24 March 2000

Chemical Physics Letters 319 (2000) 713–718

CHEMICAL PHYSICS LETTERS

www.elsevier.nl/locate/cplett

Characterization of 3D molecular structure

Ernesto Estrada [*]

Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, E-15706 Santiago de Compostela, Spain

Received 13 October 1999; in final form 1 February 2000

Abstract

A novel approach to describe the 3D structure of small/medium-sized and large molecules is introduced. A vector and an index are defined on the basis of considering second line graphs with edges weighted by the dihedral angles of the molecule. They measure the 3D 'compactness' or folding of the molecular structures, giving maximum values for the most folded structures. We have ranked five protein models according to their degree of folding. The similarity among these proteins has been determined, showing that the most folded proteins are not similar among them, while the less folded ones are similar to each other. © 2000 Elsevier Science B.V. All rights reserved.

1. Introduction

The problem of characterizing the molecular structure is of central importance for all chemistry [1]. The study of structure–property–activity relationships (QSPR/QSAR) is very dependent on an appropriate definition and selection of the molecular structure descriptors [2]. In this sense, much attention has been paid to the characterization of the 2D structure of molecules through graph-theoretical invariants [3,4]. It was only at the end of the last decade, and beginning of the present, that the first attempts to characterize the 3D molecular structure by graph-based methods were carried out [5–13]. Very recently Randic has introduced the molecular and shape profiles to quantitatively describe the molecular structure [14,15]. This approach appears to be interesting because it permits a characterization of the 3D structure of small and large molecules. It was recently applied to the characterization of the 3D sequences of proteins [16].

The other motivation for the present work comes from our proper results on the study of the spectral moments of structural matrices related to line graphs [17–20]. The importance of the use of iterated line graph sequences (ILGS) for the definition of molecular descriptors has also been claimed very recently [21]. Here we study the spectral moments of the edge matrix of the second line graph of the ILGS [22] for the generation of molecular descriptors that characterize the 3D structure of molecules. The application of this approach to the characterization of the 3D structure of small/medium-sized molecules, and to the study of protein folding, are also reported.

2. Theoretical approach

The present approach is based on the consideration of a graph representing the dihedral angles in

[*] Universidad de Santiago de Compostela, Facultad de Farmacia, Laboratorio de Quimica Farmaceutica, Campus Universitario Sur, 15706 Santiago de Compostela, Spain. Fax: +34-981-594912; e-mail: estrada66@yahoo.com

0009-2614/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0009-2614(00)00158-5

the molecule. This graph is constructed from the molecular graph following the ILGS scheme [21]. First we consider a molecular graph, G, in which the vertices represent atoms, and the edges the bonds in the molecule (see G for the graph representing biphenyl in Fig. 1). Now we identify the edges of G as the vertices of a new graph permitting that two vertices of the new graph are adjacent, if and only if, the corresponding edges of G are incident to the same vertex. In this way we obtain the first line graph $L(G)$ of the molecular graph. Since $L(G)$ is itself a graph, one can construct the line graph of $L(G)$, which will be denoted $L^2(G)$ and called the second line graph of G (see Fig. 1). The third line graph of G is denoted by $L^3(G)$ and so on. The pictorial representation of $L^3(G)$ is not shown in Fig. 1 for the sake of simplicity, it contains 26 vertices and 70 edges.

The vertices of $L(G)$ represent the bonds of the molecule, similarly, the vertices in $L^2(G)$ represent bond angles and those of $L^3(G)$ represent the dihedral angles of the corresponding molecule represented by G. Consequently, if we consider the adjacency matrix of $L^3(G)$ with the vertices weighted by a function of the corresponding dihedral angles of the molecule, we can generate graph theoretical invariants that contain 3D molecular information. The resulting matrix, $\mathbf{M}$, is defined as follows: $\mathbf{M} = \mathbf{A}(L^3(G)) + \Delta$, where $\mathbf{A}(L^3(G))$ denotes the vertex adjacency matrix of $L^3(G)$ and Δ is a diagonal matrix whose nth element on the diagonal is the cosine of the nth dihedral angle of G. The use of the cosine of the angles instead of the angles per se is only for avoiding the manipulation of very large numbers in the spectral moment calculations. The dihedral angle between two bond angles at adjacent vertices of G is the angle between the two half-planes with their boundaries being mutual along the common bond of the two bond angles, with the first half-plane containing the 3 atoms of the first bond angle, and with the second half-plane containing the 3 atoms of the second bond angle. Here we use dihedral angles with a value of 0° corresponding to the *cis* arrangement of a polyene, while the *trans* arrangement corresponds to a dihedral angle of 180°.

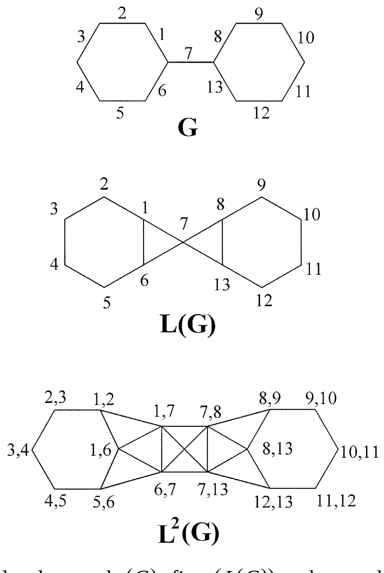

Fig. 1. Molecular graph (G), first ($L(G)$) and second line graph ($L^2(G)$) of biphenyl.

The use of the $L^3(G)$ in order to generate 3D molecular descriptors permits the generalization of ILGS as a source for characterization of the molecular structure. By this means, a series of molecular descriptors, based on the use of bond distances, bond angles and dihedral angles as diagonal entries in $\mathbf{A}(L(G))$, $\mathbf{A}(L^2(G))$ and $\mathbf{A}(L^3(G))$, respectively, can be generated. The use of other, apparently appropriate, ways to introduce the dihedral angles as weights for graph theoretical matrices is, in some cases not as general as the present way. For instance, the use of $\mathbf{A}(L(G))$ with the off-diagonal elements weighted by the cosines of the corresponding dihedral angles is appropriate neither for star graphs nor for the C_3 graph. In these cases, e.g. consider the hydrogen-depleted graph of 2,2-dimethylpropane or cyclopropane, there is no way to introduce the cosines of the dihedral angles because all the off-diagonal elements of $\mathbf{A}(L(G))$ are ones due to all the edges in these graphs which are adjacent to each other.

The 3D-structure descriptors that we consider here are the kth spectral moments, μ_k, of $\mathbf{M}$. The spectral moments of $\mathbf{M}$ are defined as: $\mu_k = \mathrm{tr}(\mathbf{M}^k)$, where tr means the trace, i.e., the sum of the diagonal entries of the corresponding matrix. These spectral moments are divided by $k!$, where k is the order of the spectral moment. By this means we obtain vectors, $\boldsymbol{v}$, composed of the different spectral moments normalized by $k!$ and containing the information on the 3D molecular structure of each molecule. A 3D-structure index, I, is then defined as the sum of the elements of the vector. In order to avoid arbitrariness in the selection of the number of elements of $\boldsymbol{v}$ that should be taken into account, I is defined as follows: $I = \mathrm{tr}(e^{\mathbf{M}})$. Consequently, I can be computed by summing the exponentials of the eigenvalues λ_i of $\mathbf{M}$:

$$I = \sum_i \exp(\lambda_i) .$$

3. Small / medium-sized molecules

With the objective of investigating the main features of the vector of spectral moments and of the 3D-structure index, we studied their changes with the conformational changes in small/medium-sized molecules. We first studied the six geometrical isomers of 1,3,5-hexatriene: three conformers of the *trans*-hexatriene (*s-trans*, *s-trans* (**1**); *s-trans*, *s-cis* (**2**) and *s-cis*, *s-cis* (**4**)) and three conformers of the *cis*-hexatriene (*s-trans*, *s-trans* (**3**); *s-trans*, *s-cis* (**5**) and *s-cis*, *s-cis* (**6**)). The numbers in parentheses correspond to numbering for these compounds in Fig. 2. A graphical representation of the spectral moment vectors for these compounds is shown in Fig. 2. As can be seen, there are significant variations in the shape of the spectral moment vectors with the changes in the 3D structure of the hexa-

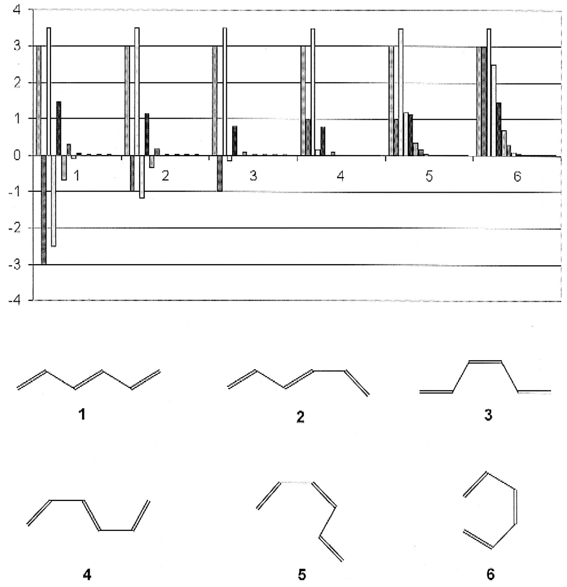

Fig. 2. Graphical representation of the 3D spectral moments normalized by $k!$ for the six isomers of 1,3,5-hexatriene represented below.

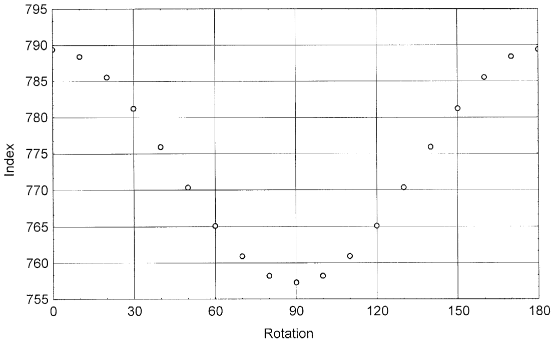

Fig. 3. Plot of the 3D-structure index vs. the dihedral angle formed by the two phenyl rings of biphenyl.

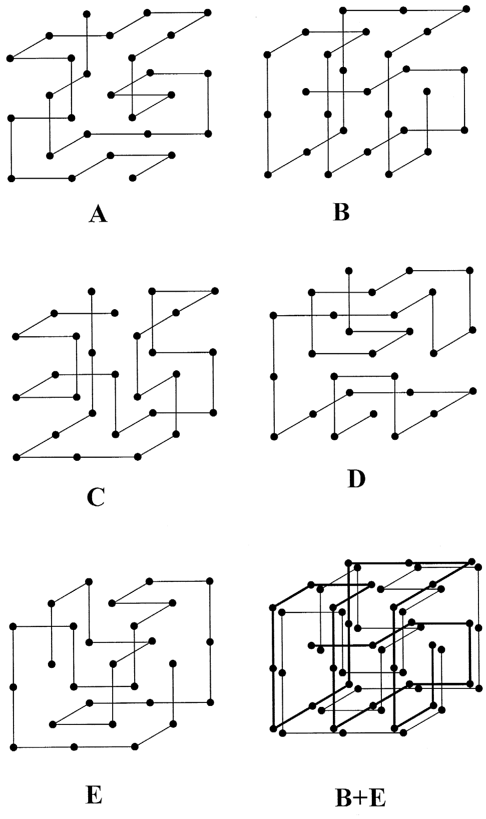

Fig. 4. Representation of the five protein models studied (A–E) as well as the superimposition of the two most similar ones (B + E).

triene isomers. The 3D-structure index takes the following values for these isomers: 1.970, 5.245, 6.197, 8.577, 10.351, and 14.560. This confirms the graphical view that the most folded isomer has the greatest values of the vector entries and consequently of the 3D-structure index. Consequently, we can consider both the vector and the sum of its elements as a measure of the ‘compactness’ of the molecular structure, taking the greater values for the most folded or ‘compact’ of the structures in a series of isomers.

Similar results are obtained for other molecular systems such as *chair*- and *boat*-cyclohexane, the first can be considered as the least folded and the second as the most folded or compact structure. By computing the 3D-structure vector for these conformers we obtain the following results: chair $\boldsymbol{v} = (6, 3, 6.75, 3.125, 2.266, 0.877, 0.387, 0.124, \ldots)$ and for boat $\boldsymbol{v} = (6, 4.464, 7.750, 4.939, 3.304, 1.577, 0.696, 0.259, \ldots)$, which produce indices with values of 22.583 and 29.114 for the chair and boat, respectively. This result confirms that the most compact structure has the greatest value of the index I.

The third example that we study with the use of the present approach for small/medium-sized molecules is the rotation of the single bond of biphenyl. We start with the planar structure and will rotate the dihedral angle 1–7–8 (see Fig. 1) with steps of 10°. In Fig. 3 we plot the values of the 3D-structure index vs. the rotation angle for the biphenyl. It is observed that the planar conformation possesses the greatest value of the index coinciding with the intuitive idea that it is the most folded structure. The smallest value of I is obtained for the *s*-90 isomer that is the least folded of all the rotamers of biphenyl.

4. Protein folding

The number of possible conformations that a protein can adopt is enormous. For instance, a protein of length 100 can take a total of about 10^{49} conformations. On the other hand, the possible number of sequences that can be built with 20 amino acids for this length is also huge: 20^{100}. These factors make the study of protein folding so complex that approaches used for its characterization are not even fully defined. However, the importance that it has for the understanding of complex biological systems makes its study a rapidly developing field in biological chemistry [23].

Table 1
Spectral moments normalized by k! of the 5 protein models given in Fig. 4

Order	*A*	*B*	*C*	*D*	*E*
0	24.0000000	24.0000000	24.0000000	24.0000000	24.0000000
1	3.0000000	3.0000000	5.0000000	2.0000000	3.0000000
2	30.5000000	25.5000000	27.5000000	26.0000000	26.5000000
3	3.5000000	3.5000000	5.3333333	2.3333333	3.0000000
4	9.8750000	6.9583333	8.1250000	7.6666667	7.5416667
5	0.9416667	0.9833333	1.5000000	0.6000000	0.8166667
6	1.4430556	0.8791667	1.1138889	1.0388889	0.9777778
7	0.1172619	0.1255952	0.1954365	0.0726190	0.1019841
8	0.1205605	0.0652530	0.0889633	0.0807540	0.0725446
9	0.0084656	0.0092841	0.0147211	0.0052221	0.0072917
10	0.0065314	0.0032341	0.0046792	0.0040735	0.0034937
11	0.0003997	0.0004512	0.0007249	0.0002503	0.0003377
12	0.0002482	0.0001157	0.0001749	0.0001443	0.0001182
13	0.0000133	0.0000156	0.0000252	0.0000086	0.0000110
14	0.0000070	0.0000031	0.0000049	0.0000038	0.0000030
15	0.0000003	0.0000004	0.0000007	0.0000002	0.0000003
I	73.5132	65.0248	72.8769	63.8020	66.0219

Table 2
The similarity/dissimilarity matrix for the five protein models by considering the spectral moments normalized by $k!$

	A	*B*	*C*	*D*	*E*
A	0.000	5.816	4.455	5.270	4.683
B		0.000	3.612	1.813	1.276
C			0.000	4.614	3.358
D				0.000	1.327

Recently, Randic and Krilov have proposed the characterization of 3D sequences of proteins based on molecular profiles [16]. They studied the 5 schematic 3D representations of proteins obtained by Li et al. as the most 'designable' proteins of length 27 composed by polar and hydrophobic amino acids occupying all sites of a $3 \times 3 \times 3$ cube [24]. Randic and Krilov found that the molecular profile approach is able to characterize the 3D structures of such protein models.

In Fig. 4 we show the forms of these 5 proteins according to the same design given in [24]. In Table 1 we give the values of $\boldsymbol{v}$ and I for each of these proteins. According to our previous results concerning the 3D structure index, the most folded protein is A and the least folded is D.

Using the elements of the $\boldsymbol{v}$ vectors and Euclidean distance we can compute the similarities among these 5 proteins. The similarity matrix is given in Table 2, where we can observe that the most similar pair of proteins is (B, E) and the most dissimilar one is (A, B). The pair (B, E) was also found to be the most similar one by Randic and Krilov [16], but from their results the most dissimilar pair is (A, E). Our results are also not coincident with those reported by these authors for the second most similar pair, because they found that it corresponds to the pair (C, D) while we found that it is (E, D).

The results of similarity reported by Randic and Krilov [16] appear to be in contradiction with those reported by them concerning the ranking of folding [1]. For instance, they reported that the second most similar pair is (C, D) but C is reported by them as the second most folded protein while D is the least folded protein of all. This contradiction does not occur in the case of the most dissimilar pair of proteins that is composed by the most folded (A) and one of the least folded proteins (E).

Our results of similarity are in complete agreement with the ranking of folding obtained with the I index. For instance, the most similar pair of proteins is composed by the two least folded proteins, D ($I = 63.802$) and B ($I = 65.025$). This pair can be observed after their superposition at the end of Fig. 4. The second pair of most similar proteins is formed by D ($I = 63.080$) and E ($I = 66.022$) which are two of the less folded proteins. The third pair of most similar proteins is (B, D) that also correspond with two of the less folded proteins. On the other hand, the most dissimilar pair is composed by the most folded protein, A ($I = 73.513$) and one of the less folded ones, B ($I = 65.025$).

These results suggest that some relationship exists between the folding of proteins and the similarity between them. These observations concerned with the 3D structure of proteins can be summarised as follows:

(1) the most folded proteins are not similar among them, for instance the two most folded proteins, A ($I = 73.513$) and C ($I = 72.877$), are ranked only as the sixth most similar pair;

(2) the less folded proteins are similar among them, for instance the three possible combinations among the less folded proteins (B, D, E) produce the three most similar pairs of proteins, (B, E), (D, E) and (B, D).

5. Conclusions

We have shown that the present approach is able to describe correctly the 3D structure of small/medium-sized and large molecules, such as proteins. The novel index defined here can be considered as a measurement of the 3D 'compactness' or folding of the molecular structures, giving maximum values for the most folded isomer. The study of protein folding of five protein models has permitted the ranking of them according to their degree of folding. We have also been able to quantify the similarity among these proteins, demonstrating that the most folded proteins are not similar among them, while the less folded proteins are similar to each other.

According to the accepted paradigm that similar compounds have similar properties/activities, and due to that, the functionality of proteins is determined by its full 3D structure, we think that the most folded proteins are more specific than the less folded ones. This is due to our finding that the most folded proteins are dissimilar to the rest of the proteins and among themselves, consequently their properties/activities are different from the rest.

References

[1] M. Randic, J. Chem. Inf. Comput. Sci. 37 (1997) 672.
[2] J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, London, 1999.
[3] N. Trinajstic, Chemical Graph Theory, CRC Press, Boca Raton, FL, 1992.
[4] M. Randic, Chem. Phys. Lett. 211 (1993) 478.
[5] M. Randic, Int. J. Quantum Chem.: Quantum Biol. Symp. 15 (1988) 201.
[6] B. Bogdanov, J. Math. Chem. 3 (1989) 299.
[7] M. Randic, B. Jerman-Blazic, N. Trinajstic, Comput. Chem. 14 (1990) 237.
[8] K. Balasubramanian, Chem. Phys. Lett. 169 (1990) 224.
[9] S. Nikolic, N. Trinajstic, Z. Mihalic, S. Carter, Chem. Phys. Lett. 176 (1991) 21.
[10] E. Estrada, L.A. Montero, Mol. Eng. 2 (1993) 363.
[11] E. Estrada, J. Chem. Inf. Comput. Sci. 35 (1995) 708.
[12] M. Randic, A.F. Kleiner, L.M. De Alba, J. Chem. Inf. Comput. Sci. 34 (1994) 277.
[13] D.J. Klein, J. Math. Chem. 18 (1995) 321.
[14] M. Randic, M. Razinger, J. Chem. Inf. Comput. Sci. 35 (1995) 140.
[15] M. Randic, J. Math. Chem. 19 (1996) 375.
[16] M. Randic, G. Krilov, Chem. Phys. Lett. 272 (1997) 115.
[17] E. Estrada, J. Chem. Inf. Comput. Sci. 36 (1996) 844.
[18] E. Estrada, J. Chem. Inf. Comput. Sci. 37 (1997) 320.
[19] E. Estrada, J. Chem. Inf. Comput. Sci. 38 (1998) 23.
[20] E. Estrada, J. Chem. Soc., Faraday Trans. 94 (1998) 1407.
[21] I. Gutman, L. Popovic, E. Estrada, S.H. Bertz, ACH Models Chem. 135 (1998) 147, and references therein.
[22] E. Estrada, J. Chem. Inf. Comput. Sci. 39 (1999) 90.
[23] C.M. Dobson, A. Sali, M. Karplus, Angew. Chem., Int. Ed. Engl. 37 (1998) 868.
[24] H. Li, R. Helling, C. Tang, N. Wingreen, Science 273 (1996) 666.