Similarity Based on Extended Basis Descriptors

Milan Randić

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

Received June 25, 1992

We consider the problem of comprehensive characterizations of chemical structures by mathematical invariants. We have illustrated a particular characterization to derive similarities between heptane isomers. Comprehensive characterizations require resolution of the following problems: (1) selection of descriptors that adequately span the structure space; (2) construction of orthogonal basis of descriptors to eliminate bias due to interconnectedness of the descriptors. The latter problem has been resolved recently by a design of a sequential orthogonalization procedure. We proceed here to outline construction of extended bases for the characterization of structures.

INTRODUCTION

By a characterization of molecular structure, we understand that to mean a list of mathematical properties independent of labeling of vertices or graphical representation of a structure. In the mathematical literature they are usually called invariants. Formally, a characterization is given as a set of descriptors: $D_k = \{d_1, d_2, d_3, d_4, \ldots d_k\}$. There is, in general, no restriction on the nature of the descriptors and their number. One can classify descriptors according to their origin as graph invariant, matrix invariant, geometrical invariant, structural invariant, quantum chemical invariant, etc. Often they have a direct structural interpretation, in which case one speaks of structure-explicit descriptors. They may have a somewhat convoluted relationship to the chemical structure, in which case one speaks of structure-implicit descriptors. Finally, one may select as descriptors a physicochemical property of a structure. The relationship of such descriptors to various structural components is obscure, in which case one speaks of structure-cryptic descriptors. Sometimes there is no sharp boundary in such a classification. Whether a descriptor has a direct or an indirect structural interpretation may be in the eye of the beholder. Nevertheless, the classification into structure-explicit, structure-implicit, and structure-cryptic descriptors[1] helps us emphasize the critical factors in selecting or designing descriptors. In Table I we illustrate the above classification for a selection of well-known descriptors.

In a typical data reduction in structure–property–activity studies, one uses a multiple regression analysis or evaluates the degree of similarity among the compounds. In both cases the goal is to predict a biological activity of compounds of interest. The first step in such considerations is to select, from the pool of available descriptors, critical descriptors. The quantitative results may critically depend on the choice of the descriptors used. Many of the past activities in structure–property–activity studies were concerned with the selection of descriptors believed to be the best for a particular application. There are no clear-cut rules on how to arrive at the best descriptors, although the principal component analysis (PCA)[2] and various statistical tests offer advantages and objectivity.[3]

An alternative to the selection of descriptors from a pool is to form a list of descriptors to be considered as a *basis*.[4] This is analogous to computations in quantum chemistry where a set of atomic orbitals are used as basis functions. Once a fixed basis is adopted, one may orthogonalize such a basis,[5] i.e., construct combinations of the initial descriptors that will result in descriptors that no longer correlate among themselves. Among the advantages of using the same basis in different applications are the easiness of comparisons of different results[4] and the possibility to identify contributions of individual descriptors.[5]

Table I. Classification of Selected Well-Known Molecular Descriptors

structure-explicit indices: graph theoretical invariants
connectivity index
Wiener index
Hosoya Z number
molecular ID number
Kier's shape index
path numbers
Pauling bond orders
Kekule structure count
structure-implicit indices: quantum chemical descriptors
Coulson's bond orders
bond overlaps
bond dipoles
MO energies
HOMO–LUMO gaps
structure-cryptic descriptors: molecular properties
bond lengths
molecular weight
molecular volume
molecular surface

Table II. List of Possible Basis Descriptors

basis descriptors	outlined in
connectivity indices $^1\chi, ^2\chi, ^3\chi, ^4\chi, \ldots$	ref 5
weighted paths $^1\pi, ^2\pi, ^3\pi, ^4\pi, \ldots$	ref 7
extended connectivities $^1\epsilon, ^2\epsilon, ^3\epsilon, ^4\epsilon, \ldots$	M. Randić, unpublished
vicinal matrix paths $^1\bar{\omega}, ^2\bar{\omega}, ^3\bar{\omega}, ^3\omega, \ldots$	ref 8
connectivity powers $^1\alpha, ^2\alpha, ^3\alpha, ^4\alpha, \ldots$	this work

BASIS DESCRIPTORS

Any set of descriptors, D_k, can be viewed as a *basis* if such descriptors are used as an ordered set. There are advantages in selecting descriptors according to some structural criteria rather than simply combining a number of ad hoc descriptors. In Table II we have listed several possible bases using descriptors introduced in the literature. A basis based on structurally related invariants may offer a simpler interpretation of the regression equations. Illustrations include the

0095-2338/92/1632-0686$03.00/0

connectivity indices,[6] path numbers,[7] the weighted paths,[8] and paths of different length derived from other than the adjacency matrix.[9]

A parallelism with quantum chemical calculations of molecular orbitals is instructive. Limitations of a basis impose limitations on the accuracy of computed molecular energies. A comprehensive basis offers a better accuracy for computed molecular properties. Typically, quantum chemical calculations are based on nonorthogonal basis functions. Whether one uses a nonorthogonal basis or orthogonal basis, the numerical outcome is the same. The results do not depend on the method of computation, even the amount of work need not be much different. However, interpretation of intermediate steps and interpretation of the final results critically depend on the orthogonality of the basis used in computations.

The situation is similar with multiple regression analysis: The correlation coefficient R and the standard error S do not depend, once a set of descriptors is chosen, on whether the descriptors are interdependent or whether they were made orthogonal. The interpretation of the results, however, does depend on the process of arriving at the regression equations. If one uses orthogonal descriptors, the regression equations display unusual numerical stability: The coefficients of the individual descriptors are independent of inclusion/exclusion of other descriptors.

The analogy with quantum chemistry extends also to the quality of calculations. If the descriptors used do not span the structure space well, i.e., some relevant structural factors are not adequately represented in the characterization of a molecule, the results will be inherently deficient. Such results could be improved by enlargement of the basis, i.e., by inclusion of additional descriptors. Thus, a good regression analysis should be based on descriptors which "cover" the structure space adequately and should at least include such components of structures that are important for the description of the properties considered. Just as a small basis in quantum chemical calculations is inherently limited, a "short" list of descriptors will be inherently limited when used as a basis. We have, therefore, to consider construction of bases that can be extended when necessary. Such an augmented basis is more likely to approach the properties of a complete basis.

EXTENDED BASES

In mathematical circles, it is generally believed that any *finite* list of simple invariants will not result in a unique representation for a structure. In other words, any such finite selection of invariants may have identical lists for two different graphs. A degree of the deficiency of a basis will be reflected by the size of the smallest pair of graphs (structures) for which such a finite list coincides.[10] If we use the coefficients of the characteristic polynomial for characterization, then $n = 6$, which is the size of the smallest isospectral graphs.[11] If we use path numbers for characterization, then $n = 9$, which is the size of the smallest pair of trees with the same path sequences. If we use the connectivity indices as a basis, then one should search among graphs with the same ID number for occurrence of structures with duplicate (same) descriptors. The identification number, ID number, is given as the sum of weighted paths in a graph.[12] Among trees, the first duplicates occur when n is at least 15.[13] The above illustrations indicate that the connectivity indices and the weighted paths are rich in structural information and may lead to bases capable of discriminating structures of relatively large size. Nevertheless, we need to consider some questions: How can one extend a basis so that it becomes even more discriminatory (among graphs of increasing size)? How close can we approach a basis which will always discriminate among structures? Can one arrive at a complete bias?

Table III. Illustration of χ-Matrix for 2,2,3-Trimethylbutane and Derived ${}^k\alpha$ Descriptors (Based on Powers of χ-Matrices) Including Some Very High Powers Which Suggest a Slow Convergence for Such Invariants

${}^k\alpha$ Invariants Derived from χ-Matrix Powers for 2,2,3-Trimethylbutane						
0	0.5000	0	0	0	0	0
0.5000	0	0.2887	0	0.5000	0.5000	0
0	0.2887	0	0.5774	0	0	0.5774
0	0	0.5774	0	0	0	0
0	0.5000	0	0	0	0	0
0	0.5000	0	0	0	0	0
0	0	0.5774	0	0	0	0

${}^k\alpha$ Invariants (Only Odd Powers)			
${}^1\alpha$	2.9434	${}^{15}\alpha$	2.1609
${}^3\alpha$	2.5490	${}^{17}\alpha$	2.1578
${}^5\alpha$	2.3519	${}^{19}\alpha$	2.1562
${}^7\alpha$	2.2532	${}^{101}\alpha$	2.1547
${}^9\alpha$	2.2040	${}^{1001}\alpha$	2.1542
${}^{11}\alpha$	2.1793	${}^{10001}\alpha$	2.1497
${}^{13}\alpha$	2.1670	${}^{100001}\alpha$	2.1051

From what has been said, one expects that a complete basis will be infinite, unless someone demonstrates that the belief in limitations of finite lists of simple descriptors is unfounded. Our interest here is in practical solutions to characterizations of structures. Hence, we will consider a modest problem: Can we arrive at a comprehensive characterization of structures (graphs) that will suffice in structure–property–activity applications and show sufficient discriminatory power between similar structures regardless of their size?

Before we consider extended bases, we will consider properties of basis functions used in other computations in chemistry and physics. Consider the well-known bases: the sine and cosine functions used in the Fourier expansion of periodic functions and the power series expansions. Both bases are complete, that is, they allow representation of the expanded functions as accurately as one desires. The sine and cosine bases are orthogonal, while the power series is not (but can be made orthogonal). In contrast, atomic orbitals, such as various Gaussian basis functions, are neither orthogonal (for functions at different atomic centers) nor complete, but for most practical purposes serve us adequately.

Our aim here is to arrive at a compromise between completeness, that requires an infinite basis, and practical accuracy, which may be based on a limited basis. The major problem with limited bases is a lack of assurances that the results are reliable and will not be dramatically altered if the basis is augmented. An apparent convergence of the results upon truncation (assuming numerical stability upon truncation) is an indication that the results might be reliable. Numerical stability can be achieved by constructing orthogonal basis.[5] Hence, we will introduce the notion of orthogonal descriptors before discussing results.

Basis implies ordering of the basis functions (descriptors). Often an order is induced naturally as is illustrated by the connectivity indices and weighted paths, where paths of longer length follow the paths of shorter length. A limitation of the connectivity basis, and the weighted path bases, is the number of descriptors that one can introduce in this way. The number is limited by the length of the longest paths in the molecules considered. It is difficult to extend such relatively small bases unless some other structural considerations are invoked. Illustration of an augmented basis based on paths was offered

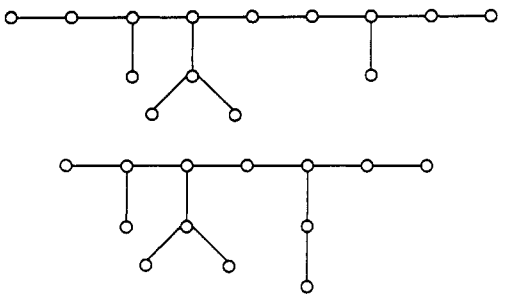


Figure 1. Smallest pair of molecules having the same molecular ID numbers.

recently[10] by considered disjoint subgraphs of paths. While such an approach increases the number of components considerably, the size of such a basis is finite because of the finite number of subgraphs that a graph can generate. Hence, to attempt to reach the completeness, that is, to augment a basis to infinite size, is in general a difficult problem. Practical requirements demand that, while the basis may be infinite, it converges fast so that one can truncate an infinite sequence after a few initial steps. Klein considered the problem of graph theoretic expansions[14] in general and outlined several choices in which the number of components in the expansions does not grow exponentially with the size of the graph considered. Difficulties of convergence, when an extensive basis based on all subgraphs is used, have been observed by McHughes and Poshusta,[15] who tried to express molecular properties in terms of all contributing subgraphs.

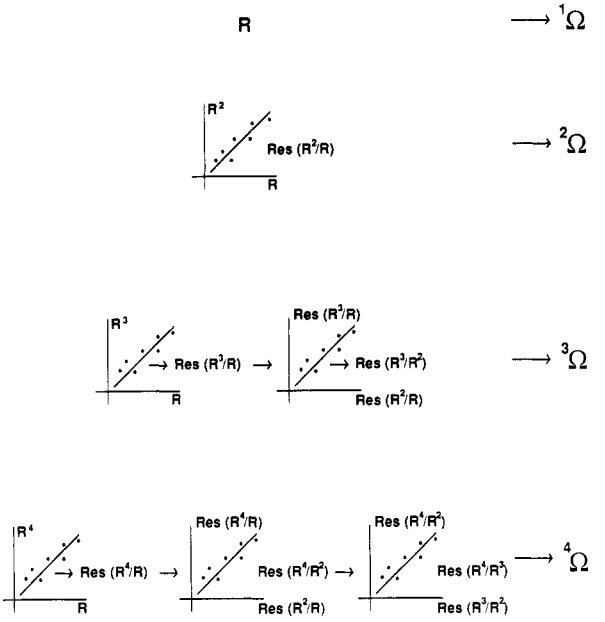
How an extensive basis is to be used in the expression of molecular properties and how to avoid the difficulties on the convergence of the expansion remain to be better understood.

INFINITE BASIS

If we could construct an infinite basis based on molecular descriptors, then one could examine effects of truncation and arrive at a practical scheme even though the initial basis is infinite. Count of paths and their modifications cannot extend indefinitely. On the other hand, such descriptors, based on paths and weighted paths, offer a useful basis for discussion of molecular properties.[16] In contrast, powers of the adjacency matrix (and other matrices) offer an *infinite* sequential buildup of invariants, but these invariants may be deficient because the matrices themselves do not encode relevant (for structure–property–activity) structural information. This is illustrated by the adjacency matrix, the powers of which count random walks in a graph.[17] Due to the Cayley–Hamilton theorem,[17] after a certain power the adjacency matrix does not introduce novel information. Thus, powers of A^n only superficially appear to lead to an infinite basis of invariants! The above illustrates difficulties which we face when considering a design of an infinite set of invariants to serve as a basis.

We will outline a construction of one such infinite basis that shows desired properties: It leads to an infinite list of

Chart I. Summary of Steps Leading to Orthogonal Descriptors



descriptors which are not restricted in their structural content. We start from (i) an assumption that the connectivity index, based on paths weighted according to the valencies of the vertices involved, is a useful molecular descriptor and (ii) an assumption that construction of powers allows building of an infinite basis (as power series itself is an infinite series). Hence, we will combine the above elements into a single procedure as follows:

We start with the χ-matrix, which is based on the adjacency matrix, but instead of the binary entries 0 and 1 we assign the value of $1/\sqrt{(m,n)}$ to nonzero matrix elements, where m and n are the valencies of the vertices involved. Thus

$$\chi_{i,j} = \begin{cases} 1/\sqrt{(m,n)} & \text{if } i,j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

This matrix is related to a more general matrix considered by Hall[18] in which entries for nonadjacent vertices instead of being zero are given by the contributions of the corresponding paths to higher connectivity indices.

In Table III we show the χ-matrix for 2,2,3-trimethylbutane, one of the heptane isomers. By summing all the nonzero entries above the main diagonal, we obtain the connectivity index χ.[6] We continue by constructing higher powers of χ-matrix using the usual matrix multiplications and arrive at $(\chi)^m$ matrices. This process can produce a large number of matrices. Now we have to select an invariant from each such matrix, just as we selected $^1\chi$ from the χ-matrix.

We will define the new invariants analogous to construction of $^1\chi$ in χ-matrix by summing the elements in the $(\chi)^m$ matrices above the main diagonal that correspond to nonzero entries

Table IV. $^k\alpha$ Descriptors (Based on Powers of χ-Matrices) for Nine Isomers of Heptane

	n-heptane	2-methyl	3-methyl	3-ethyl	2,2-dimethyl	2,3-dimethyl	2,4-dimethyl	3,3-dimethyl	2,2,3-trimethyl
$^1\alpha$	3.41421	3.27006	3.30806	3.34607	3.06066	3.18074	3.12590	3.12132	2.94338
$^3\alpha$	2.81066	2.76818	2.69548	2.63896	2.67808	2.69809	2.74099	2.51777	2.54904
$^5\alpha$	2.50889	2.50630	2.39699	2.28540	2.43896	2.41509	2.48439	2.27849	2.35187
$^7\alpha$	2.32028	2.32990	2.22547	2.10863	2.28952	2.22486	2.31332	2.17447	2.25328
$^9\alpha$	2.18825	2.20472	2.11799	2.02024	2.19611	2.15056	2.19928	2.12637	2.20399
$^{11}\alpha$	2.09159	2.11509	2.04806	1.97604	2.13773	2.09280	2.12324	2.10330	2.17934
$^{13}\alpha$	2.01968	2.05084	2.00186	1.95394	2.10125	2.05877	2.07256	2.09200	2.17602
$^{15}\alpha$	1.96590	2.00477	1.97117	1.94289	2.07844	2.03872	2.03876	2.08642	2.16086
$^{17}\alpha$	1.92559	1.97172	1.95074	1.93737	2.06419	2.02690	2.01623	2.08364	2.15778
$^{19}\alpha$	1.89538	1.94803	1.93713	1.93461	2.05528	2.01994	2.00121	2.08225	2.15623

Table V. Summary of Steps Taken in Orthogonalization Procedure[a]

operation	R	S
$^{1}\alpha$		
$^{3}\alpha$		
Res($^{3}\alpha/^{1}\alpha$)	0.6171	0.0810
$^{5}\alpha$		
Res($^{5}\alpha/^{1}\alpha$)	0.2257	0.0915
Res($^{5}\alpha/^{3}\alpha$)		
$^{7}\alpha$		
Res($^{7}\alpha/^{1}\alpha$)	0.0868	0.0782
Res($^{7}\alpha/^{3}\alpha$)	0.9081	0.0033
Res($^{7}\alpha/^{5}\alpha$)		
$^{9}\alpha$		
Res($^{9}\alpha/^{1}\alpha$)	0.4382	0.0588
Res($^{9}\alpha/^{3}\alpha$)	0.8165	0.0340
Res($^{9}\alpha/^{5}\alpha$)	0.9799	0.0068
Res($^{9}\alpha/^{7}\alpha$)	0.990	0.00095
$^{11}\alpha$		
Res($^{11}\alpha/^{1}\alpha$)	0.7503	0.0407
Res($^{11}\alpha/^{3}\alpha$)	0.6617	0.0305
Res($^{11}\alpha/^{5}\alpha$)	0.9514	0.0094
Res($^{11}\alpha/^{7}\alpha$)	0.9662	0.0024
Res($^{11}\alpha/^{9}\alpha$)	0.9926	0.00029
$^{13}\alpha$		
Res($^{13}\alpha/^{1}\alpha$)	0.9101	0.0274
Res($^{13}\alpha/^{3}\alpha$)	0.3574	0.0256
Res($^{13}\alpha/^{5}\alpha$)	0.9013	0.0111
Res($^{13}\alpha/^{7}\alpha$)	0.9334	0.0040
Res($^{13}\alpha/^{9}\alpha$)	0.9718	0.00094
Res($^{13}\alpha/^{11}\alpha$)	0.99995	0.000009

[a] Catalog of operations leading to orthogonal descriptors and the corresponding statistical information (the regression coefficients R and the standard errors S for regressions of successive descriptors)

of the adjacency matrix. Hence, the new invariants are bond additive, in view of the fact that only contributions from bonds arise. By using the higher powers of χ-matrices, one involves contributions from the more distant neighbors in a different way in different structures. In the lower part of Table III, we show the derived invariants for various odd powers of χ-matrix of 2,2,3-trimethylbutane. In acyclic structures, the even powers of $(\chi)^m$ matrices necessarily make zero contributions because one cannot arrive at an adjacent place in a graph by an even number of steps. Hence, only odd powers of $(\chi)^m$ matrices are shown. In the case of 2,2,3-trimethylbutane, we extended calculations to very high powers merely to illustrate that different values for the new invariants continually arise. One can observe an apparent slow convergence of the values of the higher invariants as n approaches infinity. However, as we will see later, the structural information contained in higher powers of the matrices gradually decreases.

AN ILLUSTRATION: HEPTANE ISOMERS

In Table IV, we give the first 10 nonzero invariants derived from powers of χ-matrices for the nine isomers of heptane, called $^{1}\alpha$–$^{m}\alpha$. All isomers show a similar behavior: a gradual numerical decrease of the $^{i}\alpha$ invariants constructed from $(\chi)^m$-matrices. The first entry for each isomer, by definition, is the connectivity index $^{1}\chi$; the remaining values show monotonically smaller values.

The smallest molecules that have the same connectivity index (none among heptanes) are 3-methylheptane and

Table VI. Orthogonal $^{k}\alpha$ Descriptors for Nine Isomers of Heptane

	n-heptane	2-methyl	3-methyl	3-ethyl	2,2-dimethyl	2,3-dimethyl	2,4-dimethyl	3,3-dimethyl	2,2,3-trimethyl
$^{1}\Omega$	3.41421	3.27006	3.30806	3.34607	3.06066	3.18074	3.12590	3.12132	2.94338
$^{3}\Omega$	–0.04723	–0.06172	0.02599	0.09754	–0.05437	–0.02693	–0.09151	0.12991	0.02832
$^{5}\Omega$	0.02109	0.02164	0.00342	–0.03480	–0.01005	–0.01968	–0.01390	0.02338	0.00892
$^{7}\Omega$	–0.00559	0.00442	0.00353	–0.00014	0.00066	–0.00175	0.00098	0.00067	–0.00277
$^{9}\Omega$	–0.00039	0.00107	–0.00107	0.00023	–0.00035	0.00133	–0.00113	–0.00017	0.00019
$^{11}\Omega$	0.0004	0.00025	–0.00031	0.00035	–0.00005	–0.00049	0.00015	–0.00013	0.00022
$^{13}\Omega$	0.00000	0.00000	0.00001	0.00000	–0.00002	0.00000	0.00001	–0.00001	0.00000

Table VII. Similarity/Dissimilarity Matrices Based on Orthogonalized $^{k}\alpha$ Descriptors (k = 1, 3, 5)

$^{1}\alpha$	2-M	3-M	3-E	2,2-M	2,3-M	2,4-M	3,3-M	2,2,3-M
				$^{1}\alpha$ Orthogonal				
n	0.144	0.106	0.068	0.354	0.234	0.288	0.293	0.471
2-M		0.038	0.076	0.209	0.089	0.144	0.149	0.327
3-M			0.038	0.247	0.127	0.182	0.187	0.365
3-E				0.285	0.165	0.220	0.225	0.403
2,2-M					0.120	0.065	0.061	0.117
2,3-M						0.055	0.059	0.237
2,4-M							0.005	0.183
2,2,3-M								0.178
				$^{3}\alpha$ Orthogonal				
n	0.144	0.129	0.160	0.354	0.234	0.292	0.342	0.477
2-M		0.096	0.177	0.210	0.096	0.147	0.243	0.339
3-M			0.081	0.260	0.138	0.217	0.214	0.365
3-E				0.323	0.206	0.290	0.227	0.409
2,2-M					0.123	0.075	0.194	0.144
2,3-M						0.085	0.168	0.244
2,4-M							0.222	0.218
2,2,3-M								0.205
				$^{5}\alpha$ Orthogonal				
n	0.145	0.130	0.170	0.355	0.238	0.294	0.342	0.477
2-M		0.097	0.185	0.212	0.104	0.151	0.243	0.339
3-M			0.090	0.261	0.140	0.218	0.215	0.365
3-E				0.324	0.208	0.291	0.234	0.411
2,2-M					0.124	0.075	0.197	0.145
2,3-M						0.085	0.173	0.245
2,4-M							0.225	0.220
2,2,3-M								0.205

4-methylheptane. The initial χ-powers for these two molecules are as follows: 3-methylheptane, 3.80806, 3.07050, 2.73200, ...; 4-methylheptane, 3.80806, 3.11068, 2.71951, ... Hence, the new descriptors discriminate among smaller molecules having the same bond type decomposition. Bond type (m,n) is defined by valencies m,n (indicating the number of neighbors for the vertices forming a connection). In the above case, we have in both molecules $2(1,2) + (1,3) + 2(2,2) + 2(2,3)$.

The smallest pair of molecules having the same ID and having the same bond decomposition are illustrated in Figure 1. The initial χ-powers for these two molecules are 6.93396, 5.82377, 5.31701, ... and 6.93396, 5.15083, 4.36061.... Again we see $(\chi)^m$ powers lead to distinct descriptors, indicating thus that $(\chi)^m$ powers are rich in structural information.

ORTHOGONALIZATION OF $(\chi)^m$-POWER DESCRIPTORS

With an extended basis, before using it in a regression analysis or similarity characterizations, we want first to construct an orthogonal basis of descriptors. A necessity for orthogonal descriptors is important in similarity testing for the following reason: Dependent descriptors introduce different weights for different components, since each time a new descriptor is used it overlaps with components with which it correlates. In multiple regression, if performed in a stepwise fashion, one can extract contributions of novel descriptors (which are given by the coefficients in the regression analysis the first time such a descriptor appears[5]). In a similarity/dissimilarity matrix, overlapping information of interdependent variables is not easy to separate. Thus, the roles of individual variables are again and again emphasized.

The orthogonalization process has been outlined elsewhere.[5,19] Briefly, we use regression analysis and the residuals between dependent descriptors in a sequential scheme which generates independent new variables. Chart I summarizes the steps that accomplish the task. The orthogonalization process is analogous to sequential orthogonalization of vectors in the Gram–Schmidt procedure for construction of orthogonal vectors in linear algebra. First, the descriptors are ordered (just as vectors), and the first descriptor D_1 is taken as the first entry of the sought orthogonal basis. The second descriptor D_2 is regressed against the first descriptor D_1, and the residual of this regression $\mathrm{Res}(D_2/D_1)$ is the second (orthogonal) descriptor. By the definition, the residual $\mathrm{Res}(D_2/D_1)$ is that part of the initial descriptor D_2 which does not correlate with D_1. Hence, it represents the orthogonal component in D_2, the part that cannot be predicted from D_1. The process continues with the third descriptor, D_3, which is the first made orthogonal to D_1, and the so-derived orthogonal component $\mathrm{Res}(D_3/D_1)$ is then made orthogonal to the second orthogonal descriptor $\mathrm{Res}(D_2/D_1)$. Thus, the third orthogonal descriptor is the residual of regressions between the residuals $\mathrm{Res}(D_3/D_1)$ and $\mathrm{Res}(D_2/D_1)$! Rather than using the cumbersome label $\mathrm{Res}\{\mathrm{Res}(D_3/D_1)/\mathrm{Res}(D_2/D_1)\}$, we will use the abbreviated notation for the new descriptor based on previous residuals: $\mathrm{Res}(D_3/D_2)$. Hence, when the subscripts of the descriptors (D_m/D_n) in a residual notation differ by one, such a residual represents the mth orthogonal descriptor, if they differ by more than one, the process of construction of the mth orthogonal descriptor is not yet complete. The procedure is illustrated in Table V for odd powers of $(\chi)^m$-matrices, the derived invariants of which we have labeled as α descriptors. As Chart I illustrates, the process continues until all initially selected descriptors are made orthogonal. Chart I shows successive steps in the construction of orthogonal descriptors.

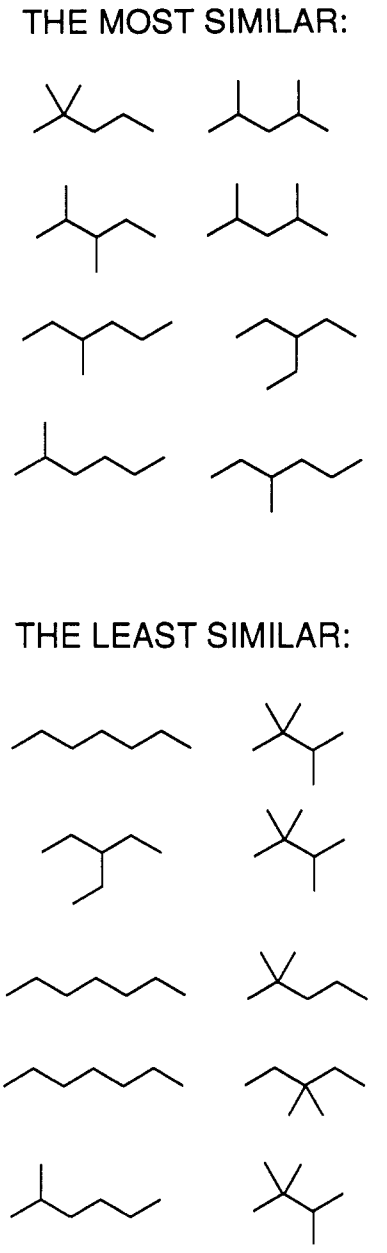


Figure 2. Most and least similar heptanes.

Each successive step requires preforming all the previous steps. Table V gives a catalog of operations (easily performed on a personal computer) with the odd χ-powers from 1 to 13. In Table V, we give the statistical information on regressions using α descriptors. Observe that although various α descriptors show initially little correlation (the correlation coefficients R are typically relatively small) as we continue construction the correlation coefficient between nonorthogonal descriptors increases and apparently approaches 1 as powers increase. Therefore, additional descriptors contribute less and less of the novel structural information. This is also reflected by the relative magnitudes of the descriptors (Tables III and IV), which differ less and less in successive steps.

SIMILARITY BASED ON χ-POWERS

Derived orthogonal α descriptors for the nine isomers of heptanes are listed in Table VI. We will use the orthogonal descriptors in Table VI to describe similarities and dissimilarities among heptane isomers in order to assess usefulness of the novel descriptors. The similarity/dissimilarity matrix based on Euclidean distance as a measure of the degree of

Table VIII. Similarity/Dissimilarity Matrices Based on $^k\alpha$ Descriptors (k = 1, 3, 5)

	2-M	3-M	3-E	2,2-M	2,3-M	2,4-M	3,3-M	2,2,3-M
				$^1\alpha$ Nonorthogonal				
n	0.144	0.106	0.068	0.354	0.234	0.288	0.293	0.471
2-M		0.038	0.076	0.209	0.089	0.144	0.149	0.327
3-M			0.038	0.247	0.127	0.182	0.187	0.365
3-E				0.285	0.165	0.220	0.225	0.403
2,2-M					0.120	0.065	0.061	0.117
2,3-M						0.055	0.059	0.237
2,4-M							0.005	0.183
2,2,3-M								0.178
				Nonorthogonal Descriptors (N = 3)				
n	0.150	0.157	0.185	0.378	0.259	0.297	0.414	0.539
2-M		0.082	0.150	0.228	0.114	0.147	0.291	0.393
3-M			0.068	0.248	0.127	0.188	0.258	0.393
3-E				0.288	0.176	0.243	0.255	0.413
2,2-M					0.122	0.091	0.171	0.174
2,3-M						0.070	0.190	0.280
2,4-M							0.223	0.265
2,2,3-M								0.181
				Nonorthogonal Descriptors (N = 5)				
n	0.150	0.192	0.290	0.384	0.276	0.298	0.474	0.561
2-M		0.137	0.267	0.238	0.146	0.148	0.370	0.423
3-M			0.131	0.252	0.129	0.207	0.284	0.396
3-E				0.326	0.218	0.314	0.255	0.418
2,2-M					0.124	0.101	0.235	0.195
2,3-M						0.098	0.234	0.287
2,4-M							0.304	0.296
2,2,3-M								0.195

similarity (thus giving the same weight to all components) is shown in Table VII for the first three cases using D_1; D_1, D_2; and D_1, D_2, D_3, respectively. The influence of the higher components increases as one can judge from apparently good convergence. As we can expect from Table VI, which lists all the descriptors to six places, higher powers hardly contribute significantly to the measure of similarity. Figure 2 shows four of the most similar as well as five of the least similar pairs of isomers among heptanes. The results from Figure 2 appear plausible and agree with the qualitative expectations based on inspection of the molecular graphs of heptane isomers.

DISCUSSION

Table VIII gives the similarity/dissimilarity matrices when we use nonorthogonal $^k\alpha$ descriptors. In the first of the three sections, each isomers is represented by a single number, $^1\alpha$; in the second section, isomers are represented as vectors having two components; and in the third section, they are represented as vectors having three components. As the number of components increases, the number of entries in the table also increases, but the changes are not dramatic. From the last of the three sections of Table VIII, we may conclude as the most similar pairs of isomers the following:

similarity index (orthogonal)	pair of isomers	similarity index (nonorthogonal)
0.075	2,2-MM and 2,4-MM	0.101
0.085	2,3-MM and 2,4-MM	0.098
0.090	3-M and 3-E	0.131
0.097	2-M and 3-M	0.137

The above results appear plausible. Similarly, plausible results are obtained when we list the least similar pairs of isomers:

similarity index (orthogonal)	pair of isomers	similarity index (nonorthogonal)
0.477	*n*-heptane and 2,2,3-MMM	0.561
0.411	3-E and 2,2,3-MMM	0.418
0.355	*n*-heptane and 2,2-MM	0.384
0.342	*n*-heptane and 3,3-MM	0.474
0.339	2-M and 2,2,3-MMM	0.423

We obtain similar results when instead of nonorthogonal $^k\alpha$ descriptors, we use orthogonal $^k\alpha$ descriptors. A comparison of parts of tables based on vectors having the same number of components gives similar results. When k = 5 (vectors having three components), again we have similar numerical results and similar ordering of pairs of isomers showing the great and the least similarity. However, a close comparison reveals significant differences, which although numerically not large do suggest a somewhat different ordering of the pairs of structures! Thus, for example, when k = 5, the most similar pair of isomers when one uses orthogonal descriptors are not 2,2-MM and 2,4-MM but 2,3-MM and 2,4-MM. Although in the case of heptane isomers, orthogonal and nonorthogonal descriptors give similar results, but comparison illustrates that nonorthogonality can influence the numerical measures of similarity. It seems, therefore, important to eliminate this unnecessary interference due to duplications in descriptors, particularly since construction of orthogonal descriptors can be performed on small computers.

CONCLUSION

We have introduced here, for the first time, an infinite basis of molecular descriptors. Such a basis appears promising in similarity/dissimilarity studies. It remains to be investigated how useful such a basis will be in other applications, particularly in multivariate regression analysis of structure–property and structure–activity studies.

The set of nine heptanes is a too small set to allow testing for completeness of the basis, that is, to see how well such a basis characterizes molecules. We cannot use too many descriptors on such a small sample size. However, the approach outlined here can be applied to octanes and nonanes, which make larger sets of compounds and will thus allow testing of a larger number of descriptors in a statistically meaningful way. We may test completeness of a basis (from practical point of view) by selecting the number of properties of structures (graphs) and search for multiple regression of these properties using as few descriptors as possible. If a set

of descriptors leads to a regression of the desired accuracy for many different data, this would suggest adequate completeness.

More complex structures, related to some rationale, should be investigated in view of limited structural (functional) properties of saturated hydrocarbons. One of the reasons for postponing such studies is the lack of optimal connectivity-type molecular descriptors for molecules containing heteroatoms. Recently, however, a scheme in which optimal connectivity indices for oxygen in alcohols and nitrogen in amines has been outlined.[20] Extension of such work to organic compounds containing other heteroatoms will make possible applications of the scheme proposed here to a wide class of molecules of interest in structure–property–activity studies.

To demonstrate incompleteness of a basis, it suffices to demonstrate a pair of nonisomorphic structures (graphs), which will have all descriptors equal. While such tests may offer useful information, we already know from applications of the multiple regressions to structure–activity and from structure–similarity studies that even an incomplete basis can span a large part of structural space of interest. Thus, the emphasis in the immediate future should be on the design and testing of alternative infinite bases rather than on viewing this first basis as the only basis of interest in QSAR. To emphasize that we may be at the beginning of development of useful bases, we deliberately designated the descriptors with the first letter of the Greek alphabet. Comparisons of similarity based on alternative bases will be instructive as they may reveal the degree of completeness or lack of it in different bases.

ACKNOWLEDGMENT

I would like to thank the Beilstein Institute for the kind invitation to contribute to the symposium and for generous financial assistance for the visit to Schloss Korb.

REFERENCES AND NOTES

(1) Trinajstić, N.; Randić, M.; Klein, D. J. On the quantitative structure–activity relationship in drug research. *Acta Pharm. Jugosl.* **1986**, *36*, 267–279.
(2) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417–441, 489–520.
(3) Malinowski, E. R. *Factor Analysis in Chemistry*; Wiley-Interscience: New York, 1991.
(4) Randić, M. Comparative structure–property studies. The connectivity basis. *Theor. Chim. Acta*, submitted.
(5) (a) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517–525. (b) Randić, M. Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
(6) (a) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615. (b) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V. Connectivity series applied to density. *J. Pharm. Sci.* **1976**, *65*, 1226–1230.
(7) Platt, J. R. Prediction of isomeric differences in paraffin properties. *J. Phys. Chem.* **1952**, *56*, 328–336.
(8) Randić, M. Graph theoretical approach to structure–activity studies: Search for optimal antitumor compounds. In *Molecular Basis for Cancer, Part A: Macromolecular Structure, Carcinogens, and Oncogenes*; Rein, R., Ed.; Alan R. Liss: New York, 1985; pp 309–318.
(9) (a) Randić, M. Generalized molecular descriptors. *J. Math. Chem.* **1991**, *7*, 155–168. (b) Randić, M.; Oxley, T. Vicinal matrices and their invariants. Preprint, to be submitted to *J. Math. Chem.* (c) Tratch, S. S.; Stankevitch, M. I.; Zefirov, N. S. Combinatorial methods and algorithms in chemistry. The expanded Wiener number—a novel topological index. *J. Comput. Chem.* **1990**, *11*, 899–908.
(10) Randić, M. Representation of molecular graphs by basic graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57–69.
(11) Baker, G. A., Jr. Drum shapes and isospectral graphs. *J. Math. Phys.* **1966**, *7*, 2238–2242.
(12) Randić, M. On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164–175.
(13) (a) Szymanski, K.; Muller, W. R.; Knop, J. V.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164–175. (b) Szymanski, K.; Muller, W. R.; Knop, J. V.; Trinajstić, N. Molecular ID numbers. *Croat. Chem. Acta* **1986**, *59*, 719–723.
(14) (a) Klein, D. J. In *Mathematical and Computational Concepts in Chemistry*; Trinajstić, N., Ed.; Harwood: Chichester, 1986; Chapter 18. (b) Klein, D. J. Chemical graph-theoretic cluster expansions. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1986**, *20*, 153–171.
(15) McHughes, M. C.; Poshusta, R. D. Graph-theoretic cluster expansions, thermochemical properties for alkanes. *J. Math. Chem.* **1990**, *4*, 227–249.
(16) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic: New York, 1976.
(17) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(18) Hall, L. H. Computational aspects of molecular connectivity and its role in structure-property modeling. In *Computational chemical graph theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990.
(19) (a) Randić, M. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45–59. (b) Randić, M. Search for optimal molecular descriptors. *Croat. Chem. Acta* **1991**, *64*, 43–54. (c) Randić, M. Chemical structure—what she is? *J. Chem. Educ.*, in press.
(20) (a) Randić, M. On Computation of Optimal Parameters for Multivariate Analysis of Structure–Property Relationship. *J. Comput. Chem.* **1991**, *12*, 970–980. (b) Randić, M.; Dobrowolski, J. Cz. Optimal Molecular Connectivity Descriptors for Nitrogen Containing Molecules. *J. Math. Chem.*, submitted.