communications in mathematical
match
and in computer chemistry

no. 42, October 2000

ISSN 0340-6253

MATCDY (42) 181-231 (2000)

ON ORDERING OF FOLDED STRUCTURES

M. Randić,*,** M. Vračko,** M. Novič,** and S. C. Basak***

* Department of Mathematics and Computer Science
Drake University, Des Moines, IA 50311, USA;

** National Institute of Chemistry,
Hajdrihova 19, 1001 Ljubljana, Slovenia;

*** Natural Resources Research Institute
University of Minnesota at Duluth. Duluth, MN 55811

e-mail: milan.randic@drake.edu
e-mail: marjan.vracko@ki.si
e-mail: marjana.novic@ki.si
e-mail: sbasak@wyle.nrri.umn.edu

Abstract

We consider partial ordering of folded structures using several numerical techniques that have been developed previously. In particular we consider a set of folded chains of equal length superimposed on the Cartesian coordinate grid, their coding and subsequent ordering. For linear structures we propose different codes which are used to derive partial orders for structures. Structure labels in partial orders obtained are subsequently replaced by numerical parameters of individual structures in order to see if there is some regularity in numerical data. In particular we considered regularities for the leading eigenvalues of the D/D matrices and the leading eigenvalues of the line-adjacency matrices selected folded curves. We have also illustrated use of partial order in structure-property activity-relationships.

Introduction

Folded structures are of considerable interest in chemistry, physics and mathematics. One of the central topics of biochemistry and molecular biology concerns folding of proteins, relationship between molecular shape and molecular structure. Besides the studies on the formation and the mechanism behind protein folding, that may yield to a prediction of the folding pattern for a known primary sequence of amino acids, of considerable interest is characterization of the geometrical patterns of already folded structures. In Fig. 1 we illustrate model proteins consisting of 27 units (amino acids) considering by Tang and coworkers [1]. Characterization of this model proteins received recently attention [2, 3]. Each structure and each pattern of folding can be described using structural invariants obtaining thus characteristic "signature" for each folded system. As long as one uses a single number to represent each structure the structure can be simply compared and can always be fully ordered, excluding occurrence of numerical degeneracies. However, when structure is represented by a sequence ordering of structures requires more attention. Typically comparison of sequences leads to partial order [4], different partial orders for different sequences. Ordering of structures is of interest as it may illuminate regularities in variations of selected properties of structures. For example, one may consider the computed folded model structures for proteins reported by Tang and

Fig. 1 Model proteins consisting of 27 units (amino acids) considering by Tang and coworkers [1].

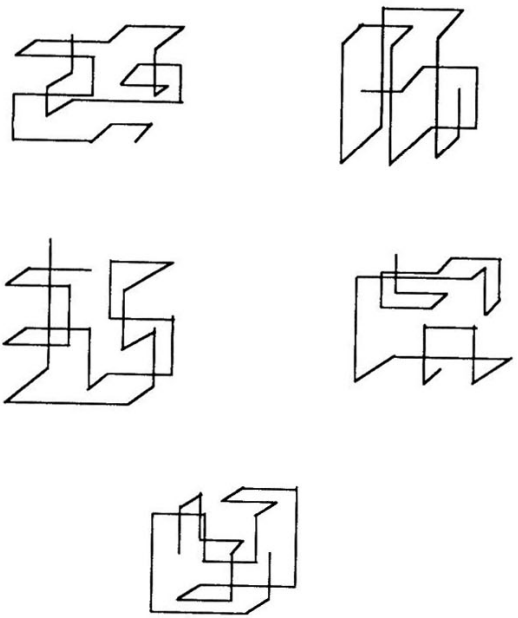
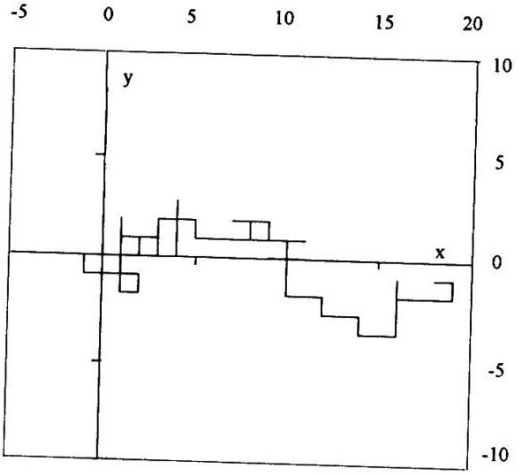


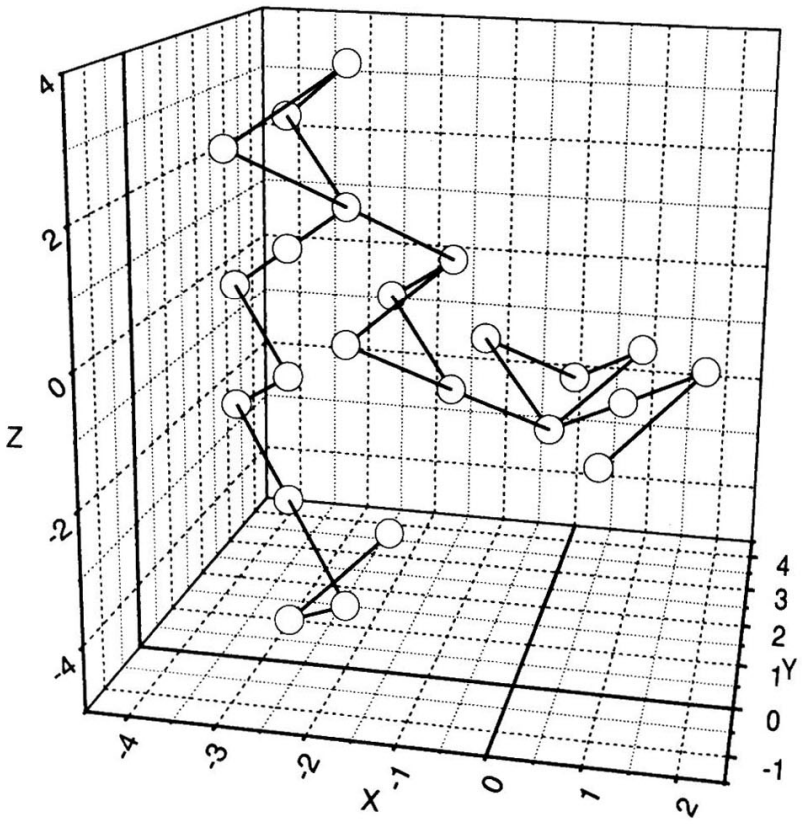
Fig. 2 The first 92 basis of β Globin Gene (having 1424 nucleic acids) as depicted by a graphical representation of DNA by A. Nandy



coworker [1]. Can these structures be ordered by some inherently structural criterion that may reflect different degree of folding of each structure?

In Fig. 2 we illustrate the first 92 basis of β globin gene (having 1424 nucleic acids) as depicted by a graphical representation of DNA by A. Nandy [5, 6]. Here the four direction of coordinate system correspond to the four nucleic acids: A and G along $\pm$ x axis and C and T along $\pm$ y axis. The problem again is that of numerical characterization of folded curves, which as we see from Fig. 2 may also overlapping itself. Different DNA primary sequences will have different pattern of folding. Moreover, alternative graphical representations may yield a variety of distinct graphical forms for the same primary sequence of DNA, Recently a 3-dimensional representation of DNA was considered [7] in which the four directions assigned to the four nucleic bases are the directions pointing from the center to the four vertices of a regular tetrahedron. An advantage of this approach is that the four directions assigned to nucleic bases are fully equivalent and no proeference is given to specific pairings, such as A - G and C - T. In Fig. 3 we show the construction of the initial steps of the 3-dimensional representation of the same DNA shown in Fig. 2 as a 2-dimensional graphical diagram. It was interesting to find that when the 3-D spatial curve of Fig. 3 is projected on the (x, y) coordinate plane one obtains as the projection precisely to folded curve of Nandy shown in Fig. 2. The problem to consider, of course, is finding structural criteria that would result in ordering of structures in some logical way, and particularly ordering of structures that

Fig. 3 Graphical 3-dimensional representation of the primary sequence of DNA already shown as 2-D diagram in Fig. 2



would parallel some of their known properties.

The conformations of normal alkane chains, that is, the conformations of chains of n carbon atoms on a 3D diamond lattice, also illustrate systems showing various degree of folding. The same is the case with the corresponding problem in 2D, the conformations of chains of n atoms on a graphite lattice. The nine conformers of 7-carbon atom chains superimposed on a graphite lattice and the 18 conformers of 8-carbon atom superimposed on a graphite lattice chains are illustrated in Fig. 4 and Fig. 5, respectively. The corresponding "degree of folding," outlined in ref. [8] and ref. [9], to be discussed later at some length, are also shown under each structure. The numerical values for the "degree of folding" are such that the extreme values "1" and "0" correspond to a straight chain (unfolded structure) and a hypothetical "mostly folded structure," respectively. Hence, the smaller values of the "degree of folding" belong to structures that are more folded. The "degree of folding" as defined is, in fact, a measure of the departure of a chain structure embedded in a space from a straight line form. Thus TTTT (the first structure of Fig. 4) is less folded than TTTC (the second structure in Fig. 4). Here T and C indicate three successive carbon-carbon bonds in *trans* and *cis* configurations, respectively. The numbers displayed in Fig. 4 and Fig. 5 under each structure appear plausible and support the interpretation of these numbers as a particular measure of the folding or the bending of chain structures. Not only the folding index is useful when comparing structures of the same size but it allows also comparison of

Fig. 4 The nine conformer of 7-carbon atom chains superimposed on a graphite lattice. The numbers displayed in under each structure are the leading eigenvalues of D/D matrices and represent a measure of folding of chain structures

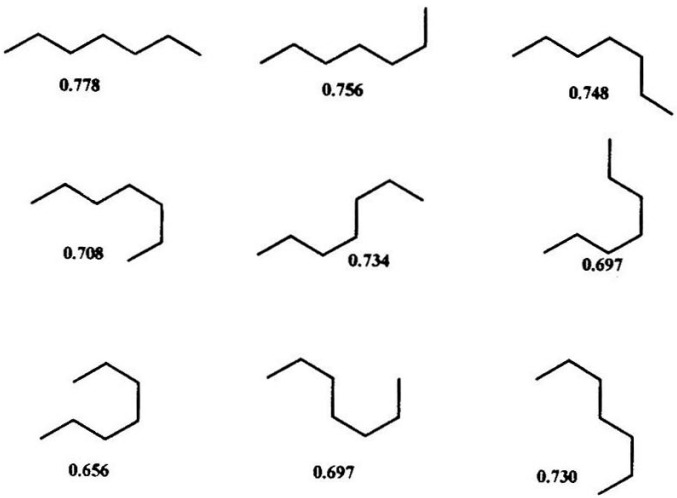
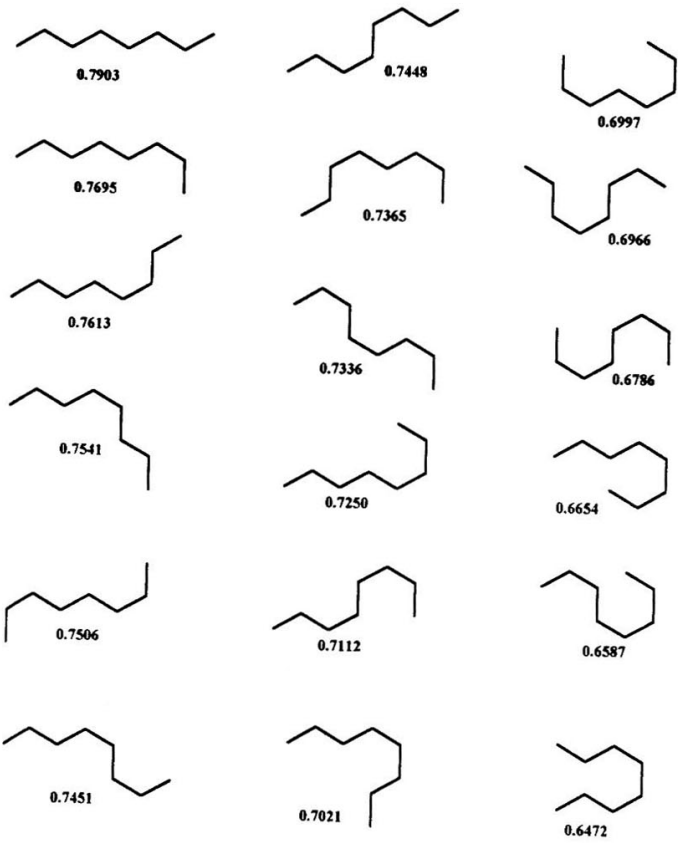


Fig. 5 The 18 conformer of 8-carbon atom chains superimposed on a graphite lattice. The numbers displayed in under each structure are the leading eigenvalues of D/D matrices and represent a measure of folding of chain structures



structures of different size. Thus we find that TTTTT (the first structure in Fig. 5) is to be viewed less folded than TTTT (the first structure in Fig. 4) and this also appears plausible, because in the limit an all *trans* chain will approaches a straight line in its appearance. Complete order, that is ordering of the structures sequentially (i. e., "1-dimensionally") is not necessarily revealing underlying structural components that may critically influence the magnitudes of various molecular properties. Can we arrive at partial order for such structures?

Fig. 6 shows several mathematical curves representing initial stages of construction of fractals illustrating different patterns of folding. Numerical characterizations of such curves, which one considers when having particular property in mind, are of some interest [10-18]. Additional mathematical curves are depicted in Fig. 7. By visual inspection qualitatively one can say that double spiral (Fig. 6 a) is more folded than simple spiral (Fig. 6 b), and that both of them are less folded than the dragon curve (Fig. 6 c). However, without a quantitative characterization it may be difficult, or at least somewhat ambiguous, to claim that the degree of folding of the worm curve (Fig. 6 d) is in between that of a spiral and double spiral. On the other hand, given the numerical values of a property, like the "degree of folding," considered here one can completely order these mathematical curves. However, just as was in the case of 7-carbon atom chains and 8-carbon atom chains, the complete (1-dimensional) ordering will not illuminate inherent

Fig. 6 Initial stages of several fractals: (a) Hilbert curve; (b) Koch's curve; (c) Sierpinski arrow

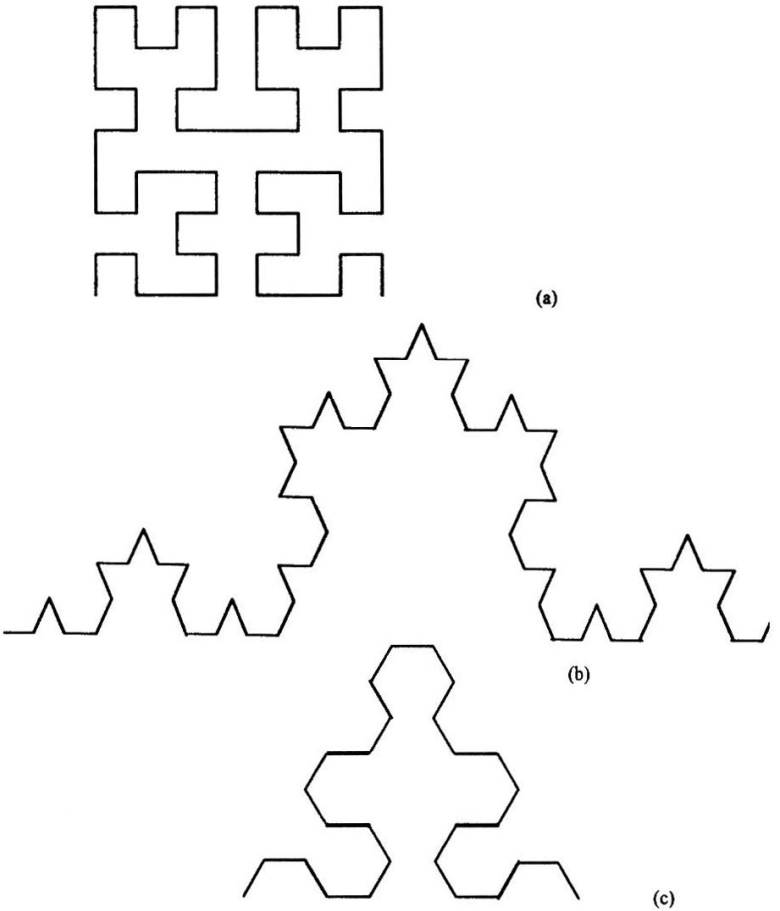
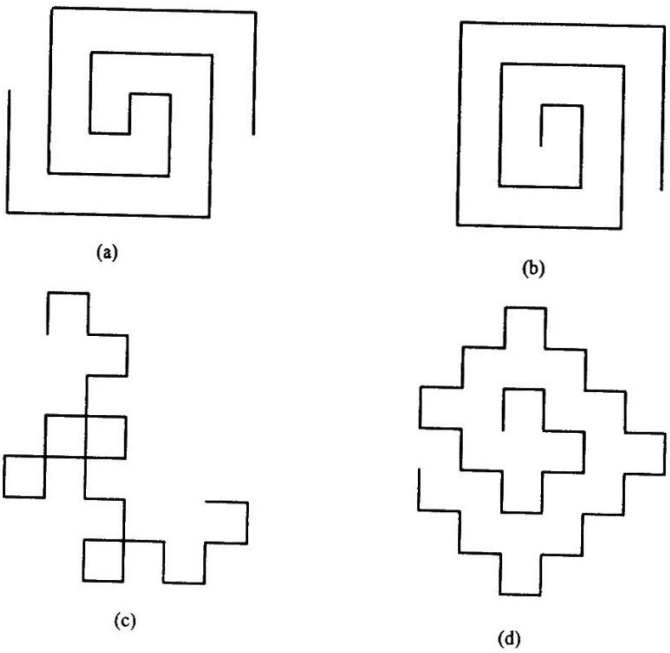


Fig. 7 Mathematical curves: (a) double spiral; (b) simple spiral; (c) the dragon curve; (d) the worm curve



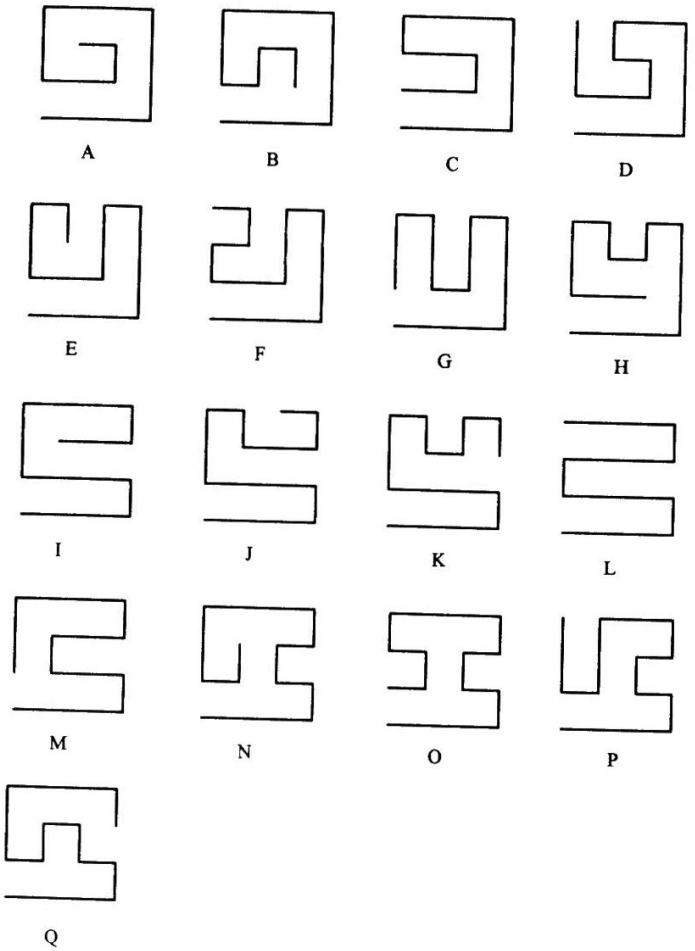
structural components that contribute to the selected fractal property.

In this article we will consider problems related to ordering of folded structures, that may be of mathematical, chemical, or physical origin. We have selected relatively simple folded curves of Fig. 8 to outline a particular mathematical analysis suitable for such cases. Although the outlined approach for this particular example need not be readily applicable to other situations it is hoped that the underlying reasoning may nevertheless be of more general interest and applicability.

Simply Folded 2D Mathematical curves

The folded curves of Fig. 8, labeled by letters from A - Q, represent all possible cases of folded curves confined to 4 x 4 Cartesian block (hence involving 16 points with integer coordinates) constrained so that the four first points form the base line. Hence, all curves of Fig. 8 start with a straight line segment given by four point, the first point being at the left lower corner of a 4 x 4 block. We selected these folded curves because of their relatively limited number (17 in all), while they apparently display fairly diverse forms of the folding. The curve A represents the initial stage of a spiral, the curve D corresponds to the initial stage of a double spiral, while the curve L represents a simple stepwise folding. Among the 17 folded curves of Fig. 8 curve L is the only case that has some symmetry (horizontal plane of reflection). Other folded curves have no apparent interpretation, but a

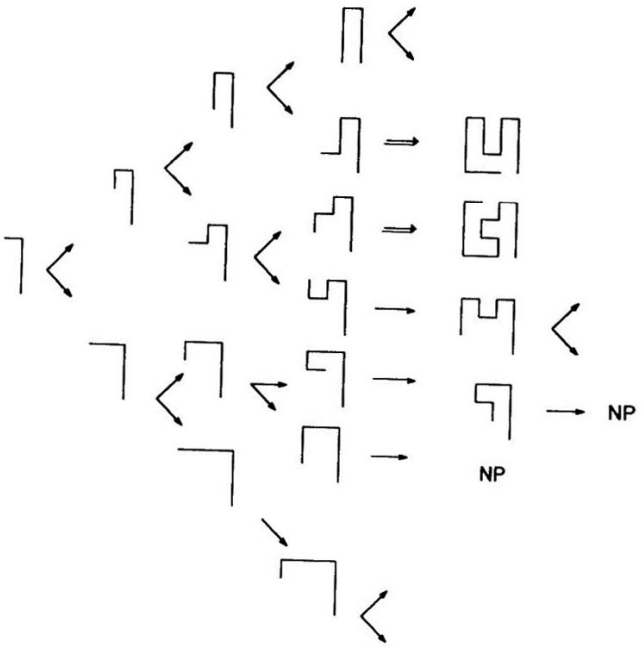
Fig. 8 The folded curves, labeled by letters from A - Q, representing all possible cases of curves confined to 4x4 block of Cartesian coordinate systems



collection as a whole exhibits different folding patterns and represents a small library of mathematical "fingerprinting" of possible folded curves. Observe that curves C, G and M have the same peripheral contour (that of letter U) and differ only by the site at which the "missing" link, that would close the curve occurs. As we will see later all these folded curves are associated with different values for the calculated "degree of folding," pointing to the sensitivity of the adopted measure of folding, which is given by the leading eigenvalue of the so called D/D matrix of a structure [8].

Immediately the question arises: How do we know that these are all the possible structures under the specified constraints? The answer follows from the use of an algorithm for generating such structures to derives such structure. In Fig. 9 we illustrate the beginning of the construction. We start from the origin (0, 0) with line involving the first four points (0, 0), (1, 0), (2, 0), (3, 0). The next point has to be (4, 1) as there are no other alternatives to continue the "growing" of the line. Now there are two possibilities: either we link point (4, 2) to (4, 1) or we link the point (3, 1) to (4, 1) as shown in Fig. 9, which illustrates also the next step that includes the available nearest neighbor points of (4, 2) and (3, 1), respectively. We will not elaborate here the particular construction algorithm which can be applied also to larger curves, and to curves without constraints used here. Observe, however, how the constraints (e. g., curves are restricted to a 4 x 4 block) limit explosive combinatorial growth of possible cases. For example, after the fourth step (the last step completely illustrated in Fig. 9) in one case we can continue

Fig. 9 The beginning of the construction of folded curves of Fig. 8



construction because there is but a single choice for added line. In additional two cases we could even complete the construction because at successive steps there was but a single choice for addition of next line segment. As we see at this early stages of the construction one can also detect unproductive constructions, indicated as NP (not possible), that do not yield folded curve subject to the constraints selected.

Below we give the count of the folded curves on Cartesian grids of increasing size to illustrate fast growth of possible forms. The first count refer to folded curves constrained so that they start with the base line of the size of the grid, and the second count gives all possible forms. Only symmetry non-equivalent folding forms were counted.

Grid size	1 x 1	2 x 2	3 x 3	3 x 4
Restricted	1	3	17	127
Unrestricted	1	3	23	282

The next question to consider is that of developing a code for such structures and subsequent characterization of such folded systems. Codes should be distinguished from characterization: A code depends on a convention adopted, which may presume some rule for labeling of vertices. On the other hand a characterization is independent of atomic labels. Characterization is based on mathematical properties of a structure, just as properties, such as the boiling points, the heats of formation, the entropy, the density, or the molar refraction offer a physicochemical characterization

of molecules, such as alkanes. For example, for the folded structure A we can write a code: 3, 3, 3, 2, 2, 1, 1 which indicates (starting with the point at the origin) the lengths of the consecutive segments as the curve is folded. In contrast, we may consider a characterization based on the length of consecutive segments, such as the sequence: 3, 2, 2 which tells that there are three segments of length 3, two segments of length 2, and two segments of length 1. Construction of the sequence 3, 2, 2 is independent of atomic labels. That is, it does not matter whether we start counting the segments from the "beginning" or from the "end" of the structure. Clearly, there is loss of information when characterization 3, 2, 2 is considered. In listing only the size of segments and their number we have lost information on the connectivity between the segments. On the other hand from the code, the sequence 3, 3, 3, 2, 2, 1, 1, one can reconstruct the shape of the folded curve completely. A good code is characterized by no loss of information, which means that reconstruction is possible, even if not straightforward and require testing alternative possibilities for feasibility.

Partial order

In Table 1 (the left column) we have listed for the 17 folded curves of Fig. 8 the segment-length codes defined in the previous section. The structures have been ordered lexicographically, that is by the magnitude of the entries in their sequence code. If the sequential entries in two sequences

Table 1 Codes for the 17 folded curves of Fig. 8

	Segment-length code	Partial Sums
A	3, 3, 3, 2, 2, 1, 1	3, 6, 9, 11, 13, 14, 15, 15, 15, 15, 15
B	3, 3, 3, 2, 1, 1, 1, 1	3, 6, 9, 11, 12, 13, 14, 15, 15, 15, 15
C	3, 3, 3, 1, 2, 1, 2	3, 6, 9, 10, 12, 13, 15, 15, 15, 15, 15
D	3, 3, 2, 1, 1, 1, 2	3, 6, 8, 9, 10, 11, 13, 15, 15, 15, 15
E	3, 3, 1, 2, 2, 2, 1, 1	3, 6, 7, 9, 11, 13, 14, 15, 15, 15, 15
F	3, 3, 1, 2, 2, 1, 1, 1, 1	3, 6, 7, 9, 11, 12, 13, 14, 15, 15, 15
G	3, 3, 1, 2, 1, 2, 1, 2	3, 6, 7, 9, 10, 12, 13, 15, 15, 15, 15
H	3, 3, 1, 1, 1, 1, 2, 2	3, 6, 7, 8, 9, 10, 11, 13, 15, 15, 15
I	3, 1, 3, 2, 3, 1, 2	3, 4, 7, 9, 12, 13, 15, 15, 15, 15, 15
J	3, 1, 3, 2, 1, 1, 2, 1, 1	3, 4, 7, 9, 10, 11, 13, 14, 15, 15, 15
K	3, 1, 3, 2, 1, 1, 1, 1, 1, 1	3, 4, 7, 9, 10, 11, 12, 13, 14, 15, 15
L	3, 1, 3, 1, 3, 1, 3	3, 4, 7, 8, 11, 12, 15, 15, 15, 15, 15
M	3, 1, 2, 1, 2, 1, 3, 2	3, 4, 6, 7, 9, 10, 13, 15, 15, 15, 15
N	3, 1, 1, 1, 1, 1, 3, 2, 1, 1	3, 4, 5, 6, 7, 8, 11, 13, 14, 15, 15
O	3, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1	3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15
P	3, 1, 1, 1, 1, 1, 2, 2, 1, 2	3, 4, 5, 6, 7, 8, 10, 12, 13, 15, 15
Q	3, 1, 1, 1, 1, 1, 1, 2, 3, 1	3, 4, 5, 6, 7, 8, 9, 11, 14, 15, 15

are the same, comparison is shifted to the next entry in the sequence until difference occurs. The column at the right side of Table 1 shows partial sums, $s_1, s_2, s_3, s_4, \ldots$ that are defined as the sequence: $s_1 = a_1$, $s_2 = a_1 + a_2$, $s_3 = a_1 + a_2 + a_3$, $s_4 = a_1 + a_2 + a_3 + a_4$, . . . To determine the dominance and the partial order for a set of sequences (such as those of Table 1), according to Muirhead [19], one compares the sequences of the corresponding partial sums. We will assume that the sequences to be compared are of the same length. When this is not the case one can make sequences to satisfy this requirement by adding as many zeros at the end of the sequence as necessary. Thus, for example, the sequence A is augmented into 3, 3, 3, 2, 2, 1, 1, 0, 0, 0, 0 in order to have the same length as the longest sequence of Fig. 8 (the sequence O of Table 1). We have left the sequences of segments in their original form, that is without listing of the additional zeros at the end (shown on the left in Table 1) but the sequences of the constructed partial sums (shown on the right in Table 1) are augmented so to be of the same length. If for two structures A (a_i) and B (b_i) the following inequalities hold:

$$a_1 \geq b_1$$

$$a_1 + a_2 \geq b_1 + b_2$$

$$a_1 + a_2 + a_3 \geq b_1 + b_2 + b_3$$

. . . . .

$$a_1 + a_2 + a_3 + \ldots + a_n = b_1 + b_2 + b_3 + \ldots + b_n$$

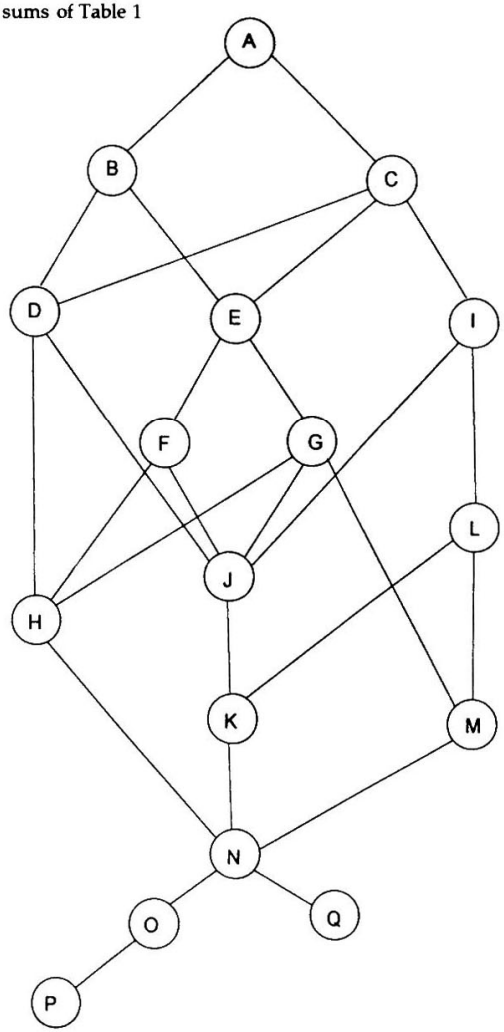
then we say that A dominates B. If any of the inequalities would not be satisfied the two structures are said not to be comparable. This means that

neither A dominates would B nor B would dominates A. Two structures that are not comparable can nevertheless be dominated by a third structure, or can dominate a third structure.

In Fig. 10 we show the resulting partial order for the 17 structures of Fig. 8. As we see A dominates all the remaining structures, while structure C also dominates all the structures except B with which it is not comparable (and of course A, by which it is dominated). The segment of the graph depicting the partial order from structure A to structure N represents a lattice. Lattice is defined as a partial order characterized by a single dominant structure (the so-called master) and a single structure dominated by all (the so called slave). However, when we consider all 17 structures we have a semi-lattice: There is a dominant master structure, but at the bottom instead of a single slave structure dominated by all there are two structures, O and Q, which are not comparable.

There is no unique graphical representation of the partial order. The diagram shown in Fig. 9 is selected for its relative simplicity. One tries as much as possible to avoid unnecessary crossing of lines in such diagrams. Once a diagram is drawn one can modify it by shifting vertices of the diagram at will as long as the dominance (the vertical relationship) is not affected. In other words, two diagrams representing a partial order are equivalent if along each path in the both diagrams the same sequence of structures are obtained.

Fig. 10 The resulting partial order for the 17 structures of Fig. 7 based on partial sums of Table 1



Alternative Codes

The codes of Table 1 represent list of segments of the folded curve starting with the longer segment first. If we reverse the ordering rule and start by listing the smaller terminal segment first we obtain the codes listed in Table 2 (left column). As we see these codes lead to a different sequences of partial sums (shown in the right column of Table 2) and result in a completely different partial order (shown in Fig. 11).

Let us consider yet another code for the folded structures of Fig. 8. Again we will start with the point (0, 0) to which we assign label zero. We proceed sequentially along the folded line listing either 0 or 1 to each integer coordinate point, depending on the orientation of the line relative to the previous point. If the line maintains the direction we write one, if it changes the direction, up or down, left or right, that is the point is the site of a "kink," we write 0. This definition gives for the structure A of Fig. 8 the following sequence:

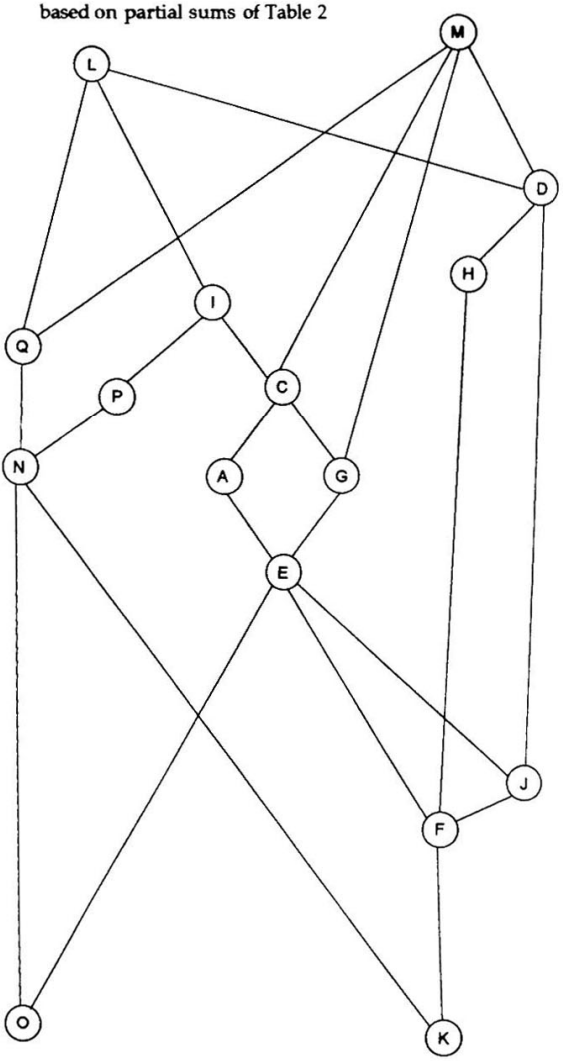
0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0

By convention we will assigned label 0 to the last point in the curve. In contrast to sequences derived from the codes based on the list of segment lengths, which may have different lengths, here all the codes have the same length, which is determined by the number of vertices involved. We will refer to this as the "line/kink" codes.

Table 2 Codes for the 17 folded curves of Fig. 8 starting with the smallest segments first

Structure	Segment-length code	Partial Sums
G	3, 1, 3, 1, 3, 1, 3	3, 4, 7, 8, 11, 12, 15
M	2, 3, 1, 2, 1, 2, 1, 3	2, 5, 6, 8, 9, 11, 12, 15
D	2, 2, 1, 1, 1, 2, 3, 3	2, 4, 5, 6, 7, 9, 12, 15
N	2, 2, 1, 1, 1, 2, 3, 3	2, 4, 5, 6, 7, 8, 9, 12, 15
I	2, 1, 3, 2, 3, 1, 3	2, 3, 6, 8, 11, 12, 15
P	2, 1, 2, 2, 1, 1, 1, 1, 3	2, 3, 5, 7, 8, 9, 10, 11, 12, 15
C	2, 1, 2, 1, 3, 3, 3	2, 3, 5, 6, 9, 12, 15
G	2, 1, 2, 1, 2, 1, 3, 3	2, 3, 5, 6, 8, 9, 12, 15
Q	1, 3, 2, 1, 1, 1, 1, 1, 3	1, 4, 6, 7, 8, 9, 10, 11, 12, 15
N	1, 1, 2, 3, 2, 1, 1, 1, 1, 1, 3	1, 2, 4, 7, 8, 9, 10, 11, 12, 15
A	1, 1, 2, 2, 3, 3, 3	1, 2, 4, 6, 9, 12, 15
E	1, 1, 2, 2, 2, 1, 3, 3	1, 2, 4, 6, 8, 9, 12, 15
J	1, 1, 2, 1, 1, 2, 3, 1, 3	1, 2, 4, 5, 6, 8, 11, 12, 15
O	1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 3	1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 15
B	1, 1, 1, 1, 2, 3, 3, 3	1, 2, 3, 4, 6, 9, 12, 15
F	1, 1, 1, 1, 2, 2, 1, 3, 3	1, 2, 3, 4, 6, 8, 9, 12, 15
K	1, 1, 1, 1, 1, 1, 2, 3, 1, 3	1, 2, 3, 4, 5, 6, 8, 11, 12, 15

Fig. 11 The resulting partial order for the 17 structures of Fig. 8 based on partial sums of Table 2



As an alternative we could have considered the complement of the above code:

1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1

obtained by reversing the assignment of zeros and ones. We decided to stay with the first choice mentioned so that the structure A is the dominant structure. In Table 3 we listed the "line/kink" codes for all the 17 folded curves of Fig. 8.

It is interesting to observe that the ordering of the curves A - Q in Table 3 agree completely with the ordering of the same curves as given of Table 1. Both codes follow the same lexical ordering. Moreover, when the dominance relations between the new codes are examined they produce identical relations to those derived from the segment/length codes of Table 1. Hence, Fig. 10 illustrates also the partial order also for the line/kink codes of Table 3, as well as for segment/length codes of Table 1. A close look at the two codes reveals that they are simply related, and that one code can be transformed easily into the other. For example, consider again the line/kink code for structure A:

0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0

which can be also written briefly as a single binary number:

0 1 1 0 1 1 0 1 1 0 1 0 1 0 0 0.

The zeros in this code indicate vertices where "kink" occurs, or where a new line segment begins. Hence, in order to get segment/length code from the line/kink code we first add the entries between adjacent zeros to obtain: 2, 2, 2,

Table 3 The "line/kink" codes for the 17 folded curves of Fig. 8

Structure	Line/kink code	Partial Sums
A	0110110110101000	0, 1, 2, 2, 3, 4, 4, 5, 6, 6, 7, 7, 8, 8, 8
B	0110110110100000	0, 1, 2, 2, 3, 4, 4, 5, 6, 6, 7, 7, 7, 7, 7
C	0110110110010010	0, 1, 2, 2, 3, 4, 4, 5, 6, 6, 6, 7, 7, 7, 8
D	0110110100001010	0, 1, 2, 2, 3, 4, 4, 5, 5, 5, 5, 6, 6, 7, 7
E	0110110010101000	0, 1, 2, 2, 3, 4, 4, 4, 5, 5, 6, 6, 7, 7, 7
F	0110110010100000	0, 1, 2, 2, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6
G	0110110010010010	0, 1, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7
H	0110110000010100	0, 1, 2, 2, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6
I	0110011010110010	0, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 7, 7, 7, 8
J	0110011010001000	0, 1, 2, 2, 2, 3, 4, 4, 5, 5, 5, 5, 6, 6, 6
K	0110011010000000	0, 1, 2, 2, 2, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5
L	0110011001100110	0, 1, 2, 2, 2, 3, 4, 4, 4, 5, 6, 6, 6, 7, 8
M	0110010010011010	0, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6, 6, 7
N	0110000001101000	0, 1, 2, 2, 2, 2, 2, 2, 2, 3, 4, 4, 5, 5, 5
O	0110000001100000	0, 1, 2, 2, 2, 2, 2, 2, 2, 3, 4, 4, 4, 4, 4
P	0110000001010010	0, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4, 4, 4, 5
Q	0110000000101100	0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5

1, 1, 0, 0, 0. Then we add +1, to each entry to acknowledge the initial the initial vertex of each segment. Thus, instead of 2, 2, 2, 1, 1, 0, 0, 0 we obtain 3, 3, 3, 2, 2, 1, 1, 1, which is the segment/length code for A in Table 1. Hence, one can view the line/kink codes of Table 3 a binary version of the segment/length codes.

The so called "line/kink" code does not contain complete information for reconstruction of the structure. It could be changed just a bit. Instead of using 0, 1 consider use of labels S, R and L for straight, the right and the left, respectively. With such choice one would obtain SRL code that preserves the relevant information on the direction of the kinks. If one chooses numerical values for SRL code: S = 1, R = 0, L = 0, the line/kink code is obtained. The SRL code is reminiscent of the 0, ± 1 codes for conformations of n-alkanes embedded on a diamond grid that lead to an interesting graphical formulation for enumeration of n-alkane conformers [20].

Before closing this section let us mention that the rules of Muirhead for the construction of sequences of partial sums can be modified. Ruch [21-23] for example, while maintaining the inequalities as formulated by Muirhead has changed the condition on the last partial sum. Instead of requesting that the last partial sums are related by equality according to Ruch the equation is replaced by the inequality:

$$a_1 + a_2 + a_3 + \ldots + a_n \geq b_1 + b_2 + b_3 + \ldots + b_n.$$

Another generalization of Muirhead inequalities outlined by one of the present authors considers a repeated use of the partial sums to resolve cases of non comparability of sequences [24]. Thus starting with

$$s_1 \geq t_1$$

$$s_1 + s_2 \geq t_1 + t_2$$

$$s_1 + s_2 + s_3 \geq t_1 + t_2 + t_3$$

. . . . .

$$s_1 + s_2 + s_3 + \ldots + a_n = t_1 + t_2 + t_3 + \ldots + t_n$$

where

$s_1 = a_1$ $\qquad$ $t_1 = b_1$

$s_2 = a_1 + a_2$ $\qquad$ and $\qquad$ $t_2 = b_1 + b_2$

$s_3 = a_1 + a_2 + a_3$ $\qquad$ $t_3 = b_1 + b_2 + b_3$

. . . . . $\qquad$ . . . . .

by substituting the partial sums at each step back into the initial set of the inequalities one obtains:

$$a_1 \geq b_1$$

$$2a_1 + a_2 \geq 2b_1 + b_2$$

$$3a_1 + 2a_2 + a_3 \geq 3b_1 + 2b_2 + b_3$$

. . . . .

$$na_1 + (n-1)a_2 + (n-2)a_3 + \ldots + a_n = nb_1 + (n-1)b_2 + (n-2)b_3 + \ldots + b_n$$

By repeating the process again and again greater weights are obtained for the initial members of the sequences under comparison. In the following step we have:

$$a_1 \geq b_1$$

$$3a_1 + a_2 \geq 3b_1 + b_2$$

$$6a_1 + 3a_2 + a_3 \geq 6b_1 + 3b_2 + b_3$$

. . . . .

and so on. One could generalize even the above procedure by introducing general (non integer) weights w_1, w_2, w_3, . . . (which may even be members of yet another sequence). We will in this article, however, use the partial ordering rules as defined by Muirhead.

Discussion

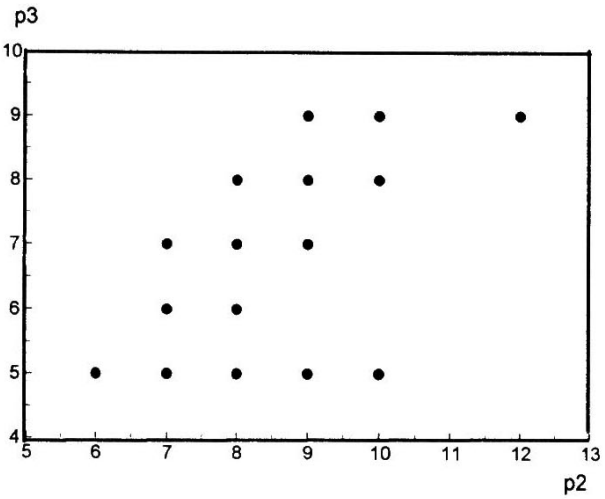
Different codes may lead to different partial ordering of the same set of structures. This has been already illustrated by Fig. 10 and Fig. 11, both of which show partial orders for the same 17 folded curves of Fig. 8. So how should one choose codes, or select one partial order over another? The situation is reminiscent of the situation characterizing consideration of molecular similarity [25, 26]. The same set of structures will yield a distinct similarity/dissimilarity table, depending on the set of invariants used for characterization of structures. Each such result is legitimate as any other, but depending on the application one set of descriptors may be found more useful

for some structure-property studies, the other may be of interest when different properties, or different structures are considered. Moreover, even when the same set of descriptors is used the entries of a similarity/dissimilarity table may very strongly depend on whether we used an orthogonalized set of descriptors or the descriptors remain interrelated [27-31].

Ordering of physicochemical properties of alkanes

Partial order, such as shown in Fig. 10 and Fig. 11 may be thought of as a "mathematical game." They become of interest in chemistry if they relate to questions and problems of chemical significance. That indeed they play an important role in discussions of isomeric variations has been shown by Randic and Wilkins [30-32]. For instance, if for octane isomers we correlate the paths of length three against the paths of length two we obtain the diagram shown in Fig. 12. Such diagrams have been introduce by Randic and Wilkins [32-34] to point to regularities in isomeric variations of selected properties of alkanes. For example, the boiling point, the critical density, the critical pressure, and the molar magnetic susceptibility show the (+, +) trend, that its, they increase with increase of p_2 and p_3. On the other hand the specific dispersion, the molar volumes, and chemical shift sums show the (+, -) trend, while the surface tension, the heat of combustion and the critical temperature have the reverse trend (-, +). The critical volume of

Fig. 12 Correlation for octane isomers of the paths of length three against paths of length two



octane was the only physicochemical property of a dozen considered that showed the (-, -) trend.

Applications of the partial order to molecular properties [35-40] showed that (p_2, p_3) coordinate system plays a role of displaying regular behavior of property of isomers and can be referred to as the Periodic Table of Isomer in analogy to the Periodic Table of Elements that displays regularities in variations of atomic properties.

Search for pharmacophore

Another illustration of use of partial order relates to search for pharmacophore. Pharmacophore is thought to be a group of atoms in a biologically active compound (not necessarily being connected and forming a fragment) believed to be responsible for its activity. In many situations molecule were found active even though identification of the atoms believed to play crucial role may be unknown. Hence, it is of considerable interest to identify pharmacophore. One way to such identification has been illustrated some time ago on nitrosamines [17, 41]. Among dozen mutagenic compounds illustrated in Fig. 13, the compounds labeled as A and B were reported to be the most potent. A glance at Fig. 13 does not reveal unusual structural features present in the most active compounds and absent in other less potent compounds. Seemingly all the compounds of Fig. 13 are reasonably similar, but as can be seen from their mutagenicities they show

enormous variation in their bio-activity. The difference between the most potent and the least potent compound is almost three orders in magnitude. How can we understand the enormous differences (three orders of magnitude) in the potency among all the dozen seemingly similar structures?

A way to attack this problem is first to introduce mathematical characterization of the structures and subsequently compare such characterizations, for a whole molecule, or only relevant parts, rather then comparing chemical structures. For example, if the structures are characterized by (weighted) path numbers one can compare derived path sequences. This is the approach outlined in ref. [41]. First one takes A and B as the standard compounds (being the most mutagenic) to which other are compared will be compared. However, rather than comparing the whole sequence for different compounds, what one would do if one is interested in the overall similarity among the compounds, one selects a *fragment* present in all compounds and consider mutual similarities for so selected fragments. As a measure of similarity one can use the Euclidean distance in n-dimensional structure space by considering the characterization of compounds by weighted path sequences as the representation of the structures by n-dimensional vectors. The relative degree of the similarity/dissimilarity will depend on the fragment selected and used for a comparison. In Fig. 14 and Fig. 15 we illustrate the partial orders derived when one uses a six or a seven atom molecular fragments. The corresponding molecular fragments are shown at the bottom of Fig 14 and

Fig. 13 Mutagenic compounds listed in Table 4 and a six and a seven atom molecular fragments used for measuring similarity

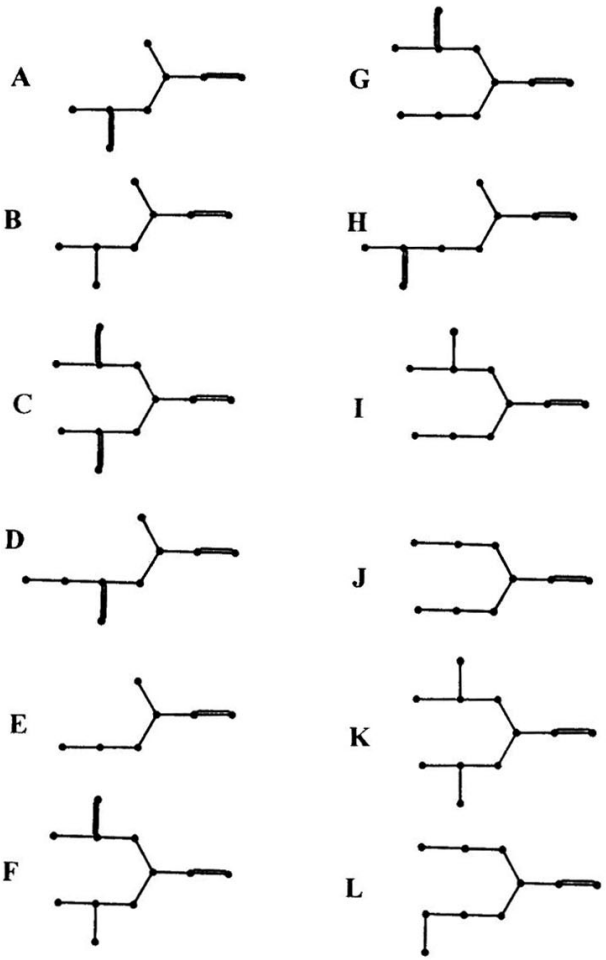


Fig. 14 Ordering of nitrosamines based on similarity towards A and B using information on six atom fragments only

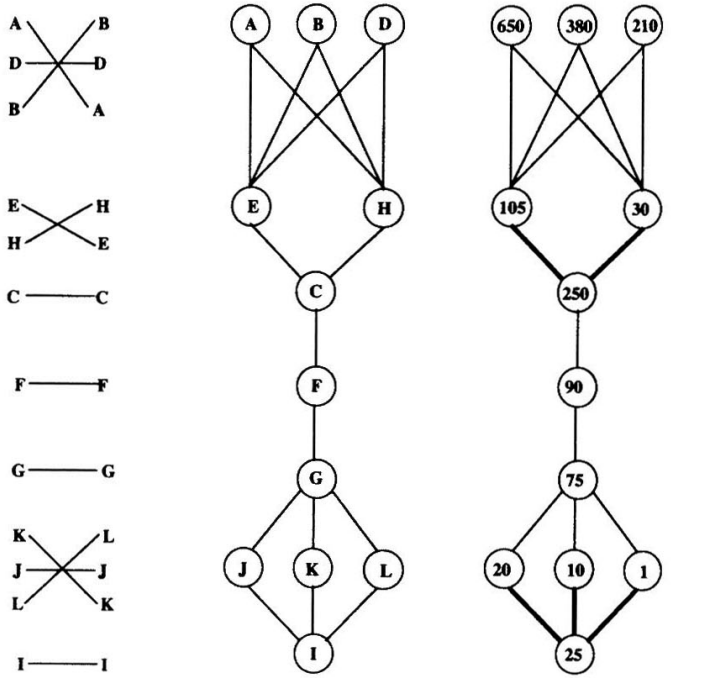
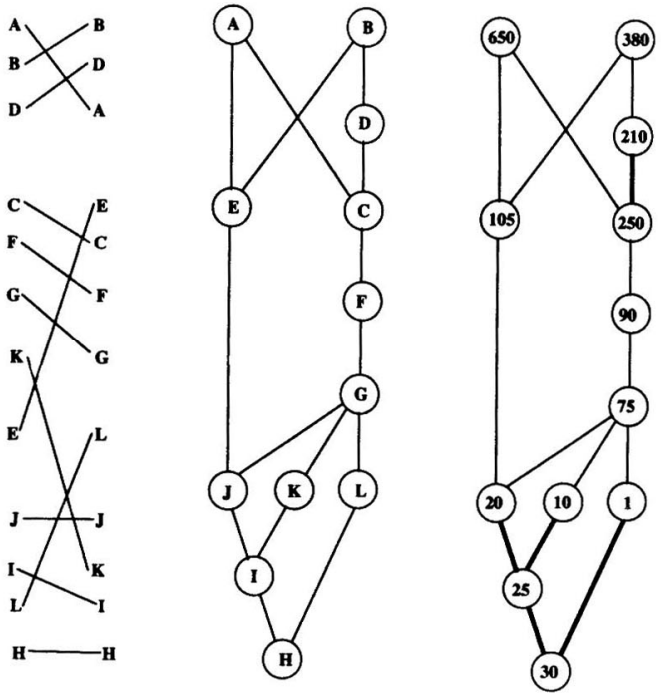
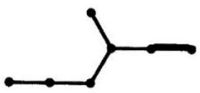


Fig. 15 Ordering of nitrosamines based on similarity towards A and B using information on seven atom fragments only





7 ATOM FRAGMENT

Fig. 15, respectively. In the next step one replaces the labels for structures by the numerical values of mutagenicity of a structure. In this way we obtain a numerical diagram which may or need not conform with the dominance hierarchy found for the structures. We have indicated by a thick lines local inconsistencies, i. e. , the local situations in which there is a reversal of the relative magnitudes for the mutagenicities for the neighboring compounds. Ideally we would like to see diagram of partial order without numerical inconsistencies, that is, without the thick lines. As we see in the case of Fig. 14 there are several major inconstancies. In particular compounds C is below E and H, both of which show considerably lesser mutagenicities. On the other hand there is no major inconsistency with Fig. 15. It is true that the compounds D is above the compounds C, but their mutagenicities are not so different as was the case with the compound D and E, H of Fig. 14. If we ignore the less important parts of the diagrams concerning the compounds of low activity, we see that in the case of the six-atom fragment we have serious contradictions but these disappear when the seven atom fragment is considered as pharmacophore. Hence, we conclude that the seven atom fragment much better characterize relative mutagenicities and can therefore be considered the sought pharmacophore for the mutagenicity of nirtosamines.

A convenient way to arrive at the diagrams for partial orders illustrated in Fig. 14 and Fig. 15 is illustrated at the left part of each figure. Nitrosamines are ordered in two columns relative to the calculated similarity

with respect to A and B based on considered fragments when only six (Fog. 14) or seven atom-fragment (Fig. 15) is considered as pharmacophore. 1-Dimensional order of compounds based on their similarity with respect to a single compound, A and B, are different for the six and the seven atom fragments. From the two 1-dimensional orders that hold for both A and B separately one has to extract the partial order. The partial order incorporates all local ordering of the compounds that satisfy the ordering that holds for A and B considered separately. Typically a partial order is represented in a form of an oriented diagram such that paths in such diagram simultaneously satisfy the relative orders for the compounds when both A and B are selected as the leading standard compound. Each crossing of lines between the same label in the two columns at the left parts of both Fig. 14 and Fig. 15 indicates structures that are not comparable.

We hope that the case discussed well illustrates the importance of the partial order for structure-property-activity studies. Additional similar application of partial ordering in QSAR (the quantitative structure-activity relationship) have been published [42-44], including modeling the mutagenicity of nitroarenes [42], and modeling of the antiviral activity of substituted benzimidazoles [44]. Before leaving this aspects of applications of partial orders in structure-property-activity studies we would like to point out to those interested in constructions of partial orders that importance of such diagrams, which we hope will increase and which we feel deserves wide publicity, lies in rationalization of experimental data. Mere diagrams of a

partial order for set of structures would be of limited interest *per se* without subsequent application. There is some parallelism here with a similar situation in designing or searching for novel topological indices. Without a clear demonstration of some advantages of novel descriptors over the existing indices, such efforts are likely to remain unappreciated, and rightly so, because they are of no consequence for chemistry.

Folding and the Degree of Folding

In this section we will illustrate use of the partial order illustrated in Fig. 10 on selected mathematical properties of the folded curves of Fig. 8. Quantitative measure of folding, an index of the "degree of folding," has been suggested by Randic, Kleiner and DeAlba [8]. The index is defined as a normalized leading eigenvalue of the so called D/D matrix of a structure. The leading eigenvalue is the largest positive eigenvalue of a matrix. The matrix elements of D/D matrix are defined as the quotient of the Euclidean distance between points (i, j) and the graph theoretical distance between the same points. In Table 4 we show D/D matrix for structure A of Fig. 8. The vertices for each structure are labeled starting with 1 at the origin (0, 0) and ending with 16 at the end of the folded curve. The leading eigenvalues λ_1 of the D/D matrices for structures A - Q are listed in Table 5 (the left column) together with normalized eigenvalues $\lambda_1/16$ which gives "the degree of

Table 4

	1	2	3	4	5	6	7	8
1	0	1/1	2/2	3/3	√10/4	√13/5	√18/6	√13/7
2		0	1/1	2/2	√5/3	√8/4	√13/5	√10/6
3			0	1/1	√2/2	√5/3	√10/4	3/5
4				0	1/1	2/2	3/3	√10/4
5					0	1/1	2/2	√5/3
6						0	1/1	√2/2
7							0	1/1
8								0

	9	10	11	12	13	14	15	16
1	√10/8	3/9	2/10	1/11	√2/12	√5/13	√8/14	√5/15
2	3/7	√10/8	√5/9	√2/10	1/11	√2/12	√5/13	2/14
3	√10/6	√13/7	√8/8	√5/9	√2/10	1/11	2/12	√5/13
4	√13/5	√18/6	√13/7	√10/8	√5/9	√2/10	√5/11	√8/12
5	√8/4	√13/5	√10/6	3/7	2/8	1/9	√2/10	√5/11
6	√5/3	√10/4	3/5	√10/6	√5/7	√2/8	1/9	2/10
7	2/2	3/3	√10/4	√13/5	√8/6	√5/7	√2/8	√5/9
8	1/1	2/2	√5/3	√8/4	√5/5	2/6	1/7	√2/8

	9	10	11	12	13	14	15	16
9	0	1/1	√2/2	√5/3	2/4	√5/5	√2/6	1/7
10		0	1/1	2/2	√5/3	√8/4	√5/5	√2/6
11			0	1/1	√2/2	√5/3	2/4	1/5
12				0	1/1	2/2	√5/3	√2/4
13					0	1/1	√2/2	1/3
14						0	1/1	√2/2
15							0	1/1
16								0

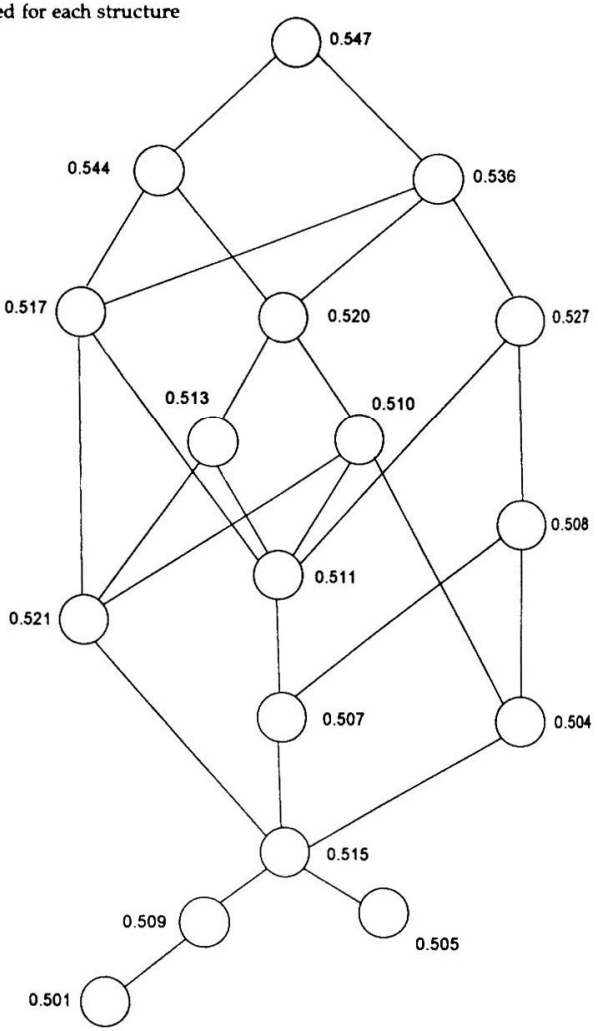
Table 5

Structure	λ_1	$\varphi(\lambda_1/16)$	λ_1^*
A	8.758812	0.547426	3.984105
B	8.707504	0.544219	3.978066
C	8.580271	0.536267	3.945648
D	8.276574	0.517286	3.754304
E	8.326970	0.520436	3.691089
F	8.208253	0.513016	3.690685
G	8.153800	0.509612	3.689218
H	8.330142	0.520634	3.682369
I	8.438829	0.527427	3.618597
J	8.177920	0.511120	3.443928
K	8.119988	0.507499	3.443615
L	8.121855	0.507616	3.472515
M	8.059103	0.503694	3.349142
N	8.240244	0.515015	3.349824
O	8.138308	0.508644	3.184506
P	8.012491	0.500781	3.098096
Q	8.077203	0.504825	3.343237

folding" (the central column). The normalization is particularly important when one wants to compare the degrees of folding in curves of different size.

In Fig. 16 we show again the partial order already depicted in Fig. 9, except that instead of using the labels A-Q for the structures we replaced each label by the numerical value of the computed leading eigenvalue λ_1 for the structure. We see the λ_1/n decreases from the top of the figure towards the bottom along the lines of the diagram, with few, but minor, exceptions. The few discrepancies that occurred concern a somewhat "higher" value of λ_1 for H and N, and possibly somewhat low value of λ_1 for structure G. The partial ordering, just as a correlation, may have occasional "outlier." Nevertheless, we may conclude that an "agreement" between the prescribed partial order of Fig. 9 and the "experimental" order shown in Fig. 16 may be viewed as very satisfactory. Hence, the segment length clearly play a dominant role in determining some properties of the folded curves. However, the minor discrepancies equally point to possible role of factors that the adopted code does not register. One such factor may be the end-to-end distance of a curve. Other features being similar, structures that have the end vertices adjacent, or at smaller separations, will appear more folded than the similar structures with the end points at larger separations. Hence, more elaborate coding that would incorporate such additional structural elements may possibly result in even better agreement between the "calculated" and the "experimental" partial orders for the considered property.

Fig. 16 The partial order of Fig. 10 with the "degree of folding" values inserted for each structure



Line Adjacency

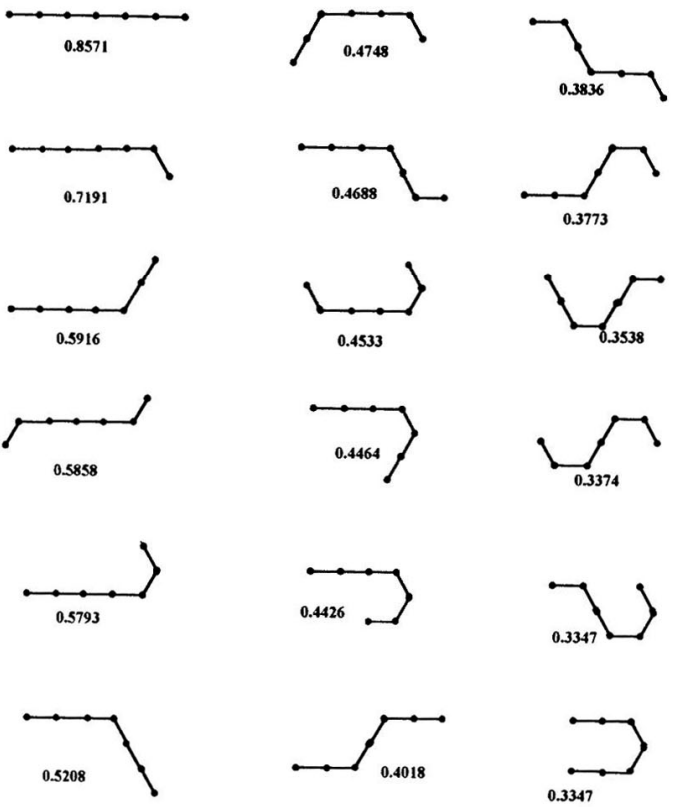
As another property of the folded curves of Fig. 8 we will consider the leading eigenvalue λ_1 of their "line adjacency" matrix. In order to differentiate this leading eigenvalue from the already considered leading eigenvalue of D/D matrices we will denote it as λ_1^*. The concept of the line-adjacency has arisen in studies of D^k/D^k matrices when one considers the limit matrix elements as k tends to infinity [2, 9, 41]. Here D^k/D^k is matrix derived from the elements of D/D matrix by raising each element of the matrix to the power k. Since all the elements of the D/D matrix are smaller than 1, or at most equal to 1, it follows that as k tends to infinity all the elements that are smaller than one will become zero in the limit, while those that are equal to 1 remain intact. The resulting matrix can be geometrically interpreted as showing "line-adjacency," that is, its elements are equal to 1 whenever vertices lie on a line. In Table 6 we show the line adjacency matrix for the folded structure A of Fig. 8.

The line-adjacency matrix is of interest in numerical characterization of DNA primary sequences [46]. The leading eigenvalue of such matrices is an invariant closely related to the measure of the degree of folding of such structures. So we decided to calculate the leading eigenvalue λ_1^* of line-adjacency matrices which are listed in Table 5 (the right column). In Fig. 17 we superimposed the line-adjacency leading eigenvalues on the partial order of Fig. 9. As we can see again a very satisfactory ordering was obtained. In

Table 6

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
6	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	1	1	1	0	1	1	1	0	0	0	0	0	0
8	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0
9	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0
11	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Fig. 17 The partial order of Fig. 6 with the leading eigenvalues of the line-adjacency matrices inserted for each structure



this instance there is not a single, even minor, discrepancy in relative positioning of the numbers in the diagram! Hence, the leading eigenvalue of the line-adjacency fully parallels the ordering based on segment/length of a folded curve. We may mention that the leading eigenvalue of the line-adjacency matrices were interpreted in one particular chemical application as an index of molecular flexibility [9]. In Fig. 17 we show the embedded line graphs for the 18 structures of Fig. 4 and their "flexibility" indices, merely to illustrate that despite some parallelism between the 'folding" and the "flexibility" indices, there are also important differences between them. Hence, the line adjacency matrix, which is in mathematical literature known as the adjacency matrix of a "Menger graph of a configuration" [47], may have wider applications in chemistry.

Acknowledgment

We acknowledge the Ministry of Science and Technology of Slovenia for partial support of this work through grant J1-8901-0104-97, and also the United States Air Force Office of Scientific Research through grant F-49620-96-1-0330. We thank D. J. Klein (Texas A&M University, Galveston, Texas) for an invitation and several useful suggestions. Also we received valuable suggestions from Professor T. Pisanski (University of Ljubljana, Slovenia) and Dr. L. Bytautas (Ames Laboratory, Ames, Iowa). This is contribution number 263 from the Center for Water and the Environment of the Natural Resources Research Institute of the state of Minnesota.

References

1 H. Li, R. Helling, C. Tang and N. Wingreen, Science, **273**, 666 (1996).

2 M. Randić and G. Krilov, Int. J. Quantum Chem. **75**, 1017 (1999).

3 M. Randić and G. Krilov, Chem. Phys. Lett., **272**, 115 (1997).

4 E. F. Beckenbach and R. Bellman, *Inequalities*, Springer Verlag, Berlin (1961).

5 A. Nandi, Current Science, **66**, 309 (1994).

6 A. Nandi, Comput. Appl. Biosci., **12**, 55 (1996).

7 M. Randić, M. Vračko, S. C. Basak, and A. Nandy, J. Theor. Biol. (submitted).

8 M. Randić, A. F. Kleiner and L. M. DeAlba, J. Chem. Inf. Comput. Sci., **34**, 277 (1994).

9 M. Randić, M. Vračko and M. Novič, *Eigenvalues as Molecular Descriptors*, in: *QSAR/QSPR by Molecular Descriptors*, (M. V. Diudea, ed.), Nova Science Publ., Commack, N. Y. (in press).

10 B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco (1982).

11 H. -O. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals*, Springer-Verlag, Berlin (1992).

12 H. von Koch, Ark. Mat., **1**, 681 (1904).

13 D. Hilbert, Math. Ann. 38, 459 (1891).

14 W. Sierpinski, C. R. Acad. Sci., 160, 302 (1915).

15 The Dragon curve was discovered by J. E. Heighway and analyzed by J. E. Heighway, W. G. Harter and B. A. Banks (see ref. 16).

16 M. Gardner, *Mathematical Magic Show*, Chapter 15, Vintage Books, New York, (1978).

17 M. Randić and M. Razinger, *On Characterization of 3D Molecular Structure*, Chapter 6 in: *From Chemical Topology to Three-Dimensional Geometry* (A. T. Balaban, ed.), Plenum Press, New York (1997).

18 L. Bytautas, D. J. Klein, M. Randic, and T. Pisanski, Proc. DIMACS Conference, Rutgers Univ., March (1998) (in press).

19 R. F. Muirhead, Edinburgh Math. Soc., **21**, 144 (1903).

20 M. Randić, Int. J. Quantum Chem: Quantum Biol. Symp. **7**, 187 (1980).

21 E. Ruch, Acc. Chem. Res., **5**, 49 (1972).

22 E. Ruch, Theor. Chim. Acta, **38**, 167 (1975).

23 E. Ruch and A. Mead, Theor. Chim. Acta, **41**, 95 (1976).

24 M. Randić, Chem. Phys. Lett., **55**, 547 (1978).

25 M. Randić, Design of Molecules with Desired Properties. *Molecular Similarity Approach to Property Optimization*, In: *Concept and Applications of Molecular Similarity* (M. A. Johnson and G. Maggiora, Eds), Wiley, New York (1990), pp. 77-145.

26 M. Randić, *Similarity Methods of Interest in Chemistry*, in: *Mathematical Methods in Contemporary Chemistry* (I. S. Kuchanov, Ed.) Gordon & Breach Science Publ., Amsterdam (1996), pp. 1-100.

27 M. Randić, New J. Chem., **15**, 517 (1991).

28 M. Randić, J. Chem. Inf. Comput. Sci., **31**, 311 (1991).

29 M. Randić, J. Comput. Chem., **14**, 363 (1993).

30 M. Randić, Int. J. Quantum Chem: Quantum Biol. Symp. **21**, 215 (1994).

31 M. Randić, J. Chem. Inf. Comput. Sci., **36**, 1092 (1996).

32 M. Randić and C. L. Wilkins, Chem. Phys. Lett., **63**, 322 (1979).

33 M. Randić and C. L. Wilkins, J. Phys. Chem., **83**, 1525 (1979).

34 M. Randić and C. L. Wilkins, Int. J. Quantum Chem. **18**, 1005 (1980).

35 M. Randić, J. Magn. Res., 39, 431 (1980).

36 M. Randić and N. Trinajstic, MATCH, **13**, 271 (1982).

37 M. Randić, Int. J. Quantum Chem. **23**, 1707 (1983).

38 M. Randić, J. Chem. Educ., **69**, 713 (1992).

39 M. Randić and N. Trinajstić, New J. Chem. 18, 179 (1994)

40 D. J. Klein and D. Babić, J. Chem. Inf. Comput. Sci., **37**, 656 (1997).

41 M. Randić, B. Jerman-Blažič, D. H. Rouvray, P. G. Seybold, and S. C. Grossman, Int. J. Quantum Chem: Quantum Biol. Symp. **14**, 245 (1987).

42 M. Randić, S. C. Grossman, B. Jerman-Blažič, D. H. Rouvray and S. El-Basil, Math. Comput. Modeling, **11**, 837 (1988).

43 M. Randić, B. Jerman-Blazic, S. C. Grossman, and D. H. Rouvray, Math. Modelling, **8**, 571 (1986).

44 M. Randić and P. C. Jurs, Quant. Struct.-Act. Relat., **8**, 39 (1989).

46 M. Randić, A. Nandy, and S. C. Basak, J. Math. Chem. (submitted).

47 H. S. M. Coxeter, Bull. Amer. Math. Soc., **56**, 413 (1950).