Full Articles

The chemical structure matrix and a new formalism for the QSPR problem

E. A. Smolenskii

N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences,
47 Leninsky prosp., 119991 Moscow, Russian Federation.
E-mail: smolensk@ioc.ac.ru

The notion of the chemical structure matrix (CSM) is introduced. The columns of the CSM represent the numbers of occurrences of different subgraphs in molecular graphs and are treated as vectors in a linear space. Any topological indices and physicochemical properties can also be treated as vectors in the same linear space. The QSPR problem is formally reduced to the search for linear correlations between vectors. A simple procedure for solving the problem is proposed. A novel method for establishing QSAR is outlined.

Key words: descriptors, matrix chemical structures, molecular graphs, structural vectors, structure—property relationships, topological indices.

Establishment of QSPR for organic compounds has long been attracted the attention of researchers. To solve this problem is to understand how the structure of organic molecules affects their properties and how to use the relationships established for prediction of the properties of hypothetical compounds. The latter aspect is of great practical value, which warrants the existence of numerous empirical and semiempirical expressions for the QSPR.

Usually, each structure is assigned a number called the topological index or descriptor. Often, this is not a one-to-one correspondence. Indeed, although each molecular structure corresponds to a unique number, this number can simultaneously correspond to different molecules (degeneration of topological indices). Clearly, in this case it is impossible to find a function that could exactly describe the dependence of a certain property on the topological index, because the same argument values correspond to different numerical values of the parameter. Nevertheless, the maximum possible squared correlation coefficient can be calculated *a priori* even in these cases.[1]

Among the first successful attempts to design the topological indices, there were studies by H. Wiener on the description of the boiling points[2] and some other thermochemical characteristics[3] of alkanes. Since then a great variety of topological indices were proposed and some of them are widely used (*e.g.*, the Hosoya index[4] and the Randič index[5]).

Many topological indices are to some extent related to graph theory (the possibility of using it in QSPR studies was demonstrated earlier.[6—9]) In this work we will (i) show that the number of independent topological indices is limited (it is much smaller than the number of the known topological indices) and (ii) derive a corresponding gen-

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 9, pp. 1447—1453, September, 2006.

eral relationship. To this end, we propose to construct a chemical structure matrix (CSM) whose columns will be called the structural vectors. Using the CSM, one can formalize the search for QSPR, reduce it to solution of a system of linear equations, and derive an expansion of a property vector over an orthonormalized basis.

The chemical structure matrix

We will consider alkanes as an example. This choice is due to not only simple molecular structures, but also specific role of alkanes in the QSPR studies.[1,10]

Let each alkane correspond to a graph whose vertices are occupied by carbon atoms (hydrogen atoms are ignored; see, *e.g.*, Ref. 11). We will respectively denote the graphs corresponding to methane, ethane, propane, *n*-butane, isobutane, *n*-pentane, isopentane, neopentane as g_1 (it has only one vertex), g_2, g_3, g_4, g_5, g_6, g_7, g_8, *etc*. In other words, all graphs with the same number of vertices are enumerated sequentially, although the enumeration order can be arbitrary. This means that if the number of carbon atoms in the *i*th alkane molecule is larger than in the *j*th alkane molecule than $i > j$.

By analogy enumerate all alkanes starting with methane up to the C_nH_{2n+2} isomers from 1 to N without imposing any restrictions on the number n. Considering the graph g_j as a subgraph of the graph g_i, denote the number of occurrences of g_j in g_i as $[g_j]_i$. If $i < j$, then by definition one has $[g_j]_i \equiv 0$ (number of vertices in a subgraph is no larger than the number of vertices in the corresponding graph). Now we can construct a matrix S of dimension $N \times N$ with the matrix elements $s_{ij} = [g_j]_i$. The matrix S is called the chemical structure matrix.

Consider the CSM of alkanes C_1—C_6 as an example.

$$S = \left|\begin{matrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 3 & 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 4 & 3 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 4 & 4 & 2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 4 & 6 & 0 & 4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
6 & 5 & 4 & 3 & 0 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
6 & 5 & 5 & 3 & 1 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
6 & 5 & 5 & 4 & 1 & 1 & 2 & 0 & 0 & 0 & 1 & 0 & 0 \\
6 & 5 & 6 & 4 & 2 & 0 & 4 & 0 & 0 & 0 & 0 & 1 & 0 \\
6 & 5 & 7 & 3 & 4 & 0 & 3 & 1 & 0 & 0 & 0 & 0 & 1
\end{matrix}\right| \qquad (1)$$

In the case of hexane isomers the graphs g_9, g_{10}, g_{11}, g_{12}, and g_{13} correspond to *n*-hexane, 2-methylpentane, 3-methylpentane, 2,3-dimethylbutane, and 2,2-dimethylbutane, respectivly. The matrix (1) is a lower triangular matrix with all diagonal elements equal to unity. This corresponds to the fact that each graph is the only subgraph of its own. In matrix (1) we show the submatrices for the butane, pentane, and hexane isomers; here, the all diagonal elements are also equal to unity. If we increase the number of alkanes and include nine isomers of heptane, the matrix elements s_{ij} ($14 \le i, j \le 22$) are also equal to unity. This reasoning is valid for any n when the number of vertices increases by unity; therefore, it can be proved by induction that in the general case the determinant of the matrix S equals unity. This means that if we consider the columns of the matrix S as structural vectors, all of them are linearly independent in the N-dimensional linear space generated by N chemical structures. The properties of the CSM are independent of the type of the set of compounds under study. If we add cyclic alkanes, then for, *e.g.*, pentane one should also allow for cyclopentane, methylcyclobutane, ethylcyclopropane, and two dimethylcyclopropane isomers. Similarly, one can add compounds with the multiple bonds, heteroatoms, *etc*. At any extension the determinant of the matrix (1) equals unity and the structural vectors remain linearly independent despite an increase in their number.

Topological indices

Denote the *j*th column-vector of the matrix S as $\bar{g}_j = [g_j]_i$. Then, the *i*th element of this vector is the number of occurrences of the subgraph g_j in graph g_i and corresponds to the *i*th molecule in the set of the molecules under study ($1 \le i \le N$). We can say that an arbitrary (linear) space of dimensionality N is generated by the set including N molecules if each element of the vector $\bar{g}_j = [g_j]_i$ is related in some way to the corresponding molecule.

Since any topological index (descriptor I) for the *i*th molecule equals I_i, it can be treated as a column-vector in the same N-dimensional space generated by the set of N compounds (as an example, we consider alkanes, although this limitation is insignificant). Because the linearly independent (see above) vectors $\bar{g}_j$ form the basis of this space, any vector in this space can be represented by a linear combination

$$\bar{I} = \sum_{j=1}^{N} a_j \bar{g}_j. \qquad (2)$$

Each vector $\bar{g}_j$ is also a topological index. This assumption was first implicitly put forward in a study[12] of chains of different length (unbranched molecular fragments). Of course, the basis $\{\bar{g}_j\}$ is not unique, because, *e.g.*, N different linearly independent topological indices can also be chosen as a basis. It should be emphasized that the number of linearly independent topological indices can not be larger than N. But we consider the basis $\{\bar{g}_j\}$ as being a natural, simplest, and having a clear and unambiguous chemical sense, namely, all subgraphs g_i correspond to all possible alkanes (with allowance for the limitation imposed on the number of atoms n). The chemical sense of all other bases of topological indices will differ from that of the basis $\{\bar{g}_j\}$.

Now we will consider some examples. For the Wiener index[2] one can choose the numbers of occurrences of unbranched chains of length 2, 3, *etc.* up to the maximum value $\rho(m, n)$ as the basis vectors and the corresponding chain lengths as coefficients:

$$W = \sum_{j=1}^{\max \rho} j[g'_j]. \qquad (3)$$

In the case of the matrix (1) the vectors $g_j{}'$ are the numbers of occurrences of the subgraphs g_2, g_3, g_4, g_6, g_9, *etc.*

The Randič index[5] is given by

$$\chi = \sum_{j=1}^{9} b_j[t_j], \qquad (4)$$

if we use the following notations for the subgraphs with labeled vertices, namely, t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, and t_9 are the subgraphs $C_1—C_2$; $C_1—C_3$; $C_1—C_4$; $C_2—C_2$; $C_2—C_3$; $C_2—C_4$; $C_3—C_3$; $C_3—C_4$; and $C_4—C_4$, respectively. A labeled vertex is assumed to be a vertex with the corresponding index, which in this case means the multiplicity of a given carbon atom, *i.e.*, the number of other carbon atoms bonded to it (ignoring H atoms). All C_iC_j subgraphs have two vertices that can be different. The coefficients b_j can be expressed through the multiplicities of the atoms C_i and C_k, namely, $b_j = 1/\sqrt{ik}$ for the subgraphs $C_i—C_k$. The figures in square brackets in expressions (3) and (4) denote the numbers of occurrences of corresponding subgraphs in the molecular graph.

Derivation of relationships (3) and (4) presents no difficulties. Mention may be made that the subgraphs t_j are subgraphs with labeled vertices and the CSM is formed by simpler graphs with non-labeled vertices. Therefore, it is of interest to express the Randič index through the basis vectors $\bar{g}_j$. Of course, we can take into account the fact that the parameters $[t_j]$ are topological indices and can be expressed through the vectors $\bar{g}_j$, but we will derive an explicit expression for the Randič index in order to clearly demonstrate the essence of the approcah proposed.

Let us denote the contributions of the subgraphs $C_i—C_j$ as α_i (Table 1) in order to retain the meaning of the notations and to solve the equations without using a computer. We thus obtain the coefficients in the form of integer numbers because a computer-assisted procedure using standard programs can not give us clear and readily understandable results.

Table 1. Contributions of subgraphs $C_i—C_j$ to the Randič index

Subgraph	Notation of contribution	Contribution magnitude
$C_1—C_2$	$\alpha_1 = 1/\sqrt{1\cdot 2}$	$\alpha_1 = \sqrt{2}/2$
$C_1—C_3$	$\alpha_2 = 1/\sqrt{1\cdot 3}$	$\alpha_2 = \sqrt{3}/3$
$C_1—C_4$	$\alpha_3 = 1/\sqrt{1\cdot 4}$	$\alpha_3 = 1/2$
$C_2—C_2$	$\alpha_4 = 1/\sqrt{2\cdot 2}$	$\alpha_4 = 1/2$
$C_2—C_3$	$\alpha_5 = 1/\sqrt{2\cdot 3}$	$\alpha_5 = \sqrt{6}/6$
$C_2—C_4$	$\alpha_6 = 1/\sqrt{2\cdot 4}$	$\alpha_6 = \sqrt{2}/4$
$C_3—C_3$	$\alpha_7 = 1/\sqrt{3\cdot 3}$	$\alpha_7 = 1/3$
$C_3—C_4$	$\alpha_8 = 1/\sqrt{3\cdot 4}$	$\alpha_8 = \sqrt{3}/6$
$C_4—C_4$	$\alpha_9 = 1/\sqrt{4\cdot 4}$	$\alpha_9 = 1/4$

The Randič indices of twenty-two alkanes $C_1—C_7$ expressed using the notations given in Table 1 are listed in Table 2 along with one octane isomer that is included in the basis for the Randič index (this basis is complete).

The contributions a_j (see Table 2) can be found by solving the system of linear equations:

$$\sum_{j=1}^{N} a_j g_{ij} = \chi_i, \qquad (5)$$

where χ_i is the Randič index for the ith molecule and g_{ij} are the elements of the CSM. Since S is the triangular matrix, solution of the first two equations in system (5) gives $a_1 = a_2 = 0$, because the values χ_1 (for methane) and χ_2 (for ethane) can be set equal to zero. In the former case this is obvious while in the latter case a formal value of unity $(1/\sqrt{1\cdot 1})$ for the $C_1—C_1$ subgraph is senseless, because this subgraph is included only in the molecular graph of ethane. From the third equation of system (5) it follows that

$$a_3 = 2\alpha_1,$$

because the χ_3 value for propane equals $2\alpha_1$.

The fourth equation of system (5) has the form

$$2a_3 + a_4 = 2\alpha_1 + \alpha_4.$$

Substituting the a_3 value gives

$$a_4 = \alpha_4 - 2\alpha_1.$$

For the fifth equation one has

$$3a_3 + a_5 = 3\alpha_2,$$

and from here it follows that $a_5 = 3\alpha_2 - 6\alpha_1$. Following this way, one gets $a_6 = 0$, $a_7 = \alpha_5 - 2\alpha_4 - \alpha_2 + 3\alpha_1$, *etc.* All subgraphs with nonzero elements in the third column of Table 2 form the basis for the Randič index. The total

Table 2. Randič indices and coefficients a_i for the basis vectors

Compound	Index Randič	Contributions of the basis vectors a_i	Contribution magnitude
Methane	0	0	0
Ethane	0	0	0
Propane	$2\alpha_1$	$2\alpha_1$	1.4142
Butane	$2\alpha_1 + \alpha_4$	$\alpha_4 - 2\alpha_1$	–0.9142
2-Methylpropane	$3\alpha_2$	$3\alpha_2 - 6\alpha_1$	–2.5106
Pentane	$2\alpha_1 + 2\alpha_4$	0	0
2-Methylbutane	$\alpha_1 + 2\alpha_2 + \alpha_5$	$\alpha_5 - 2\alpha_4 - \alpha_2 + 3\alpha_1$	0.9522
2,2-Dimethylpropane	$4\alpha_3$	$4\alpha_3 - 12\alpha_2 + 12\alpha_1$	3.5571
Hexane	$2\alpha_1 + 3\alpha_4$	0	0
2-Methylpentane	$\alpha_1 + 2\alpha_2 + \alpha_4 + \alpha_5$	0	0
3-Methylpentane	$2\alpha_1 + \alpha_2 + 2\alpha_5$	0	0
2,2-Dimethylbutane	$\alpha_1 + 3\alpha_3 + \alpha_6$	$\alpha_6 - 3\alpha_5 + 3\alpha_4 - \alpha_3 + 3\alpha_2 - 4\alpha_1$	–0.9676
2,3-Dimethylbutane	$4\alpha_2 + \alpha_7$	$\alpha_7 - 4\alpha_5 + 4\alpha_4 + 2\alpha_2 - 4\alpha_1$	–0.9734
Heptane	$2\alpha_1 + 4\alpha_4$	0	0
2-Methylhexane	$\alpha_1 + 2\alpha_4 + 2\alpha_2 + \alpha_5$	0	0
3-Methylhexane	$2\alpha_1 + \alpha_4 + \alpha_2 + 2\alpha_5$	0	0
3-Ethylpentane	$3\alpha_1 + 3\alpha_5$	0	0
2,2-Dimethylpentane	$\alpha_1 + 3\alpha_3 + \alpha_4 + \alpha_6$	0	0
2,3-Dimethylpentane	$\alpha_1 + 3\alpha_2 + \alpha_5 + \alpha_7$	0	0
2,4-Dimethylpentane	$4\alpha_2 + 2\alpha_5$	0	0
3,3-Dimethylpentane	$2\alpha_1 + 2\alpha_3 + 2\alpha_6$	0	0
2,2,3-Trimethylbutane	$2\alpha_2 + 3\alpha_3 + \alpha_8$	$\alpha_8 - 3\alpha_7 - 2\alpha_6 + 9\alpha_5 - 6\alpha_4 + \alpha_3 - 4\alpha_2 + 5\alpha_1$	0.9819
2,2,3,3-Tetramethylbutane	$6\alpha_3 + \alpha_9$	$\alpha_9 - 6\alpha_8 + 9\alpha_7 + 6\alpha_6 - 18\alpha_5 + 9\alpha_4 - 2\alpha_3 + 6\alpha_2 - 6\alpha_1$	–0.9877

number of such subgraphs is nine. Renaming the indices (rejecting zero contributions), one gets

$$\bar{\chi} = \sum_{j=1}^{9} a'_j \bar{g}'_j. \qquad (6)$$

The meaning of the terms in expression (6) is clear (see Table 2). Note that the basis vectors thus obtained and relationship (6) in the new basis also describe the Tatevskii method[13] upon renaming the Tatevskii parameters P_{ij} as α_i. It should also be noted that the Randič index[5], the Tatevskii method,[13] and more complex methods[14] were both formally and in essence foreshadowed by Taylor *et al.*[15] in one of the first serious QSPR studies.

Seemingly, the expressions for the Hosoya index should be more complicated, but this is not the case. Repeating the procedure described above, one can readily obtain the solution and the expansion coefficients for the Hosoya index[4] are much simpler than for the Randič index. For the set including thirrteen alkanes (see matrix (1)) the basis includes four vectors; therefore, one has

$$Z = [g_4] + [g_6] + 2[g_9] + [g_{11}].$$

When considering the heptane isomers, the following expression is added

$$3[g_{14}] + [g_{16}] + [g_{17}],$$

where g_{14}, g_{16}, and g_{17} correspond to *n*-heptane, 3-methylhexane, and 3-ethylpentane, respectively; *i.e.*, for $N = 22$ the basis includes seven vectors. The basis vectors for the Hosoya index and the corresponding coefficients for the set including alkanes up to nonane ($N = 75$) are listed in Table 3; in this case the number of the basis vectors is twenty-three.

Thus, any topological index can be expanded over a corresponding basis:

$$I = \sum a_j \left[g_j \right]. \qquad (7)$$

The simplest form has the so-called complexity index K;[16] here the basis includes all subgraphs of the molecular graph and all coefficients a_j are equal to unity. This means that the coordinates of the complexity index are equal to the sums of all elements of the matrix S in the rows:

$$K_i = \sum_{j=1}^{N} s_{ij}.$$

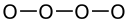
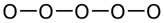
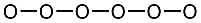
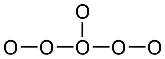
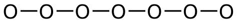
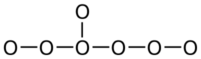
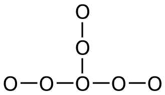
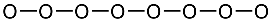
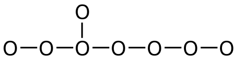
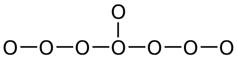
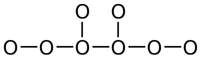
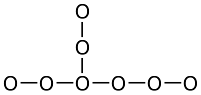
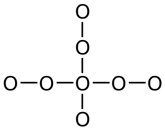
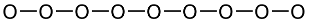
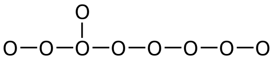
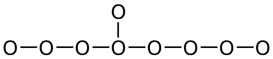
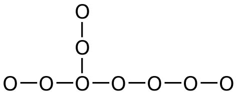
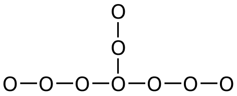
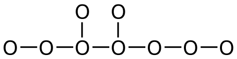
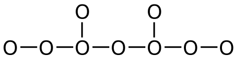
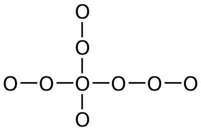
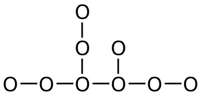
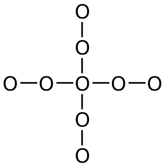
Physicochemical properties

Relationship (7) was proposed[8] for representation of the physicochemical properties rather than topological indices (notion "topological index" was not coined at that time):

$$P = \sum a_j \left[g_j \right]. \qquad (8)$$

Since at present it is clear that there is no index suitable for correct representation of an arbitrary property P even for alkanes, this is achieved using a linear combination of a number of topological indices.[10]

Table 3. The basis subgraphs for the Hosoya index

C_n	Subgraphs	Contributions to $Z_k = p(G, k)$			Contribution to Z
		Z_2	Z_3	Z_4	
C_4		1	0	0	1
C_5		1	0	0	1
C_6		1	1	0	2
		0	1	0	1
C_7		1	2	0	3
		0	1	0	1
		0	1	0	1
C_8		1	3	1	5
		0	1	1	2
		0	1	0	1
		0	0	1	1
		0	0	1	1
		0	0	1	1
C_9		1	4	3	8
		0	1	2	3
		0	1	1	2
		0	1	2	3
		0	1	2	3
		0	0	1	1
		0	0	1	1
		0	0	1	1
		0	0	1	1
		0	0	1	1

As mentioned above, for any topological index one can construct a corresponding vector and expand it over structural vectors using expression (7). Formally, the same vector can be constructed for any physicochemical property, but there is a fundamental distinction between these two cases. If for the topological indices we can "calculate" all elements of the vector, the determination of the elements of the property vector requires experimental studies. Therefore, some elements of the corresponding vector can be unknown, the more so measurements should be preceded by the synthesis of one or a group of compounds, which can present a complicated problem. In essence, the problem in question had posed due to the need of doing quite reliable calculations (although not exact ones) rather than synthesizing compounds and measuring their characteristics.

Now we will turn to formalization of the QSPR problem. First of all we shall reject[17] a senseless extension of the variety of topological indices and use the matrix S for deriving, for each set including N compounds with the known property P, the following system of equations

$$S\overline{A} = \overline{P}, \qquad (9)$$

$$\overline{A} = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_N \end{pmatrix}, \qquad \overline{P} = \begin{pmatrix} P_1 \\ P_2 \\ \cdot \\ \cdot \\ \cdot \\ P_N \end{pmatrix},$$

where $\overline{A}$ is the solution vector and $\overline{P}$ is the property vector. Emphasize: this does not mean that topological indices are no longer required (see, *e.g.*, Ref. 18).

Unfortunately, but most likely, no simple solutions similar to those mentioned above will be obtained. Here, simplicity assumes vanishing of a large number of coefficients a_i and, as a consequence, retention of a relatively small number of the basis vectors. Probably, almost all coefficients a_i will differ from zero due to approximate character of the computer representation of numbers. The problem is to choose a minimum number of the basis vectors using the solution to the system (9) and at the same time to ensure the maximum possible value of the squared correlation coefficient, R^2 or $R^2_{\text{c.v.}}$. However, when choosing, *e.g.*, twenty basis vectors for $N = 100$ (see Ref. 19) the computational cost of the solution will be high. Therefore, we propose a simpler method of solution. Using orthogonalization according to Schmidt, one can obtain an orthonormalized basis $\{g_i{'}\}$. We thus no longer work with expansion (8) but get the following one

$$P' = \sum_{i=1}^{N} c_i g_i', \qquad (10)$$

where P' is the normalized property vector. Here one has

$$\sum_{i=1}^{N} c_i^2 = 1,$$

and the coefficients c_i do not depend on the number of vectors in the sum (10) (one can use a smaller number m instead of N terms) and on the order of their choice. In essence, this is an expansion over the orthonormalized basis:

$$P' = \sum_{i=1}^{m} c_i g_i'. \qquad (11)$$

For instance, to ensure $R^2 = 0.95$, one should use the largest coefficients c_i in order to provide $\sum_{i=1}^{m} c_i^2 \approx 0.95$. This approximately gives $R^2 = 0.95$. A sufficiently small number m suggests that the we have solved the problem posed and related the properties P to the molecular structure. This is analogous to the situation with the inclusion of electron correlation energy in quantum chemical calculations carried out by the configurational interaction method.

At the same time if for a certain property all coefficients c_i are similar and it is impossible to choose an optimum proportion between m and $\sum_{i=1}^{m} c_i^2$, we can state that this property can not be expressed as a function of the molecular structure. In these cases there are no topological indices or other chemical descriptors that could provide reasonable approximate relationships. This first of all can concern the melting points.[1,10,20]

Nevertheless, the solution to Eq. (9) will make it possible to draw important conclusions about the role of particular "substructures" (subgraphs of molecular graph). Large coefficients a_i mean that the corresponding substructures make large (positive or negative) contributions to the magnitude of a property under consideration. *Vice versa*, the subgraphs with small absolute values of the coefficients a_i can be ignored. One can also combine the subgraphs characterized by similar or multiple values of the coefficients a_i and thus design new topological indices.

Operation in the orthonormalized basis is convenient from the computational standpoint, although it can be accompanied by the loss of clarity. Namely, vectors will be associated with linear combinations rather than numbers of subgraphs (not necessarily with integer coefficients in the former case). This means the loss of simple chemical meaning of the parameters, but can be useful for improvement of the accuracy of prediction.

Equation (9) can be treated as the main equation relating the structure to the properties. Formally, it generalizes all situations in which the index approaches are employed. Of course, this only is a linear approximation. It is possible to find another types of solutions, *e.g.*, nonlinear ones that are suitable for obtaining[21] good correlations for certain types of biological activity of some steroids.

* * *

Thus, we reported on fundamental progress in attacking the QSPR problem, which can be considered solved at least in the linear approximation. Obtaining particular data mainly requires calculatons concerned with solution of Eq. (9). However, the structure—activity relationships (QSAR) are of much greater practical importance. It is also possible to construct a corresponding activity vector. In this case the choice of compounds plays a crucial role. On the one hand, structurally similar compounds that are both active and (to some extent) inactive should be taken into account. On the other hand, it is clear that the matrix S (it can be called the topological matrix) as well as versatile topological indices can hardly be useful in solving the general QSAR problem.

Nevertheless, some progress can be made in the field. First, by analogy with the studes[20,22] one can construct the topographic CSM in which particular substructures will correspond to actual spatial arrangement of atoms, conformational states of molecules, and other features, and solve the QSAR problem using the method described above. Second, going beyond the linear approximation is highly expected (see above),[21] but in any case this approximation should serve as a basis for subsequent refinement.

Additionally, in carrying out particular calculations one should carefully follow that the number of the basis vectors be no larger than the number of compounds (in particular, when certain structures should be excluded due to the lack of corresponding experimental data). For instance, linear dependences between the structural vectors can appear in studies of the molecular refraction of n-alkanes (data for the light alkanes cannot be obtained because of too low boiling points), *e.g.*,

$$2[g_2] - [g_1] = [g_3].$$

This feature will play a particular role in the search for QSAR. In this case the number of different "topographic subgraphs" can appear to be larger than the number of compounds in a particular set and "representativeness" of the set will preclude construction of even a linear approximation. By the way, the approach described above will allow one to take control over the situation and to avoid obtaining artifacts.

Appendix

Importance of the interrelations between different topological indices has long been reported.[17] The mutual correlation between the topological indices was presented in the matrix form.[10] Graph invariants and the physicochemical properties were expressed[11,23] using various structural fragments. This interplay was more thoroughly considered in other publications.[24—26] All these studies used the notion of the invariants of molecular graphs and therefore many propositions required additional theoretical substantiation. A number of theorems were rigorously proved[24,25] and, in particular, the theorem on the existence of the basis of molecular graph invariants. Some interesting examples were also reported, which demonstrate the dependence of the molecular graph invariants on the number of occurrences of different subgraphs in the molecular graph.

The approach proposed in this work uses vectors instead of graph invariants, is simpler and clearer. Indeed, since any topological index can be multiplied by an arbitrary real number α, both vectors, I and αI, are also topological indices. The numerical value of any property P can also be multiplied by an arbitrary number β; thus we obtain a formally different value of *the same* property. This means a change in the scale or in the units of measurement. We can introduce the notion "generalized property" that can be obtained by adding the vector corresponding to the topological index with the property vector. Formally, both conventional topological indices and the properties can be considered as generalized properties. Then we can state that if the vectors of topological indices (I) and the property vectors (P) belong to a certain space, the vector $\alpha I + \beta P$ also belongs to this space. Mathematically, this means that the space in question is a linear space with well-studied properties. Therefore, in this work in contrast to studies[24,25] we don´t have to proof the existence of the basis and the possibility of expressing any index and property as a linear combination of the basis vectors. Thus, we can use not only these, but also all other properties of linear spaces, *e.g.*, the fact that any $N + 1$ vectors V_i in the N-dimensional space are related as follows:

$$\sum_{i=1}^{N+1} \alpha_i V_i = 0,$$

where $\sum\limits_{i=1}^{N+1} \alpha_i^2 > 0$. This proves that the number of linearly independent topological indices does not exceed N.

Of course at first glance adding the vectors of topological indices and the property vectors seems to be unnatural, because it is similar to the request to add three apples to four oranges. But this is impossible in arithmetics, being quite probable in algebra. The formulation of the QSPR problem also implicitly contains this idea. Indeed, from the algebraic point of view establishment of linear correlations between the topological indices and properties is nothing but this kind of addition.

We believe that it is somewhat alogical transition from graph invariants to vectors allows one to formalize the QSPR problem and to derive a linear equation for solving it.

References

1. E. A. Smolenskii, *Dokl. Akad. Nauk*, 1999, **365**, 767 [*Dokl. Chem.*, 1999 (Engl. Transl.)].
2. H. Wiener, *J. Am. Chem. Soc.*, 1947, **69**, 17.
3. H. Wiener, *J. Am. Chem. Soc.*, 1947, **69**, 2636.
4. H. Hosoya, *Bull. Chem. Soc. Jpn*, 1971, **44**, 2332.
5. M. Randič, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
6. E. A. Smolenskii, *Zh. Vychisl. Mat. Mat. Fiz.* [*J. Comput. Math. Math. Phys.*], 1962, **2**, 371 (in Russian).
7. E. A. Smolenskii and A. L. Seifer, *Zh. Fiz. Khim.*, 1963, **37**, 2657 [*J. Phys. Chem. USSR*, 1963, **37** (Engl. Transl.)].
8. E. A. Smolenskii, *Zh. Fiz. Khim.*, 1964, **38**, 1288 [*J. Phys. Chem. USSR*, 1964, **38** (Engl. Transl.)].
9. E. A. Smolenskii and A. L. Seifer, *Zh. Fiz. Khim.*, 1964, **38**, 1548 [*J. Phys. Chem. USSR*, 1964, **38** (Engl. Transl.)].
10. D. E. Needham, I. C. Wei, and P. C. Seybold, *J. Am. Chem. Soc.*, 1988, **110**, 4186.
11. T. G. Schmalz, D. J. Klein, and D. L. Sandleback, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 54.
12. J. R. Platt, *J. Phys. Chem.*, 1952, **56**, 328.
13. V. M. Tatevsky and Yu. G. Papulov, *Zh. Fiz. Khim.*, 1962, **36**, 189 [*J. Phys. Chem. USSR*, 1962, **36** (Engl. Transl.)].
14. E. A. Smolenskii, *Dokl. Akad. Nauk SSSR*, 1976, **230**, 373 [*Dokl. Chem.*, 1976 (Engl. Transl.)].
15. W. J. Taylor, J. M. Pignocco, and F. D. Rossini, *J. Res. Natl. Bur. St.*, 1945, **34**, 413.
16. D. Bonchev, E. Marcel, and A. Dekmezian, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1274.
17. D. H. Rouvray, in *Chemical Applications of Topology and Graph Theory*, Ed. R. B. King, Elsevier, Amsterdam, 1983, 159.
18. E. A. Smolenskii, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 522.
19. J. G. Topliss and R. P. Edwards, *J. Med. Chem.*, 1979, **22**, 1238.
20. E. A. Smolenskii, A. N. Ryzhov, A. L. Lapidus, and N. S. Zefirov, *Dokl. Akad. Nauk*, 2002, **387**, 69 [*Dokl. Chem.*, 2002 (Engl. Transl.)].
21. A. V. Kamernitskii, E. A. Smolenskii, G. M. Makeev, I. V. Vesela, N. M. Mirsalikhova, A. M. Turuta, and N. S. Zefirov, *Bioorg. Khim.*, 2002, **28**, 269 [*Russ. J. Bioorg. Chem.*, 2002, **28** (Engl. Transl.)].
22. Z. Mihalić and N. Trinajstić, *J. Mol. Struct. (THEOCHEM)*, 1991, **232**, 65.
23. D. J. Klein, *Int. J. Quantum Chem.*, 1986, **20**, 153.
24. I. I. Baskin, M. I. Skvortsova, I. V. Stankevich, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 527.
25. M. I. Skvortsova, I. I. Baskin, L. A. Skvortsov, V. A. Palyulin, N. S. Zefirov, and I. V. Stankevich, *J. Mol. Struct. (THEOCHEM)*, 1999, **466**, 211.
26. I. I. Baskin, *Dokl. Akad. Nauk*, 1994, **339**, 346 [*Dokl. Chem.*, 1994 (Engl. Transl.)].

Received November 7, 2005;
in revised form May 15, 2006