Full Articles

Modeling of structure—property relationships for hydrocarbons involving the basis topological descriptors

M. I. Skvortsova,[a] *K. S. Fedyaev,*[b] *V. A. Palyulin,*[c] *and N. S. Zefirov*[c]★

[a]*N. D. Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences,*
47 Leninsky prosp., 119991 Moscow, Russian Federation.
Fax: +7 (095) 135 5328
[b]*Institute of Physiologically Active Compounds, Russian Academy of Sciences,*
142432 Chernogolovka, Moscow Region, Russian Federation.
Fax: +7 (095) 785 7024
[c]*Department of Chemistry, M. V. Lomonosov Moscow State University,*
Leninskie Gory, 119992 Moscow, Russian Federation.
Fax: +7 (095) 939 0290. E-mail: vap@org.chem.msu.su, zefirov@org.chem.msu.ru

The article concerns the problem of selection of topological molecular descriptors to be used in constructing mathematical models of structure—property relationships for hydrocarbons. A new method of solving this problem is proposed. The approach involves construction of parameter-dependent basis topological descriptors and selection of the descriptors suitable for constructing particular models. A qualitative substantiation of the method proposed is given and some examples of its application to modeling of the structure—property relationships for hydrocarbons are presented.

Key words: molecular graphs, hydrocarbons, models for structure—property relationships, basis of topological molecular descriptors.

Constructing quantitative structure—property/activity relationships (QSPR/QSAR) is one of the most important avenues of research in modern theoretical chemistry.[1,2] The relations obtained are used for estimating certain characteristics of those compounds for which experimental data are unavailable and for constructing mathematical models of various physicochemical processes or action mechanisms of biologically active compounds. Usually, a QSPR/QSAR model has the form of an equation

$$y = f(x_1, ..., x_n),$$

relating particular topological, electronic, physicochemical, *etc.* molecular parameters $x_1, ..., x_n$ to a property y by means of a certain function of n variables, $f = f(x_1, ..., x_n)$.

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 8, pp. 1527—1535, August, 2004.

The main problems in the field of QSPR/QSAR studies involve the choice and calculations of the parameters x_1, ..., x_n, the choice of the function f, determination of the applicability range of the mathematical model constructed, and interpretation of the model and possible implications.

The most popular molecular parameters used in the QSPR/QSAR studies are the topological descriptors (TDs), *i.e.*, invariants of the molecular graphs representing the chemical structures.[1,2] The advantages of the TDs are relative simplicity and the possibility of being rapidly calculated (compared to certain quantum-chemical descriptors), which is of particular importance in processing of large databases. In addition, the TDs can even be calculated for hypothetical structures (in contrast to, *e.g.*, the physicochemical descriptors) and constructed with inclusion of specific features of the electronic and spatial structure. Numerous structure—property relationships obtained using the TDs confirm the efficiency of this approach.

However, it is possible to construct an infinite number of different TDs suitable for the use in the QSPR/QSAR studies; however, most of them usually have no clear structural or physicochemical interpretation. Therefore, the problem poses of selection of a finite number of descriptors belonging to this infinite set. Often, good correlations are the fruits of researcher´s intuition. One should also keep in mind that many descriptors are interrelated by some rigorous or approximate mathematical relations; as a consequence, they contain the same structural information.

The problem, which naturally poses in this connection, can be formulated as follows: it is necessary to find a family of universal descriptors suitable for unambiguous expression of any TD for any set of molecular graphs. We will call descriptors belonging to this family the *basis descriptors*. Given the basis descriptors, one would use them instead of an infinite set of TDs employed in the QSPR/QSAR studies, so that corresponding mathematical models could be constructed using the basis descriptors only.

Construction of a basis of graph invariants has been a subject of some earlier studies.[3—8] Often, the numbers of occurrences of particular subgraphs (*basis structural fragments*) in a graph were used as the basis descriptors. The approaches employed can be divided into two groups. The first group comprises methods which include a mathematically rigorous proof of the basis property of corresponding descriptors or subgraphs (see, *e.g.*, Refs 6 and 7). The second group includes methods in which the basis property of particular sets of descriptors is intuitively determined by a researcher or simply postulated. The first group of approaches appeared to be inefficient in solving specific chemical problems due to either the use of subgraphs with relatively large numbers of vertices or very complicated procedure of definition of these subgraphs. The descriptors proposed for the second group of methods were (i) the connectivity indices of different order; (ii) the indices defined as the sums of the elements of the mth degree (m = 1, 2, ...) of some matrix of a graph (*e.g.*, connectivity matrix); and (iii) the numbers of occurrences of chains of different length and their unconnected subgraphs in the graph.[3—5,8] Often, particular sets of descriptors are chosen as the basis descriptors only due to the fact that there is a large number of structure—property relationships involving these descriptors or due to the possibility of unambiguous encoding of the structures of some small class of compounds using these descriptors.

In this work, we introduce a new family of the basis TDs for simple molecular graphs. These descriptors are constructed using particular structural fragments and depend on certain parameters. The choice of these invariants for constructing structure—property relationships is substantiated. Application of the method proposed is exemplified by constructing QSPR-equations for those hydrocarbons whose structures can be represented by simple molecular graphs. In the examples given below both various physicochemical characteristics and a number of widely used TDs of compounds are considered as the "property."

Description of Method

Definition of the basis descriptors. Consider the following types of structural fragments of a simple graph G: (a) a chain of length L (L = 1, 2, ...); (b) a ring with r vertices (r = 3, 4, ...); and (c) a ring with r vertices (r = 3, 4, ...) and one more vertex attached to some of its vertices (we consider all possible types of attachment provided that at most one vertex is attached to a vertex). Let a subgraph F_k of the graph G with k vertices be the join of several unconnected components represented by fragments of the types (a), (b), or (c). It is possible that some kind of fragments will be absent in F_k. Notice that a number of different (non-isomorphous) subgraphs F_k of the graph G can exist for one k value. Enumerate all non-isomorphous subgraphs F_k corresponding to a preset k value arbitrarily and denote the mth subgraph as $F_{k,m}$. Figure 1 presents the subgraphs $F_{k,m}$ for k = 2 (m = 1), k = 3 (m = 1, 2), k = 4 (m = 1—4) and k = 5 (m = 1—6). The procedure for generation and enumeration of the subgraphs F_k at a preset k value was reported earlier.[9] In that study[9] we also introduced two sets of TDs, $\{X_{k,m}\}$ and $\{\varphi_{k,m}\}$, based on the $\{F_{k,m}\}$ subgraphs and applied them in the QSPR/QSAR studies. The $\{X_{k,m}\}$ and $\{\varphi_{k,m}\}$ TDs were defined as follows.

Definition. An occurrence of the subgraph $F_{k,m}$ in the graph G is a subgraph of G, which is isomorphous to the $F_{k,m}$ fragment. Denote the number of occurrences of $F_{k,m}$ in G as $X_{k,m}$. Enumerate all subgraphs of the graph G,

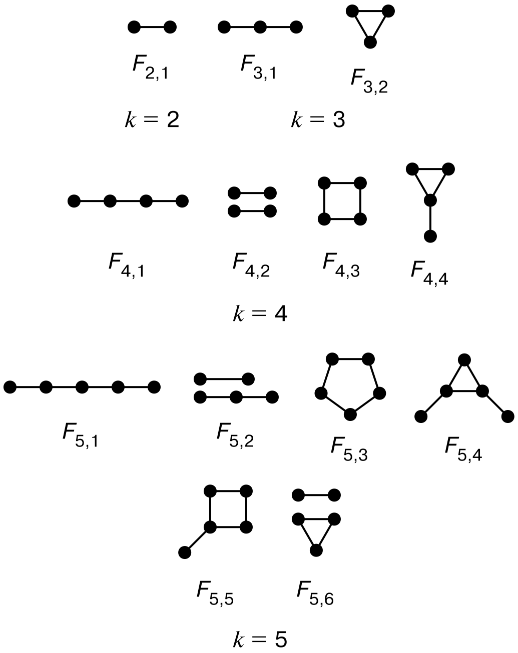

Fig. 1. Basis fragments for $k = 2$—5.

which are isomorphous to $F_{k,m}$, denote the jth subgraph as $F_{k,m,j}$, and assign $F_{k,m,j}$ a number

$$\mu_{k,m,j} = \sum_i \left(v_1^{(j)} v_2^{(j)} \ldots v_{n_i}^{(j)}\right)^{-1/n_i},$$

where n_i is the number of vertices in the ith connectivity component $F_{k,m,j}$, $v^{(j)}{}_p$ ($p = 1, 2, \ldots$) are the degrees of the vertices of $F_{k,m,j}$ in G, and summation is performed over all the connectivity components $F_{k,m,j}$. Then, the descriptor $\varphi_{k,m}$ is given by the relation

$$\varphi_{k,m} = \sum_{j=1}^{X_{k,m}} \mu_{k,m,j}.$$

Here, we propose the following modification of the descriptors $\varphi_{k,m}$. Consider the descriptors $\psi_{k,m}(\alpha) = \varphi_{k,m}/n^{\alpha}$, where n is the number of vertices of the graph G and $\alpha = \alpha(k, m) \geq 0$ is an arbitrary parameter which can take any value for any pair (k, m). Therefore, in the case of $\psi_{k,m}(\alpha)$ any pair (k, m) corresponds to a one-parametric family of descriptors. Let us include yet another descriptor, $\psi_0(\alpha) = n^{\alpha}$, in the family of the $\{\psi_{k,m}(\alpha)\}$ descriptors.

Definition. A family of descriptors $\{Z_k(\alpha)\}$ ($k = 1, 2, \ldots$) depending on the parameter α is called the *basis descriptors* if for any invariant J of any set of non-isomorphous graphs $\{G_1, \ldots, G_N\}$ there exist N descriptors $Z_{j1}(\alpha), \ldots, Z_{jN}(\alpha)$ belonging to this set and N numbers $\alpha_1{}', \ldots, \alpha_N{}'$ such that

$$J = \sum_{i=1}^{N} c_i Z_{ji}(\alpha_i{}'),$$

where c_i are constants ($i = 1, \ldots, N$). To check that this family of descriptors possesses the property of being a basis, it is sufficient to establish the correctness of the following statement: for any set of non-isomorphous graphs $\{G_1, \ldots, G_N\}$ there exist N descriptors $Z_{j1}(\alpha_1{}'), \ldots, Z_{jN}(\alpha_N{}')$ at certain values $\alpha_1{}', \ldots, \alpha_N{}'$ of the parameters $\alpha_1, \ldots, \alpha_N$ such that the matrix $A = (Z_{jp}(\alpha_{jp}{}', G_i))$ ($i, jp = 1, \ldots, N$; $Z_{jp}(\alpha_{jp}{}', G_i)$ is the value of the $Z_{jp}(\alpha_{jp}{}')$ descriptor for the graph G_i) is nondegenerate.

It is possible to prove that the $\{\psi_{k,m}(\alpha), \psi_0(\alpha)\}$ descriptors possess the property of being a basis for, *e.g.*, the following types of sets of the graphs $\{G_1, \ldots, G_N\}$: (A) the graphs $G_1, \ldots, G_N$ have different numbers of vertices (in this case a total of N desired descriptors can be provided by $\psi_{2,1}(\alpha)$, $\alpha = 1, \ldots, N$) and (B) all the graphs G_i ($i = 1, \ldots, N$) belong to the $\{F_{k,m}\}$ set at certain k and m values (in this case the desired number, N, of descriptors can be provided by $\psi_{k,m}(\alpha)$ at corresponding N pairs of the numbers (k, m) and at $\alpha = \alpha(k, m) = 0$).

Notice that the number of sets of the types described above is very large even at relatively small N values and number of graph vertices, n. For instance, the number of the sets of the type (A), which include only the chemical trees with $2 \leq n \leq 11$ and $N = 10$, equals $T(2) \cdot T(3) \cdot \ldots \cdot T(11) = 2028442500$. (Here, $T(n)$ is the number of chemical trees with n vertices; $T(2) = 1$, $T(3) = 1$, $T(4) = 2$, $T(5) = 3$, $T(6) = 5$, $T(7) = 9$, $T(8) = 18$, $T(9) = 35$, $T(10) = 75$, and $T(11) = 159$). The total number of sets of the type (B) for the connected graphs with $2 \leq n \leq 10$ and $2 \leq N \leq 56$ equals $C_{56}{}^2 + C_{56}{}^3 + \ldots + C_{56}{}^{56} = 2^{56} - C_{56}{}^1 - C_{56}{}^0 = 2^{56} - 57$ (the total number of such graphs equals 56; the number of sets including N graphs equals the number of combinations of 56 things N at a time, $C_{56}{}^N$).

Descriptors $\{\psi_{k,m}(\alpha), \psi_0(\alpha)\}$: substantiation of choice. Let us clarify the reasons why the subgraphs $\{F_{k,m}\}$ and descriptors $\{X_{k.m}\}$ were chosen. It is known that the spectrum of a graph with n vertices can be unambiguously determined using the Sachs subgraphs $\{S_k\}$ ($k = 1, 2, \ldots, n$) with k vertices and the chains of length $L = 1$ and rings with r vertices ($r = 3, 4, \ldots$) as the connectivity components.[10] As a consequence, two graphs characterized by the same $\{S_k\}$ sets have the same spectra. However, a graph is not uniquely defined by its spectrum (*i.e.*, the $\{S_k\}$ set). Non-isomorphous graphs having the same spectra are called isospectral graphs (IGs). A reasonable question arises as of how we can distinguish between the IGs? We analyzed the structures of the isospectral trees with $n = 2, \ldots, 10$ reported in Ref. 10 (among a total of two hundred trees with $2 \leq n \leq 10$, there are nine pairs of isospectral trees; among one hundred and forty-nine chemical trees, there are six pairs of isospectral trees) and found that these graphs differ from one another in number of chains of different length. By comparing some ring-containing IGs we found that they have at least different number of rings of particular size with one more vertex attached to some ring vertices. Based on this fact, the set $\{S_k\}$ was augmented with subgraphs representing chains of different length and the rings mentioned above. Taken altogether, these subgraphs form the set $\{F_k\}$.

To be precise, a set of subgraphs is called the *basis set of subgraphs* for a given set of graphs if these graphs can be unambiguously encoded using vectors with the components equal to the numbers of occurrences of corresponding fragments in a graph. Note that it is sufficient to establish the basis property of the set $\{F_{k,m}\}$ for a class of graphs only using the IGs belonging to this class (if exist). We considered[9] a number of the IGs reported in the literature and constructed different pairs of IGs or groups including four IGs. In all cases the subgraphs we introduced were found to be the basis subgraphs. In addition, the basis property of these subgraphs was checked taking some classes of chemical graphs (*e.g.*, trees with $n = 11$ and the graphs representing benzenoid hydrocarbons with seven six-membered rings) as examples.

Now let us substantiate the expedience of introducing the descriptors $\{\varphi_{k,m}\}$ instead of $\{X_{k,m}\}$. Notice that the descriptors equal to the numbers of occurrences of particular structural fragments in the structure (the so-called fragmental descriptors, FDs) are usually characterized by low resolving ability. Therefore, unambiguous encoding of structures using the $\{X_{k,m}\}$ descriptors requires a rather large set of subgraphs. For the same reason, in this case good structure—property correlations can usually be obtained using a rather large number (compared to the number of structures in the training set) of the $\{X_{k,m}\}$ descriptors. Earlier,[7] we proved that for any set including N structures and any property there exist N fragments such that they provide an ideal correlation between the property and the numbers of occurrences of these fragments (*i.e.*, a correlation is transformed in a strict equality using the starting set). From this it follows that the problem of constructing an exact model for structure—property relationship can be considered solved. However, the main goal of constructing the correlation equations is to predict the properties of compounds that are not included in the initial set. As applied to chemistry, of particular interest are those correlation equations which depend on a small number of parameters. Because of this, yet another goal is to construct a model with a reduced number of parameters without loss of accuracy.

A possible way of improving the FDs is as follows. Let M be the number of occurrences of a specified fragment in a graph G (*i.e.*, M is the value of a FD on the graph G). Enumerate all these occurrences from 1 to M. Let a weight w be a function of, *e.g.*, the degrees of the vertices of an occurrence in the graph G. Assign the weight w_p ($p = 1, ..., M$) to the pth occurrence. Define a new descriptor I using the relation $I = \Sigma w_p$ (summation is performed over all M occurrences of the fragment in G). Here, the weights w_p are chosen to be different non-integers. By denoting the number of occurrences having the same weight, s_j, as n_j it is possible to represent I in the form $I = \sum_j s_j n_j$.

Consider the FD $X_{2,1} = q$, which equals the number of edges of a graph, as the simplest example. Let us show that the so-called Randić index

$$\chi = \Sigma(v_i v_j)^{-0.5},$$

is constructed on its basis using the procedure mentioned above; here v_i and v_j are the degrees of the vertices i and j of the graph, respectively, and summation is performed over all the (i, j) edges of the graph.[1,2,11] The type of the (i, j) edge is determined by two numbers (v_i, v_j), $v_i \le v_j$ and the edge weight is given by the number $(v_i v_j)^{-0.5}$. Denote the number of the (k, l) edges as n_{kl}. Then, χ can be written in the form

$$\chi = \Sigma n_{kl}(kl)^{-0.5}$$

(summation is performed over all (k, l) pairs, $k \le l$).

The Randić index offers some advantages over the FD q.

First, the resolving ability of the χ index is much higher than that of the FD q. For instance, the q descriptor takes the same value for any series of C_3—C_7 alkane isomers, whereas the Randić index is non-degenerate in this case (twofold degeneration of χ occurs for the C_8 alkanes).[11]

Second, the Randić index is characterized by much higher informativity compared to the FD q. Consider simple graphs representing the carbon frameworks of catacondensed benzenoid hydrocarbons. Denote the number of six-membered rings in a graph as r and the number of the (3, 3) edges lying on the perimeter of the graph as n_{33}'. According to Ref. 12,

$$\chi = an_{33}' + br + c,$$

$$a = (5\cdot 6^{1/2} - 12)/(6\cdot 6^{1/2}),\ b = (12 + 6^{1/2})/(3\cdot 6^{1/2}),$$
$$c = (8\cdot 6^{1/2} - 12)/(3\cdot 6^{1/2}),$$

and, in addition,

$$n_{33} = n_{33}' + r - 1,\ n_{23} = 4r - 4 - 2n_{33}',\ n_{22} = n_{33}' + 6.$$

Using irrationality of the numbers a, b, and c, it can be shown that two such graphs characterized by the same χ value necessarily have equal parameters r and n_{33}'. From this it follows that the number χ provides a means of unambiguous and simultaneous determination of two independent parameters, r and n_{33}', which are then used to determine the quantities n_{22}, n_{33}, and n_{23}. At the same time the descriptor q (here, $q = 5r + 1$) is suitable for the determination of a single parameter, r. Generally, the following four topological indices can be unambiguously reconstructed given the χ value:

$$2n_{22} + n_{24},\ 2n_{13} + n_{34},\ n_{23},\ 6(n_{22} + n_{14}) + 4n_{23} + 3n_{44}.$$

The proof of this fact is also based on irrationality of the Randić index χ.

Third, the structure—property correlations involving the index χ are better than those based on the FD q. Consider the C_3—C_7 alkanes with known boiling points (T_{boil}/°C, the property y) and construct the following two correlations based on these data:

$$y = -129.732 + 68.33\chi,$$
$$\delta_{\max} = 9.5\ ^\circ\text{C},\ s = 4.6\ ^\circ\text{C},\ R = 0.9934,\ F = 1454.5;$$

$$y = -102.352 + 31.885q,$$
$$\delta_{\max} = 15.7\ ^\circ\text{C},\ s = 7.6\ ^\circ\text{C},\ R = 0.9827,\ F = 505.4;$$

where $\delta_{\max}$ is the maximum absolute error for the starting set, s is the root-mean-square deviation, R is the correlation coefficient, and F is the Fisher test. As can be seen, all the characteristics of the model with the Randić index χ are better than those of the model with the FD q.

It should be emphasized that the $\{\varphi_{k,m}\}$ descriptors introduced in this work are a generalization of the index χ and that $\chi = \varphi_{2,1}$. These descriptors are constructed using the subgraphs $\{F_{k,m}\}$ and the procedure described above, so the reasoning similar to that mentioned above for χ is also valid in this case.

Let us discuss the expedience of passage from the $\{\varphi_{k,m}\}$ descriptors to the family of $\{\psi_{k,m}(\alpha), \psi_0(\alpha)\}$ descriptors. Often, TD-based QSPR/QSAR studies involve normalization of the descriptor, *i.e.*, division of the descriptor by some function, $f(n)$, of the number, n, of vertices of a molecular graph (*e.g.*, $f(n) = n^{1/2}$, lg$n$, $(n - 1)(n - 2)^2$, n^2, n).[1,13—15] Experience shows that often this does improve correlations. For instance, the authors of Ref. 15 constructed two kinds (type I and type II) of models for structure—property relationships for six physicochemical properties (boiling point, specific heat, Gibbs free energy, vaporization enthalpy, refractive index, and density) of C_2—C_{10} alkanes

$$y = a_0 + a_1n + a_2I, \qquad \text{(I)}$$

$$y = b_0 + b_1n + b_2I/n. \qquad \text{(II)}$$

Five different descriptors were used as I; then, the statistical characteristics of the models (R, s, F) were compared. A total of thirty equations of each type were obtained. All the three characteristics of the type II models were better than those of the corresponding type I models in twenty-eight out of thirty cases; in two cases the parameters R and s were equal while the parameters F were somewhat better for the type I models.

Thus, the use of descriptor normalization procedure can favor the obtaining of better results. When constructing correlation equations, it is *a priori* unknown which descriptors must be selected and which functions $f(n)$ should be chosen for each descriptor. Because of this we only use $f(n) = n^\alpha$ at $\alpha \geq 0$ (for all descriptors) and choose the best α value for each descriptor. The α values for a particular problem can be chosen, *e.g.*, as follows. Consider a finite set of α values ($\alpha = t/h;\ t = 0, 1, 2, ..., T$, where T is a specified integer and h is a preset natural number). This gives a total of $T + 1$ values of the parameter α at any constant h. Then, the $\psi_{k,m}(\alpha)$ and $\psi_0(\alpha)$ descriptors are constructed using all the $F_{k,m}$ fragments present in the training set at all values of the parameter α. Then, the best descriptors are selected in constructing a correlation equation using a conventional step-wise linear regression. If, for any reason, the correlation thus obtained is unsatisfactory, the procedure can be repeated using an increased T or h value. In the examples given below we used $h = 10$ and $T = 20, 30, 40, 50, 60$ (depending on a particular example), which gave satisfactory results.

Passage from $\{\varphi_{k,m}\}$ to $\{\psi_{k,m}(\alpha)\}$ can be considered as a possible way of extension of the set of the $\{\varphi_{k,m}\}$ descriptors, because $\psi_{k,m}(0) = \varphi_{k,m}$.

Now let us discuss the expedience of inclusion of the descriptor $\psi_0(\alpha) = n^\alpha$ in the $\{\psi_{k,m}(\alpha)\}$ set. It is known that many properties of alkanes are strongly dependent on n.[15] For instance, a correlation obtained for the boiling point of C_3—C_7 alkanes has the form

$$y = an + b,$$
$$\delta_{\max} = 15.7\ ^\circ\text{C},\ s = 7.6\ ^\circ\text{C},\ R = 0.983.$$

At the same time, the TDs can also strongly depend on n. In particular, for the same set and the Wiener index (W) one has

$$W = cn^2 + d,$$
$$\delta_{\max} = 8.6,\ s = 3.4,\ R = 0.978.$$

Therefore, it is quite logical to extend the family of the descriptors constructed by the descriptor $\psi_0(\alpha) = n^\alpha$ and to choose an appropriate α value using the procedure described above.

Examples

In this Section we present the results obtained using the method proposed in this work.

We used a number of databases containing information on the physicochemical properties of various classes of hydrocarbons and on the values of some well-known topological indices. The characteristics considered were the boiling point (T_{boil}/°C), critical temperature (T_{cr}/°C), molar refraction (MR/cm^3 mol^{-1}), enthalpy of formation ($\Delta H^\circ_{298.16}$/kcal mol^{-1}), heat of combustion ($\Delta_{\text{comb}}H^\circ$/kcal mol^{-1}), critical pressure (P_{cr}/atm), molar volume (V_{mol}/cm^3 mol^{-1}) at 20 °C, heat of vaporization (ΔH_{vap}/kJ mol^{-1}), surface tension (γ/dyn cm^{-1}), density (d/kg m^{-3}), enthalpy of formation, melting point (T_{m}/°C), Gibbs free energy (ΔG/kJ mol^{-1}), and the specific heat (C_p/J mol^{-1} K^{-1}). The topological indices employed were the Wiener (W), Hosoya (Z), second-order molecular connectivity ($^2\chi$), and the Kier $^1\kappa$ and $^2\kappa$ molecular shape indices (see a

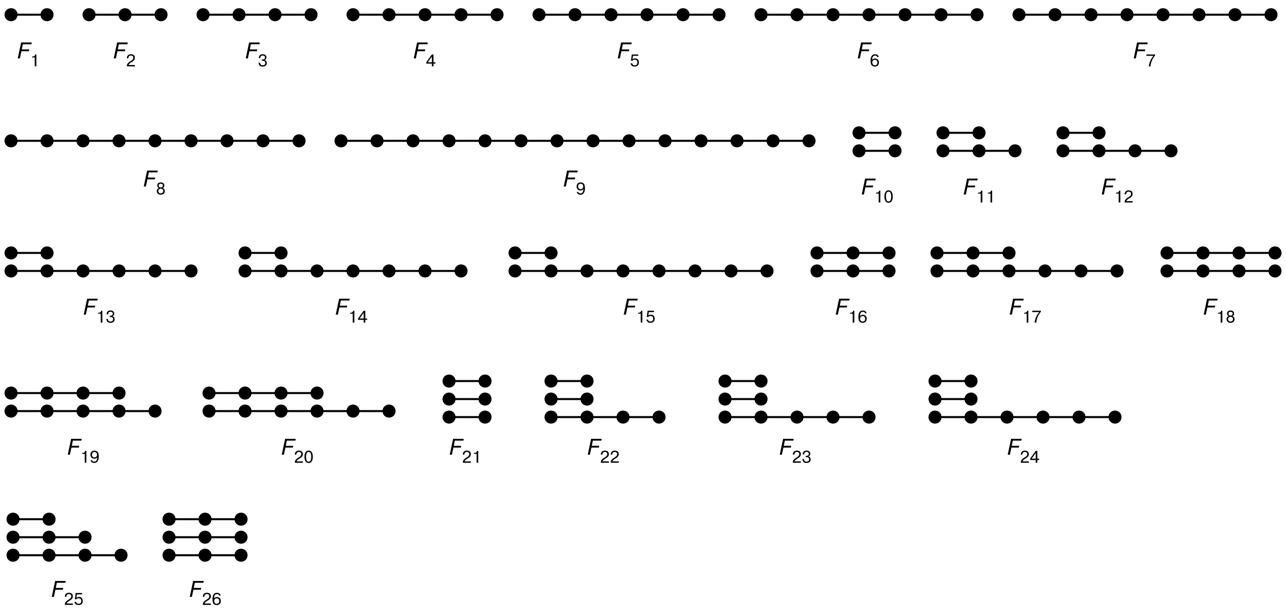

Fig. 2. Basis fragments used in Eqs (1)—(19).

review[1]) and the total π-electron energy E_π (in β units). The corresponding training sets and test sets were obtained by randomly dividing the content of the databases in such a way that the number of structures in the test sets be nearly 10% of the total number of structures in the databases. The QSPR-equations were obtained using the training sets and then used for calculating the corresponding characteristics of the compounds in the test sets. Parameters of the correlation equations given below are as follows: N is the number of structures in the training set; m is the number of the descriptors used in the equation; R, F, and s are the correlation coefficient, Fisher test, and standard deviation for the training set, respectively; N_p is the number of compounds in the test set; and *prms* and R_p are the root-mean-square error and the correlation coefficient for the test set, respectively. For simplicity, the basis descriptors appeared in the correlation equations and the corresponding fragments shown in Figs 1—5 are enumerated 1, 2, ..., *etc.*

The fragments corresponding to the descriptors used in Eqs (1)—(19) are shown in Fig. 2.

Alkanes C_2—C_9. The correlation equations obtained for a number of sets of C_2—C_9 alkanes with known physicochemical parameters[16] and the Randić, Wiener, Hosoya, and Kier[1] topological indices are as follows:

$$T_{\text{boil}} = -150.5 + 63.1\psi_1(0) + 75.5\psi_2(1.3) + $$
$$+ 1527.6\psi_3(2.9) + 8606.4\psi_7(3.7) - 0.62\psi_{13}(0), \quad (1)$$

$N = 67$, $m = 5$, $R = 0.998$, $F = 3721$, $s = 2.8$, $N_p = 7$, $prms = 3.6$, $R_p = 0.995$;

$$P_{\text{cr}} = 12.25 + 125.6\psi_1(1.8) + 133.81\psi_2(2.8) + $$
$$+ 241.88\psi_3(2.7) - 2.11\psi_6(0.5) - 989.33\psi_{10}(4.6), \quad (2)$$

$N = 67$, $m = 5$, $R = 0.99$, $F = 551$, $s = 0.67$, $N_p = 7$, $prms = 0.68$, $R_p = 0.97$;

$$T_{\text{cr}} = 189.01 + 71.9\psi_1(0) - 436.2\psi_1(1) + $$
$$+ 1685.8\psi_3(2.5) + 12558\psi_4(4.2), \quad (3)$$

$N = 67$, $m = 4$, $R = 0.995$, $F = 1449$, $s = 6.2$, $N_p = 7$, $prms = 6.2$, $R_p = 0.989$;

$$MR = 27.18 + 16.47\psi_3(1.4) + 47.69\psi_4(2.8) + $$
$$+ 0.77\psi_8(0) - 3299.6\psi_{10}(4.9) + 0.7\psi_{11}(0.3) + $$
$$+ 4907.4\psi_{21}(5) - 51.12\psi_{26}(2.9), \quad (4)$$

$N = 62$, $m = 7$, $R = 0.996$, $F = 1029$, $s = 0.47$, $N_p = 7$, $prms = 0.61$, $R_p = 0.995$;

$$V_{\text{mol}} = 143.261 - 7720.5\psi_1(4.1) - 197.5\psi_3(2.5) - $$
$$- 6417.2\psi_4(4.2) + 3.173\psi_{11}(0.5) + 13248\psi_{21}(5), \quad (5)$$

$N = 62$, $m = 5$, $R = 0.997$, $F = 2070$, $s = 1.28$, $N_p = 7$, $prms = 1.56$, $R_p = 0.997$;

$$\Delta H_{\text{vap}} = 2.87 + 9.42\psi_1(0) + 32.35\psi_3(2.5) - $$
$$- 27.8\psi_4(2.3) - 0.53\psi_6(0.3) + 1070.2\psi_7(3.7) + $$
$$+ 78.78\psi_{11}(3.8) + 3.32\psi_{12}(3.1) + 689.7\psi_{14}(4.2) + $$
$$+ 0.003\psi_{23}(1), \quad (6)$$

$N = 62$, $m = 9$, $R = 0.998$, $F = 1712$, $s = 0.33$, $N_p = 7$, $prms = 0.41$, $R_p = 0.997$;

$$\gamma = 24.98 - 31.13\psi_1(1.2) - 0.73\psi_5(0) + $$
$$+ 353.95\psi_{11}(3.9) + 3.39\psi_{12}(1.3) + 1219.5\psi_{12}(4.5) + $$
$$+ 7.27\psi_{14}(1.4) + 0.48\psi_{17}(0.1) - 4.66\psi_{23}(1.9), \quad (7)$$

$N = 61$, $m = 8$, $R = 0.984$, $F = 197$, $s = 0.36$, $N_p = 7$, $prms = 0.45$, $R_p = 0.98$;

$$^2\chi = 0.087 + 0.94\psi_2(0.1) - 3\psi_3(1.8), \tag{8}$$

$N = 67$, $m = 2$, $R = 0.9988$, $F = 14480$, $s = 0.043$, $N_p = 7$, $prms = 0.03$, $R_p = 0.9993$;

$$W = -1.736 + 40.59\psi_2(1.4) - 3965.1\psi_4(4) + 3.74\psi_6(0) + 2.567\psi_7(0) + 4.03\psi_{10}(0) + 1243.7\psi_{16}(3.4), \tag{9}$$

$N = 67$, $m = 6$, $R = 0.9988$, $F = 4047$, $s = 1.563$, $N_p = 7$, $prms = 1.311$, $R_p = 0.9989$;

$$\ln Z = -0.502 + 1.296\psi_1(1) - 0.8\psi_2(2.3) - 86.55\psi_4(4.3) - 119.72\psi_{16}(4.9), \tag{10}$$

$N = 67$, $m = 4$, $R = 0.9997$, $F = 36720$, $s = 0.015$, $N_p = 7$, $prms = 0.022$, $R_p = 0.9993$;

$$^1\kappa = 8.806 + 2.03\psi_1(0) - 17.48\psi_1(1) - 10.81\psi_2(2.5) + 0.0055\psi_{16}(0), \tag{11}$$

$N = 67$, $m = 4$, $R = 0.9996$, $F = 21865$, $s = 0.044$, $N_p = 7$, $prms = 0.02$, $R_p = 0.9998$;

$$^2\kappa = -20.12 + 43.52\psi_1(0.9) - 207.03\psi_1(6) - 106.54\psi_3(3.1) + 2280.8\psi_7(3.7) + 655.9\psi_{16}(4.9) - 0.37\psi_{21}(1.7) + 0.05\psi_{23}(0.5), \tag{12}$$

$N = 67$, $m = 7$, $R = 0.9785$, $F = 190$, $s = 0.32$, $N_p = 7$, $prms = 0.38$, $R_p = 0.956$.

Alkanes C_6—C_{10}. The following correlation equations were obtained for sets of C_6—C_{10} alkanes with known physicochemical characteristics[17] :

$$d_4^{25} = 639.33 + 19.35\psi_3(0) - 2810.8\psi_{11}(3.8) + 41156\psi_{13}(4.9) + 3458.8\psi_{16}(3.5) - 0.68\psi_{18}(0), \tag{13}$$

where d_4^{25} is the density (in kg m^{-3}) at 25 °C, $N = 121$, $m = 5$, $R = 0.943$, $F = 184$, $s = 9.38$, $N_p = 13$, $prms = 6.63$, and $R_p = 0.97$;

$$\Delta H_f = 2.73 + 10.41\psi_1(1) + 1.91\psi_2(0.7) - 20.88\psi_3(1.8) + 2207.9\psi_7(3.8) + 2799.4\psi_{14}(4), \tag{14}$$

where ΔH_f is the enthalpy of formation (in kJ mol^{-1}) at 300 K, $N = 121$, $m = 5$, $R = 0.987$, $F = 897$, $s = 0.69$, $N_p = 13$, $prms = 0.65$, and $R_p = 0.991$;

$$\Delta G = -45.9 + 22.9\psi_2(0.8) + 8.93\psi_3(0) + 62.2\psi_4(1.1) + 18254\psi_7(3.8) - 1.66\psi_{18}(0.4), \tag{15}$$

where ΔG is the Gibbs free energy (in kJ mol^{-1}) at 300 K, $N = 121$, $m = 5$, $R = 0.969$, $F = 359$, $s = 3.7$, $N_p = 13$, $prms = 2.6$, and $R_p = 0.981$;

$$C_p = 263.9 - 7111.2\psi_1(2.8) - 4612\psi_6(3.4) + 0.74\psi_{13}(0) + 0.61\psi_{19}(0) + 1.13\psi_{24}(0.2), \tag{16}$$

where C_p is the specific heat (in J mol^{-1} K^{-1}) at 300 K, $N = 121$, $m = 5$, $R = 0.983$, $F = 653$, $s = 4.68$, $N_p = 13$, $prms = 4.77$, and $R_p = 0.987$;

$$n_D^{25} = 1.34 + 0.0091\psi_3(0) + 0.45\psi_{11}(2.9) + 11.75\psi_{13}(4.9) + 0.067\psi_{15}(1.5) - 0.0002\psi_{18}(0) + 0.00006\psi_{19}(0) + 0.016\psi_{20}(1) - 0.00063\psi_{24}(0.2), \tag{17}$$

where n_D^{25} is the refractive index at 25 °C, $N = 120$, $m = 8$, $R = 0.987$, $F = 505$, $s = 0.0021$, $N_p = 13$, $prms = 0.0022$, and $R_p = 0.989$.

***n*-Alkanes.** A correlation equation obtained for a set of fifty-four *n*-alkanes with known T_m values[18,19] has the form

$$T_m = -184.16 + 3403.7\psi_3(4.6) + 9065.2\psi_9(2.9) + 403.64\psi_{10}(1.9) - 22636\psi_{11}(5.1), \tag{18}$$

$N = 48$, $m = 4$, $R = 0.9995$, $F = 10145$, $s = 2.82$, $N_p = 6$, $prms = 2.12$, $R_p = 0.9982$.

Note. We also carried out a QSPR study of a set comprising fifty-six branched alkanes.[18,19] Two cases considered were (a) $N = 49$ and $N_p = 7$ and (b) $N = 56$ and $N_p = 0$ (no test set). The model for the case (a) was characterized by $N = 49$, $m = 38$, $R = 0.9778$, $F = 5.739$, $s = 16.5$, $N_p = 7$, $prms = 14.6$, and $R_p = 0.7601$. The model for the case (b) was characterized by the parameters $N = 56$, $m = 41$, $R = 0.973$, $F = 6.242$, and $s = 15.6$.

Notice that attempts to construct adequate linear models for the T_m of the branched alkanes with relatively small number of descriptors have failed. This confirms the results obtained earlier.[20]

Benzenoid hydrocarbons. The correlation equation for a set of 118 benzenoid hydrocarbons with seven six-membered rings and characterized by known E_π values obtained from EHT calculations[21] has the form

$$E_\pi = -29.026 + 0.0003\psi_{20}(1.3) - 2.91\psi_{22}(2.7) - 60.409\psi_{25}(4.3), \tag{19}$$

$N = 107$, $m = 3$, $R = 0.972$, $F = 577$, $s = 0.032$, $N_p = 11$, $prms = 0.031$, and $R_p = 0.951$.

Alkylbenzenes. A set of forty-eight alkylbenzenes with known values of some physicochemical characteristics[8,22] is described by the following correlation equations (the fragments corresponding to the descriptors appeared in Eqs (20)—(24) are shown in Fig. 3):

$$\Delta H^\circ_{298.16} = 350.3 - 214.64\psi_1(0.8) + 149.67\psi_3(0.9) - 7840.7\psi_6(2.4) - 164.6\psi_7(0.9) + 82945\psi_{10}(4.6) - 27649\psi_{12}(3.6), \tag{20}$$

where $\Delta H^\circ_{298.16}$ is the enthalpy of formation of gaseous alkylbenzenes from atoms (in kcal mol^{-1}) at 25 °C, $N = 43$, $m = 6$,

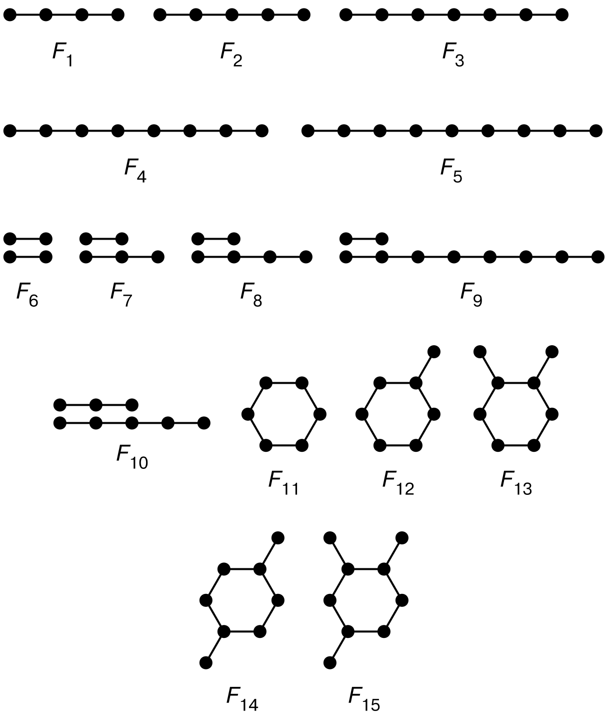

Fig. 3. Basis fragments used in Eqs (20)—(24).

R = 0.99997, F = 87832, s = 7.96, *av.dist* = 5.42, N_p = 5, *prms* = 4.1, and R_p = 0.99998;

$$\Delta H°_{C_nH_{2n-6}} = 32.91 - 1678.1\psi_4(3.8) - 9.54\psi_7(1.2) + \\ + 2827.6\psi_{10}(4.3) - 1.35\psi_{12}(0) + 1058.9\psi_{14}(3.7), \quad (21)$$

where $\Delta H°_{C_nH_{2n-6}}$ is the enthalpy of formation of gaseous alkylbenzenes from elements (in kcal mol^{-1}) at 25 °C, N = 43, m = 5, R = 0.9998, F = 15233, s = 0.43, N_p = 5, *prms* = 0.38, and R_p = 0.9994;

$$\Delta_{comb}H° = 10.1 - 4628.6\psi_6(2.4) - 110\psi_7(0.9) + \\ + 64450\psi_{10}(4.6) - 56063\psi_{13}(4), \quad (22)$$

where $\Delta_{comb}H°$ is the heat of combustion of gaseous alkylbenzenes (in kcal mol^{-1}) at 25 °C, N = 43, m = 4, R = 0.9998, F = 46783, s = 8.95, N_p = 5, *prms* = 8.04, and R_p = 0.9997;

$$\Delta_{comb}H°_{298.16} = - 486.89 - 182.13\psi_2(1.5) - \\ - 37796\psi_5(3.8) - 42.71\psi_7(0.6) - 39933\psi_8(4) - \\ - 3.97\psi_9(0) + 21855\psi_{11}(2.6) + 1.89\psi_{15}(0), \quad (23)$$

where $\Delta_{comb}H°_{298.16}$ is the heat of combustion of liquid alkylbenzenes (in kcal mol^{-1}) at 25 °C, N = 32, m = 7, R = 0.9997, F = 5633, s = 4.48, N_p = 4, *prms* = 3.26, and R_p = 0.9988;

$$MR = 13.56 + 302.12\psi_4(3) + 1443.4\psi_5(3.8) + \\ + 1.8\psi_7(0.7) + 1413.8\psi_8(4), \quad (24)$$

N = 32, m = 4, R = 0.9987, F = 2647, s = 0.26, N_p = 3, *prms* = 0.24, and R_p = 0.9943.

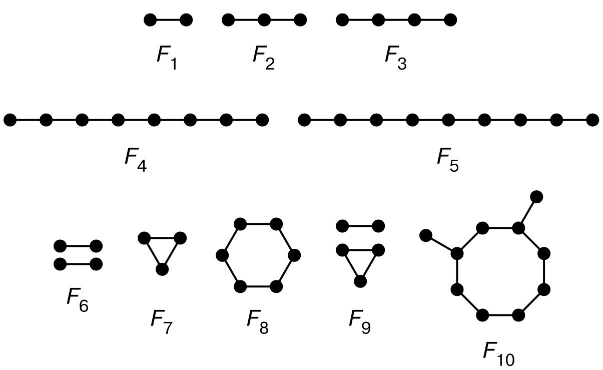

Fig. 4. Basis fragments used in Eqs (25) and (26).

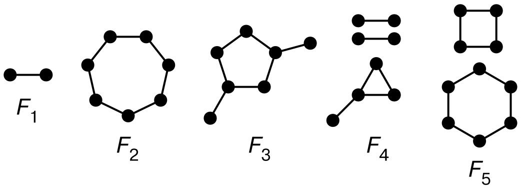

Fig. 5. Basis fragments used in Eq. (27).

Mixed sets of hydrocarbons. The following correlation equations were obtained for a set of hydrocarbons with known enthalpies of formation[23] (fragments used in Eqs (25) and (26) are shown in Fig. 4 and those used in Eq. (27) are presented in Fig. 5):

$$\Delta H°_{f,g} = -78.32 + 3698.8\psi_1(6) + 44223\psi_2(6) + \\ + 1302.1\psi_3(2.5) + 111770\psi_4(5.2) + 7.61\psi_9(0) - \\ - 2418.6\psi_{10}(3.5), \quad (25)$$

where $\Delta H°_{f,g}$ is the enthalpy of formation of gaseous hydrocarbons (in kJ mol^{-1}) at 298.15 K, N = 173, m = 6, R = 0.968, F = 417, s = 7.18, N_p = 15, *prms* = 7.68, and R_p = 0.94;

$$\Delta H°_{f,l} = -260.12 + 651.8\psi_1(1.4) + 168\psi_2(1.3) + \\ + 175.96\psi_5(2.2) + 478.6\psi_6(2.8) + 104.45\psi_7(0.7) - \\ - 15850\psi_8(3.5), \quad (26)$$

where $\Delta H°_{f,l}$ is the enthalpy of formation of liquid hydrocarbons (in kJ mol^{-1}) at 298.15 K, N = 85, m = 6, R = 0.982, F = 344, s = 4.0, N_p = 7, *prms* = 3.1, and R_p = 0.958.

The correlation equation obtained for a set of ring-containing hydrocarbons (a total of 381 compounds) with known T_{boil} values[24] has the form:

$$T_{boil} = -145.04 + 106\psi_1(0.2) + 58.5\psi_2(0.7) - \\ - 39.3\psi_3(1.9) - 0.4\psi_4(0) + 195.6\psi_5(0.7), \quad (27)$$

N = 353, m = 5, R = 0.987, F = 2684, s = 6.3, N_p = 28, *prms* = 5.1, and R_p = 0.989.

Thus, in this work we proposed a new method of solving the problem of selection of a relatively small, finite set of molecular topological descriptors for constructing structure—property correlations. The method is based on the concept of basis property of a family of parameter-dependent descriptors. This approach offers the advantages of (i) algorithmization and subsequent full automation and (ii) the use of parameter-dependent descriptors, which extends the field of possible applications. Besides, the basis descriptors introduced in this work permit a uniform description of the structure—property relationships for various properties of different classes of hydrocarbons; the models obtained are characterized by high predictive power. The examples presented above demonstrate that the method elaborated is of great practical value.

This work was financially supported by the INTAS (Grant 00-0363).

References

1. M. I. Stankevich, I. V. Stankevich, and N. S. Zefirov, *Usp. Khim.*, 1988, **57**, 337 [*Russ. Chem. Rev.*, 1988, **57**, 191 (Engl. Transl.)].
2. O. A. Raevskii, *Usp. Khim.*, 1999, **68**, 555 [*Russ. Chem. Rev.*, 1999, **68**, 505 (Engl. Transl.)].
3. M. Randić, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 686.
4. M. Randić, *J. Chem. Educ.*, 1992, **69**, 713.
5. M. Randić, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 57.
6. A. K. Dewdney, *Aequat. Math.*, 1970, **4**, 326.
7. I. I. Baskin, M. I. Skvortsova, I. V. Stankevich, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 527.
8. M. Randić, *New J. Chem.*, 1991, **15**, 517.
9. M. I. Skvortsova, K. S. Fedyaev, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 2002, **382**, 645 [*Dokl. Chem.*, 2002 (Engl. Transl.)].
10. D. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs. Theory and Applications*, Academic Press, New York, 1980.
11. M. Randić, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
12. I. V. Stankevich, M. I. Skvortsova, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1992, **324**, 133 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
13. S. S. Tratch, M. I. Stankevitch, and N. S. Zefirov, *J. Comput. Chem.*, 1990, **11**, 899.
14. A. T. Balaban, D. Ciubotariu, and M. Medeleanu, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 517.
15. O. Ivanciuc, *Rev. Roum. Chim.*, 2001, **46**, 129.
16. D. E. Needham, I. C. Wei, and R. G. Seybold, *J. Am. Chem. Soc.*, 1988, **110**, 4186.
17. A. A. Gakh, E. G. Gakh, B. G. Sumpter, and D. W. Noid, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 832.
18. R. D. Obolentsev, *Fizicheskie konstanty uglevodorodov* [*Physical Constants of Hydrocarbons*], State Publ. House for Literature on Petroleum, Mining, and Fuel Industry, Moscow—Leningrad, 1953 (in Russian).
19. *Fiziko-khimicheskie svoistva individual´nykh uglevodorodov* [*Physicochemical Properties of Individual Hydrocarbons*], Ed. V. M. Tatevskii, State Publ. House for Literature on Petroleum, Mining, and Fuel Industry, Moscow, 1960, 412 pp. (in Russian).
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. J. D. Roberts, *Notes on Molecular Orbital Calculations*, Benjamin, New York, 1962.
22. V. M. Tatevskii, *Khimicheskoe stroenie uglevodorodov and zakonomernosti v ikh fiziko-khimicheskikh svoistvakh* [*The Chemical Structure of Hydrocarbons and Trends of Their Physicochemical Properties*], Izd. Moscow State Univ., Moscow, 1953, 320 pp. (in Russian).
23. E. S. Domalski and E. D. Hearing, *J. Phys. Chem. Ref. Data*, 1988, **17**, 1637.
24. G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 788.

Received July 1, 2003;
in revised form January 5, 2004