the heptane surface areas. The results in Table VI should be compared with those of Table II and Table V to appreciate the visible improvements in correlations. The connectivity ${}^1\chi/{}^1\chi$ and ID based id/ID descriptors give a better regression that previously obtained regressions using two connectivity indices. Even the novel index z/Z derived from Hosoya's Z topological index,[32] which shows limited success in this application, gives still a better regression than the initial regression based on the connectivity index ${}^1\chi$. However, the novel indices based on the Wiener number[33] and Balaban's J index, w/W and j/J, respectively, lead to an impressive improvement of single variable regressions with standard errors of 2.12 and even 1.70, respectively. It is interesting to note that in each case the novel x/X descriptors derived from graph theoretical bond orders outperformed the corresponding descriptor X. The best results were obtained when al-

ready descriptor X showed a better, if not the best performance. Accordingly, since from the work of Labanowski, Motoc, and Dammkoehler[13] we know that the Wiener number W and Balaban's J index gave excellent single-variable regression for the isomeric variations of heptane surface areas, we should not be surprised to see that w/W and j/J resulted in a regression of exceedingly high quality. The standard errors of these best single descriptor regressions are by a factor of five better than the initial single-variable regression using the connectivity index $^1\chi$!

Labanowski and colleagues[13] considered also a number of information theoretic indices whose information content $I(W, D)$, based to the partition of the distance in a graph, and its mean $I(W, D)/W$ produced excellent correlation with $R=0.98$, $S=2.82$ and $R=0.99$, $S=1.63$, respectively. The information content $I(W,D)$ is given by:

$$I(WD) = W \log W - \Sigma \, \mathrm{f} \, d \log d$$

where f indicates the incidence of distance d in graph G. Note that both the w/W descriptor and the information content $I(W, D)$ index can be viewed as a function of the Wiener index W, each involving a different functional dependence. Thus, there is no doubt that the success of both descriptors ultimately follows from the relevance of the Wiener number in characterization of molecular properties.

A word of caution is in place when deciding which of the »finalists« among regression equations is to be favored. The statistical parameters R and, particularly, S surely offer very important criteria, but other factors have to be considered as well. First, one needs some assurance that some of the better regressions do not result from a chance correlation. A way to remove such doubts would to be extend the analysis to octanes and other alkanes and, thus, by increasing the pool of data minimize accidental coincidence that chance correlation represents. Alternatively, when the pool of compounds is limited, standard statistical methods, such as cross validation,[34] boot-strapping,[35] the partial least square method,[36] should be employed. Another important factor to consider is the interpretation of descriptors. Informational theoretical indices are primarily based on the partitioning of an invariant, which may be influenced by symmetry (equivalence) and other nonlocal molecular characteristics. As such, their interpretation may be somewhat convoluted in comparison with graph invariants with a direct (local) structural interpretation. For example, it is not apparent why 3-ethylpentane and 2,2-dimethylpentane should have almost the same mean information content. In contrast, the relative magnitudes of w/W and j/J show a regularity that parallels Balaban's centric index[36] in that more »compact« isomers are associated with larger w/W or smaller j/J. A similar behavior can be detected for the relative magnitudes of id/ID and $^1\chi$ $^1\chi$. It is this possibility of visualizing the dominant structural features that makes graph invariants so attractive.

MOLECULAR PROPERTIES AS DESCRIPTORS

A number of high quality property-property correlations are well known, such as, in the case of alkanes, correlations of the heats of formation and the boiling points, correlation of the indices of refraction and liquid densities, and, finally, correlation between the chromatographic retention volumes and the boiling points, as pointed out by Kovats.[37] In an important study based on the principal component analysis, Cramer[38] examined the following physical properties for a set of over 100 diverse molecules: aqueous solvation energy, the partition coefficients, molar refractivity, boil-

ing points, liquid state molar volumes and the heats of vaporization. He found that most variance in the properties is associated with two principal components, to which he attributed the labels »bulk« and »cohesiveness«. More recently, Seybold and coworkers[39] examined in a similar manner several properties of alkanes and found considerable mutual interrelatedness, although a more recent work suggest that this interrelatedness does not necessarily hold when attention is restricted to isomers.[11] Such studies are of considerable interest, particularly with respect to QSAR (quantitative structure-activity relationship) as they offer insights into the diversities of properties. Properties as descriptors are of interest in the traditional QSAR[40] where the philosophy is, in part, to represent more convoluted molecular properties in terms of less complicated and better understood properties. Hence, the interest in considering the question: Can we arrive at novel property-property relationship in an orderly manner, not by accident or by inefficient exhaustive searching?

We will illustrate how structure-property correlations may, as a by-product, suggest a novel property-property correlation. In Table VI, we identified several successful regressions for heptanes between the van der Waals areas and various molecular indices. On the other hand, we know from the literature[41] that Balaban's J index correlates well with the octane numbers in alkanes. Until the recently introduced index p/P,[29] based on the count of molecular path numbers, Balaban's J index was the best descriptor for octane numbers. The fact that the same graph theoretical descriptor correlates well with two different properties suggest that the properties themselves may be highly correlated. Thus, by recognizing J and j/J as very successful descriptors for the molecular surface areas (at least in heptanes) we anticipate interdependence of octane numbers and the molecular surface, arriving thus at a novel property-property correlation:

Octane number:	ON = –2.319 CA + 788.14	with $R = 0.89$ and $S = 5.77$
Cavity Area	CA = –35.07 ON + 334.47	with $R = 0.89$ and $S = 14.87$

This well illustrates the benefits of studying the structure-property for property-property relationship, where new correlations and new insights can thus be obtained. We hope that such studies may be of interest to traditionally oriented chemometricians and medicinal chemists interested in identifying novel properties as potential molecular descriptors for the use in QSAR.

REFERENCES

1. M. Randić, *J. Amer. Chem. Soc.* **97** (1975) 6609.
2. L. B. Kier, W. J. Murray, M. Randić, and L. H. Hall, *J. Pharm. Sci.* **65** (1976) 1225.
3. M. Randić, *New J. Chem.*, (in press).
4. M. Randić, *J. Chem. Inf. Comp. Sci.*, (in print).
5. R. B. Hermann, *J. Phys. Chem.* **76** (1972) 2754.
6. L. B. Kier, L.H. Hall, W. J. Murray, and M. Randić, *J. Pharm. Sci.* **64** (1975) 1971.
7. R. H. Rohrbaugh and P.C. Jurs, *Anal. Chim. Acta* **199** (1987) 99.
8. P. G. Mezey, in: *Computational Chemical Graph Theory*, (D. H. Rouvray, ed.), Nova Sci. Publ., New York (1990), pp. 176–197.
9. L. B. Kier, in: *Computational Chemical Graph Theory*, (D.H. Rouvray, ed.), Nova Sci. Publ., New York (1990), pp. 152–174.
10. I. Motoc, *Topics Curr. Chem.* **114** (1983) 93.
11. M. Randić and P.G. Seybold, *J. Amer. Chem. Soc.*, (submitted).

12. L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York (1976).
13. J. K. Labanowski, I. Motoc, and R. A. Dammkoehler, *Comput.&Chem.*, (in print).
14. R. S. Pearlman, in : *Physical Chemical Properties of Drugs*, (S.H. Yalkowski, A. A. Sinkula, and S. C. Valvani, eds), M. Dekker, New York (1980), pp. 321–347.
15. I. Motoc and G. G. Marshall, *Chem. Phys. Lett.*, **116** (1985) 415.
16. H. Hotelling, *J. Educ. Psychology*, **24** (1933) 417.
17. M. Randić, work in progress.
18. M. Randić and C. L. Wilkins, *J. Phys C hem.*, **83** (1979) 1525.
19. M. Randić, *J. Magn. Res.*, **39** (1980) 431.
20. M. Randić and N. Trinajstić, *Theor. Chim. Acta*, **73** (1988) 233.
21. Y. Miyashita, T. Okuyama, H. Ohsako, and S. Sasaki, *J. Amer. Chem. Soc.* **111** (1989) 3469.
Y. Miyashita, H. Ohsako, T. Okuyama, and M. Randić, *Magn. Res. in Chem.*, (in press).
22. L. H. Hall, in: *Computational Chemical Graph Theory*, (D.H. Rouvray, ed.), Nova Sci. Publ., New York (1990), pp. 202–233.
23. M. Randić, *New J. Chem.*, in press.
24. J. Topliss and R. Costello, *J. Med. Chem.* **15** (1972) 1066.
T. R. Stouch and P. C. Jurs, *Quant. Struct. — Act. Relat.* **5** (1986) 57.
25. M. Randić, *J. Chem. Inf. Comput. Sci.*, **24** (1984) 164.
26. M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Comput. & Chem.* **3** (1979) 5. Listing of the modified program (in BASIC) is available from: Graph Theory Center, Dept. of Mathematics and Computer Sci. Drake University, Des Moines, IA 50311.
27. M. Randić, *Int. J. Quant. Chem: Quant. Biol. Symp.*, **11** (1984) 137.
28. M. Randić, P. J. Hansen, and P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **28** (1988) 60.
29. M. Randić, (to be published).
30. A. T. Balaban, *Theor. Chim. Acta* **53** (1979) 355.
31. M. Randić, *J. Math. Chem.*, (in print).
32. H. Hosoya, *Bull. Chem. Soc. Jpn.* **44** (1971) 2332.
33. H. Wiener, *J. Amer. Chem. Soc.* **69** (1947) 17.
34. S. Giesser, *J. Am. Stat. Assoc.*, **70** (1975) 328.
35. P. Diaconis and B. Efron, *Sci. Am.*, (1984) 116.
36. R. W. Gerlach, B. R. Kowalski, and H. O. A. Wold, *Anal. Chim. Acta (Comp. Tech. Optim.)*, **112** (1979) 417.
37. E. Kovats, *Z. Anal. Chim.* **181** (1961) 351.
38. R. D. Cramer, III, *J. Amer. Chem. Soc.* **102** (1980) 1837.
39. D. E. Needham, I-C. Wei, and P. G. Seybold, *J. Amer. Chem. Soc.* **110** (1988) 4186.
40. C. Hansch, *Acc. Chem. Res.* **2** (1969) 232.
41. A. T. Balaban and I. Motoc, *MATCH* **5** (1979) 197.

SAŽETAK

Iznalaženje optimalnih molekularnih deskriptora

Milan Randić

Pokazano je kako se mogu pronalaziti optimalni deskriptori molekula. Kao temeljni deskriptor upotrijebljen je index povezanosti. Postupak se temelji na ideji da se skup deskriptora ortogonalizira. Ortogonalizirani deskriptori daju regresijske jednadžbe odnosa strukture i svojstava s boljim statističkim značajkama od onih koje su izdrađene od neortogonaliziranih deskriptora. Osim indeksa povezanosti upotrijebljeni su i molekularni *ID*-broj, Wienerov broj i Balabanov indeks. Također je predložen i novi model odnosa strukture i svojstava između oktanskog broja i površine molekule.