Chemical Physics Letters 599 (2014) 1–6

Contents lists available at ScienceDirect

Chemical Physics Letters

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

On of molecular similarity based on a single molecular descriptor

Milan Randić

National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 7 January 2014
In final form 10 March 2014
Available online 15 March 2014

ABSTRACT

We consider the characterization of molecular similarity through a single molecular descriptor, instead of the customary use of sets of structural invariants to characterize individual molecules. Moreover, we require that the 'similarity' descriptor be conceptually and computationally simple so that it is suitable for screening huge combinatorial libraries in search for target compounds. We have outlined one such general approach for construction of 'similarity' descriptors, which is illustrated on the set of 35 nonane constitutional isomers.

1. Introduction

Numerical characterizations of molecular graphs, which are often taken as models of molecules, by suitable mathematical invariants which serve as molecular descriptors, is underlying most studies of structure–property and structure–activity relationships. Most often such invariants are extracted from pictorial representations of molecules as graphs, or from constructed structural matrices. The pioneering work on use of mathematical descriptors for structure–property studies goes back to Wiener [1] and Platt [2] in 1947. However, their work on introducing mathematical descriptors for characterizing molecules was overlooked for well over 25 years, and would have remained overlooked were it not for fast growth of Chemical Graph Theory in mid 1970s. Then additional numerical characterizations of molecular structure emerged [3–5] giving boost quantitative structure–property studies. Over the next 25 years, till recent times, the number of mathematical descriptors for use in structure–property–activity proliferated beyond expectations and possibly beyond the need, resulting in several hundreds an more of documented new, so called topological indices, and their variations [6–8]. It is to be noted that a selection of topological indices have interesting and distinctive properties, while remaining conceptually simple and straightforward. Among these we should mention the Balaban's *J* index [9], introduced already in 1982, which has been one of the early topological index showing a relatively high power of discrimination between molecules. While this is not a property that is considered important for structure–property correlations, the ability to discriminate very similar molecules became of considerable interest in recent times with interest in screening of combinatorial libraries in search for molecules highly similar to selected target structures. The importance of this novel avenue in search for useful drugs came to attention mostly through the pioneering work of Lahana and coworkers [10], who were able, after screening a combinatorial library of many thousands of virtual compounds, to extract about 20, five of which have been after further scrutiny selected for synthesis. After testing it was found that one of the new compounds has almost 100 times higher immune suppressive action from the parent structure. This is a remarkable finding in drug research and illustrates the power of the screening combinatorial libraries in searching for new drugs.

2. Molecular descriptors for screening combinatorial libraries

Hitherto construction of new topological indices has been primary aiming at improving the structure–property–activity correlations. Occasionally there have been reports on molecular descriptors of unusually high discriminatory power. One such descriptor is known as molecular ID number, which has been constructed by adding the connectivity index and all possible higher order connectivity numbers of a molecule [11]. In the case of alkanes (acyclic graphs with maximal degree or valence of four) Balaban's index *J* has six pairs of duplicates among 335 dodecanes $C_{12}H_{26}$ [12], while molecular ID has one pair of duplicates among 4347 $C_{15}H_{32}$ alkane isomers. This is quite impressive, but molecular ID numbers are not suitable for screening combinatorial libraries in view that they are computational intensive. Namely, they require computing all paths in a molecule, which in the case of cyclic graphs can is generally very time-consuming.

Since combinatorial libraries can have 100 000 or more virtual compounds, molecular descriptors for screening such libraries have to satisfy the following conditions:

E-mail address: mrandic@msn.com

http://dx.doi.org/10.1016/j.cplett.2014.03.022
0009-2614/

(i) They should have high discriminatory power.
(ii) They have to be computationally simple.
(iii) They have to be conceptually simple.

Computationally simple means here to be calculated readily. A way to have 'simple' calculations is to use integers in the construction of the descriptors as much as possible, and only in the last step of calculation to allow the use of real numbers (e.g., rational or irrational). Conceptually 'simple' means here to involve some elementary mathematical notions that are related and allow the interpretation of the derived molecular descriptors in terms of structural elementary concepts. Under these conditions one may expect that structurally closely related molecules will have numerically indices of similar magnitudes. In the next section we will outline an approach that apparently satisfies our constraints and appears promising not only for use in screening combinatorial libraries but also as a molecular 'similarity' index.

3. Generalized connectivity indices

We will refer to novel a kind of molecular descriptors as 'generalized connectivity indices' in view that there is some formal similarity in their construction and a way in which one can construct the connectivity index χ. Recall that the connectivity index is bond additive, the bond contributions being $1/\sqrt{(m \times n)}$, where m, n are the vertex degrees of each bond end points in the hydrogen depleted graph. Recall that Balaban's J index is also bond additive, the bond contributions being $1/\sqrt{(R_i \times R_j)}$, where R_i and R_j are the row sums of the corresponding rows in the distance matrix. Because the row sums of the adjacency matrix are the vertex degrees one can construct the connectivity index χ by summing the bond contributions written as $1/\sqrt{(R_i \times R_j)}$, where R_i, and R_j are the row sums of the corresponding rows of the adjacency matrix. Thus for the connectivity index χ one can use symbol χ(A), where A stands for the adjacency matrix. Using the same formalism one can represent the Balaban's J index analogously as χ(D), where D stands for the distance matrix.

Both χ(A) and χ(D) can be viewed as special cases of the 'generalized' connectivity indices χ(X), which all are derived using different matrices X, the bond contributions of which use the same algorithm: $1/\sqrt{(R_i \times R_j)}$, where R_i, and R_j are the row sums of the corresponding rows of chosen matrix X. In the next section we will outline construction of a sparse distance matrix obtained by considering the columns of the adjacency matrix as vectors in n-dimensional vector space, and considering only distances between adjacent vertices of a graph. We are using the Hamming distance [13], defined for strings of numbers of the same length. It is given by the number of places at which the corresponding elements in two strings (sequences) are different. Such approach has been introduced in Ref. [14] by considering full distance matrix, however, as already mentioned, here we consider only matrix elements belonging to distances between column vectors of adjacent vertices, because sparse matrices are computational far less intensive than full dense matrices.

4. The Hamming Distance Matrix (HD)

We will illustrate construction of the n-dimensional distance matrix on a pair of nonane isomers 3,4-dimethylheptane and 3-ethyl-2methylhexane illustrated in 1, the adjacency matrices of which are listed at the top in Table 1. Strictly speaking the n-dimensional Hamming Distance Matrix is in general case a pseudo-distance matrix, as it may have zero elements off the main diagonal. This will happen whenever the corresponding columns of adjacency matrix are identical. For pseudo-distance matrices the first axiom on distance, that distance is positive, has been relaxed in the requirement that the distance be non-negative. To construct the n-dimensional distance matrix one views the nine columns in adjacency matrices of each nonane isomer as nine vectors in 9-dimensional space. Next, we consider, for each isomer separately, only the distances corresponding to adjacent vertices of each molecule and calculate the Hamming distance between all pairs of vectors (columns) corresponding to each CC bond, which for binary sequences is given by the number of entries which are different in two columns of adjacency matrix. Thus in the case of 3,4-dimethylhexane the element (1, 2), the distance between columns one and two (vectors v_2 and v_3):

$$(0,1,0,0,0,0,0,0,0) \text{ and } (1,0,1,0,0,0,0,0,0)$$

is equal 3, because the two sequences have different elements only in three locations, the first three elements. Similarly the element (2, 3), the distance between vectors v_2 and v_3 is 5, the two sequences have different elements in five locations, and so on. Continuing in this way one obtains the HD matrix of 3,4-dimethylhexane shown at the left bottom part of Table 1. Similarly follows the construction of the HD matrix of 3-ethyl-2-metylhexane isomer, shown at the bottom right of Table 1. At the right of each distance matrix we have listed the corresponding row sums (RS).

Table 1
Top: The adjacency matrices (A) for the two nonane isomers illustrated in Figure 1 and the corresponding row sums (RS); Bottom: The Hamond Distance matrices (HD) obtained for the columns vectors of A corresponding to the adjacent vertices of nonane isomers graphs.

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

	1	2	3	4	5	6	7	8	9	RS
1	0	3	0	0	0	0	0	0	0	3
2	3	0	5	0	0	0	0	0	0	8
3	0	5	0	6	0	0	0	4	0	15
4	0	0	6	0	5	0	0	0	4	15
5	0	0	0	5	0	4	0	0	0	9
6	0	0	0	0	4	0	3	0	0	7
7	0	0	0	0	0	3	0	0	0	3
8	0	0	4	0	0	0	0	0	0	4
9	0	0	0	4	0	0	0	0	0	4

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

	1	2	3	4	5	6	7	8	9	RS
1	0	4	0	0	0	0	0	0	0	4
2	4	0	6	0	0	0	4	0	0	14
3	0	6	0	5	0	0	0	5	0	16
4	0	0	5	0	4	0	0	0	0	9
5	0	0	0	4	0	3	0	0	0	7
6	0	0	0	0	3	0	0	0	0	3
7	0	4	0	0	0	0	0	0	0	4
8	0	0	5	0	0	0	0	0	3	8
9	0	0	0	0	0	0	0	3	0	3