Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 290–305
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

On representation of proteins by star-like graphs

Milan Randić [a,*], Jure Zupan [a], Dražen Vikić-Topić [b]

[a] National Institute of Chemistry, Ljubljana, Slovenia
[b] Institute Rudjer Bošković, Zagreb, Croatia

Received 4 September 2006; received in revised form 8 December 2006; accepted 8 December 2006
Available online 15 December 2006

Abstract

To arrive at graphical representations of proteins one is confronted with number of arbitrary decisions how to assign the 20 natural amino acids to equivalent or non-equivalent sites of underlying geometrical objects used for construction of their graphical representation. Here we consider representation of proteins based on generalized star graphs, which are graphs with one vertex of maximal degree in the center to which are attached other vertices of either degree one or two. The matrix representation of proteins based on star-like graphs has an important advantage in that, while its pictorial representation depends on selected assignment of amino acids to various branches of star graph, its properties do not depend on the adopted assignment of vertices to amino acids. Hence, the derived graph invariants, devoid of artifacts associated with graphical representations of biosequences, will better reflect upon the inherent properties of protein structure. We describe several graph invariants, mostly extracted from distance matrices of star-like graphs, which can serve as protein descriptors. The approach is illustrated on strand A of the human insulin.

Keywords: Star-like graphs; Graphical representation of proteins; Human insulin; Line distance matrix

1. Introduction

Graphical representations of DNA were initiated in 1985 by Hamori [1] and Gates [2], whose pioneering work was soon followed by introduction of alternative such representations [3–5]. Graphical representations offer visual qualitative inspection of similarities and dissimilarities among DNA sequences. An important advance followed some 15 years later when graphical representations were supplemented by development of accompanying quantitative analysis [6,7], which resulted in numerical characterization of DNA. The numerical characterizations of DNA sequences that followed have been based on properties various matrices constructed for considered 2-D graphical representation of DNA. A bonus of these developments is arrival of 3-D graphical representations of DNA [6,8] as well as non-graphical representations of DNA [9–11], both of which has offered additional mathematical descriptors for DNA. Most of these approaches involve arbitrary choices for the assignment of nucleic bases to vertices of the selected underlying graphical object. This is the case with the assignment of the four bases to the four corners of a square in the approach of Jeffrey [3], the four directions of the x, y coordinate axes in the approach of Gates [2], Leong and Morgenthaler [4] and Nandy [5], the order in which the four horizontal lines are assigned to four nucleotides [12,13], or the binary labels used to differentiate individual bases [14]. Although any selection is equally legitimate, such graphical representation, while reflecting some inherent properties of DNA sequences, are not unique, because alternative graphical representations lead to different numerical characterizations.

In the case of DNA it is possible to arrive at unique (up to symmetry operations) graphical and non-graphical representations, although this is achieved at loss of the simplicity of accompanying 2-D graphical representations [9,15–26]. For instance, if one assigns the four nucleotides A, C, T and G to the four corners of tetrahedron, instead of the four corner of a square, then because the four labeled vertices of tetrahedron are fully equivalent one obtains a unique 3-D graphical representation of DNA [13]. The same is true with 4-D representation of DNA, where the four unit vectors of 4-D space are all mutually equivalent [9]. However, graphical representations and visual inspection of line in 3-D space is more tedious and impossible in the case of 4-D representation of DNA.

* Corresponding author. Permanent address: 3225 Kingman Road, Ames, IA 50014, United States. Tel.: +1 515 292 7411; fax: +1 515 292 8629.
E-mail address: mrandic@msn.com (M. Randić).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.12.006

It is significant that it took 20 years since the introduction of graphical representations of DNA for the arrival of the first graphical representation of protein sequences [19]. This ought not to be surprising in view that in the case of proteins instead of the 4-letter sequences (based on A, C, G, T) for the 4 nucleotides we have 20-letter sequences for the 20 natural amino acids. Hence, instead of considering one choice among 4! possibilities, which at most mean 24 alternatives, or less in the case of symmetry, we would have 20! possibilities. Even if one employs highly symmetrical object (e.g., regular dodecahedron) this would still leaves too many steps in which arbitrary decisions are to be made in order to arrive at a graphical representation of protein structure. Soon after the first graphical representation of proteins followed additional alternative representations of proteins [27–32], including a representation in 20-D space, which surprising as it may sound, also allows graphical representation of proteins—this time by depicting set of computed proteins invariants [33]. We will in this contribution describe a particular construction of graphical representation of proteins, which is remarkable in that it is not sensitive to arbitrary assignments of amino acid to parts of associated graphical representation of graph used to represent protein.

2. Graphical representation of proteins

First we will briefly outline currently available graphical representations of proteins, and will follow with a description of a novel graphical and numerical representation of proteins, which as we will see does not depend on arbitrary assignment of the 20 natural amino acids to vertices of graphs considered. The approach will be illustrated on strand A of human insulin. We will introduce also several novel protein invariants, some of which appears particularly attractive in that it allows representation of proteins by bar graphs. We will refer to such representation, derived from the distance matrices of generalized star graphs, as protein ''profile'' in analogy with similar graphical constructions for small molecules, which have been referred to as molecular ''profiles.'' The term ''profile'' may have not been the most fortunate as it may insinuate the outline of a molecular model of a protein, but in a broader sense the word ''profile'' relates to ''any outline or contour'' [34], and this is the sense in which we have used this term.

It will be clear from the short overview of the four existing graphical representations of proteins that all of them are based on one particular selection among too many for the assignment of amino acids to vertices of graphs or geometrical objects considered. It almost appears inconceivable that one can arrive at graphical representation of proteins that will not involve arbitrary choices among enormous number of possible alternatives. However, if such representation is possible it would reflect the innate structure of the protein sequence, rather then the apparent structure of a legitimate but arbitrary graphical representation of the protein. As we will see in the next section, even if it may sound impossible, it is possible to arrive at a graphical representation of proteins that is free of arbitrary selections! The significance of this is difficult to overlook. The ''secret'' of our approach is that we took advantage of the well-known properties of graphs, that they have no unique graphical representation! While this feature of graphs continues to represent a source of difficulties, particularly when generating or enumerating graphs, which requires testing graphs for graph isomorphism, the problem where one has to establish whether two apparently different graph are isomorphic or not, here this very undesirable property allows one to associate different forms of the same graph with a single protein.

The four existing graphical representations of proteins were all introduced within the last year. The most straightforward of these graphical representations of proteins, which is limited to proteins with known m-RNA, is based on 2-D graphical representations of DNA. When one knows the underlying codons (triplets of DNA that code amino acids for a protein), construction of graphical representation of proteins follows directly from corresponding graphical representation of DNA. Hence, it involves the same degree of arbitrariness as does the corresponding DNA graphical representation. The first such graphical representation of proteins was described by Randić and Zupan [19], is based on a modification of the highly compact 2-D graphical representation of DNA considered by Jeffrey. This graphical representation of DNA introduced in 1990 by Jeffrey, is based on the Chaos Game [35], an algorithm originally described by Barnsley in 1988, in which one picks a point at random inside a regular polygon and follows by drawing the next point at a fraction of the distance between the selected point and one of n-vertices of the polygon selected at random. One continues this process indefinitely, obtaining at the end various fractal-like geometrical structures with higher or lesser symmetry [35]. Jeffrey has selected a square as a polygon of choice and assigned to its four corners the four bases A, C, G and T. Instead of picking point and random here one starts at the center of the square and moved half way towards the corner having the labels of the first nucleic base in a DNA sequence. One then continues from this point half way towards the corner having the label of the second base in the DNA sequence, and so on. In this way, DNA sequence having N nucleic acids is represented by N points within the interior of the square. If now one associate to each trio of nucleotides that define a triangle a single amino acid one can represent it by a spot placed in the center of the triangle. In this way, one immediately obtains graphical representation of the protein associated with DNA codons. Although the above construction is straightforward and elegant its application is rather limited because it requires knowledge of codons involved in production of proteins, which is in many cases not known.

This limitation, however, has been lifted by construction of a fictitious virtual genetic code [36], which assigns to each amino acid for which there are alternative codons a *single* codon. In this way, a DNA–protein mapping becomes unique, which allows use of graphical representations of DNA to be extended to proteins. It is true that the selected virtual genetic code is one of $6^3\,4^5\,3\,3^9$ possible such virtual codes in view of there being three amino acids associated with six codons, five with four codons, one with three and nine with two codons. However,

once the notion of the virtual genetic code is accepted the construction of graphical representations of proteins is reduced to that of modifying graphical construction of accompanied virtual DNA sequence.

Another promising direction for graphical representation of proteins, which does not require use of the virtual genetic code, is based on tables of codons, in which the 64 triplets of nucleic bases are arranged in a 8×8 table [17,28–32]. The assignment of individual triplets within the table is *uniquely* determined by following the algorithm of Jeffrey for construction of highly condensed representation of DNA. However, in contrast to Jeffrey, who considered sequences of DNA having 100,000 bases and more here one uses the same algorithm, but one terminates it deliberately after each third nucleotide.

The third recent representation of proteins is based on an alphabetic, which is essentially arbitrary ordering because the alphabet represents a convention, arrangement of 20 amino acids on a circumference of a circle. Graphical representation is followed by construction that is somewhat analogous to that of Jeffrey used for highly compact representation DNA, however, now instead of square with labels of 4 bases we have 20 amino acids forming a circle [27].

Finally, we have already mentioned the 20 vertices of regular dodecahedron. Bai and Wang [37] used 3-D Cartesian coordinate system to represent protein sequences and assigned 20 amino acids to the 20 directions in 3-D space defined by vectors from the center of the dodecahedron to its 20 vertices. As a result they obtained graphical representation of a protein as spatial zigzag curve, which is then numerically represented by a selected distance matrix.

3. Representation of proteins by star graphs

In Graph Theory, a star graph is defined as a graph on N vertices, one of degree $N - 1$ and $N - 1$ vertices of degree one [38]. We will generalize the concept of star graph by allowing in addition to the central vertex and the terminal vertices that ordinary star graphs possess presence of any number of vertices of degree two. In order to avoid confusion with the standard ‘‘star’’ graphs one could refer to here considered generalized form of the star graphs as the ‘‘shining star’’ graphs, but when no confusion is likely we will simply refer to our graphs simply as star graphs. As is well-known trees (of which star graphs are special case) have a property that there is a unique path between any pair of vertices. Moreover (which may be not so widely known), trees can be fully represented by a matrix involving only the distances between the terminal vertices. We will use both these properties for construction of various distance matrices associated with our star graphs.

In Fig. 1, we illustrate three alternative forms of a generalized star graph having central vertex and 21 vertices distributed in its 11 branches. These are three out of thousands different possible graphical representations of the same graph. Although there are multitude of graphical representations of this graph, most of which will have different adjacency and different distance matrices, all these matrices will nevertheless produce identical matrix invariants, because by definition the graph invariants do not depend on labeling of vertices or on the pictorial representation of the graph. Observe that all the 3 graphs of Fig. 1 have 11 terminal vertices, they have 3 branches of length one, 7 branches of length two, and 1 branch of length four, hence, they represent the same graph.

The graph of Fig. 1 allows $(3!) \times (7!)$ or 30,240 different assignments of amino acid to its vertices if one knows which 3 amino acids occurs once, which 7 amino acids occur twice and which single amino acids occurs 4 times. Hence, the same graph depicts over 30,000 proteins. This point to a colossal loss of information that typically accompanies various graphical representations of biosequences. However, as will be described later in more details, some of the lost information can be recovered when labels are assigned to vertices of such graphs. The graph of Fig. 1 stands for graphical representation of the A strand of human insulin, which involves the following 21 amino acids:

Gly-Ile-Val-Glu-Gln-Cys-Cys-Thr-Ser-Ile-Cys-Ser-Leu-Tyr-Gln-Leu-Glu-Asn-Tyr-Cys-Asn.

As we can see here Cys appears four times, Asn, Gln, Glu, Ile, Leu, Ser and Tyr appear twice, while Gly, Thr and Val

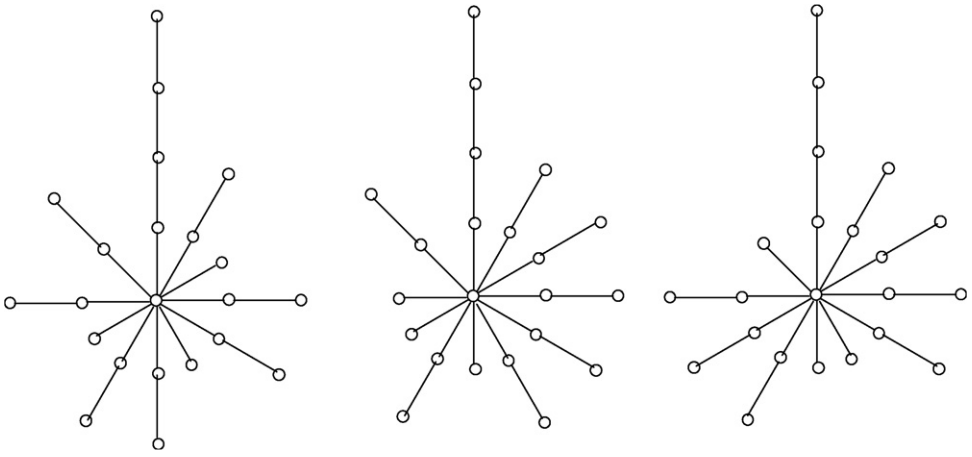

Fig. 1. A star graph depicting three out of many possible geometrical realizations of the same graph.

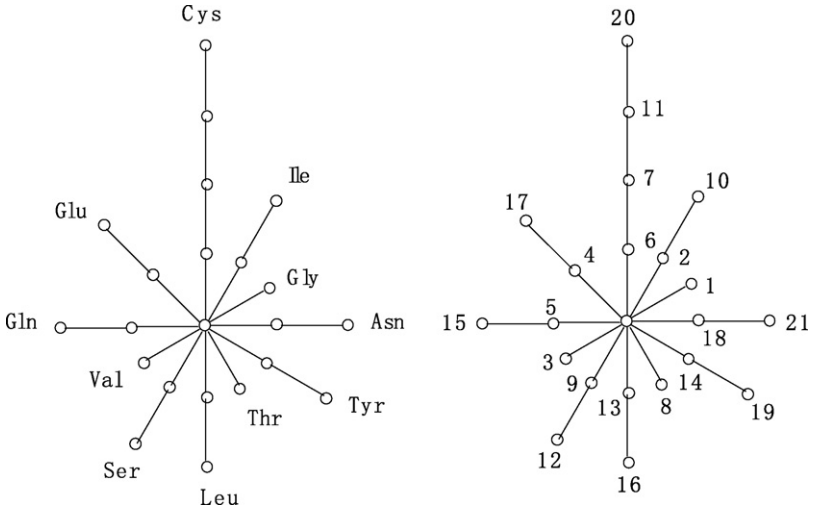

Fig. 2. Left: The star graph of Fig. 1 which represents a protein having the following amino acids: one Gly, Thr and Val, two Asn, Gln, Glu, Ile, Leu, Ser and Tyr, and four Cys (which are amino acids of the A strand of human insulin). Right: Labeled star graph by the sequential labels in which amino acids occur in the A strand of human insulin.

appear once. In Fig. 2 (left), we have depicted the first graph of Fig. 1 with the labels for the 11 branches starting with Gly and continuing labeling branches in order in which the 11 amino acids of A strand of human insulin appear, ending with Asn. Observe also that the central vertex of the star, which represents the ''origin'' of the star, does not belong to any of the amino acids. Fig. 2 (right) again depicts the same graph but this time we have indicated by labels 1–21 vertices of the graph, where labels indicate the location of individual amino acids in the sequence.

4. Matrix invariants

Before continuing let us comment on matrix invariants to be used as descriptors of proteins. By depicting proteins by star-like graphs we will arrive at representation of proteins, which will be numerically represented by matrices belonging to the accompanied graphs. Properties of these matrices thus became the subject of exploration and eventually serve for characterizations of proteins. Graphs and matrices have many properties but of particular interest are those properties of matrices that do not depend on labeling of vertices of graphs, which are known as matrix invariants. Such are eigenvalues, including in particular the leading eigenvalue of a matrix. In addition one can consider construction of a set of ''higher order'' matrices and their leading eigenvalues which thus lead to additional matrix invariants. We will outline later in particular a set of invariants, which in the case of smaller molecules have been referred to as molecular ''profile,'' and which in the case of star graphs representing proteins lead to analogous ''profiles'' that can be used as protein description.

The eigenvalues of a matrix represent a unique diagonal form of a matrix obtained when eigenvectors are selected as the basis for matrix representation. In general it is not easy and need not be possible to interpret the eigenvalues of a graph in terms of simple straightforward structural concepts. In the case of regular graphs (i.e., graphs in which all vertices have the same degree), the leading eigenvalue equal to the degree of vertices of the graph. A theorem of Matrix Theory [39] tells that the leading eigenvalue of a symmetric matrix is bounded by the largest and the smallest row sums of the matrix. Thus, if the row sums do not vary considerably, the leading eigenvalue approximates the average row sum, which differs from the average matrix element or the ''Wiener number'' of the matrix [40] only by a normalization factor. This is not apparent from the definition of the Wiener index, which according to Wiener is calculated by multiplying the number of atoms on each side of a bond and adding contributions of all bonds. However, Hosoya [41] has shown that Wiener number (for trees) can also be calculated by adding all entries above the main diagonal of the distance matrix of a graph. In the case of trees Lovasz and Pelikan [42] have interpreted the leading eigenvalue of the adjacency matrix of graphs as an index of ''branching'' of graphs. More recently it was suggested that the leading eigenvalue of a graphical matrix in which matrix element (i, j) is the leading eigenvalues of the path between vertices i and j may be even a better index of branching of skeletal graphs [43,44].

In the case of graphs embedded in 2-D or 3-D space such that they have fixed geometry (usually all edges have the same unit length and angles between the edges are either the trigonal angle of 120° or tetrahedral angle of 109°28′ associated with the graphite lattice in 2-D and the diamond lattice in 3-D, respectively) one can construct the so-called ''distance/distance'' or D/D matrix. The matrix elements (i, j) in a D/D matrix are given as a quotient of the Euclidian (''through the space'') distance between vertices i and j and the graph theoretical (''through the bonds) distance between the same two vertices. The leading eigenvalue of the D/D matrix of chain-like structures has been interpreted as a measure of the degree of bending, or the degree of folding, of the chain structures [45–50]. In Fig. 3, we have illustrated 18 possible conformers for a chain having 8 carbon atoms which are superimposed on a graphite lattice. Next to each structure we show the leading eigenvalue. Already visual inspection of

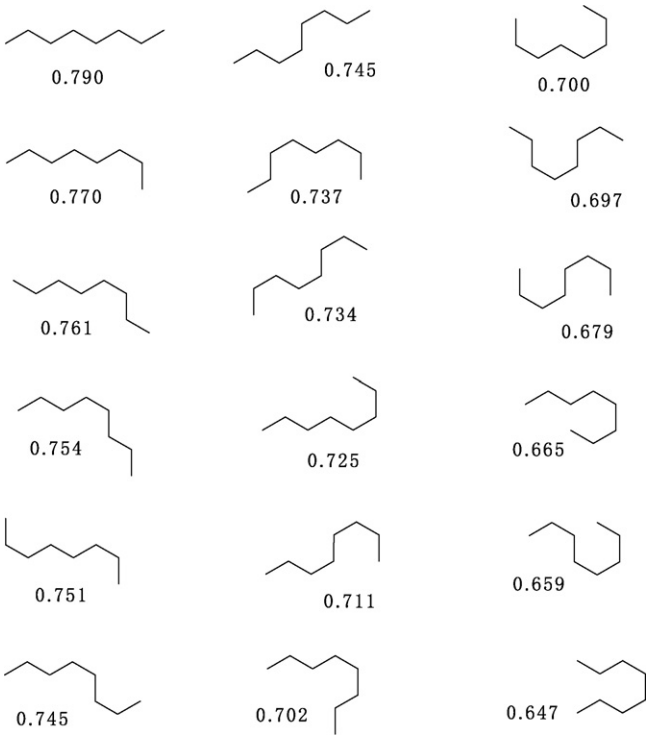

Fig. 3. The normalized leading eigenvalues for 18 configurations of a chain having 8 vertices obtained by their superimposition on the hexagonal (graphite) lattice illustrating the leading eigenvalue as an index of the degree of folding of a structure.

the relative magnitudes of the leading eigenvalues of the proposed index offer good numerical characterization of the degree of folding of shown chain structures. Theoretical justification for the interpretation of the leading value of D/D matrix as an index of folding of a structures comes from the already mention theorem of Matrix Algebra that bounds the leading eigenvalues by the largest and the smallest matrix row sums. Because as chain folds the Euclidean distances between vertices decrease, while the graph theoretical distances remain the same, thus necessarily the row sums will decrease and as a consequence the bounds on the leading eigenvalue will decrease.

In applications of D/D matrices additional molecular descriptors were obtained by considering the ''higher order'' ${}^{k}\mathrm{D}/{}^{k}\mathrm{D}$ matrices, which are obtained from D/D matrix by raising the individual matrix elements to higher powers, the operation known as element-by-element exponentiation [51,52]. In this way, one can construct k additional leading eigenvalues as protein descriptors. Because matrix elements of D/D matrix are smaller than one, or at most equal to one, when considering higher powers of their elements there is no need to introduce appropriate normalizing factor to ensure convergence of its eigenvalues. The ${}^{k}\mathrm{D}/{}^{k}\mathrm{D}$ matrices were also applied to non-chain structures, such as smaller molecules, ring systems, polyhedra and molecular surfaces [53–55]. In these systems, the leading eigenvalue can be interpreted as a measure of the compactness of a system, because in more compact systems the Euclidean distances will be relatively small and consequently the bounds on the leading eigenvalue will be decreased, implying thus that smaller magnitudes for the leading eigenvalue will characterize more compact systems.

5. Reduced matrices

Novel graphical representation of proteins by star graphs serve as an auxiliary stage for construction of accompanying numerical characterizations of proteins. We will now outline construction of graph invariants derived from reduced distance matrices associated with star graphs. There are two such matrices: (1) the TD matrix or the ''terminal distance'' matrix, in which only the distances between the terminal vertices have been considered; and (2) the LD matrix or the ''line distance'' matrix, in which distances between successive points that partition a line are matrix elements. We will first consider TD matrix of the graph representing the A strand of human insulin represented by Fig. 2 and will continue with LD matrix associated with the same graph by explaining how can one associate with star-like graph a partitioned line, which then leads to the line distance matrix.

5.1. Terminal distance matrix

In Table 1, we show the upper triangular part of the terminal distance matrix for the star-like graph of Fig. 2. The rows and columns of the matrix are labeled with labels of terminal vertices: 1, 3, 8 (for branches of length one), 10, 12, 15, 16, 17, 19, 21 (for branches of length two) and 20 (for branch of length four). The sequence 2, 2, 3, 3, 3, 3, 3, 3, 5, 3, which is the first row of the 11 × 11 matrix of Table 1 lists the distances from the vertex 1 to the remaining 10 terminal vertices. The sequence 2, 2, 3, 4, 4, 4, 4, 4, 6, 6, which are the matrix entries that appears just above the main diagonal indicates the separations between successive terminal vertices.

In Table 2, we show the same matrix augmented by replacing the zeros on the main diagonal by the distances of the corresponding terminal vertices to the central vertex. Augmented matrices have been occasionally considered in Chemical Graph Theory. For example, the variable connectivity index [56,57] has been constructed from the connectivity index [58] by replacing the zeros on the main diagonal of the

Table 1
The upper triangular part of the symmetric distance matrix for terminal vertices (TD) of the star graph of Fig. 2

	1	3	8	10	12	15	16	17	19	20	21	Row sum
1	0	2	2	3	3	3	3	3	3	5	3	30
3		0	2	3	3	3	3	3	3	5	3	30
8			0	3	3	3	3	3	3	5	3	30
10				0	4	4	4	4	4	6	4	39
12					0	4	4	4	4	6	4	39
15						0	4	4	4	6	4	39
16							0	4	4	6	4	39
17								0	4	6	4	39
19									0	6	4	39
20										0	6	57
21											0	39

Table 2
The upper triangular part of the augmented distance matrix for terminal vertices of the star graph of Fig. 2

	1	3	8	10	12	15	16	17	19	20	21	Row sum
1	1	2	2	3	3	3	3	3	3	5	3	31
3		1	2	3	3	3	3	3	3	5	3	31
8			1	3	3	3	3	3	3	5	3	31
10				2	4	4	4	4	4	6	4	41
12					2	4	4	4	4	6	4	41
15						2	4	4	4	6	4	41
16							2	4	4	6	4	41
17								2	4	6	4	41
19									2	6	4	41
20										4	6	61
21											2	41

adjacency matrix of molecular graphs by variables that could discriminate atoms of different kind [59–66], and even atoms of the same kind in different environment [67]. The so augmented matrices offer construction of additional invariants for characterization of complex systems. Table 2 reveals better than Table 1 an interesting underlying feature of the TD matrices of star-like graphs in that they are fully defined by the information given by entries of the main diagonal, which suffices for construction of full matrix. In this respect, TD matrices bear some similarity with LD matrices, and Toepliz and Hankel matrices of Matrix Theory, which are fully determined by entries of the first row. There is an additional parallelism between the TD and LD matrices: they both have all eigenvalues negative except for the leading eigenvalue, which is positive. This has been proved for LD matrices [68] and is a conjecture for the TD matrices.

In Table 3, in the columns TD and ATD we have listed the eigenvalues of the terminal distance matrix TD of Table 1 and the augmented terminal distance matrix ATD of Table 2. The row sums for the individual rows of TD matrix are in the range $30 < \lambda_1 < 59$, the actual value being $\lambda_1 = 39.3125$, which is not very different from the average row sum, which is 420/11 = 38.1818. The row sums for individual rows of the ATD matrix are in the range $31 < \lambda_1 < 61$, the actual value being $\lambda_1 = 41.5062$, which is not very different from the average row sum, which is 441/11 = 40.0909. In view of the proximity of the

Table 3
The eigenvalues of the distance matrix for terminal vertices (TD) of the graph of Fig. 2, the augmented distance matrix (ADT) and the eigenvalues of the line distance (LD) matrix of the same graph (to be described later)

	TD	ADT	LD
λ_1	39.3125	41.5062	162.9460
λ_2	−2	−1	−1.2473
λ_3	−2	−1	−2.0278
λ_4	−2.9603	−1.6607	−2.2397
λ_5	−4	−2	−2.7232
λ_6	−4	−2	−3.5583
λ_7	−4	−2	−4.5158
λ_8	−4	−2	−6.5320
λ_9	−4	−2	−11.0694
λ_{10}	−4	−2	−30.7853
λ_{11}	−8.3523	−4.8455	−98.2472

Table 4
Construction of protein profiles for A strand of human insulin

k	Reduced profile	Full profile
1	38.18	55.85
2	78.55	99.40
3	115.82	127.70
4	136.91	134.93
5	137.14	122.91
6	119.95	98.93
7	93.23	71.43
8	65.13	46.73
9	41.25	27.91
10	23.84	15.32
11	12.65	7.77
12	6.20	3.66
13	2.82	1.61
14	1.19	0.66
15	0.47	0.26

leading eigenvalues and the average row sums, which is conceptually easy to understand and computationally easy to obtain, one may adopt characterization of proteins (here the A strand of human insulin) by considering the average row sums of the higher order matrices as preferred invariants. In Table 4, we have illustrated normalized average row sums for the case of the 11 × 11 distance matrix of Table 1, which results in a convergent sequence, which is graphically illustrated in Fig. 3 as a bar graphs, referred to as protein ''profile.''

5.2. Line distance matrix

The sequence 2, 4, 7, 11, 15, 19, 23, 27, 33, 39, which is the first row of the 11 × 11 matrix of Table 5 is constructed from the sequence: 2, 2, 3, 4, 4, 4, 4, 4, 6, 6, which represents the sequential separations between the terminal vertices. This sequence has appeared as the entries above the main diagonal of the TD matrix of Table 1 and also it appears as the entries above the main diagonal of the LD matrix of Table 5. Hence, the TD and LD matrices have common entries just above (and below) the main diagonal. One can view this sequence to define a partition of a line, the entries of the sequence giving the length of individual segments on the line. The line distance matrix, which has only recently received some attention [69–71], is

Table 5
The upper part of the symmetric line distance matrix for partial sums of ordered terminal distances

	1	3	8	10	12	15	16	17	19	20	21
1	0	2	4	7	11	15	19	23	27	33	39
3		0	2	5	9	13	17	21	25	31	37
8			0	3	7	11	15	19	23	29	35
10				0	4	8	12	16	20	26	32
12					0	4	8	12	16	22	28
15						0	4	8	12	18	24
16							0	4	8	14	20
17								0	4	10	16
19									0	6	12
20										0	6
21											0

fully defined for any line or a curve (including zigzag lines), that is partitioned in segments. These matrices have some interesting properties: they are fully determined once their first row has been defined, while all their eigenvalues except for the leading eigenvalue are negative. The line distance matrix is the graph theoretical distance matrix constructed for a graph, the vertices of which lie on a straight line. In Table 5, we have illustrated the line distance matrix belonging to the star graph of Fig. 1, while in Table 3 we have already listed the eigenvalues of this matrix. The line distance matrix offers construction of additional invariants for star graphs representing proteins by considering ''higher order'' distance matrices analogous to the already outlined construction of the bar graphs based on the leading eigenvalues for TD matrices illustrated in the upper part of Fig. 3, which is shown in the lower part of Fig. 3. As one can see the two ''profiles'' are of similar shape and magnitude, though showing some variation in magnitudes in individual ''higher order'' leading eigenvalues.

6. Recovery of lost information

Recall that there is considerable loss of information accompanying protein sequences represented by unlabeled star graphs. The same extends to representation of star graphs by reduced distance matrices, which contain only information on terminal vertices, from which not only one does not know which vertex corresponds to which amino acid but also one has no information on the sequential occurrence of amino acids. The situation is similar with several 2-D graphical representations of DNA [5,19,20] in which the resulting graphical patterns are devoid of labels. However, some lost information is recovered when sequential labels for the bases of DNA sequence are introduced. In the case of reduced distance matrices of star-like graphs it is also possible to recover much of the lost information accompanying construction of TD matrix. There are three steps that one can consider to gradually recover portions of structural information not taken onto account when considering the reduced distance matrices of proteins:

(1) Instead of considering the reduced distance matrix (the matrix of 11 terminal vertices in the case of the star-like graph of Fig. 2) one can consider the full distance matrix (D) on all vertices of a star-like graph (21 vertices for the graph of Fig. 2).
(2) Distance matrices (the reduced TD matrix or the full D matrix) do not take into account the sequential distribution of amino acids in a protein. Thus, for example, if we change labels on the vertices of the graph of Fig. 2, which represent A strand of human insulin, we will obtain another protein which will have the same overall composition of amino acids but which will occur in different sequential order. Any arbitrary labeling of vertices, while producing different protein and different form for D matrix, will generate the same graph matrix invariants. Possible route to include information on sequential distributions of amino acids in a protein is to take advantage of knowing the sequential labels of amino acids of a protein. One way of doing this is illustrated in Fig. 5, where starting from the star graph of Fig. 2 we created family of smaller graphs, each one obtained one from the other by deleting one vertex at a time in a sequential order.
(3) Finally, there is a way to incorporate into distance matrices information on the kind of amino acid involved by replacing the zeros on the main diagonal of the matrices with numerical parameters characterizing individual amino acids. Such are, for example, the molecular weight, the hydrophobicities, or some other molecular property. While the so augmented matrices will not differentiate between proteins having identical number of each of 20 natural

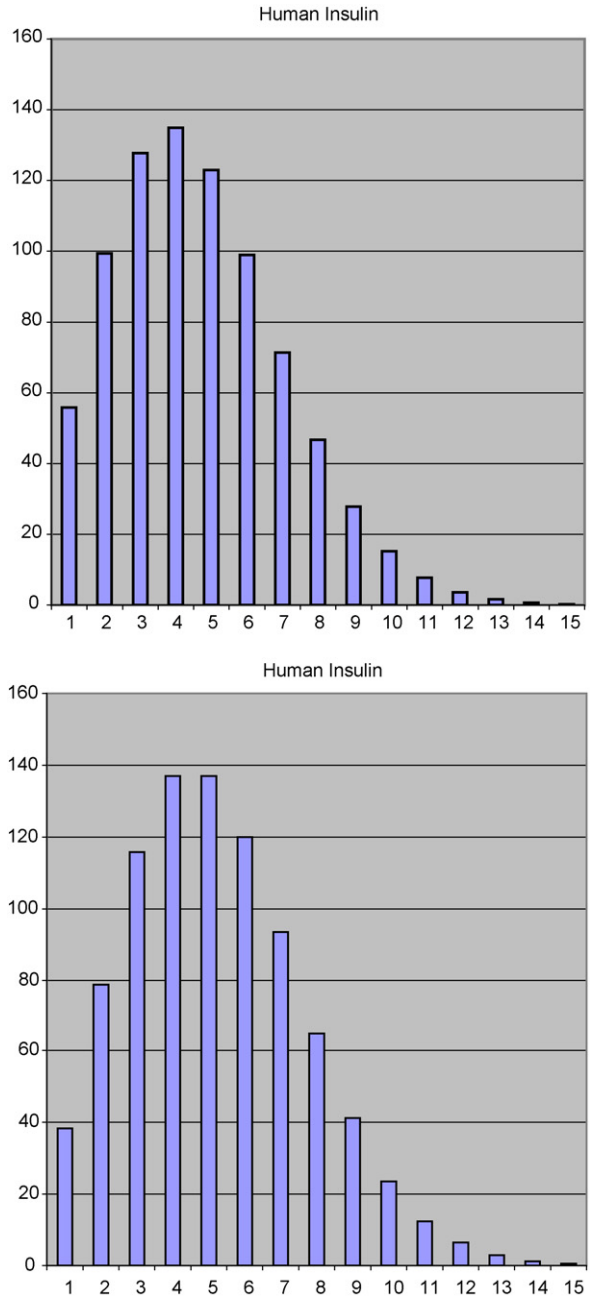

Fig. 4. The ''profile'' of A strand of human insulin based on the reduced distance matrix TD (top) and the full distance matrix (bottom).

amino acids it may discriminate between proteins having identical ''frequency'' number but belonging to different amino acids.

In this report, we will consider only the first two steps mentioned above and we hope to demonstrate that they offer useful additional tool for characterization of proteins. We will first consider the complete distance matrix and will follow with a briefly outline of use of sequential labels for recovering the lost information on proteins when they are characterized by reduced distance matrices. Before that let us comment on some of the ''steps'' considered for a characterization of a protein. One may ask: ''Why should one consider reduced distance matrices when alternatives that involve information on all amino acids are also available?'' Let us try to offer an answer. Most of the current numerical characterizations of proteins or DNA have been initiated in order to facilitate a quantitative comparative study of such systems. The important aspect of a ''Table of Similarity and Dissimilarity'' is that the presence of large entries in such tables *definitely* points to a lack of similarity of objects considered, while small entries indicate a *necessary* but not a *sufficient* condition that two objects are similar. In view of this, when screening a large pool of objects it may be advantageous and more efficient to employ simple characterization (of otherwise complex systems, such as proteins) and thus in pre-screening eliminates large number of structurally dissimilar cases. This, in the case of proteins, can be accomplished by using protein profiles based on matrices involving only terminal distances. Once a large pool of objects is reduced one can then employ other more elaborate characterizations of proteins, such as use of the full distance matrix or invariants derived from sequential elimination of vertices of a star graph. The efficiency of such strategy has been well illustrated in application of topological indices to QSAR (quantitative structure–activity relationship). For example, Lahana and coworkers [72] have screened combinatorial library having about 280,000 virtual compounds (decapeptides) in a in search for novel potential immunosuppressive drug compounds. By using a dozen molecular descriptors (half of which were various topological indices) they were able to narrow attention to some two dozen compounds, which were further more closely investigated. Finally they focus on five compounds that were synthesized, one of which turned out to have outstanding immunosuppressive activity.

7. Characterization of proteins by complete distance matrix

In Table 6, we have illustrated the complete distance matrix for the star graph of Fig. 2. This matrix offer construction of protein profile in parallel with the protein profile shown in Fig. 4, which has been based on the reduced distance matrix. We have included in Table 4 (the right end column) and Fig. 4 (the bottom) the numerical and graphical representation of the A strand human insulin protein as computed from the full distance matrix. As one can see the both profiles are of a similar form, though they show minor differences in the position of the maximum and the tail, which for the case of complete distance matrix decreases slightly faster. This follows because by inclusion of all vertices in the distance matrix the average row sum slightly decreases, because the ''inside'' vertices (vertices of degree two) on average are at smaller distances than are the terminal vertices. The advantage of the profiles computed using only terminal matrices is that they involve smaller matrices, which at most can be of 20×20 size, while use of the complete distance matrix will result in a matrix of $N \times N$ size, where N indicates the number of amino acids in a protein. Thus, use of terminal matrices will involve in general situations much less computation.

Table 6
The distance matrix for the star graph of Fig. 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	0	2	2	2	2	2	3	2	2	3	4	3	2	2	3	3	3	2	3	5	3
2		0	2	2	2	2	3	2	2	1	4	3	2	2	3	3	3	2	3	5	3
3			0	2	2	2	3	2	2	3	4	3	2	2	3	3	3	2	3	5	3
4				0	2	2	3	2	2	3	4	3	2	2	3	3	1	2	3	5	3
5					0	2	3	2	2	3	4	3	2	2	1	3	3	2	3	5	3
6						0	1	2	2	3	2	3	2	2	3	3	3	2	3	3	3
7							0	3	3	4	1	2	3	3	4	4	4	3	4	2	4
8								0	2	3	4	3	2	2	3	3	3	2	3	5	3
9									0	3	4	1	2	2	3	3	3	2	3	5	3
10										0	5	4	3	3	4	4	4	3	4	6	4
11											0	5	4	4	5	5	5	4	5	1	5
12												0	3	3	4	4	4	3	4	6	4
13													0	2	3	1	3	2	3	5	3
14														0	3	3	3	2	1	5	3
15															0	4	4	3	4	6	4
16																0	4	3	4	6	4
17																	0	3	4	6	4
18																		0	3	5	1
19																			0	6	4
20																				0	6
21																					0

8. On sequential order of amino acids

Use of information on the sequential occurrence of individual amino acids as they appear in a protein when constructing star graphs will further reduce the possibility that different proteins are represented by the same labeled star graphs. The vertices of the star graph of Fig. 2 have been labeled from 1 to 21, where labels indicate the order in which amino acids appear in the A strand of human insulin. Observe that there is no label for the central vertex that represents the hub from which the ''rays'' of the star graph are drawn and which does not correspond to amino acid. Recall that while the form of the distance matrix depends on labeling of vertices invariants of such matrix are independent of labels and hence will belong to all proteins that are represented by the same unlabeled star graph. However, one can construct a set of matrices that will have information on the sequential appearance of amino acids in a protein. Start with the 21×21 distance matrix of the graph of Fig. 2 and follow with construction of 20 additional distance matrices obtained from the graph of Fig. 2 with a successively deletion of the last amino acid in the sequence. In this way, we would produce 21 ordered graphs and hence 21 ordered matrices, each having fewer rows and columns, associated with the same protein. The initial graphs of the outlined construction are illustrated in

Table 7
The eigenvalues, the partial sums and the leading λ_1 value of the sequential distance matrices of the insulin star graph

	Eigenvalues	Partial sum	Sequential λ_1	Dimension
1	63.5823	63.5823	63.5823	21
2	0.3051	63.8874	59.5696	20
3	−0.7639	63.1235	52.2498	19
4	−0.7639	62.3596	48.2765	18
5	−0.7639	61.5957	46.0165	17
6	−0.7639	60.8318	42.0516	16
7	−0.7639	60.0679	38.0954	15
8	−0.8187	59.2492	34.1511	14
9	−1.0000	58.2492	31.8758	13
10	−1.5569	56.6923	29.5657	12
11	−2.0000	54.6923	26.0709	11
12	−2.0000	52.6923	21.1212	10
13	−2.2391	50.4532	17.4942	9
14	−3.3744	47.0792	15.3936	8
15	−5.2361	41.8431	13.2692	7
16	−5.2361	36.6070	10.0000	6
17	−5.2361	31.3709	8.0000	5
18	−5.2361	26.1348	6.0000	4
19	−5.2361	20.8987	4.0000	3
20	−5.7404	15.1583	2.0000	2
21	−15.1575	0	0	1

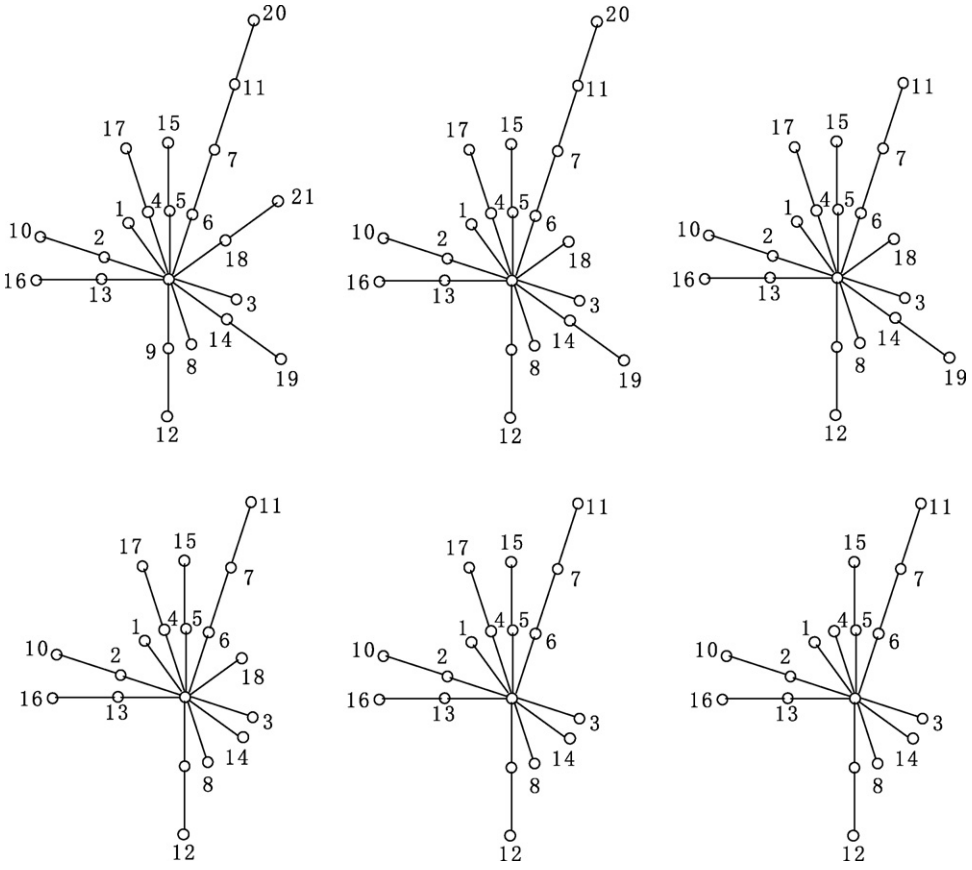

Fig. 5. The gradual sequential reduction of the star graph of Fig. 2 that leads to 20 additional leading eigenvalues that offer an alternative characterization for the A strand of human insulin.

Table 8
Ten random constructed proteins having 21 amino acids

1	Asn-Phe-Arg-Arg-Glu-His-Phe-Cys-Gln-Phe-Cys-Asn-Met-Val-Gly-Asp-Pro-Cys-Arg-Pro-Arg
2	Gly-Asp-Cys-Leu-Cys-Lys-Ile-Glu-Trp-Ala-Cys-Phe-Ala-Phe-His-Glu-Arg-Val-Cys-Gln-Ser
3	Ala-Pro-Ile-Trp-Val-Leu-Ile-Trp-Ser-Arg-Tyr-Glu-Asp-Lys-Ala-Glu-Asp-Ala-Tyr-Ala-Cys
4	His-Thr-Gln-Ser-Gln-Leu-Asn-Leu-Gln-Gln-Lys-Gln-Asp-Arg-Ser-Trp-Asn-His-Trp-Asp-Pro
5	Thr-Met-Cys-Asn-Lys-Arg-Lys-Tyr-Asn-Trp-Gly-His-Met-Ser-Gly-Glu-Ser-Gly-Gln-Ala-Cys
6	Cys-Ser-Glu-Lys-Gly-His-Lys-Gly-Lys-Tyr-Ser-Glu-Trp-Val-Pro-Asp-Val-Val-Pro-Gln-Cys
7	Trp-Asp-Ala-Gln-Val-Asn-Ile-Cys-Gly-Asp-Ser-Glu-Thr-Asp-Ile-Phe-Phe-Leu-Ile-Glu-Ser
8	Lys-Met-Ser-Phe-Glu-Cys-Thr-His-Met-Arg-Leu-Tyr-Ser-Val-Thr-Glu-Gly-Glu-Leu-Gly-Glu
9	Met-Asp-Leu-Ser-Pro-Ile-Ile-Phe-Asp-Ile-Gln-Tyr-Ala-Asp-Leu-Gly-Leu-Asp-Thr-Asn-Asp
10	Val-Lys-Ala-Asn-Cys-Arg-Tyr-Ser-Val-Phe-His-Asn-Gln-His-Asp-Val-Phe-Ser-Trp-Val-Thr

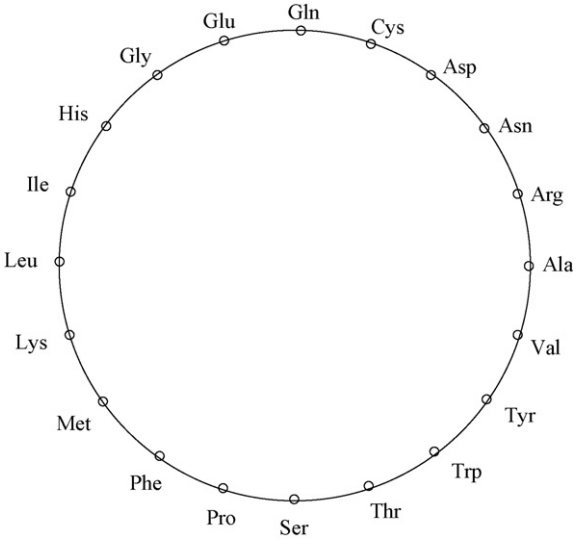

Fig. 6. Twenty amino acids uniformly placed on the circumference of unit circle.

Fig. 5. For each graph of Fig. 5 one can evaluate the distance matrix, the line distance matrix and the reduced distance matrix TD, which involved only distance between terminal vertices. The leading eigenvalue of these matrices represent additional protein descriptors (sequence invariants), which involve additional information not considered in the construction of protein profiles of Fig. 4.

Table 10
"Walk around" codes for the 10 random constructed proteins having 21 amino acids

1	0 0 0 0 1 1 1 1 0 0 1 1 0 1 0 1 0 0 0 1 1 1 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 0 0 1 1 0 1
2	0 0 1 1 0 1 0 1 0 0 0 0 1 1 1 1 0 1 0 0 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 1 1 0 1 0 1 0 1
3	0 0 0 0 1 1 1 1 0 1 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 1 0 1 0 1 0 1 0 1 0 0 1 1 0 0 1 1 0 1
4	0 1 0 0 1 1 0 0 1 1 0 0 0 0 0 1 1 1 1 1 0 0 1 1 0 0 1 1 0 1 0 1 0 1 0 0 1 1 0 1 0 1 0 1
5	0 1 0 1 0 0 1 1 0 0 1 1 0 1 0 1 0 0 0 1 1 1 0 1 0 0 1 1 0 0 1 1 0 0 1 1 0 1 0 1 0 1 0 1
6	0 1 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 1 0 0 0 1 1 1 0 0 1 1 0 0 1 1 0 1 0 1 0 0 0 1 1 1
7	0 1 0 1 0 0 0 1 1 1 0 1 0 1 0 0 1 1 0 1 0 0 0 1 1 1 0 1 0 0 1 1 0 0 1 1 0 1 0 1 0 1
8	0 1 0 1 0 0 0 0 1 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 1 1 0 1
9	0 1 0 1 0 0 0 0 0 1 1 1 1 1 0 1 0 1 0 0 0 1 1 1 0 0 0 0 1 1 1 0 1 0 1 0 1 0 1 0 1 0 1
10	0 1 0 1 0 0 1 1 0 1 0 1 0 1 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 1 0 1 0 1 0 0 0 0 1 1 1 1

In Table 7, we have listed the 21 eigenvalues of the full distance matrix (shown in Table 6). In the adjacent column, entitled "partial sum," we show the corresponding partial sums of the eigenvalues. They contain the same information but have an advantage as being expressed by positive numbers. In the next column, we have listed the leading eigenvalues of the 21 submatrices obtained by reducing the star graph by deleting vertex belonging to the last amino acid, thus gradually decreasing the size of submatrices from 21×21 to 1×1. Although the numbers in the two columns appear similar there is a significant distinction in the information content between the two columns. The leading eigenvalues of the sequentially reduced distance matrices belongs to a smaller class of proteins than the corresponding partial sums of the eigenvalues of the initial distance matrix. The number of proteins having the same *sequential* distribution of the 20 letters alphabet is considerably smaller than the class of proteins having the same distribution of 20 amino acids which may appear in different sequential order. Hence, the column of Table 7, which list the

Table 9
Construction of profile for random protein 1, which has 11 branches ($N = 11$)

	$k=1$	$k=2$	$k=3$	$k=3$	$k=3$	$k=6$
Σ	462	1.901×10^4	8.387×10^5	3.949×10^7	1.967×10^9	1.024×10^{11}
Σ/N^k	38.5	132.03	485.34	1.9045×10^3	7.9040×10^3	3.4302×10^4
$(\Sigma/N^k)/k!$	38.500	66.014	80.889	79.356	65.867	47.642
	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$	$k=12$
Σ	5.520×10^{12}	3.052×10^{14}	1.720×10^{16}	9.840×10^{17}	5.693×10^{19}	3.324×10^{21}
Σ/N^k	1.5404×10^5	7.0976×10^5	3.3339×10^6	1.5892×10^7	7.6627×10^7	3.7284×10^8
$(\Sigma/N^k)/k!$	30.564	17.603	9.187	4.379	1.9197	0.778

Σ: the average row sum; Σ/N^k: the scaling factor; $(\Sigma/N^k)/k!$: the normalization of scaled values by factorials.

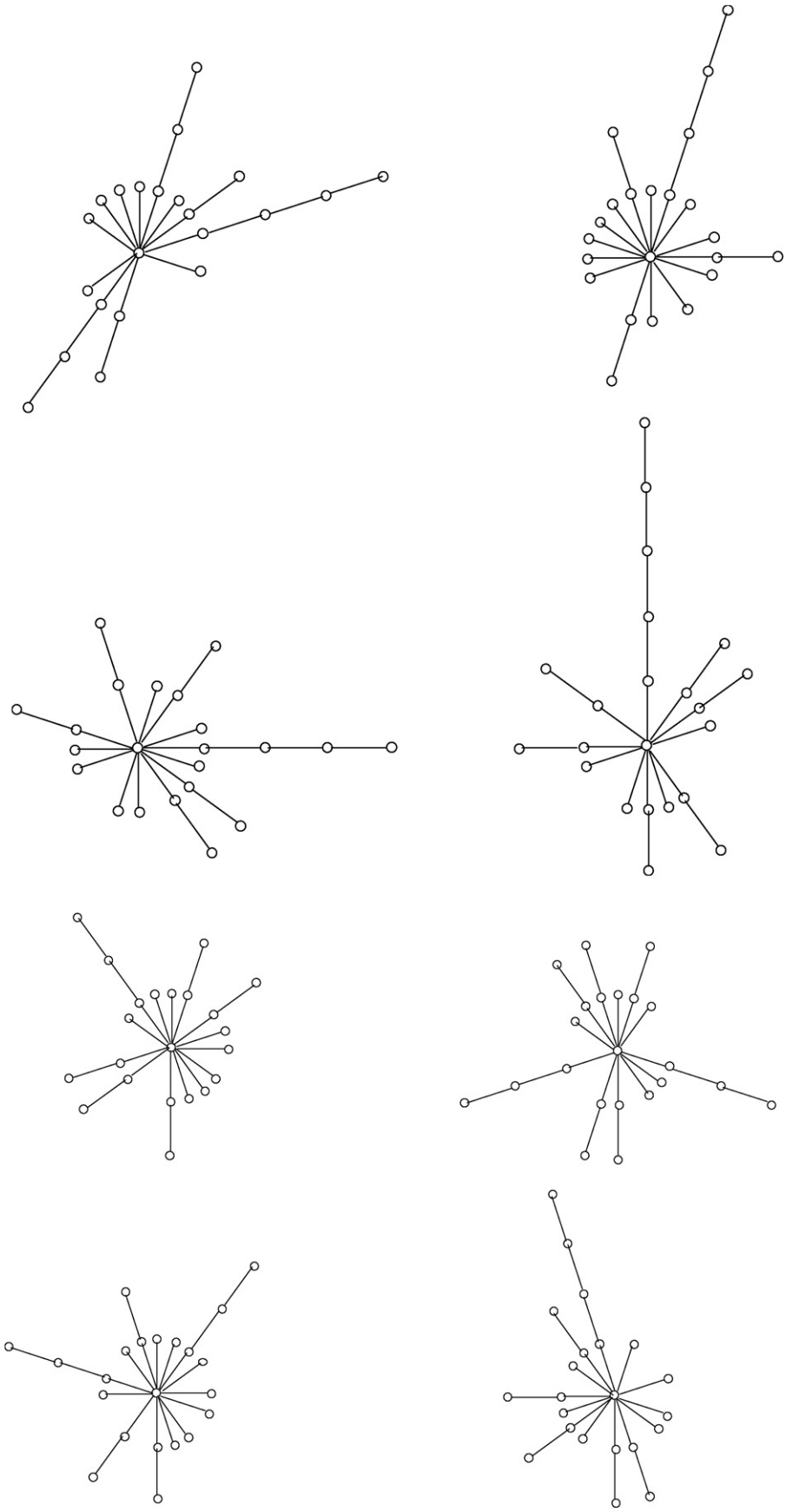

Fig. 7. The standard representation of star-like graphs corresponding to 10 proteins having 21 amino acid constructed at random.

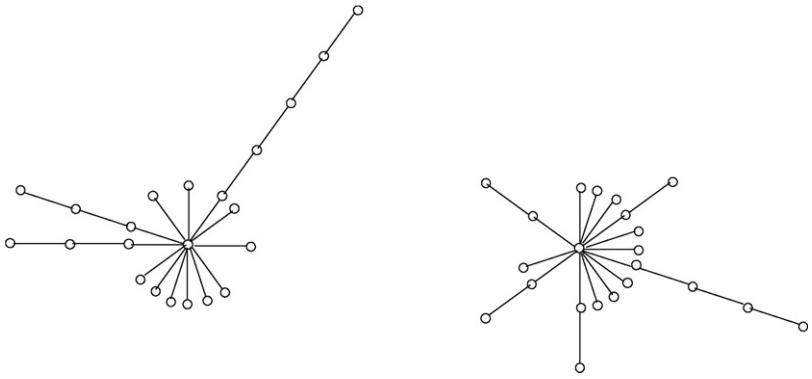

Fig. 7. (*Continued*).

sequential eigenvalues for the star graph of Fig. 2, offers a more specific characterization of A the strand of human insulin, even though there could be other proteins with the same characterization. A likelihood that this will occur for proteins of interest is decreasing dramatically with an increase in the length of proteins examined. However, when need arises for further differentiation among proteins one can still consider, as mentioned earlier, selected property of amino acids as additional parameter to be included in diagonal elements of the distance matrix.

9. On the standard representation of proteins by star-like graphs

The outlined approach of graphical representation and numerical characterizations of proteins, illustrated on strand A of human insulin would appear to be particularly suitable for smaller proteins, those having hundreds rather than thousands of amino acids. When drawing star-like graph, just as when drawing any graph, one always faces dilemma in selecting the sites for the vertices, which involves arbitrary choices. It is unlikely thus that the same protein will be represented by same graph without further rules for construction of graphical representations of star-like graphs. In order to facilitate visual comparison of different protein sequences we are suggesting that a standard format for star-like graphs of proteins be adopted. We propose that the standard format for graphical representation of proteins by star-like graphs is based on arranging the 20 natural amino acids alphabetically (based on 3-letter codes) on a periphery of a circle (Fig. 6) in anti-clockwise order by starting at the positive x-axis and moving in steps of $2\pi/20$ (or 18°). In this way, the sites of the 20 amino acids will be uniformly distributed on the periphery of the unit circle, which will also define uniform orientation for the 20 branches on which vertices depicting multiple occurrence of any amino acid will be located. If this advice is followed not only that one will arrive at unique geometrical representation of proteins but such representation will also define uniquely polar coordinates for amino acids of a protein.

In Fig. 7, we have illustrated the proposed standard representation on a set of 10 random graphs representing random proteins (listed in Table 8), all having 21 amino acids. Fig. 7 immediately visually shows considerable variations among the 10 random proteins. It is obvious that similar proteins will have similar amino acid composition. Therefore, it is necessary, though not sufficient condition for two proteins to be similar that the corresponding unlabeled star-like graphs are similar. For two proteins to be similar the sufficient condition is not only that unlabeled star-like graph are similar but also that labeled star-like graph, where labels correspond to sequential location of individual amino acids, are similar—in which case the associated distance and line distance matrices will be similar and will display similar properties. In Fig. 5 (left top), we illustrate the proposed ''standard'' representation of A strand of human protein. As one can see, and as could have been expected, none of the random graphs show much similarity with the A strand of human protein (and among themselves) even in the content of amino acids. The inclusion of the random graphs was made mainly to illustrate that, at least for smaller proteins, the outlined approach is likely to represent different proteins by different unlabeled star graphs. In the case of larger proteins, which may have similar composition of amino acid it will be important to construct labeled star graphs, which will be sensitive not only on the composition of amino acid but also on their distribution within the primary sequence of proteins. In Fig. 8 we show the profiles for the ten random proteins of Fig. 7. Construction of the profiles of Fig. 8 is outlined in Table 9 for the first random graph.

One of the advantages of adopting the standard representation of star-like graphs for proteins is that by doing this one has obtained graphical objects. In Table 10 for the ten random graphs we have listed the ''Walk Around'' binary codes [73] based on their standard graphical representation. The codes allow one to catalogue star-like graphs in lexicographic order, which will facilitate search for similar proteins. Having graphs with definite geometry allow one to construct the D/D matrices for proteins. The leading eigenvalue of D/D matrices of general graphs offers a measure of a ''compactness'' of the considered object, which in the case of star-like graphs indicates a measure of similar frequency of occurrence for all amino acids present in the protein. In addition to so constructed D/D matrix one can consider the 21 sites of the 21 amino acids of the A strand of human insulin as points in the (x, y) plane and connect them by a zigzag line with which another D/D matrix can be constructed.

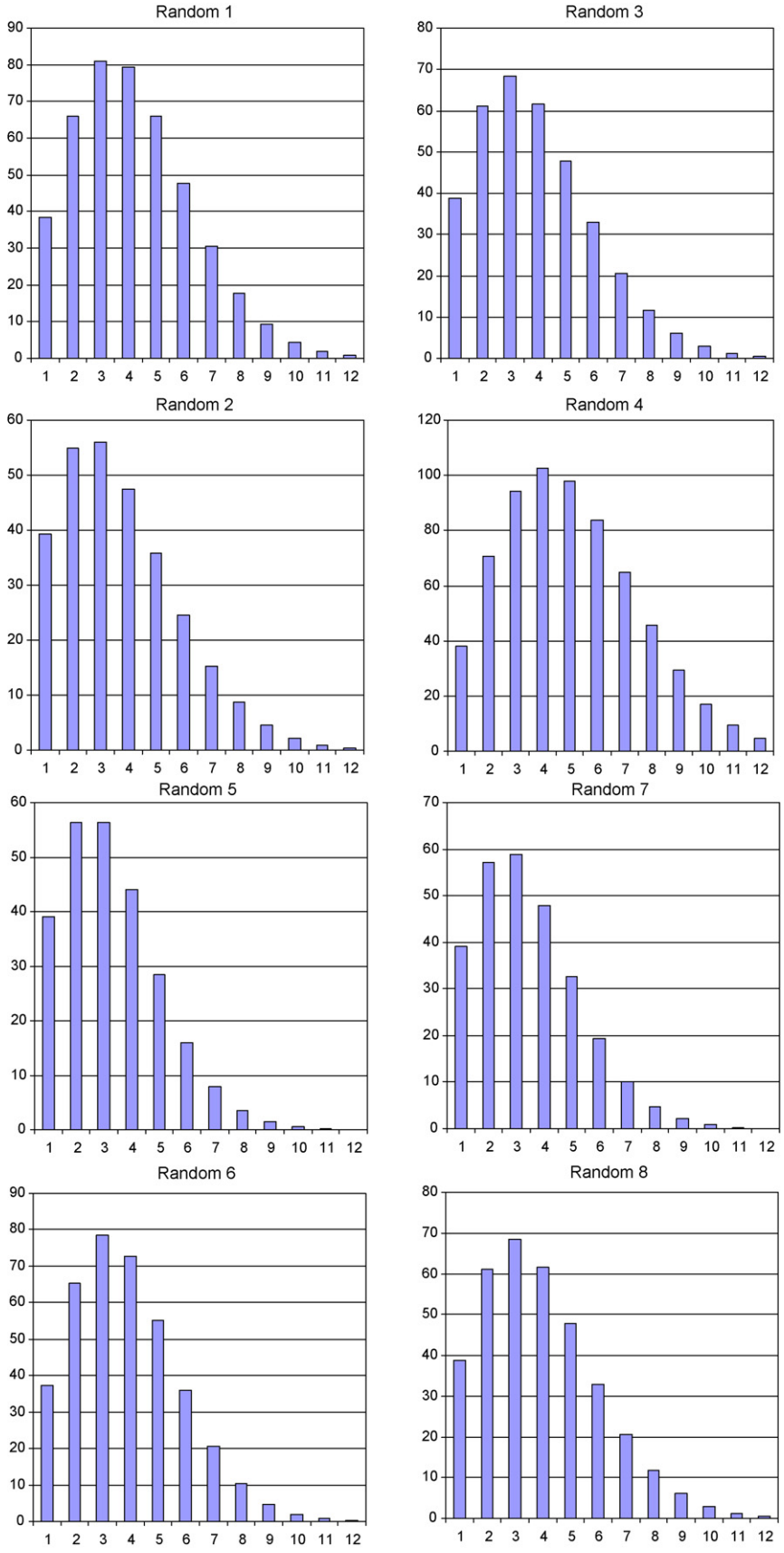

Fig. 8. The profiles of 10 random proteins illustrated in Fig. 7 based on the reduced distance matrix. Observe different scale for different random protein profiles.

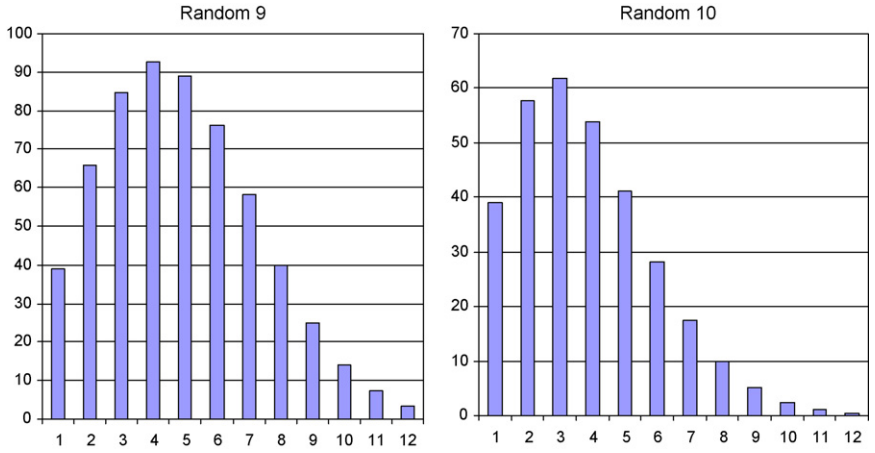

Fig. 8. (*Continued*).

Finally, in the case of large proteins (having thousands amino acids) one can modify the proposed standard graphical representation and instead of using equidistant concentric circles of linearly increasing radius select circles the radius of which increases by fraction. This would be analogous to the ''Chaos Game'' of Barsney [35] and Jeffrey's algorithm [3] for graphical representation of lengthy DNA structures.

10. Concluding remarks

We have outlined a novel graphical representation of proteins based on star-like graphs, which have an important advantage in that the accompanying matrices used for extraction of protein sequence invariants are free of arbitrary assignment of amino acids to particular branches of the star-like graphs. In particular we have outlined several distance-based matrices and used the leading eigenvalues and the average row sums as descriptors of choice for characterization of proteins. One can also arrive at ''standard graphical representations'' by adopting simple rules using alphabetical order of amino acids (based on three-letter codes), which facilitate visual comparison of star graphs representing different proteins.

The question can be raised: Why should one use the leading eigenvalues, or the average row sums of matrices to represent protein characterization? What is the foundation of the theory behind this approach to characterization of proteins? Before answering these questions let us make comparison with few approaches of Quantum Chemistry to calculations of the wave functions of molecules, which involve exact ab initio calculations as well as semi-empirical approaches. Why are people using Gaussian orbitals and not Slater orbitals, which have better asymptotic behavior? The answer is simple: four center molecular integrals can be calculated (as shown by Boys [74], over 50 years ago!) for Gaussian functions, but cannot be solved (yet?) when electrons are described by Slater orbitals.

The situation to some degree is similar with applications of Discrete Mathematics to Chemical Structure, even though there are no the first principle (Axioms of Quantum Theory) in this case. The task is to represent chemical structures by numbers, rather than representing chemical structures by molecular properties, as has been the case with number of structure–property–activity studies in physical chemistry and medicinal chemistry. The numbers that we are seeking, which represent mathematical properties of molecules, should bear some structural interpretation, if possible. The first such meaningful approach was due to Platt [75], who suggested (for the case of molecules of alkanes) that the sequence, the entries of which count the number of paths of length k in molecular graph, be used as molecular descriptors. The Wiener number [40], the Hosoya Z topological index [41], and the connectivity index of one of present authors [58] together with the ''higher order'' connectivity indices [76] have been the first such descriptors that continue to offer useful characterization of molecules for structure–property–activity studies. Over the past two decades the pool of molecular descriptors has widely increased [77] facilitating thus expansion of structure–property–activity studies to all types of compounds, including also virtual libraries. Observe that the alternative approaches to those based on mathematical descriptors are limited only to known molecules (and often molecules of known properties) and such approaches cannot consider screening of virtual libraries—which is a serious limitation. On the other hand, even though it is desirable to have structural interpretation of various mathematical invariants used as descriptors of complex systems, this is not necessary conditions for screening or comparative study of such systems.

In the case of DNA and proteins we are in a similar situation. Here the effort is to arrive at mathematical descriptors for biological sequences, which would than allow one to ''replace'' the actual biological sequences (DNA, RNA or proteins) by suitable numerical sequences, which can be used not only for calculation of the degree of similarity or dissimilarity between two and more sequences quantitatively, but may also allow some arithmetic manipulations. Very recently it has been demonstrated, first for proteins [78] and then also for DNA [79], that some graphical representations, when ''translated'' into numerical format, allow one to find optimal alignment of two protein or two DNA sequences. The alignments are found by simply subtracting the corresponding numerical sequences and looking at spots having zero difference. This novel approach to

the old problem of sequence alignment has yet to grow and develop to apply to more general areas of sequence comparisons, but preliminary results appear encouraging [80]. However, even though we may be long way from approaching Basic Local Alignment Search Tool (BLAST) [81] and other computer algorithms, the graphical–geometrical approach to sequence alignment has one important advantage over computer-oriented algorithms currently in use, in that here one is in a position to characterize a *single* protein or DNA sequence. Computer-oriented approaches are restricted to sequence comparisons and can only be used for comparison of a pair or more sequences, and when used on a single sequence again can only compare two or more portions of such sequences, while we can characterize single sequence or parts of a single sequence. The present results outlined in this article offer additional descriptors for such *single* sequence characterization, which of course, can also be used for comparison of sequences. However, we would like to emphasize that this approach has no intention to replace some existing tool even if it ever is further developed but it is hoped that it may supplement existing approaches, and perhaps be even a part of some such schemes.

References

[1] E. Hamori, Novel DNA sequence representation, Nature 314 (1985) 585–586.
[2] M. Gates, A simple way to look at DNA, J. Theor. Biol. 119 (1986) 319–328.
[3] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acid Res. 18 (1990) 2163–2170.
[4] P.M. Leong, S. Morgenthaler, Random-walk and gap plots of DNA sequences, Comput. Appl. Biosci. 11 (1995) 503–507.
[5] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I. Methodology and application to globin gene, Curr. Sci. 66 (1994) 309–313.
[6] M. Randić, M. Vračko, A. Nandy, S.C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, J. Chem. Inf. Comput. Sci. 40 (2000) 1235–1244.
[7] M. Randić, M. Vračko, On the similarity of DNA primary sequences, J. Chem. Inf. Comput. Sci. 40 (2000) 599–606.
[8] C. Li, J. Wang, On a 3-D representation of DNA primary sequences, Comb. Chem. High Throughput Screen 7 (2004) 23.
[9] M. Randić, A.T. Balaban, On a four-dimensional representation of DNA primary sequences, J. Chem. Inf. Comput. Sci. 43 (2003) 532–539.
[10] M. Randić, On characterization of DNA primary sequences by a condensed matrix, Chem. Phys. Lett. 317 (2003) 29–34.
[11] M. Randić, Condensed representation of DNA primary sequences, J. Chem. Inf. Comput. Sci. 40 (2000) 50–56.
[12] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chem. Phys. Lett. 368 (2003) 1–6.
[13] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (2003) 202–207.
[14] M. Randić, M. Vračko, J. Zupan, M. Novič, Compact 2-D graphical representation of DNA, Chem. Phys. Lett. 373 (2003) 558–562.
[15] M. Randić, X.F. Guo, S.C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, J. Chem. Inf. Comput. Sci. 41 (2001) 619–626.
[16] X.F. Guo, M. Randić, S.C. Basak, A novel 2-D graphical representation of DNA sequences of low degeneracy, Chem. Phys. Lett. 350 (2001) 106.
[17] M. Randić, J. Zupan, A.T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, Chem. Phys. Lett. 397 (2004) 247–252.
[18] B. Liao, T.-M Wang, 4-D representation of RNA secondary structure and their numerical characterization, Private information from Professor Bo Liao, Department of Applied Mathematics, Dalian University of Technology, Dalian, China, May 2006.
[19] M. Randić, J. Zupan, Highly compact 2-D graphical representation of DNA sequences, SAR QSAR Environ. Res. 15 (2004) 191–205.
[20] M. Randić, Graphical representation of DNA as a 2-D map, Chem. Phys. Lett. 386 (2004) 468–471.
[21] J. Zupan, M. Randić, Algorithm for coding DNA sequences into "spectrum-like" and "zigzag" representations, J. Chem. Inf. Model. 45 (2005) 309–313.
[22] X. Liu, Q. Dai, T. Wang, A novel 2-D graphical representation of DNA sequences and its application, J. Mol. Graph. Model., in press.
[23] M. Randić, D. Vikić-Topić, A. Graovac, N. Lerš, D. Plavšić, Novel graphical and numerical representation of DNA, Periodicum Biologorum 107 (2005) 437–444.
[24] M. Randić, Novel 1-dimensional representation of DNA, Periodicum Biologorum 107 (2005) 415–422.
[25] B. Liao, Y.S. Liu, R.F. Li, W. Zhu, Coronavirus phylogeny based on triplets of nucleic acids bases, Chem. Phys. Lett. 421 (2006) 313–318.
[26] B. Liao, J.W. Luo, R.F. Li, W. Zhu, RNA secondary structure 2D graphical representation without degeneracy, Int. J. Quantum Chem. 106 (2006) 1749–1755.
[27] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, Chem. Phys. Lett. 419 (2006) 528–532.
[28] M. Randić, A.T. Balaban, M. Novič, A. Založnik, T. Pisanski, A novel graphical representation of proteins, Periodicum Biologorum 107 (2005) 403–414.
[29] M. Randić, M. Novič, D. Vikić-Topić, D. Plavšić, Novel numerical and graphical representation of DNA sequences and proteins, SAR QSAR Environ. Res. 17 (2006) 1–13.
[30] M. Randić, Spectrum-like graphical representation of DNA based on codons, Acta Chim. Slovenica 53 (2006) 477–485.
[31] M. Randić, D. Vikić-Topić, N. Lerš, D. Plavšić, Graphical and numerical representation of DNA based on codons, J. Mol. Graph. Model., submitted for publication.
[32] M. Novič, M. Randić, Novel invariant representation of proteins in 20-D space, J. Comput. Biol., submitted for publication.
[33] A.C. Guillermin, G.D. Humberto, R. Molina, V.S. Javier, E. Uriarte, G.D. Yenny, Novel 2D maps and coupling numbers for protein sequences, The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L., FEBS Lett. 580 (2006) 723–730.
[34] Funk & Wagnalls Standard Desk Dictionary, Harper & Row Publishers, Inc., 1984.
[35] M.F. Barnsley, H. Rising, H. Fractals, Everywhere, second ed., Academic Press, Boston, MA, 1993.
[36] M. Randić, 2-D graphical representation of proteins based on virtual genetic code, SAR QSAR Environ. Res. 15 (2004) 147–157.
[37] F. Bai, T. Wang, On graphical and numerical representation of protein sequences, J. Biomol. Struct. Dyn. 23 (2006) 537–545.
[38] F. Harary, Graph Theory, Addison-Wesley, Reading, MA, 1969.
[39] K.H. Rosen, Discrete Mathematics and its Applications, fifth ed., McGraw-Hills, New York, 2003.
[40] H. Wiener, Structural determination of paraffin boiling points, J. Am. Chem. Soc. 69 (1947) 17–20.
[41] H. Hosoya, Topological index A newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons, Bull. Chem. Soc. Jpn. 44 (1971) 2332–2339.
[42] L. Lovasz, J. Pelikan, On the eigenvalues of trees, Periodica Math. Hung. 3 (1973) 175–182.
[43] M. Randić, On structural ordering and branching of acyclic saturated hydrocarbons, J. Math. Chem. 24 (1998) 345–358.
[44] M. Randić, X.F. Guo, S. Bobst, Use of path matrices for a characterization of molecular structures, DIMACS Ser. Discr. Math. Theor. Comput. Sci. 51 (2000) 305–322.

[45] M. Randić, A.F. Kleiner, L.M. De Alba, Distance/distance matrices, J. Chem. Inf. Comput. Sci. 34 (1994) 277–286.
[46] M. Randić, On characterization of three-dimensional structures, Int. J. Quantum Chem.: Quantum Biol. Symp. 15 (1988) 201–208.
[47] M. Randić, G. Krilov, On characterization of 3-D structure of proteins, Chem. Phys. Lett. 272 (1997) 115–119.
[48] M. Randić, G. Krilov, On a characterization of the folding of proteins, Int. J. Quantum Chem. 75 (1999) 1017–1026.
[49] G. Krilov, M. Randić, Quantitative characterization of proteins structures: application to a novel a/b fold, New J. Chem. 28 (2004) 1608–1614.
[50] L. Bytautas, D.J. Klein, M. Randić, T. Pisanski, Foldedness in linear polymers: a difference between graphical and Euclidean distances, DIMACS Ser. Discr. Math. Theor. Comput. Sci. 51 (2000) 39–61.
[51] M. Randić, Molecular profiles Novel geometry-dependent molecular descriptors, New J. Chem. 19 (1995) 781–791.
[52] M. Randić, Molecular shape profiles, J. Chem. Inf. Comput. Sci. 35 (1995) 373–382.
[53] M. Randić, G. Krilov, Bond profiles for cuboctahedron and twist cuboctahedron, Int. J. Quantum Chem.: Quantum Biol. Symp. 23 (1996) 127–139.
[54] M. Randić, On characterization of the conformations of nine-membered rings, Int. J. Quantum Chem.: Quantum Biol. Symp. 22 (1995) 61–73.
[55] M. Randić, G. Krilov, On characterization of molecular surfaces, Int. J. Quantum Chem. 65 (1998) 1065–1076.
[56] M. Randić, Novel graph theoretical approach to heteroatoms in quantitative structure–activity relationship, Chemom. Intel. Lab. Syst. 10 (1991) 213–227.
[57] M. Randić, On computation of optimal parameters for multivariate analysis of structure–property relationship, J. Comput. Chem. 12 (1991) 970–980.
[58] M. Randić, On characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[59] M. Randić, J.Cz. Dobrowolski, Optimal molecular connectivity descriptors for nitrogen-containing molecules, Int. J. Quantum Chem. 70 (1998) 1209–1215.
[60] M. Randić, S.C. Basak, Construction of high-quality structure–property–activity regressions: the boiling points of sulfides, J. Chem. Inf. Comput. Sci. 40 (2000) 899–905.
[61] M. Randić, S.C. Basak, M. Pompe, M. Novič, Prediction of gas chromatographic retention indices using variable connectivity index, Acta Chim. Slovenica 48 (2001) 169–180.
[62] M. Randić, High quality structure–property regressions. Boiling points of smaller alkanes, New J. Chem. 24 (2000) 165–171.
[63] M. Randić, D. Mills, S.C. Basak, On characterization of physical properties of amino acids, Int. J. Quantum Chem. 80 (2000) 1199–1209.
[64] M. Randić, S.C. Basak, On use of the variable connectivity index $^1\chi^f$ in QSAR: toxicity of aliphatic ethers, J. Chem. Inf. Comput. Sci. 41 (2001) 614–618.
[65] M. Randić, M. Pompe, The variable connectivity index $^1\chi^f$ versus traditional molecular descriptors: a comparative study of $^1\chi^f$ against descriptors of CODESSA, J. Chem. Inf. Comput. Sci. 41 (2001) 631–638.
[66] M. Pompe, M. Veber, M. Randić, A.T. Balaban, Using variable and fixed topological indices for the prediction of reaction rate constants of volatile unsaturated hydrocarbons with OH radicals, Molecules 9 (2004) 1160–1176.
[67] M. Randić, D. Plavšić, N. Lerš, Variable connectivity index for cycle-containing structures, J. Chem. Inf. Comput. Sci. 41 (2001) 657–662.
[68] M. Randić, J. Zupan, T. Pisanski, On representation of DNA by line distance matrix, J. Math. Chem., in press.
[69] M. Randić, S.C. Basak, Characterization of DNA primary sequences based on the average distances between bases, J. Chem. Inf. Comput. Sci. 41 (2001) 561–568.
[70] G. Jaklič, T. Pisanski, M. Randić, Visualizing Cauchy's interlacing property for line distance matrices, in: Presented at Pascal Workshop: Complex Objects Visualization 2005, Koper, Slovenia, November 16–19, 2005.
[71] G. Jaklič, T. Pisanski, M. Randić, J. Comput. Biol., in press.
[72] G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h, R. Buelov, Computer-assisted rational design of immunosuppressive compounds, Nat. Biotechnol. 16 (1998) 748–752.
[73] R.C. Read, The coding of various kinds of unlabelled trees, in: R.C. Read (Ed.), Graph Theory and Computing, Academic Press, New York, 1972, pp. 153–182.
[74] S.F. Boys, Electronic wavefunction I. A general method of calculation for stationary states of any molecular system, Proc. Roy. Soc. [London] A200 (1950) 542–554.
[75] J.R. Platt, Prediction of isomeric differences in paraffin properties, J. Phys. Chem. 56 (1952) 328–336.
[76] L.B. Kier, W.J. Murray, M. Randić, L.H. Hall, Molecular connectivity. V: connectivity series concept applied to density, J. Pharm. Sci. 65 (1976) 1226–1230.
[77] R. Todeschini, V. Consonni, in: R. Mannhold, H. Kubinyi, H. Timmerman (Eds.), Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry, vol.11, Wiley-VCH, 2000.
[78] M. Randić, On geometry-based approach to protein sequence alignment, J. Math. Chem., in press.
[79] M. Randić, J. Zupan, D. Vikić-Topić, D. Plavšić, A novel unexpected use of a graphical representation of DNA: graphical alignment of DNA sequences, Chem. Phys. Lett. 431 (2006) 375–379.
[80] M. Randić, 2-D graphical representation of proteins based on physicochemical properties of amino acids, Chem. Phys. Lett., submitted for publication.
[81] S.F. Altschul, W. Gish, W.W. Miller, G. Myers, D. Lipman, Basic local alignment tool, J. Mol. Biol. 215 (1990) 403–410.