J Math Chem (2009) 45:1004–1020
DOI 10.1007/s10910-008-9393-4

ORIGINAL PAPER

Reduced matrix of topological distances with a minimum number of independent parameters: distance vectors and molecular codes

E. A. Smolenskii · E. V. Shuvalova · L. K. Maslova · I. V. Chuvaeva · M. S. Molchanova

Received: 28 January 2008 / Accepted: 3 March 2008 / Published online: 13 June 2008

Abstract Entries of the topological distance matrix are shown to be functions of entries of the reduced distance matrix, which has a smaller size. The latter entries are expressed through a minimum number m of independent parameters ($m \leq 2n-3$). The expanded distance matrices, whose sum constitutes the reduced matrix, are defined. The reduced vectors have lower degeneracy than the corresponding vectors proposed by Randić as molecular codes. Nondegenerate sets of reduced-matrix entries are proposed as molecular codes.

Keywords Topological distances · Tree graphs · Reduced distance matrix · Molecular coding · Nondegenerate vectors

1 Introduction

From the mathematical point of view, structural formulas of organic compounds are multigraphs, usually termed molecular graphs. Their vertices represent atoms, and edges represent chemical bonds.

One of the best known characteristics of a molecular graph (**G**) is its distance matrix $\mathbf{D}(\mathbf{G}) = ||d_{ij}||$, where each integer number d_{ij} is the number of graph edges between vertices i and j. A multitude of topological indices are based on matrix $\mathbf{D}(\mathbf{G})$ [1–3]. The best known among them is the Wiener index W, which was proposed in 1947 [4,5] as a structural parameter for describing the dependences of physicochemical properties of alkanes on their structure. By definition, W is the half-sum of all entries d_{ij} of matrix $\mathbf{D}(\mathbf{G})$ [6]:

E. A. Smolenskii (✉) · E. V. Shuvalova · L. K. Maslova · I. V. Chuvaeva · M. S. Molchanova
Zelinskii Institute of Organic Chemistry, Russian Academy of Sciences, Leninskii pr. 47, Moscow 119991, Russia
e-mail: smolensk@ioc.ac.ru

$$W = \frac{1}{2}\sum_{(i,j)} d_{ij}.$$

The value of W is associated with alkane branching: it is maximum for linear alkanes and minimum for the most branched ones. In [4,5], Wiener proposed another index p_3, which is equal to the half-number of all "threes" $d_{ij} = 3$ in $\mathbf{D}(\mathbf{G})$:

$$p_3 = \frac{1}{2}\sum_{d_{ij}=3} \frac{1}{3} d_{ij}.$$

Index f suggested by Platt in his key paper [7] is the half-sum of the numbers of all "twos" in matrix $\mathbf{D}(\mathbf{G})$:

$$f = \frac{1}{2}\sum_{d_{ij}=2} \frac{1}{2} d_{ij}.$$

Note that Platt was the first to use the term *index* in this context. Moreover, the same paper by Platt logically supplements Wiener's idea of using p_i values as independent variables (Wiener considered the sum of p_i values with odd i: $k_1 p_1 + k_3 p_3 + k_5 p_5 \ldots$). However, neither Wiener nor Platt knew at that time that their parameters were related to graph theory and would later be termed topological indices [6,8,9].

Graph theory was first explicitly applied to establishment of relationships between properties of chemical compounds and their molecular structures in [10–13]. Algebraic operations with adjacency matrices aimed at calculating various structural parameters are reviewed in [10]. In [11,12], the physicochemical properties of alkanes are introduced as functions of entries of another matrix, which was earlier proposed by the same author for linear coding of chemical graph structures [14]. This matrix is the matrix of distances only between vertices of degree 1. We will refer to it as the *reduced distance matrix*; evidently, its size is usually much smaller than that of the initial distance matrix. Later, in [15], it was proved that there are many linear dependences between entries of the reduced matrix: among its $C_n^2 = \frac{n(n-1)}{2}$ entries, no more than $(2n - 3)$ are linearly independent.

In this study, we analyze the structure of the reduced matrix and show how one can express it using a minimum number of independent parameters.

Despite its wide use in various applications, the distance matrix is redundant [16], i.e., most of its entries are dependent on others and do not carry any useful information. The latter fact is important for consideration of topological indices based on this matrix. Obviously, since the overwhelming part of the matrix entries are "information garbage," its actual information content is low and search of structure–property regularities on this basis is somewhat hindered. As an especially evident example confirming the latter assertion, one can take the ordered sequence of p_i values proposed by Wiener and Platt. Randić [17] presented them in the form of vector (p_1, p_2, … p_k) and suggested its use as the "molecular code." However, even Randić himself mentioned examples of its degeneration for isomers starting from C_9H_{20}. Naturally, the degeneration becomes increasingly significant for heavier alkanes.

We propose the term *distance vector* for the vector constructed on the basis of the distance matrix. The vector constructed in the same way from the reduced distance matrix can be named the *reduced distance vector*. Below, we prove that the reduced distance vector has a lesser degree of degeneration despite its smaller length. Finally, using the reduced distance matrix, one can build a fairly simple set of integers (for example, the top row of this matrix together with the maximum nonzero diagonal) that is completely nondegenerate. It can be regarded as the molecular code [15].

2 Structure and properties of reduced distance matrix

For simplicity, here we will consider alkanes (tree graphs), because generalization of the problem for arbitrary graphs is obvious. All N vertices of an alkane will be numbered in the following way: numbers from 1 to n_1 correspond to first-degree vertices (methyl groups in an alkane), and other vertices are numbered from $n_1 + 1$ to N. Then the distance matrix is

$$\mathbf{D}(\mathbf{G}) = \left|\begin{array}{cccc|ccc} 0 & d_{12} & \ldots & d_{1,n1} & d_{1,n1+1} & \ldots & d_{1N} \\ d_{21} & 0 & \ldots & d_{2,n1} & d_{2,n1+1} & \ldots & d_{2N} \\ \ldots & \ldots & 0 & \ldots & \ldots & \ldots & \ldots \\ d_{n1,1} & d_{n1,2} & \ldots & 0 & d_{n1,n1+1} & \ldots & d_{n1,N} \\ \hline d_{n1+1,1} & d_{n1+1,2} & \ldots & d_{n1+1,n1} & 0 & \ldots & d_{n1+1,N} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{N,1} & d_{N,2} & \ldots & d_{N,n1} & d_{N,n1+1} & \ldots & 0 \end{array}\right| . \quad (1)$$

Since $N = n_1 + n_2 + n_3 + n_4$ (here n_1, n_2, n_3, and n_4 are the numbers of primary, secondary, tertiary and quaternary C atoms, respectively), we usually have $n_1 \ll N$ (only for ethane $n_1 = N = 2$). It was shown in [15] that all entries of matrix (1) not included in the selected top-left-corner matrix are dependent and do not carry any actual information on the structure of tree **G**. The selected part of matrix **D**(**G**) is hereafter termed the reduced distance matrix $\mathbf{D_0}(\mathbf{G})$:

$$\mathbf{D}_0(\mathbf{G}) = \begin{vmatrix} 0 & r_{12} & \ldots & r_{1n} \\ r_{21} & 0 & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & 0 \end{vmatrix}, \quad (2)$$

where $n = n_1$, $r_{ij} = d_{ij}$. In [14], matrix (2) in the form of ordered sequence

$$\mathbf{R}(\mathbf{G}) = (r_{12}, r_{13}, \ldots, r_{1n}; r_{23}, \ldots, r_{2n}; \ldots; r_{n-1,n}) \quad (3)$$

was proposed for linear coding of graphs. It has been shown that one can unambiguously restore tree **G** from sequence (3), and relationship to symmetry group S_3 has been pointed out. However, set (3) has too many entries, i.e., $C_n^2 = \frac{n(n-1)}{2}$, and therefore is not of practical interest for linear coding either.

As is mentioned above, matrix $\mathbf{D_0(G)}$ contains no more than $(2n-3)$ independent entries [15], i.e., one can skip at least $\frac{n^2}{2}-\frac{3n}{2}+3$ entries in sequence (3), whereas the remaining $(2n-3)$ entries can be used in practical coding of chemical structures. As the basis in (3), one can take entries of the diagonal adjacent to the main (zero) one together the first row or last column of the matrix, etc.

If m is the minimum number of parameters that can describe the topological structure of the tree, we have

$$m \leq 2n-3.$$

One can define a *branch* as the set of vertices starting from a first-degree vertex $i\,(i \leq n_1)$ and extending to the nearest vertex of degree 3 or 4 [16]. For alkanes, it is a chain of the form $CH_3-(CH_2)_{k1}$ ($k_1=0, 1, 2, \ldots$). Denoting the branch length by ρ_i, we obtain

$$\rho_i = 1 + k_i.$$

Similarly, the length of any segment between vertices of degree 3 or 4, denoted ρ_{ij} ($n_1 < i, j \leq N$), is the number of vertices of degree 2 (CH_2 groups) between them plus one. Obviously, the set of all lengths ρ_i and ρ_{ij} is a combination of independent parameters from which the structure of tree **G** can be restored, and the number of such parameters is just m. But here it is essential to take into account the order of these parameters as well, not just their set.

Let us determine m. As we have already found, the number of CH_2 groups (vertices of degree 2) affects only the values of parameters ρ_i and ρ_{ij} but not their number. Therefore, one can take $n_2 = 0$ and thus arrive at an obvious relationship

$$n_1 + n_3 + n_4 = N, \tag{4}$$

i.e., assume that the alkane tree in question contains only vertices of degrees 1, 3, and 4. Thus, the number of its edges is $N-1$. On the other hand, we have a well-known relationship $n_1 + 2n_2 + 3n_3 + 4n_4 = 2(N-1)$. So for $n_2 = 0$ we obtain

$$\frac{n_1}{2} + \frac{3}{2}n + 2n_4 = N - 1. \tag{5}$$

Obviously, $m = N-1$, and replacement of N by $m+1$ in relationship (4) yields

$$n_3 = m + 1 - n_1 - n_4.$$

Substitution into (5) yields

$$N - 1 = m = \frac{n_1}{2} + \frac{3}{2}(m + 1 - n_1 - n_4) + 2n_4 = \frac{3}{2}m + \frac{3}{2} - n_1 + \frac{1}{2}n_4,$$

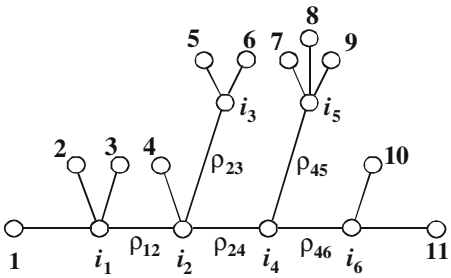
whence we arrive at $\frac{3}{2}m - m = n_1 - \frac{1}{2}n_4 - \frac{3}{2}$, i.e., $m = 2n_1 - 3 - n_4$.

Substituting $n_1 = n$, we finally obtain

$$m = 2n - 3 - n_4. \tag{6}$$

So, addition of each quaternary atom C reduces the number of independent parameters by unity and this number assumes the highest value for structures free of quaternary atoms C.

It is easy to express all entries of reduced distance matrix (2) in terms of ρ_i and ρ_{ij}. Let us consider the following graph as an example:



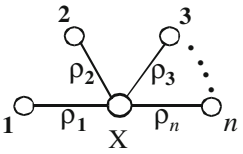
Since $n = 11$, the size of matrix $\mathbf{D_0}$ is 11×11:

$$\mathbf{D}_0(\mathbf{A}_1) = \begin{vmatrix} 0 & \rho_1+\rho_2 & \rho_1+\rho_3 & \rho_1+\rho_{12}+\rho_4 & \rho'_{15} & \dots & \rho'_{1,11} \\ \rho_2+\rho_1 & 0 & \rho_2+\rho_3 & \rho_2+\rho_{12}+\rho_4 & \rho'_{25} & \dots & \rho'_{1,11} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \rho'_{11,1} & \rho'_{11,2} & \dots & \dots & \dots & \dots & 0 \end{vmatrix}, \tag{7}$$

where, for simplicity, we have introduced the following designations:

$\rho'_{15} = \rho_1 + \rho_{12} + \rho_{23} + \rho_5$,
$\rho'_{25} = \rho_2 + \rho_{12} + \rho_{23} + \rho_5$,
$\rho'_{1,11} = \rho'_{11,1} = \rho_1 + \rho_{12} + \rho_{24} + \rho_{46} + \rho_{11}$, and
$\rho'_{11,2} = \rho'_{2,11} = \rho_{11} + \rho_{12} + \rho_{24} + \rho_{46} + \rho_2$.

This matrix is very cumbersome; therefore, let us represent it as a sum of simpler ones, making the meaning of separate parameters in Eq. 7 more obvious. To begin with, let us consider the following graph:



$\mathbf{A}_2$

We can construct a diagonal matrix $\mathbf{R}_n$ for graph $\mathbf{A_2}$ (so-called star graph in graph theory):

$$\mathbf{R}_n = \begin{pmatrix} \rho_1 & \dots & 0 \\ \dots & \rho_2 & \dots \\ 0 & \dots & \rho_n \end{pmatrix}. \tag{8}$$

All its entries except for those in the main diagonal are zeros, whereas the entries in the main diagonal are the branch lengths ρ_i. Also let us define the *topological matrix* $\boldsymbol{\tau}_n^{\mathbf{0}}$ of order $n \times n$:

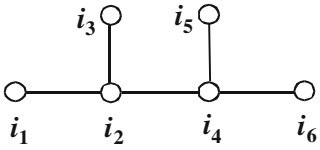
$$\boldsymbol{\tau}_n^0 = \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 0 \end{pmatrix}, \tag{9}$$

where all entries, except for the zero main diagonal, are equal to unity. Using matrices (8) and (9), we can express the reduced matrix of graph $\mathbf{A_2}$ as

$$\mathbf{D_0}(\mathbf{A_2}) = \boldsymbol{\tau}_n^0 \mathbf{R}_n + (\boldsymbol{\tau}_n^0 \mathbf{R}_n)^{\mathrm{T}}, \tag{10}$$

where symbol T means transposition.

Let us assume that vertex X of graph $\mathbf{A_2}$ has an internal structure; e.g., for graph $\mathbf{A_1}$ it is the graph obtained after truncation of all its eleven branches [18]:

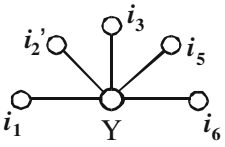


$\mathbf{A_3}$

In the case of $n = 11$, Eq. 10 yields

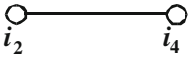
$$\mathbf{D_0}(\mathbf{A_2}) = \boldsymbol{\tau}_{11}^0 \mathbf{R}_{11} + (\boldsymbol{\tau}_{11}^0 \mathbf{R}_{11})^{\mathrm{T}}. \tag{11}$$

In turn, graph $\mathbf{A_3}$ can be represented as



$\mathbf{A_4}$

In this case, vertex Y of the "truncated" graph $\mathbf{A_4}$ is graph



$\mathbf{A_5}$

All vertices from which branches 1–11 grow in graph $\mathbf{A_1}$ are now joined to Y in graph $\mathbf{A_4}$. That is why vertex i_2' from which branch 4 grows is considered as a separate one but the distance between i_2' and Y is assumed to be zero. This vertex is as if virtual, and it enables us to make allowance for the fact that the branch extending from the real vertex i_2 was truncated earlier. That is why we have introduced this zero distance. Thus, matrix $\mathbf{R_5}$ for $\mathbf{A_4}$ has the form

$$\mathbf{R_5} = \begin{pmatrix} \rho_{12} & & & & \\ & 0 & & & \\ & & \rho_{23} & & \\ & & & \rho_{45} & \\ & & & & \rho_{16} \end{pmatrix},$$

and the reduced matrix of size 5×5 is

$$\mathbf{D_0}(\mathbf{A_4}) = \boldsymbol{\tau}_5^0 \mathbf{R_5} + (\boldsymbol{\tau}_5^0 \mathbf{R_5})^{\mathrm{T}}. \tag{12}$$

For graph $\mathbf{A_5}$ we have

$$\mathbf{D_0}(\mathbf{A_5}) = \begin{pmatrix} 0 & \rho_{24} \\ \rho_{24} & 0 \end{pmatrix}. \tag{13}$$

So, matrices (7) and (11) are of size 11×11, matrix (12) is of size 5×5, and matrix (13) is of size 2×2. They seem incommensurable, and therefore their comparison (not to speak of summation) seems impossible. In this regard, we introduce the following construction. Let us use the designation $[i_k]$ for the number of branches extending from vertex i_k, and then entries (k, j) in matrix (12), which are equal to the distances between vertices i_k and i_j, (sums of the corresponding distances ρ_{ij}), will be replaced by constant matrices of size $[i_k] \times [i_j]$. After that, we obtain a square matrix of size $\left\{\sum_k [i_k]\right\} \times \left\{\sum_j [i_j]\right\}$:

$$\overline{\mathbf{D}}_{\mathbf{0}} = \begin{vmatrix} [i_1 \times i_1] & [i_1 \times i_2] & [i_1 \times i_3] & \cdots & [i_1 \times i_4] & [i_1 \times i_5] & [i_1 \times i_6] \\ [i_2 \times i_1] & [i_2 \times i_2] & [i_2 \times i_3] & & [i_2 \times i_4] & [i_2 \times i_5] & [i_2 \times i_6] \\ \cdots & \cdots & & \cdots & & \cdots & \cdots \\ [i_6 \times i_1] & [i_6 \times i_2] & [i_6 \times i_3] & & [i_6 \times i_4] & [i_6 \times i_5] & [i_6 \times i_6] \end{vmatrix}, \tag{14}$$

where $[i_k \times i_j]$ are constant matrices of size $[i_k] \times [i_j]$. One can term this matrix the *expanded distance matrix*. Here, instead of the zero main diagonal, we have a quasi-diagonal consisting of zero square matrices $[i_k] \times [i_k]$, and instead of the symmetry property, we have its generalization

$$[i_k \times i_j] = [i_j \times i_k]^{\mathrm{T}}.$$

Having defined the expanded distance matrix $\overline{\mathbf{D}}_\mathbf{0}$ in such a way, we obtain

$$\mathbf{D_0}(\mathbf{A_1}) = \mathbf{D_0}(\mathbf{A_2}) + \overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_4}) + \overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_5}), \tag{15}$$

where matrix $\overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_4})$ has assumed the following look instead of (14):

$$\overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_4}) = \begin{array}{|ccccc|c|} \hline 3 & 1 & 2 & 3 & 2 & \\ \hline 0 & \rho_{12} & \rho_{12}+\rho_{23} & \rho_{12}+\rho_{45} & \rho_{12}+\rho_{46} & 3 \\ \rho_{12} & 0 & \rho_{23} & \rho_{45} & \rho_{46} & 1 \\ \rho_{23}+\rho_{12} & \rho_{23} & 0 & \rho_{23}+\rho_{45} & \rho_{23}+\rho_{46} & 2 \\ \rho_{45}+\rho_{12} & \rho_{45} & \rho_{45}+\rho_{23} & 0 & \rho_{45}+\rho_{46} & 3 \\ \rho_{46}+\rho_{12} & \rho_{46} & \rho_{46}+\rho_{23} & \rho_{46}+\rho_{45} & 0 & 2 \\ \hline \end{array}$$

Here, the numbers at the top and on the right of the 5×5 matrix are the sizes of the constant rectangular matrices. For $\overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_5})$, instead of (13), we obtain an even simpler matrix:

$$\overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_5}) = \begin{array}{|cc|c|} \hline 6 & 5 & \\ \hline 0 & \rho_{24} & 6 \\ \rho_{24} & 0 & 5 \\ \hline \end{array}$$

So, matrix $\overline{\mathbf{D}}_\mathbf{0}(\mathbf{A_4})$ and matrix (13) have been expanded from size 5×5 to 11×11 and from 2×2 to 11×11, respectively.

According to [18], any tree after several truncations becomes a simple graph with no more than 2 vertices of degree 1, that is, a normal alkane or methane. This means that any tree can be characterized by an expression like (15) for its reduced matrix. Separate matrices constituting the sum characterize the parameters of "shells" formed by a succession of truncations [18], and the number of terms in such an expansion characterizes the structural complexity and, in a certain sense, the branching of alkanes.

3 Topological types of trees

The topological structure of a tree (alkane in our case) is determined by the set of its segments and branches without consideration of their length, i.e., in the first approximation, by the numbers of primary (n_1), tertiary (n_3) and quaternary (n_4) atoms. Relationships between them are derived by excluding N from Eqs. 4 and 5:

$$n_3 + 2n_4 = n_1 - 2.$$

By the way, if we consider $n_2 \neq 0$ in these equations, the n_2 value is excluded together with N; this means that one can increase or decrease the number of vertices

of degree 2 without changing the topological type. We state that the main value determining the type of an alkane is n_1 (i.e., the number of methyl groups). Hereafter we will denote it by n, and the last equation turns into

$$n_3 + 2n_4 = n - 2. \tag{16}$$

The n and n_4 values determine the size of the reduced matrix and the number of independent parameters m in formula (6), respectively. Note that, for given n and n_4, the n_3 value is strictly determined and invariable. Let us consider all possible topological types for different n. We will not analyze $n = 0$ or $n = 1$, since there are no trees for $n = 0$ and the tree for $n = 1$ has only one vertex (methane). Let us take $n = 2$; then, as follows from (16),

$$n_3 + 2n_4 = 0.$$

There is only one solution: $n_3 = n_4 = 0$. It is obvious that one can consider only integer nonnegative numbers as solutions to Eq. 16. Therefore, in the case of $n = 2$, we have two CH_3 groups and an arbitrary number of CH_2 groups, i.e. normal unbranched alkanes. This type of alkanes is hereafter termed type 2.

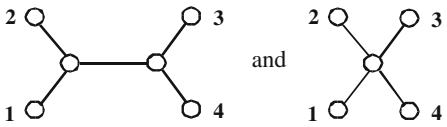
For n = 3 we have

$$n_3 + 2n_4 = 1.$$

There is only one solution: $n_3 = 1$, $n_4 = 0$. This topological type (i.e., type 3) corresponds to singly branched alkanes, or monoalkylalkanes. Note that, as follows from Eq. 16, the parities of integer numbers n_1 and n_3 are always equal for all alkanes. For $n = 4$ we have

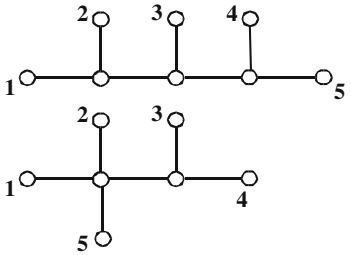
$$n_3 + 2n_4 = 2.$$

There are two solutions: $n_3 = 2$, $n_4 = 0$ and $n_3 = 0$, $n_4 = 1$. Thus, one can distinguish two subtypes of solutions:

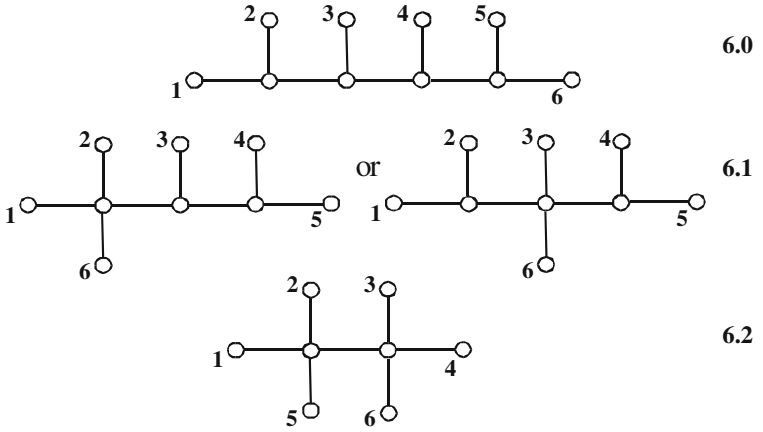


4.0 **4.1**

For $n = 5$ we have $n_3 + 2n_4 = 3$ and also two solutions: $n_3 = 3$, $n_4 = 0$ and $n_3 = 1$, $n_4 = 1$; i.e., there are subtypes 5.0 and 5.1:

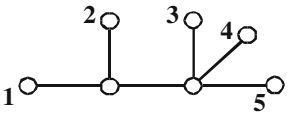


For $n = 6$ we have three solutions: $n_3 = 4$, $n_4 = 0$; $n_3 = 2$, $n_4 = 1$; and $n_3 = 0$, $n_4 = 2$; i.e., there are three subtypes:



So, the type and subtype are again determined by the numbers of primary (methyl groups) and quaternary atoms, respectively. For $n = 7$ we also have three solutions; for $n = 8$ we have four solutions (accordingly, four subtypes); etc.

Thus, for each type n, the size of the reduced matrix is $n \times n$. Let us consider $\mathbf{r}_{ij}$ values as vectors of infinite length whose coordinates $\mathbf{r}_{ij} = \left\{r_{ij}^{m}\right\}$ are generated by the alkane (tree) with number m; then there are specific linear dependences between these vectors for each type and the basis consists of $(2n - 3)$ vectors. However, if one takes not all alkanes of type n but only those of a certain subtype $n.k$, then the basis is reduced to $m = 2n-3-n_4 = 2n-3-k$ vectors. For example, see the following alkane:



$\mathbf{A_6}$

For this alkane $\mathbf{A_6}$ of type 5 and subtype 5.1, we have $m = 2n-3-k = 2\cdot5-3-1 = 6$. For example, we can choose the following vectors as the basis:

$$\{\mathbf{r}_{12}, \mathbf{r}_{13}, \mathbf{r}_{14}, \mathbf{r}_{15}, \mathbf{r}_{23}, \mathbf{r}_{34}\}. \tag{17}$$

The remaining vectors are expressed in basis (17) as follows:

$$\begin{aligned} \mathbf{r}_{24} &= \mathbf{r}_{14} + \mathbf{r}_{23} - \mathbf{r}_{13} \\ \mathbf{r}_{25} &= \mathbf{r}_{15} + \mathbf{r}_{23} - \mathbf{r}_{13} \\ \mathbf{r}_{35} &= \mathbf{r}_{34} + \mathbf{r}_{15} - \mathbf{r}_{14} \\ \mathbf{r}_{45} &= \mathbf{r}_{34} + \mathbf{r}_{15} - \mathbf{r}_{13} \end{aligned} . \tag{18}$$

Expressions (18) are not too symmetrical: they do not contain basis vector $\mathbf{r}_{12}$. But this depends on the way of vertex numbering, since vertices 1 and 2 in graph $\mathbf{A}_6$ are structurally different from vertices 3–5. If the lengths of branches ρ_1 to ρ_5 are not equal to each other, graph $\mathbf{A}_6$ has $n! = 5! = 120$ ways of vertex numbering and, hence, 120 different variants of the reduced matrix. We will consider this issue in more detail in the next section.

4 Distance vectors: molecular codes

The set of p_k values mentioned in the Introduction was termed *molecular code* by Randić [17]. In fact, the term is not quite appropriate, since it is generally implied that a code means a set of numbers unambiguously corresponding to the coded structure, whereas the set

$$\mathbf{P} = \{p_1, p_2, \ldots, p_k\}$$

is degenerated even for C_9H_{20} (to say nothing of larger molecules), as was pointed out by Randić himself. As is mentioned above, one can define the *reduced distance vector* by analogy with Randić's vector using the reduced distance matrix instead of the ordinary distance matrix:

$$\mathbf{P}^0 = \{p_2^0, \ldots, p_k^0\}.$$

The length of vector $\mathbf{P^0}$ is always smaller by unity than that of vector $\mathbf{P}$, since the minimum distance between first-degree vertices is 2 (remember that p_1 is the number of edges, i.e., the number of C–C bonds in an alkane). At the same time, the last coordinates p_k and p_k^0 of these vectors are always equal, since the maximum distances for any tree are the distances between vertices of degree 1. Although vectors $\mathbf{P}$ and $\mathbf{P^0}$ are constructed according to the same algorithm, one can expect the reduced vector (built using a simpler but more "meaningful" matrix) to contain more information as a whole. And, indeed, $\mathbf{P^0}$ is first degenerated only for $C_{10}H_{22}$, and its degeneracy for more complex alkanes is also lower than that of vector $\mathbf{P}$. Table 1 lists the degrees of degeneracy for alkanes from C_9H_{20} to $C_{15}H_{32}$ (remember that neither $\mathbf{P}$ nor $\mathbf{P^0}$ is degenerated for alkanes under C_9H_{20}).

As it is seen from Table 1, vectors $\mathbf{P}$ and $\mathbf{P^0}$ are degenerated not only in pairs but also in triples etc. Also, it is obvious that neither $\mathbf{P}$ nor $\mathbf{P^0}$ (the latter is degenerated to a somewhat lower extent) can be used as the molecular code. However, using matrix

Table 1 Numbers of degenerated structures for vectors **P** and $\mathbf{P^0}$ in alkanes

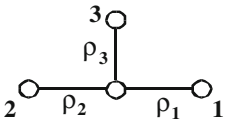
Alkanes	Number of isomers	Number of isomers with degeneration	
		For vector **P**	For vector $\mathbf{P^0}$
C_9H_{20}	35	2	0
$C_{10}H_{22}$	75	2	2
$C_{11}H_{24}$	159	25	12
$C_{12}H_{26}$	355	34	28
$C_{13}H_{28}$	802	77	55
$C_{14}H_{30}$	1858	161	138
$C_{15}H_{32}$	4347	638	419

$\mathbf{D_0}$ enables one to propose a fairly simple and effective nondegenerate set, which is quite suitable to be a molecular code.

For a graph of type 2, matrix $\mathbf{D_0}$ contains only one entry r_{12}. The case is trivial, and we may skip it. For type 3 we have

$$\mathbf{D_0} = \begin{pmatrix} 0 & r_{12} & r_{13} \\ & 0 & r_{23} \\ & & 0 \end{pmatrix}.$$

For simplicity we did not write out the entries under the main diagonal. Clearly, three numbers r_{ij} are enough to restore a structure of type 3:



$\mathbf{A_7}$

since we have simple relationships

$$\rho_1 = \frac{r_{12} + r_{13} - r_{23}}{2}, \quad \rho_2 = \frac{r_{12} + r_{23} - r_{13}}{2}, \quad \rho_3 = \frac{r_{13} + r_{23} - r_{12}}{2}.$$

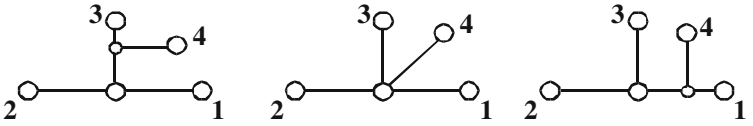
In this case, the numbering of vertices does not matter. However, for more complicated compounds, the order of numbering becomes important, since some entries of the $\mathbf{D_0}$ matrix are linearly dependent. Here we will apply the "cyclic" numbering, where the numbers at vertices of degree 1 increase as we walk around the structural formula in the clockwise direction, as it is described in detail in [15]. Notice that vertices in graph $\mathbf{A_1}$ are numbered in just the same manner. For alkanes of type 4 we have

$$\mathbf{D_0} = \begin{pmatrix} 0 & r_{12} & r_{13} & r_{14} \\ & 0 & r_{23} & r_{24} \\ & & 0 & r_{34} \\ & & & 0 \end{pmatrix}.$$

Let us successively write down entries of the first row (from right to left) and then entries of the diagonal (downwards):

$$\mathbf{R}_4^0 = (r_{14}, r_{13}, r_{12}, r_{23}, r_{34}). \tag{19}$$

Since the number of these entries is odd (namely, $2n - 3$), there is always a "middle element" in such a set. Here it is r_{12}, whose neighbors are r_{13} and r_{23}. These three numbers are sufficient to construct tree $\mathbf{A_7}$. Owing to cyclic numbering, vertex 4 will be between vertices 3 and 1; so, replacing the middle "triplet" by the r_{13} value and using the left and right neighbors r_{14} and r_{34}, we construct one of the three graphs



The choice between them is determined by numbers r_{ij} themselves, as is mentioned above for restoration of tree $\mathbf{A_7}$. Thus, the r_{24} element is not used (it can be expressed as a linear combination of others). For type 5, instead of (19), we obtain the set

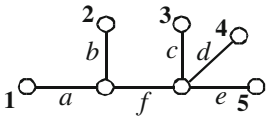
$$\mathbf{R}_5^0 = (r_{15}, r_{14}, r_{13}, r_{12}, r_{23}, r_{34}, r_{45}).$$

One can unambiguously restore a tree of type 5 from this set, starting from the central triplet and using the above method. So, the set of $(2n - 3)$ numbers constituting the top row and the maximum nonzero diagonal

$$\mathbf{R}_n^0 = (r_{1n}, r_{1,n-1}, \ldots, r_{13}, r_{12}, r_{23}, \ldots, r_{n-2,n-1}, r_{n-1,n}) \tag{20}$$

always corresponds to only one tree and can be used as its molecular code if the numbering of vertices of degree 1 is cyclic. Actually, a set like (20) can be selected in many ways. For example, the set proposed in [15] is actually more convenient from the standpoint of coding and decoding.

Let us consider an example. This is an alkane of type 5.1:



$\mathbf{A_8}$

Here *a, b, c, d* and *e* are the lengths of branches 1–5, respectively, and *f* is the length of the internal segment. Then the reduced matrix according to formula (15) looks as follows:

$$\mathbf{D_0(A_8)} = \boldsymbol{\tau}_5^0 \begin{pmatrix} a & & & & \\ & b & & & \\ & & c & & \\ & & & d & \\ & & & & e \end{pmatrix} + \left[\boldsymbol{\tau}_5^0 \begin{pmatrix} a & & & & \\ & b & & & \\ & & c & & \\ & & & d & \\ & & & & e \end{pmatrix} \right]^{\mathrm{T}} + \begin{pmatrix} 0 & 0 & f & f & f \\ 0 & 0 & f & f & f \\ f & f & 0 & 0 & 0 \\ f & f & 0 & 0 & 0 \\ f & f & 0 & 0 & 0 \end{pmatrix}.$$

Table 2 Numberings for tree $\mathbf{A_8}$ with $a = b$ and $c = d = e$

Arbitrary numbering	Cyclic numbering
aaccc	aaccc
acacc	–
accac	–
accca	accca
caacc	caacc
cacac	–
cacca	–
ccaac	ccaac
ccaca	–
cccaa	cccaa

If a is not equal to b in $\mathbf{A_8}$ and c, d, and e are different from each other, the vertices can be numbered in 5! = 120 ways. Each numbering is determined by the sequence of these 5 letters. The condition of numbering cyclicity reduces the number of combinations to 60. For a simpler case, where $a = b$ and $c = d = e$, we obtain 10 arbitrary numberings and 5 cyclic numberings. These variants are presented in Table 2.

As is seen from Table 2, only one-half of arbitrary numberings are cyclic. Let us consider one of the simplest cases for $\mathbf{A_8}$, where $a = b = f = 1$ and $c = d = e = 2$, i.e., alkane 2-methyl-3,3-diethylpentane:

$$\mathbf{D_0} = \begin{pmatrix} 0 & 2 & 4 & 4 & 4 \\ & 0 & 4 & 4 & 4 \\ & & 0 & 4 & 4 \\ & & & 0 & 4 \\ & & & & 0 \end{pmatrix}$$

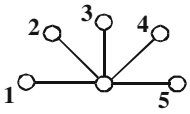
The reduced matrix has 10 entries. At the same time, we can notice that

$$\mathbf{P} = \{9,\ 12,\ 15,\ 9\} \quad \text{and} \quad \mathbf{P^0} = \{1,\ 0,\ 9\}.$$

If the numbering coincides with the one shown at tree $\mathbf{A_8}$, we obtain

$$\mathbf{R_5^0} = (4, 4, 4, 2, 4, 4, 4).$$

Sequence $\mathbf{R_5^0}$ enables one to restore the alkane. At the same time, note that, irrespective of the actual numbering (if its cyclicity is preserved), number "2" is always present in set $\mathbf{R^0}$. This is due to the fact that 2 is either the distance between vertices with adjacent numbers, i.e., this distance is present in the matrix diagonal, or it is equal to r_{15} (at the end of the first row). If we try to decode set $\mathbf{R^0} = (4, 4, 4, 4, 4, 4, 4)$, we will obtain the following tree through the aforementioned algorithm:



The lengths of all branches in this tree are equal: $a = 2$. There is no alkane of this type, but the tree itself does exist.

5 Conclusions

The structure of the reduced distance matrix considered in this paper makes it possible to divide this matrix into several parts, which correspond to "shells" of the tree obtained by successive truncation of its branches [18]. These components enable us to design new types of topological indices and to simplify the analysis of dependences between various properties and the molecular structure.

Linearly independent entries of the reduced matrix can be regarded as constituting a nondegenerate molecular code applicable for complicated organic compounds [15]. At the same time, one has to specify the conditions which sequence (20) must satisfy so that it would correspond to real chemical structures. A similar problem for sequence (3) was solved in [19].

As an example of using independent parameters that make up the reduced matrix, let us take the expression for the Wiener index in the case of alkanes. Normal alkanes of type 2 (that is, n-C_nH_{2n+2}) have a chain of length $n-1$ (a total of $n-1$ C–C bonds) and the value of the Wiener index equal to

$$W = \frac{(n+1)n(n-1)}{3!} = C_{n+1}^3.$$

These numbers for $n = 1, 2, 3, \ldots$ are 1, 4, 10, 20, 35, 56, etc. Let us define the *Wiener function* for an integer nonnegative x as follows:

$$W(x) = \frac{(x+2)(x+1)x}{3!}.$$

Note that $W(0) = 0$. Then, for example, an alkane of type 3 with formula $\mathbf{A_7}$ has

$$W(\mathbf{A_7}) = \sum_{(i,j)} W(\rho_i + \rho_j) - \sum_{i=1}^{3} W(\rho_i),$$

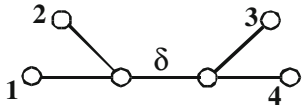
For tree $\mathbf{A_2}$ we obtain

$$W(\mathbf{A_2}) = \sum_{(i,j)} W(\rho_i + \rho_j) - (n-2) \sum_{i=1}^{n} W(\rho_i).$$

For alkanes of subtype 4.1 we get

$$W = \sum_{(i,j)} W(\rho_i + \rho_j) - 2 \sum_{i=1}^{4} W(\rho_i),$$

and for an alkane of subtype 4.0



where δ is the distance between the tertiary atoms, we have

$$W = \sum_{(ij)} W(\rho_{ij}) - \sum_{i=1}^{4} W(\rho_i) - \sum_{i=1}^{4} W(\rho_i + \delta) + W(\delta).$$

Here $\rho_{ij} = \rho_i + \rho_j + \delta$ for any (i, j) except for pairs $(i, j)=(1, 2)$ and $(3, 4)$, for which δ is not included in the distance between vertices i and j.

Obviously, the latter formula for the Wiener index coincides with the previous formula for type 4.1 if $\delta = 0$. Similar expressions for arbitrary alkanes make it possible to estimate the Wiener index via linear combinations of Wiener functions of parameters constituting the reduced matrix. Notice that a similar problem is considered in [20] in more detail but in a less generalized form, recurrent relationships are obtained in [21], and more general formulas for the Wiener index are proposed in [22].

Evidently, for any topological index $I(\mathbf{D})$ constructed on the basis of the distance matrix, one can construct the corresponding index $I(\mathbf{D_0})$ using the same formula but on the basis of the reduced matrix, i.e., with $\mathbf{D}$ replaced by $\mathbf{D_0}$. Then, for all known relationships [1,2,17], we obtain

$$P = f[I(\mathbf{D})] \Rightarrow P = f[I(\mathbf{D_0})],$$

That it, each dependence of property P on index $I(\mathbf{D})$ can formally be put in correspondence with a similar dependence on $I(\mathbf{D_0})$. Naturally, the parameters of function f may change but the general form remains the same. In addition, it is interesting to investigate such dependences not only for $\mathbf{D_0}$ but also for its separate components, i.e., components of (15).

Dependences of physicochemical properties on the structure of alkanes, usually represented as

$$P = a_0 + \sum_{i=1}^{k} a_i p_k, \tag{16}$$

where p_k are entries of the Randić distance vector [12,13], can also be modified using our reduced vector:

$$P = a_0^0 + \sum_{i=2}^{k} a_i^0 p_k^0. \tag{17}$$

There are grounds to hope that formula (17) will describe the aforementioned dependences more precisely than (16) and therefore will be useful in solution of the structure–property problem.

References

1. A.T. Balaban, Pure Appl. Chem. **55**, 199 (1983)
2. Z. Mihalić, S. Nicolić, N. Trinajstić, J. Am. Chem. Soc. **32**, 28 (1992)
3. M.I. Stankevich, I.V. Stankevich, N.S. Zefirov, Usp. Khim. **57**, 337 (1988)
4. H. Wiener, J. Am. Chem. Soc. **69**, 17 (1947)
5. H. Wiener, J. Am. Chem. Soc. **69**, 2636 (1947)
6. H. Hosoya, Bull. Chem. Soc. Jpn. **44**, 2332 (1971)
7. J.R. Platt, J. Phys. Chem. **56**, 328 (1952)
8. D.H. Rouvray, CHEMTECH **6**, 379 (1973)
9. D.H. Rouvray, Amer. Sci. **61**, 729 (1973)
10. E.A. Smolenskii, Zh. Fiz. Khim. **38**, 1288 (1964)
11. A.L. Safer, E.A. Smolenskii, Zh. Fiz. Khim. **37**, 2657 (1963)
12. A.L. Safer, E.A. Smolenskii, Zh. Fiz. Khim. **38**, 202 (1964)
13. A.L. Safer, E.A. Smolenskii, Zh. Fiz. Khim. **38**, 2230 (1964)
14. E.A. Smolenskii, Zh. Vychisl. Mat Mat Fiz. **2**, 371 (1962)
15. E.A. Smolenskii, Dokl. Akad. Nauk. **380**, 60 (2001)
16. S.V. Yushmanov, Dokl. Akad. Nauk SSSR **259**, 49 (1981)
17. M. Randić, C.L. Wilkins, J. Phys. Chem. **83**, 1525 (1979)
18. E.A. Smolenskii, N.S. Zefirov, Dokl. Akad. Nauk. **379**, 774 (2001)
19. K.A. Zaretskii, Usp. Matem. Nauk. **20**, 90 (1965)
20. D. Bonchev, N. Trinajstić, J. Chem. Phys. **67**, 4517 (1977)
21. E.R. Canfield, R.W. Robinson, D.H. Rouvray, J. Comput. Chem. **6**, 598 (1985)
22. I. Gutman, J. Mol. Struc. (THEOCHEM) **104**, 137 (1993)