Journal of Molecular Graphics and Modelling 27 (2009) 637–641

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Graphical representation of proteins as four-color maps and their numerical characterization

Milan Randić [a,*], Ketij Mehulić [b], Damir Vukičević [c], Tomaž Pisanski [d,e], Dražen Vikić-Topić [f], Dejan Plavšić [f,*]

[a] National Institute of Chemistry, P.O. Box 3430, 1001 Ljubljana, Slovenia
[b] School of Dental Medicine, University of Zagreb, Gundulićeva 5, 10000 Zagreb, Croatia
[c] Department of Mathematics, University of Split, Nikole Tesle 12, 21000 Split, Croatia
[d] IMFM, Department of Theoretical Computer Science, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia
[e] UP-PINT University of Primorska, Koper, Slovenia
[f] Ruđer Bošković Institute, NMR Center, P.O. Box 180, HR-10002 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 6 June 2008
Received in revised form 13 October 2008
Accepted 15 October 2008
Available online 1 November 2008

Keywords:
Protein structure
Graphical representation
Virtual genetic code
Four-color map
Structure matrix **S**
Protein descriptor

ABSTRACT

We put forward a novel compact 2-D graphical representation of proteins based on the concept of virtual genetic code and a four-color map. The novel graphical representation uniquely represents proteins and allows one to easily and quickly visually observe and inspect similarity/dissimilarity between them. It also leads to a novel protein descriptor, a 10-dimensional vector derived from a novel structure matrix **S** associated with the map. The introduced numerical characterization of proteins is not only useful for their comparative study, but also for cataloguing information on a single protein. The approach is illustrated with the A chain of human insulin and the A chain of human insulin analogue glargine.

1. Introduction

Like DNA and RNA, proteins are linear polymers. Unlike DNA whose graphical representations were initiated over 20 years ago [1–4], graphical representations of proteins have only recently been proposed [5–17]. The main reason for this delay is the great complexity of the primary structure of proteins being formed from a selection of 20 building blocks rather than 4 as is the case with DNA and RNA. Moreover, the direct extension of graphical representations of DNA to proteins results in complicated representations of proteins whose numerical characterization are computationally involved. For example, if one extends the most simple, most elementary and straightforward representation of DNA based on four horizontal lines [18,19] to proteins, then the vector characterizing a protein have 20!/2 components (a horrendous number $1.21645100 \times 10^{18}$) and not 4!/2 or 12 components as is the case with DNA and RNA sequences. In order to surmount the aforementioned difficulties and to construct simple and useful graphical representations of proteins, the concept of virtual genetic code [5] (VGC) was introduced (*vide infra*). A useful graphical representation of proteins should meet the following requirements: (1) to allow one to easily visually observe and inspect similarity/dissimilarity between proteins; (2) to be compact or rather compact with regard to the spatial requirement for display of a protein; and (3) to be the basis for a simple numerical characterization of proteins.

The characterization of molecular structure by invariants of molecular graph (topological indices) turned out to be highly useful in studies of molecules and examining similarity/dissimilarity between them, QSPR (quantitative structure property relationship), and QSAR (quantitative structure activity relationship) modeling [20–27], as well as in current trends in drug discovery, including QPTR (quantitative proteome-toxicity relationship) modeling [28–31]. It stands to reason that a similar way of characterizing proteins, e.g. by invariants of a graphical representation of proteins (protein descriptors), and the proteome can be useful in their study. For a review of the research on

* Corresponding authors.
E-mail addresses: mrandic@msn.com (M. Randić), mehulic@sfzg.hr (K. Mehulić), vukicevi@pmfst.hr (D. Vukičević), tomaz.pisanski@fmf.uni-lj.si (T. Pisanski), vikic@irb.hr (D. Vikić-Topić), dplavsic@irb.hr (D. Plavšić).

1093-3263/$ – see front matter © 2008 Published by Elsevier Inc.
doi:10.1016/j.jmgm.2008.10.004

proteomics, networks and connectivity indices consult a recent review by González-Diaz et al. [32], and for a review of quantitative characterizations of proteome maps see Ref. [33].

As a rule, the characterization of a chemical object by a mathematical invariant entails some loss of information about the structure of the object. To reduce the loss of information one can characterize the object by a set of invariants encoding different information about its structure. Different graphical representations of a protein give different views of its structure. Therefore, different graphical representations of proteins and on them based protein descriptors encode different information about the structure of proteins. Clearly, it is justifiable and desirable to construct novel graphical representations of proteins and novel protein descriptors because they can provide one with new information about the structure of proteins and enable one to better characterize proteins. Therefore, here we propose a novel compact 2-D graphical representation of proteins based on a four-color map and VGC. This representation also leads to a novel numerical characterization of proteins based on a 10-dimensional vector, which is not only useful for comparative study of proteins, but also for cataloguing information about a single protein as well as computer screening of proteins.

2. Virtual genetic code

A subset of the standard genetic code consisting of 20 codons coding for the 20 naturally occurring amino acids is called virtual genetic code [5]. One should note that VGC is a mathematical construct like graphical or matrix representations of DNA and not a biological concept. The number of possible VGCs is 339 738 624 = $1^2 \cdot 2^9 \cdot 3^1 \cdot 4^5 \cdot 6^3$ because two amino acids are specified by one codon, nine amino acids are coded by two codons, one amino acid is specified by three codons, five amino acids are coded by four codons, and three amino acids are specified by six codons. When one of VGCs is selected, then there exists a bijection between the set of (biosynthesized) proteins and the set of sequences of codons of the selected VGC that would produce these proteins if they were actually present in mRNAs. This bijection enables one to represent a protein by the corresponding hypothetical mRNA sequence and in this way the difficult problem of graphical representation of proteins, a "problem involving 20 letters," is transformed into a "problem involving 4 letters." In Table 1 we list the VGC, proposed in Ref. [5], which will be used in this article and whose codons were selected in such a way that each base appears with the same frequency.

Table 1
The virtual genetic code.

Codon	Amino acid (three-letter symbol)
GCG	Ala
CGG	Arg
AAC	Asn
GAU	Asp
UGC	Cys
CAU	His
CAG	Gln
GAA	Glu
GGC	Gly
AUC	Ile
CUC	Leu
AAG	Lys
AUG	Met
UUC	Phe
CCA	Pro
UCU	Ser
ACA	Thr
UGG	Trp
UAU	Tyr
GUU	Val

3. Outline of the novel graphical representation of proteins

The novel graphical representation of proteins is based on the concept of VGC and a four-color map. The basic information about the construction of a four-color map representing a DNA sequence has been given in an earlier paper [34]. The construction of four-color maps representing proteins will be illustrated with the A chain of human insulin whose primary structure is Gly–Ile–Val–Glu–Gln–Cys–Cys–Thr–Ser–Ile–Cys–Ser–Leu–Tyr–Gln–Leu–Glu–Asn–Tyr–Cys–Asn [35].

The first step in constructing the four-color map representing the A chain of human insulin is to convert its amino acid residues into the corresponding codons listed in Table 1. The obtained hypothetical sequence of nucleotide bases GGCAUCGUUGAACA-GUGCUGCACAUCUAUCUGCUCUCUCUAUCAGCUCGAAAACUAUUG-CAAC uniquely represents the A chain of human insulin. The next step is to represent this sequence of nucleotide bases graphically. As a template for that a spiral of fused square cells is taken. The spiral whose one edge is emphasized for its better visibility is shown on the left in Fig. 1. It starts with the center cell containing G and ends in the last cell at the periphery containing C. The spiral can also be viewed as a part of the square grid in which individual cells have one of the four letters, G, C, U, and A. When one replaces the four letters by four colors, e.g. G with red, C with green, U with blue, and A with yellow as we do in this article, and deletes the sides between adjacent cells (two cells are adjacent if and only if they have a common side) having the same color, then one obtains the four-color map representing the A chain of human insulin. The map is shown on the right in Fig. 1. One should note that the four-color map representing a protein differs from an arbitrary four-color map: (1) its regions are uniquely colored (up to arbitrary choice of four colors) because of assigning the same color to all regions belonging to the same kind of nucleotide base; and (2) boundaries between its regions are based on square cells.

It is clear that the four-color maps represent proteins uniquely and rather compactly. Moreover, two proteins having similar primary structures will generate similar four-color maps because except for the region(s) associated with the bases in which the sequences of nucleotide bases representing the two proteins differ, most of the remaining part of the two maps will be identical. We will illustrate that with the A chain of human insulin and the A chain of glargine being a new long-acting human insulin analog produced by recombinant DNA technology utilizing a non-pathogenic laboratory strain of *Escherichia coli* as a production organism [36,37]. The primary structures of the A chain of human insulin and the A chain of glargine differ only in position 21, where the former has Asn and the latter Gly [37]. The four-color map representing the A chain of glargine is shown in Fig. 2. The visual comparison of this map with the one associated with the A chain of human insulin clearly reveals that they are very similar. To wit, the corresponding regions of these two maps are identical apart from the red region labeled 6 (belonging to GG of the triplet GGC specifying Gly) on the map representing the A chain of glargine that appears on the map of the A chain of human insulin as the yellow region labeled 41 (belonging to AA of the triplet AAC coding for Asn). In this case a single codon change (a single amino acid change) has brought about the change of color of a single region on the map. In general, a single codon change results in changing color of region(s) associated with the codon as well as changing boundaries between this (these) and neighboring regions.

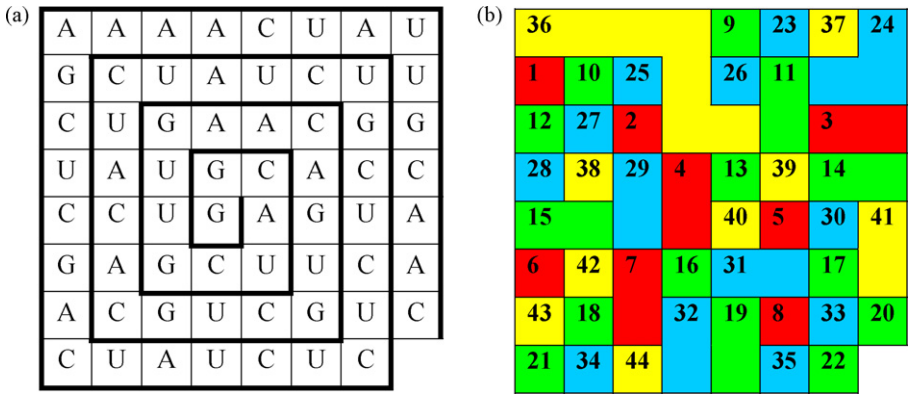

Fig. 1. (a) Representation of the hypothetical sequence of nucleotide bases, which uniquely represents the A chain of human insulin, as a spiral of square cells. (b) Four-color map representation of the A chain of human insulin. The red regions belonging to G are labeled 1–8, the green regions belonging to C are labeled 9–22, the blue regions belonging to U are labeled 23–35, and the yellow regions belonging to A are labeled 36–44.

4. Numerical characterization of proteins

Our objective is also to arrive at a numerical characterization of four-color maps depicting proteins because it allows a quantitative comparison of proteins and obtaining an important additional insight into their similarity/dissimilarity. A possibility to achieve that is to characterize the map by mathematical invariants. In order to find some of the invariants suitable for characterizing proteins, we will transform the four-color map, being essentially a nonnumerical mathematical object, into another mathematical object, a matrix. Once one has a matrix representing a protein, one can use it to construct protein descriptors.

A matrix associated with the four-color map should contain information about essential characteristics of its structure. They are (1) the number of regions; (2) the size of each region; (3) the color of each region; and (4) distances between regions. In the preliminary outline of the four-color map representation of DNA the used structure matrix encodes only a part of the required information about the map [34]. Here we introduce a novel matrix, the structure matrix **S**, satisfying all of the aforementioned requirements. The **S** matrix of the four-color map consisting of n regions is the symmetric matrix of order n whose entry in the i-th row and j-th column, $[\mathbf{S}]_{ij}$, is defined by the expression

$$[\boldsymbol{S}]_{ij} = \begin{cases} d_{ij} & \text{if } i \neq j \\ s_i & \text{if } i = j \end{cases}$$

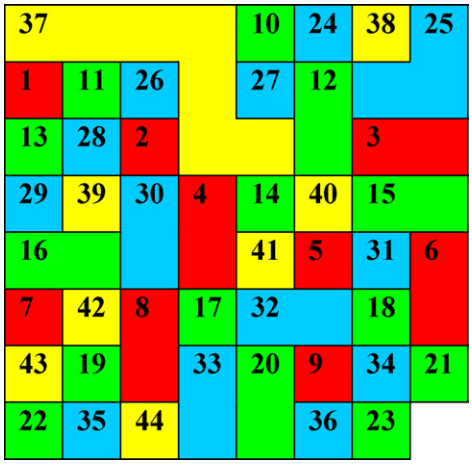

Fig. 2. The four-color map representing the A chain of insulin glargine. Regions on the map are numerically labeled.

where d_{ij} denotes the distance between regions r_i and r_j defined as the minimum number of borders that one has to cross in going from region r_i to region r_j. The symbol s_i represents the size of region r_i defined by the area of the region r_i given by the number of square cells constituting it. The number of the regions, sizes of the regions, and distances between the regions are incorporated in **S** through its order, diagonal entries, and off-diagonal entries, respectively. Information about colors of regions is included in the **S** matrix by its partitioning into submatrices belonging to different colors and pairs of different colors. In Table 2 such a partition is schematically exemplified by the **S** matrix associated with the map shown on the right in Fig. 1 whose red, green, blue, and yellow regions are labeled 1–8, 9–22, 23–35, and 36–44, respectively. In this way, one gets four square submatrices **GG**, **CC**, **UU**, and **AA** that lie along the diagonal of **S** and are associated with pairs GG, CC, UU, and AA of the same color regions, respectively. The dimensions of these submatrices are 8×8, 14×14, 13×13, and 9×9 respectively. One also obtains 12 off-diagonal submatrices. The six of them **GC**, **GU**, **GA**, **CU**, **CA**, and **UA** are in the upper triangle of **S** and are associated with pairs GC, GU, GA, CU, CA, and UA of the differently colored regions respectively. Their dimensions are respectively 8×14, 8×13, 8×9, 14×13, 14×9, and 13×9. The remaining six submatrices **CG**, **UG**, **UC**, **AG**, **AC**, and **AU** are in the lower triangle of **S** and each of them is the transpose of the corresponding submatrix in the upper triangle, e.g. **CG** = (**GC**)$^{\mathrm{T}}$.

It might be thought that the construction of the **S** matrix is involved but it is not. To wit, each map in a plane can be represented by its dual graph being a planar graph whose vertices represent regions and edges connect two vertices if the regions represented by these vertices are adjacent [38]. Two regions that touch at only one point are not considered adjacent. As our definition of distance between regions on the four-color map is equivalent to the definition of distance between vertices in a graph, the **S** matrix of the map can easily be constructed using its dual

Table 2
The schematic representation of the partition of the **S** matrix associated with the four-color map shown in Fig. 1 into the square submatrices **GG**, **CC**, **UU**, and **AA** lying along the main diagonal and the off-diagonal submatrices **GC**, **GU**, **GA**, **CG**, **CU**, **CA**, **UG**, **UC**, **UA**, **AG**, **AC**, and **AU**.

	1–8	9–22	23–35	36–44
1–8	**GG**	**GC**	**GU**	**GA**
9–22	**CG**	**CC**	**CU**	**CA**
23–35	**UG**	**UC**	**UU**	**UA**
36–44	**AG**	**AC**	**AU**	**AA**

graph and one of the standard procedures for constructing the distance matrix of a graph [39–42].

We will now construct a novel protein descriptor using the **S** matrix. The characterization of the four-color map by a descriptor being a single number entails some loss of information about the structure of the map. For instance, if the four-color map is characterized by the leading eigenvalue of **S** or by the average matrix element of **S**, then the information about colors of regions is lost. Therefore, the novel protein descriptor will be a 10-dimensional vector whose components are respectively suitable invariants of the **GG**, **GC**, **GU**, **GA**, **CC**, **CU**, **CA**, **UU**, **UA**, and **AA** submatrices of the **S** matrix. As the off-diagonal submatrices of **S** are not generally square matrices, one cannot use eigenvalues for their characterization. We will, therefore, use the average matrix elements of the off-diagonal submatrices, $\left(\sum_{i=1}^{i=k}\sum_{j=1}^{j=l}[\boldsymbol{M}]_{ij}\right)/(k\cdot l)$, **M** = **GC**, **GU**, **GA**, **CU**, **CA**, **UA**, and the average matrix elements of the upper triangles of the submatrices lying along the diagonal of **S**, $2\left(\sum_{i=1}^{i=m}\sum_{j=i}^{j=m}[\boldsymbol{M}]_{ij}\right)/(m^2+m)$, **M** = **GG**, **CC**, **UU**, **AA**, as the components. If one out of the four colors does not appear on the map, then the components of the 10-dimensional vector that correspond to the missing color are by definition zero. For example, if the "four-color" map does not contain green regions, then the second, fifth, sixth, and seventh component of the 10-dimensional vector characterizing the map are zero. The 10-dimensional vector encodes information on the number of colors appearing on the map and if this number is less than 4 then the vector reflects which of the four colors is (are) missing. It also encodes information about the number of regions of a given color, the average value of the sum of sizes of regions of a given color and distances between them, and the average distances between differently colored regions.

The 10-dimensional vector characterizing the A chain of human insulin is (212/72, 392/112, 359/104, 243/72, 776/210, 700/182, 476/126, 672/182, 441/117, 326/90). A comparison of this vector with the one associated with the A chain of glargine (278/90, 444/126, 412/117, 247/72, 776/210, 700/182, 424/112, 672/182, 388/104, 252/72) reveals that they are different. The differences between their corresponding components are either zero or are small as expected. The largest difference is between the first components and it is about 5%. The components derived from **CC**, **CU**, and **UU** are identical for both vectors because no changes were introduced concerning the respective regions. These findings show that the four-color map representation of proteins, the **S** matrix, and the constructed 10-dimensional vector have captured important features of the proteins considered.

The question poses itself: Is the characterization of the four-color maps (proteins) by the 10-dimensional vector unique? We will give the answer in the form of the following claim and an example proving the claim.

Claim. There exist at least two different four-color maps depicting proteins that are characterized by the same 10-dimensional vector.

Proof. Let P_1 and P_2 be two different proteins whose primary structures are respectively Glu–Ala–Leu–Cys–Ile–Phe–Trp–Gly–Thr–Ala–Thr–Gly–Arg–Phe–Trp–Glu–Leu–Leu–Cys–Ala and Glu–Ala–Leu–Cys–Ile–Phe–Trp–Gly–Thr–Ala–Thr–Gly–Arg–Ser–Trp–Glu–Leu–Leu–Cys–Ala and let M_1 and M_2 be the four-color maps representing them respectively. The maps M_1 and M_2, shown in Fig. 3, differ just in a green region labeled 19 and two blue regions labeled 27 and 28 whose sizes on M_1 and M_2 are 1, 3, 1 and 1, 2, 2, respectively. These regions on both maps have the same neighbors. Therefore, the structure matrixes $\mathbf{S}_1$ and $\mathbf{S}_2$ associated with the maps M_1 and M_2, respectively, differ only in the two diagonal entries, $[\mathbf{S}_1]_{27\,27} \neq [\mathbf{S}_2]_{27\,27}$ and $[\mathbf{S}_1]_{28\,28} \neq [\mathbf{S}_2]_{28\,28}$, belonging to their submatrices **UU**. Since $[\mathbf{S}_1]_{27\,27} + [\mathbf{S}_1]_{28\,28} = [\mathbf{S}_2]_{27\,27} + [\mathbf{S}_2]_{28\,28} = 4$, it follows that the 10-dimensional vectors characterizing the four-color maps M_1 and M_2 are equal.

The nonuniqueness of the characterization by the 10-dimensional vector may seem to be a serious drawback of the approach, but this is not the case in practice. We have demonstrated that by an *in silico* experiment. Using the Monte Carlo method we generated a family of 50 000 very similar proteins of length 200 aa. The leading protein, P_1, of the family was obtained by selecting amino acids at random. The remaining proteins, P_n, $2 \leq n \leq 5 \times 10^4$, of the family were obtained from P_1 as follows: the amino acid at a random position in P_1 is replaced by a randomly chosen different amino acid. For each of the generated proteins the corresponding four-color map and the corresponding 10-dimensional vector were constructed using in-house developed computer program [43]. Among the 1 259 975 000 pairs of the so constructed 10-dimensional vectors there were only 16 pairs of equal vectors. Half of these 16 pairs arose from the pairs of distinct proteins (nondiscriminatory mates), while the other half of pairs resulted from duplicate proteins. We repeated the same computational experiment two more times. In the first repetition 23 pairs of equal 10-dimensional vectors were found and 14 pairs out of them arose from the distinct proteins (nondiscriminatory mates) while in the second repetition there were 12 pairs of equal 10-dimensional vectors and only 4 pairs out of them resulted from

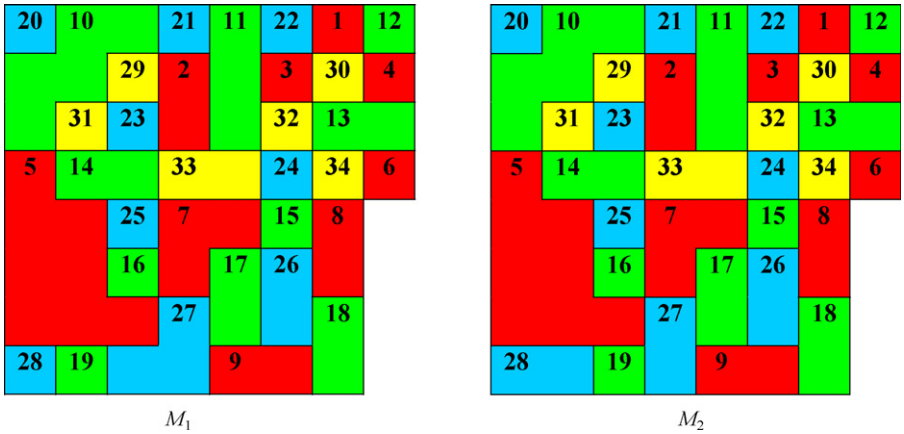

Fig. 3. The four-color maps M_1 and M_2 representing proteins P_1 (Glu–Ala–Leu–Cys–Ile–Phe–Trp–Gly–Thr–Ala–Thr–Gly–Arg–Phe–Trp–Glu–Leu–Leu–Cys–Ala) and P_2 (Glu–Ala–Leu–Cys–Ile–Phe–Trp–Gly–Thr–Ala–Thr–Gly–Arg–Ser–Trp–Glu–Leu–Leu–Cys–Ala), respectively. Regions on the maps M_1 and M_2 are numerically labeled.

the distinct pairs of protein. Thus, on average the probability that two different proteins of length 200 aa are characterized by the same 10-dimensional vector is about 10^{-8}. In the case of longer proteins this negligible probability will be even lower.

5. Concluding remarks

In this article, we have outlined a novel graphical representation of proteins enabling one to quickly and easily visually observe and inspect similarity/dissimilarity between proteins. It is also the basis for the construction of a novel protein descriptor being a 10-dimensional vector allowing one to make a quantitative comparison of proteins. The components of the vector are invariants of the submatrices of the novel structure matrix **S**, associated with the four-color map, which is easy to construct. The four-color map representation of proteins and the structure matrix **S** as well as the protein descriptor 10-dimensional vector capture important features of proteins. Owing to the conceptual simplicity, computational efficiency and efficacy we feel that the outlined approach will be of interest to the researcher in the field of proteins.

Acknowledgements

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under the Projects 098-0982929-2917, 065-0650446-0435, 037-0000000-2779, and 177-0000000-0884 and by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia through the Projects P1-017, ARRS 1000-08-210382, P1-0294, JI-6062, and LI-7230. M.R. thanks the National Institute of Chemistry, Ljubljana, Slovenia for warm hospitality.

References

[1] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 285 (1983) 1318–1327.
[2] E. Hamori, Novel DNA sequence representations, Nature 314 (1985) 585–586.
[3] M.A. Gates, Simpler DNA sequence representations, Nature 316 (1985) 219.
[4] M.A. Gates, A simple way to look at DNA, J. Theor. Biol. 119 (1986) 319–328.
[5] M. Randić, 2-D graphical representation of proteins based on virtual genetic code, SAR QSAR Environ. Res. 15 (2004) 147–157.
[6] 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.
[7] M. Randić, A.T. Balaban, M. Novič, A. Založnik, T. Pisanski, A novel graphical representation of proteins, Period. Biol. 107 (2005) 403–414.
[8] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, Chem. Phys. Lett. 419 (2005) 528–532.
[9] F. Bai, T. Wang, On graphical and numerical representation of protein sequences, J. Biomol. Struct. Dyn. 25 (2006) 537–545.
[10] 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) 583–595.
[11] M. Randić, J. Zupan, D. Vikić-Topić, On representation of proteins by star-like graphs, J. Mol. Graphics Modell. 26 (2007) 290–305.
[12] B. Horvat, T. Pisanski, M. Randić, Terminal polynomials and star-like graphs, MATCH-Commun. Math. Comput. Chem. 60 (2008) 493–512.
[13] M. Randić, M. Novič, Representation of proteins as walks in 20-D space, SAR QSAR Environ. Res. 19 (2008) 317–337.
[14] M. Randić, M. Novič, M. Vračko, On novel representation of proteins based on amino acid adjacency matrix, SAR QSAR Environ. Res. 19 (2008) 339–349.
[15] M. Randić, 2-D graphical representation of proteins based on physico-chemical properties of amino acids, Chem. Phys. Lett. 444 (2007) 176–180.
[16] M. Randić, D. Plavšić, Novel matrix and graphical representation of proteins: amino acid adjacency matrix, Chem. Phys. Lett., submitted for publication.
[17] M. Randić, J. Zupan, D. Vikić-Topić, A.T. Balaban, D. Plavšić, On graphical representation of proteins, Chem. Rev., submitted for publication.
[18] 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.
[19] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (2003) 202–207.
[20] M. Randić, Topological indices, in: P.v.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer, III, P.R. Schreiner (Eds.), Encyclopedia of Computational Chemistry, Wiley, Chichester, 1998, pp. 3018–3032.
[21] M. Reinhard, A. Drefahl, Handbook for Estimating Physicochemical Properties of Organic Compounds, Wiley, New York, 1999.
[22] J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, Amsterdam, 1999.
[23] M.V. Diudea (Ed.), QSPR/QSAR Studies by Molecular Descriptors, Nova Science Publishers, Huntington, 2001.
[24] D.H. Rouvray, R.B. King (Eds.), Topology in Chemistry: Discrete Mathematics of Molecules, Horwood, Chichester, 2002.
[25] M.V. Diudea, M.S. Florescu, P.V. Khadikar, Molecular Topology and Its Applications, EfiCon, Bucharest, 2006.
[26] A.M. Johnson, G.M. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.
[27] M. Randić, Similarity methods of interest in chemistry, in: S.I. Kuchanov (Ed.), Mathematical Methods in Contemporary Chemistry, Gordon and Breach, Amsterdam, 1996, pp. 1–100.
[28] H. González-Diaz, S. Vilar, L. Santana, E. Uriarte, Medicinal chemistry and bioinformatics—current trends in drug discovery with networks topological indices, Curr. Top. Med. Chem. 7 (2007) 1015–1029.
[29] S. Vilar, H. González-Diaz, L. Santana, E. Uriarte, QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice network, J. Comput. Chem., in press.
[30] M. Cruz-Monteagudo, H. González-Diaz, F. Borges, E.R. Dominguez, M. Natália, D.S. Cordeiro, 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. Quantitative proteome-toxicity relationships (QPTR) based on mass spectrum spiral entropy, Chem. Res. Toxicol. 21 (2008) 619–632.
[31] G. Ferino, H. González-Diaz, G. Delogu, G. Podda, E. Uriarte, Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer, Biochem. Biophys. Res. Commun. 372 (2008) 320–325.
[32] H. González-Diaz, Y. González-Diaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics, networks and connectivity indices, Proteomics 8 (2008) 750–778.
[33] M. Randić, Quantitative characterization of proteomics maps by matrix invariants, in: P.M. Conn (Ed.), Handbook of Proteomics Methods, Humana Press, INC. Totowa, NY, 2003, pp. 429–450.
[34] M. Randić, N. Lerš, D. Plavšić, S.C. Basak, A.T. Balaban, Four-color map representation of DNA or RNA sequences and their numerical characterization, Chem. Phys. Lett. 407 (2005) 205–208.
[35] http://www.expasy.org/cgi-bin/sprot-ft-details.pl?P01308@PEPTIDE@90@110.
[36] K. McKeage, K.L. Goa, Insulin glargine: a review of its therapeutic use as a long-acting agent for the management of type 1 and 2 diabetes mellitus, Drugs 61 (2001) 1599–1624.
[37] H. Thisted, S.P. Johnsen, J. Rungby, An update on the long-acting insulin analogue glargine, Basic Clin. Pharmacol. Toxicol. 99 (2006) 1–11.
[38] K.H. Rosen, Discrete Mathematics and Its Applications, 6th ed., McGraw-Hill, Boston, 2007.
[39] M. Barysz, D. Plavšić, N. Trinajstić, A note on topological indices, MATCH-Commun. Math. Comput. Chem. 19 (1986) 89–116.
[40] Z. Mihalić, D. Veljan, D. Amić, S. Nikolić, D. Plavšić, N. Trinajstić, The distance matrix in chemistry, J. Math. Chem. 11 (1992) 223–258.
[41] H.P. Schultz, Topological organic chemistry. 13. Transformation of graph adjacency matrixes to distance matrixes, J. Chem. Inf. Comput. Sci. 40 (2000) 1158–1159.
[42] D. Janežič, A. Miličević, S. Nikolić, N. Trinajstić, Graph Theoretical Matrices in Chemistry, Mathematical Chemistry Monographs 3, Series Editor Ivan Gutman. University of Kragujevac and Faculty of Science, Kragujevac, 2007.
[43] The program is available freely to non-commercial users on request from D.V. or T.P.