LINEAR NOTATION FOR BENZENOID AROMATIC HYDROCARBONS. MOLECULAR SIMILARITY BASED ON NOTATION SIMILARITY

W.C. HERNDON and A.J. BRUCE

Department of Chemistry, The University of Texas at El Paso, El Paso, Texas 79968, USA

Received 1 August 1986
(in final form 20 November 1987)

Abstract

Two succinct linear notation systems to encode the structure of polybenzenoid aromatic hydrocarbons are exemplified. Both notation systems use a labeled dual inner graph to represent the hydrocarbon. A molecular similarity index ranging from unity (identical molecules) to zero (completely different molecules) is defined based on a comparison of the linear notations for a pair of compounds. The similarity index procedure is applied to a correlation of the carcinogenic properties of the benzenoid hydrocarbons.

1. Introduction

A number of specific procedures have been suggested to designate polycyclic benzenoid aromatic hydrocarbon (PBAH) structures using codes or alphanumeric notations of various types [1–15]. The standard Chemical Abstracts nomenclature rules for completely unsaturated fused ring carbon compounds comprise, of course, one such system [16]. However, the CA system makes use of a large number of trivial names, hierarchal rules, and recondite numbering and fusion methods based on preferred orientations of the molecular structural graphs. It is not, therefore, very useful for solving problems of classification and enumeration or for investigating questions related to numerical comparisons of structure. The more abstract notation systems based on graph-theoretical formalisms [1–15] prove to have more utility in these regards.

We have previously advanced completely general linear notation systems for inorganic and organic structures [17–20]. These procedures involve unique canonical numberings of the chemical systems, use only standard chemical symbols, and have been implemented in the form of microcomputer programs [21]. A string comparison technique was also adopted to estimate the similarity of two molecular linear string

notations [14,20]. The principal part of the analysis simply involves counting the number of insertions and/or deletions required to convert one molecular notation to the other. This technique allows one to define a molecular similarity index with values that range from zero to unity, the zero value characterizing complete dissimilarity and the value of unity denoting identity. Comparison of the similarity index values with distance measures of similarity based on subgraph enumeration for a small set of aliphatic alcohols [22] showed that the two concepts give rise to parallel estimates.

The quantitative elucidation of molecular structural similarity derives significance from the basic premise that molecules with similar chemical structures will exhibit similar physical and chemical properties. Perhaps even more important, they could also exhibit similar biological or pharmacological activities. Practical use of a molecular similarity concept would involve QSAR (Quantitative Structure Activity Relationships) methodologies, where various statistical techniques are used to seek for correlations between an observed biological or chemical activity and a set of arbitrarily selected chemical or molecular descriptors. Molecular descriptors used in the past have included various kinds of physical properties, both theoretical and experimental reactivity parameters, and many kinds of structural descriptors [23–30]. In principle, a similarity term, perhaps expressed relative to the most active compound in a data set, could be included as an independent variable in QSAR analyses.

This present paper, which will be concerned with applications of the foregoing concepts to PBAH, will have the following format. First, a new, succinct structural representation of PBAH will be presented. Then, codings of the PBAH structural representations will be exemplified using two linear notation systems presented in previous work [17–20], and notations for several PBAH will be used to illustrate the calculation of molecular similarity indices. Finally, in a test of the procedures, we will attempt the characterization of the well-known [31] carcinogenic properties of PBAH in terms of the similarity index concept.

2. Structural representations of PBAH

PBAHs are commonly depicted by their polyhex [32] graphs. The polyhex graph represents the molecular structure as a planar figure constructed by connecting regular hexagons so that any two hexagons are either disjoint or have a single common edge. The number of polyhex graphs may be restricted by requiring the polyhex to be superimposable on the two-dimensional tessellation of regular hexagons, i.e. the graph of the graphite lattice. The benzenoid polyhex graph can also be reduced to a smaller and even more abstract representation by joining the centers of neighboring hexagons and then deleting the edges of the original benzenoid graph. The resulting graph is variously called the dual inner graph, the dualist graph, or the characteristic graph [1,32–35]. The properties of the dualist graphs of PBAH have been the subject of numerous studies in chemical applications of graph theory [36–40].

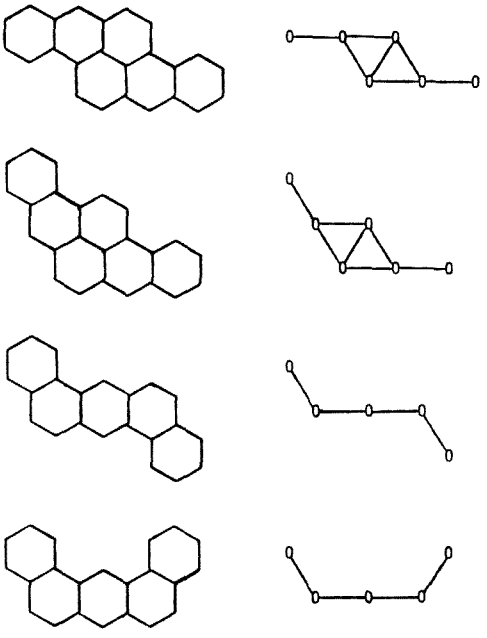


Fig. 1.

The polyhex to dualist graph conversion is illustrated in fig. 1 for the molecular graphs corresponding to several representative PBAHs. No structural information is lost if the relative orientations of the dualist graph vertices are presumed to reflect the relative orientations of the hexagonal rings in the original molecular graph. However, the representation of this pictorial information presents obvious difficulties for notation schemes. Most previous complete solutions to this coding problem prescribe a prior reference placement of the molecular structure on the graphite or a related lattice, and/or designate the angular orientation information with a sequence of numerals [41].

Our approach to the relative orientation problem is bounded by the fact that computer programs we have developed used molecular graph structural information in the form of connection tables or the equivalent adjacency matrix array. Diagonal elements of this array correspond to the vertices of the molecular graph, and off-diagonal elements represent the graph edges (bonds). Both types of elements are stored as alphanumeric strings rather than in numerical form, and this presupposes that all structural information be expressed as vertex and edge labels of the molecular graph. The structures shown in fig. 2 demonstrate that it is possible to convey information in the dualist graph within these constraints with the use of the following definitions:

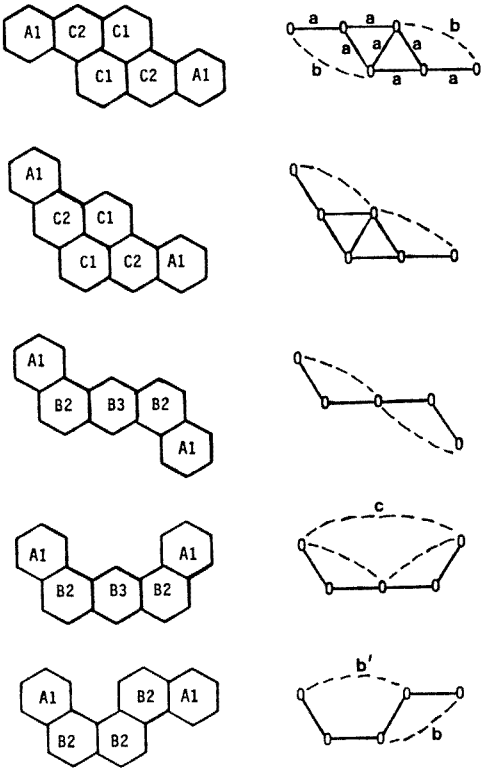


Fig. 2.

(a) Benzenoid rings with one to six adjacent rings are labeled A through F, respectively [42]. Numerical suffixes 1, 2, and 3 refer to the substitution pattern and are defined in the context of the figures.

(b) A graph line that corresponds to the common edge of two adjacent rings is labeled with a lower case a [43]. The interactions between rings that correspond to the so-called bay-regions and pseudo-bay-regions are labeled b and b', respectively. Longer range cisoid relationships are indicated by lower case c.

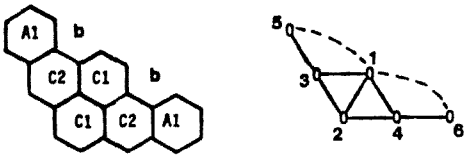
The resulting graph is a completely labeled dual inner graph, augmented with lines representing secondary ring interactions. It requires no predetermined orientation convention or additional assumptions to completely represent the original PBAH molecular graph structure.

A certain amount of redundant structural information is present in these labeled, augmented dualist graphs. For example, in every case one could deduce the original PBAH structure without making use of the vertex labels, the edge labels being sufficient for this purpose if the underlying graphite lattice is assumed. However, the

vertex labels greatly simplify the process of drawing the original PBAH, given the linear notation to be described in the next section. The very same kind of advantage arises when saturated hydrocarbons are represented as labeled graphs (i.e. vertex labels CH, CH_2, and CH_3) rather than by carbon skeleton line drawings. With or without the vertex labels, every PBAH has a unique graph representation of this type, which is, naturally, the minimal requirement for a molecular graph-based nomenclature or notation system. One should also note that other kinds of condensed benzenoid molecules can be denoted within this system including, for example, helicenes and molecules with interior holes (coronae-type [36]). In fact, if ring size is included in the vertex labels, nonbenzenoids can also be encompassed, giving a general labeled dualist graph representation for condensed π-system molecules of all types [44].

3. Linear notations for PBAH

Two linear notation systems (LN-1 and LN-2) are to be used in the present work, and both start with a canonical numbering of the underlying simple graph corresponding to the molecular graph. Explicit rules have been given previously [17,18] for the ancillary canonical assignment that leads to LN-1. Extended connectivity is used as the basic algorithmic numbering tool, and high connectivity and centrality in the graph are the main factors giving numerical priority. As a result, vertices of the highest degree are numbered first, ensuing numbers tend to cluster,



GC-1 (01)0203040506(02)0304(03)05(04)06(05)(06)

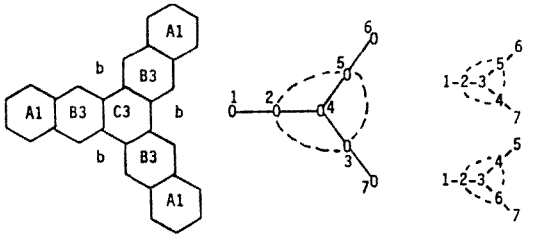
LN-1 (C1)02a 03a 04a 05b 06b (C1)03a 04a (C2)05a (C2)06a (A1) (A1)

Fig. 3.

and terminal graph nodes will appear last in the order of citation. In many cases, these factors allow the assignment of the correct canonical labels without actually proceeding through the numbering algorithm or using the computer program that gives the final notation upon input of an arbitrarily numbered molecular adjacency matrix. The resulting graph code (GC-1) cites each vertex in parentheses, followed by citations to neighboring vertices, all given in ascending numerical order, as shown in fig. 3 for the dualist graph of dibenzo[*a*, *i*]pyrene. Note that bay-region interactions are explicitly depicted and that this code is a linear equivalent of the adjacency matrix in which each bond, however, receives only one designation.

The conversion of the GC-1 code to LN-1 linear notation is straightforward, as demonstrated in fig. 3. The parenthetical numerals in the graph code are non-essential, so they are discarded and replaced by the vertex labels, and edge or interaction labels are introduced in the appropriate notation position. As stated previously, the long range interaction labels (b, b', and c) are necessary in PBAH to distinguish isomeric structures. In applications to normal molecular graph structures, one simply requires an edge or vertex label corresponding to every chemical bond and atom (group) in the molecular structure, respectively [45]. When two or more alternative codes or notations can be derived for the same graph, then that code or notation which is lexicographically superior at the first point of difference is chosen. Of course, symmetry will lead to alternative equivalent notations, one of which is to be chosen arbitrarily.

The procedure used to obtain the graph code (CG-2) for LN-2 takes the longest path in a graph as the single structural element to initiate canonical numbering. Path branches, which are paths that emanate from previously numbered paths, are subsequently numbered in order of decreasing path branch length. Then, the code is completed by adding locants for the path branches and double locants for single-edge path bridges that define the graph cyclicity. The components of GC-2 are written in the prescribed order: longest chain length, locants and path branch sizes in decreasing



GC-2 06(03)01(0204)(0205)(0305)

LN-2 A1-a-B3-b-B3-a-C3-a-B3-a-A1(03)-a-A1(0204)-a-(0205)-b-(0305)-b-

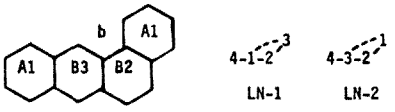
Fig. 4.

order of length, locants for bridges. All locants are enclosed in parentheses and, again, a lexicographic criterion is used to choose among otherwise equivalent derived codes and notations. A branched PBAH structure and its dualist graph are given in fig. 4 to explicate the assignment of GC-2 and LN-2. Two alternative incorrect numberings of the dualist graph are also shown.

The first element of the correct code, 06, denotes a path of length six. The first locant and the following elements, (03)01, designate a path branch of length unity at vertex 3 of the main path. The cyclicity is delimited by the path bridges (0204), (0205), and (0305), and the vertex numbering given is one of three possible optimal numberings, due to the threefold symmetry of the molecular graph. The

first alternative numbering in fig. 4 gives the code 06(04)01(0204)(0205)(0405), which is inferior to the correct code. The second alternative numbering would give 05(03)02(0204)(0206)(0405), which does not make use of a longest path to derive the code. Linear notation LN-2 is obtained from GC-2 by replacing each path length with the molecular graph vertex and edge labels that comprise the path. The edge or bond types that correspond to the cyclic path bridges are also designated, as demonstrated in fig. 4. Neither the hyphens in LN-2 nor the blank spaces in LN-1 are required, but both are added to improve readability of the notations.

LN-1 notations are generally longer than those derived in the LN-2 system, principally because more locants are necessary. This fact is illustrated for the dualist graph of benz[*a*]anthracene in fig. 5, where the notation lengths are 12 for LN-1 and



LN-1 (B3)02a 03b 04a (B2)03a (A1) (A1)

LN-2 A1-a-B2-a-B3-a-A1(0103)-b-

Fig. 5.

10 for LN-2. However, in our experience, obtaining either notation is a facile process, either by hand or with the use of the available computer programs. We do find that the reverse process, obtaining a structure from a linear notation, seems to be somewhat easier in the LN-2 system. This may be because chain structures and cyclicity can be directly discerned from a single scan of the notation.

4. Definition of similarity

In either of the two notation systems, a compound is represented by a string of alphanumeric symbols. Our procedures for obtaining quantitative metrics related to similarity involve comparison of the two sequences of symbols for two different molecules [19,20]. This procedure has been anticipated in work by Adamson and Bowden, where noncanonical Wiswesser notations were used in the same fashion [46]. The general approach that is used here is to determine the number of insertions and/or deletions that are required to convert the linear string notation of one molecule to that of the second molecule [47–50]. The larger this number, called the distance between the two strings, the less similar are the two notations, and by implication, the less similar are the two structures that gave rise to the notations.

Three PBAH examples are given in fig. 6 for both LN-1 and LN-2 notation systems. Each numerical locant and each vertex or edge label is considered to be a structural term of the linear notation. Matching terms in pairs of sequences can be

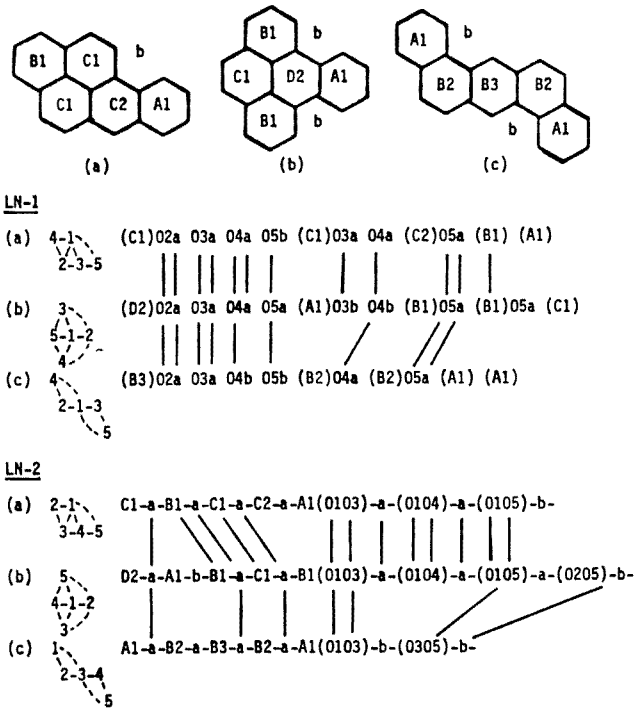


Fig. 6.

indicated by lines drawn from one sequence to the other, and such a drawing constitutes a "trace". In a trace comparing two sequences, no terms in either sequence can be connected by more than one line, and the lines are not allowed to cross each other. The optimal trace has the largest number of allowed lines. Four different traces, i.e. those of dualist graph (b) with each of (a) and (c) in both notations, are also illustrated in fig. 6.

The total of the terms without lines in the optimal trace is the number of insertions/deletions required to convert one sequence to the other and thus constitutes a distance between the sequences. The similarity (S) is then calculated as unity minus the quotient of the distance (d) required for conversion divided by the total number of terms in the two sequences, cf. eqs. (1)–(5).

$$S(i,j) = 1 - d(i,j)/(N_i + N_j) \tag{1}$$

$$S_{\mathrm{LN}-1}(6a,6b) = 1 - 16/(19+21) = 0.600 \tag{2}$$

$$S_{\mathrm{LN}-2}(6a,6b) = 1 - 13/(18+21) = 0.667 \tag{3}$$

$$S_{\mathrm{LN}-1}(6b,6c) = 1 - 20/(21+17) = 0.474 \tag{4}$$

$$S_{\mathrm{LN}-2}(6b,6c) = 1 - 22/(21+15) = 0.389. \tag{5}$$

If the reader is interested, traces for the pairs (a) and (c) in fig. 6 can be examined to finally obtain the calculated similarities as $S_{LN-1}(6a, 6c) = 0.667$ and $S_{LN-2}(6a, 6c) = 0.545$.

5. A test of the similarity concept and discussion

The usual methods for quantitative investigations of a molecular property and its relationship to molecular structure involve linear correlations of the dependent observed property with structural descriptors, the number or values of which act as independent variables. As mentioned in the introduction, structural descriptors of many types have been used as independent variables; the sets of topological indices employed by Kier and Hall [24], path lengths as defined by Randić and coworkers [51–56], and the molecular graph fragments used by Klopman and coworkers [29,57–60] constitute good examples of this kind of approach. The number of parameters, their types, and their forced or unforced inclusion in the multi-linear regression are all matters of judgment. Our recent study of PBAH carcinogenicities lies in this category [42].

We have now elected to proceed in a different way, which devolves from the ability to accurately represent the molecular structure of an organic compound by means of succinct linear notations. Thus, it is our goal first to define a metric of molecular similarity based on the notations, and then to test the chemical consequences of the previously defined concept. This approach is certainly in keeping with that of several other recent investigators who have attempted to obtain definitions of similarity or related ideas [61–68]. The extreme simplicity of our similarity definition may constitute an advantage. Whether or not this approach and the similarity concept have validity can only be determined by testing and by comparison with other definitions.

As an initial test, the procedures developed herein will be applied to correlations of the carcinogenic properties of the 16 PBAH compounds whose structures are represented in fig. 7. We choose this example partly because of our previous interest in this problem [42,69,78]. However, PBAH carcinogenicities constitute one of the most important aspects of PBAH chemistry, and the continuing interest in theoretical models for PBAH carcinogenicity is manifest by several pertinent studies published only in the last two years [71–79].

The numerical value given in fig. 7 for each compound is the experimentally obtained Iball index of carcinogenicity [31,80,81]. This index is defined as the percentage of skin cancer or papilloma-developing mice (skin painting experiments) divided by the average latent period in days for affected animals (times 100). A few other PBAH compounds demonstrate weak carcinogenic activity in other types of experiments, and these compounds are listed in Dipple's excellent review articles [31]. Most other PBAHs are generally accepted to be inactive. The values of the Iball indices

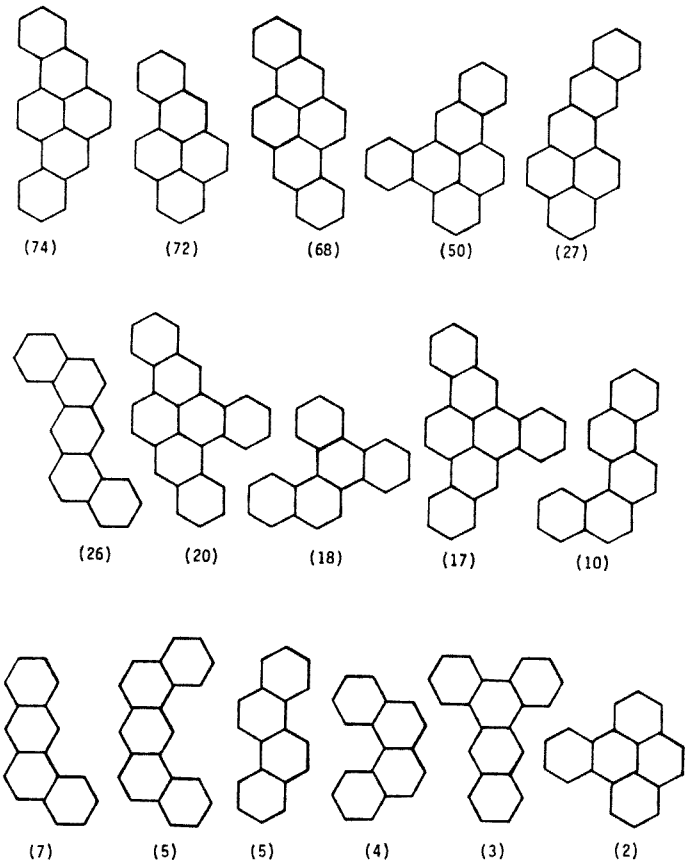


Fig. 7.

are normally taken to provide an acceptable partitioning of the PBAH carcinogens into weak (<20), moderate ($20-40$), and strongly active (>40) compounds, even though the reliable determination of a carcinogenicity index is hampered by several possible experimental difficulties. However, nearly all previous studies have adopted the Iball indices or the 3-category partition as the carcinogenicity variable. Since the discontinuous nature of the latter classification would be inconvenient for a simple correlative study (depending on linear regression), we will use the Iball values given in fig. 7 for the present study, aware that shortcomings in this approach do exist.

The results of the similarity analyses and Iball indices, along with some other structural parameter values, are given in table 1. The similarities are, of course, defined relative to the most carcinogenic compound dibenzo[*a*, *i*]pyrene. The localization energies given in the table are quantum-mechanical parameters for reactions at molecular sites that have been considered to be involved in carcinogenesis in many previous theoretical analyses [82]. All reactivity parameters were calculated using

Table 1

Experimental Iball indices I; similarities S(LN-1) and S(LN-2) Bay-region, M-region and K-region atom localization energy indices[a]

Compound	I	S(LN-1)	S(LN-2)	$B(a)$	$M(a)$	$K(a)$
Dibenzo[*a*, *i*]pyrene	74	1.000	1.000	1.674	1.404	1.253
Benzo[*a*]pyrene	72	0.837	0.829	1.504	1.299	1.170
Dibenzo[*a*, *h*]pyrene	68	0.917	0.957	1.653	1.460	1.495
Dibenzo[*a*, *e*]pyrene	50	0.720	0.673	1.534	1.294	1.326
Naptho[2, 3-*a*]pyrene	27	0.696	0.791	1.466	1.386	1.299
Dibenzo[*a*, *h*]anthracene	26	0.634	0.553	1.421	1.099	1.204
Tribenzo[*a*, *e*, *h*]pyrene	20	0.717	0.823	1.621	1.457	1.504
Benzo[*g*]chrysene	18	0.512	0.585	1.447	1.099	1.232
Tribenzo[*a*, *e*, *i*]pyrene	17	0.691	0.778	1.655	1.368	1.269
Benzo[*c*]chrysene	10	0.439	0.474	1.322	1.073	1.219
Benz[*a*]anthracene	7	0.500	0.485	1.386	1.050	1.099
Dibenzo[*a*, *j*]anthracene	5	0.465	0.341	1.386	1.070	1.153
Chrysene	5	0.579	0.500	1.281	1.056	1.099
Benzo[*c*]phenanthrene	4	0.444	0.424	1.224	1.012	1.099
Dibenz[*a*, *c*]anthracene	3	0.512	0.536	1.386	0.990	1.124
Benzo[*e*]pyrene	2	0.533	0.591	1.344	0.969	1.068
Correlation coefficient with I		0.914	0.812	0.704	0.725	0.510

[a] See refs. [31] and [42] for descriptions of these parameters.

valence bond structure resonance theory, which gives reactivity parameters that are highly correlated to the results of experiments and to the results of molecular orbital calculations [83]. The correlation coefficients of the listed parameters with the Iball index are also given in table 1.

The similarity indices and each one of the reactivity parameters correlate to a significant degree in table 1. In fact, the correlation of similarity based on the LN-1 notation with the Iball indices is the highest correlation for a single parameter of which we have knowledge. The correlation coefficient of 0.914 indicates that S(LN-1) accounts for 84% of the variance in table 1. A stepwise linear regression model would select either similarity index ahead of any of the reactivity indices. The significance of the higher correlation coefficient for $M(a)$ rather than $B(a)$ has been discussed previously [42]. The more common extended analysis utilizing multi-linear regression of size, reactivity, and structural factors will naturally lead to even more improved correlations of the carcinogenicity data, but is outside the scope of the present work.

The two similarity indices also correlate with each other, correlation coefficient = 0.939, as might be expected due to the parallel procedures for obtaining

the similarity values. This correlation coefficient is low enough that one can infer that a rational preference for one or the other definition may arise for particular types of compounds exhibiting particular activities. This is certainly the case for the PBAH carcinogenicities, where the similarity indices based on the LN-1 notation give a very reasonable correlation with the Iball index.

Whether or not the indices actually correlate with "similarity" may be an unanswerable question. The difficulties in obtaining precise numerical indices for qualitative molecular structural concepts should not be underestimated. A moderate degree of success has been achieved in the present case in correlating the PBAH property of carcinogenesis, but this success lies a long way from validating the accuracy of these similarity concepts. Finally, the question of a mechanistic basis for understanding the correlations arises, and in the present investigation it is difficult to delineate the particular structural features that are responsible. It may be that analysis of the sequences within the linear notations would help to answer this question.

The two constructs for defining a similarity index that we have presented do have the practical advantages of conciseness and simplicity. However, we must point out that the similarity values obtained must be perceived as only *locally* valid. For a property dissimilar to the present one, a different molecule would have to serve as the standard of comparison, since one must necessarily choose the standard molecule to have either the maximum or the minimum value of the property under investigation. It is, of course, also possible that more complex definitions of similarity will be necessary in other applications, or that many different definitions will prove to be useful. We plan to compare different types of similarity definitions in future work, along with carrying out further tests of validity and applicability.

Acknowledgement

The authors acknowledge the financial support of the Robert A. Welch Foundation of Houston, Texas.

References and footnotes

[1] A.T. Balaban and F. Harary, Tetrahedron 24(1968)2505.
[2] A.T. Balaban, Tetrahedron 25(1969)2949.
[3] F. Harary and R.C. Read, Proc. Edinburgh Math. Soc. 17(1970)1
[4] A.T. Balaban, Rev. Roum. Chim. 15(1970)1251.
[5] I. Gutman, Theor. Chim. Acta 45(1977)309.
[6] K. Balasubramanian, J.F. Kaufman, W.S. Koski and A.T. Balaban, J. Comput. Chem. 1 (1980)149.
[7] D. Bonchev and A.T. Balaban, J. Chem. Inf. Comput. Sci. 21(1981)223.
[8] I. Gutman, Z. Naturforsch. 37a(1982)69.
[9] J.V. Knop, K. Szymanski, Ž. Jeričević and N. Trinajstić, J. Comput. Chem. 4(1983)23.

[10] D. Bonchev, Pure and Appl. Chem. 55(1983)221.
[11] N. Trinajstić, Ž. Jeričević, J.V. Knop, W.R. Müller and K. Szymanski, Pure and Appl. Chem. 55(1983)379.
[12] S. El-Basil, Croat. Chem. Acta 57(1984)21.
[13] J. Cioslowski and A.M. Turek, Tetrahedron 40(1984)2161; Computers and Chemistry 9 (1985)247.
[14] H. Wenchen and H. Wenjie, Theor. Chim. Acta 68(1985)301.
[15] R. Tošić, R. Doroslovački and I. Gutman, Match 19(1986)219.
[16] *Chemical Substance Name Selection Manual,* 1982 edition, Vol. 1, Chemical Abstracts Service, pp. B34 – B87.
[17] W.C. Herndon and J.E. Leonard, Inorg. Chem. 22(1984)554.
[18] W.C. Herndon, in: *Chemical Applications of Topology and Graph Theory,* ed. R.B. King (Elsevier, Amsterdam, 1983) pp. 231 – 242.
[19] W.C. Herndon, Computers and Math. with Appl., in press.
[20] W.C. Herndon and S.H. Bertz, J. Comput. Chem. 8(1987)367.
[21] Compiled Basic language microcomputer programs are available upon request to implement procedures described herein.
[22] S.H. Bertz and W.C. Herndon, in: *Artificial Intelligence Applications in Chemistry,* ed. T.H. Pierce and B.A. Hohne, ACS Symposium Series 306 (American Chemical Society, Washington, D.C., 1986) pp. 169 – 175.
[23] W.V. Valkenburg, ed., *Biological Correlations – The Hansch Approach* (American Chemical Society, Washington, D.C., 1972).
[24] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
[25] C. Hansch and A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology* (Wiley, New York, 1979).
[26] A.J. Stuper, W.E. Brugger and P.C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function* (Wiley-Interscience, New York, 1979).
[27] R. Osman, H. Weinstein and J.P. Green, in: *Computer-Assisted Drug Design,* ed. E.C. Olson and R.E. Cristofferson, ACS Symposium Series 112 (American Chemical Society, Washington, D.C., 1979) pp. 21 – 77.
[28] V.E. Golander and A.B. Rozenblit, *Logical and Combinatorial Algorithms for Drug Design* (Research Studies Press, Ltd., Letchworth, England, 1983).
[29] G. Klopman, J. Amer. Chem. Soc. 106(1984)7315.
[30] P.C. Jurs, T.S. Stouch, M. Czerwinski and J.N. Narvaez, J. Chem. Inf. Comput. Sci. 25 (1985)296.
[31] A. Dipple, R.C. Moschel and C.A.H. Bigger, in: *Chemical Carcinogens,* 2nd ed., ed. C.E. Searle, ACS Monograph 182 (1984) Ch. 2;
A. Dipple, in: *Polycyclic Hydrocarbons and Carcinogenesis,* ed. R.G. Harvey, ACS Symposium Series 283 (1985) Ch. 1.
[32] A.T. Balaban, in: *Chemical Applications of Graph Theory,* ed. A.T. Balaban (Academic Press, New York, 1976) Ch. 5, p. 74.
[33] A.T. Balaban, Pure and Appl. Chem. 54(1982)1075.
[34] N. Trinajstić, *Chemical Graph Theory,* Vol. I (CRC Press, Inc., Boca Raton, Florida, 1983) p. 23.
[35] A.T. Balaban, J. Chem. Inf. Comput. Sci. 25(1985)334.
[36] O.E. Polansky and D.H. Rouvray, Match 2(1976)63; ibid. 2(1976)91; ibid. 3(1977)97.
[37] A.T. Balaban, J. Mol. Struct. (Theochem) 120(1985)117.
[38] I. Gutman, Theor. Chim. Acta 45(1977)309.
[39] J. Ciosłowski, Chem. Phys. Lett. 122(1985)234; Theor. Chim. Acta 68(1985)315; Match 19 (1986)1963.

[40] S. El-Basil, P. Krivka and N. Trinajstić, Croat. Chem. Acta 57(1984)339.
[41] The Balaban approach (best explicated in ref. [33]) relies on providing the angular orientation information in the body of the code, whereas the Ciosłowski and Turek method (ref. [13]) typifies the requirement for canonical placement of the molecular structure on a reference lattice.
[42] W.C. Herndon and L. v. Szentpály, J. Mol. Struct. (Theochem) 148(1986)141.
[43] In storing molecular structure information, we have tentatively reserved the lower case letters s, d, t, and a to denote single, double, triple, and aromatic (delocalized) bonds, respectively. The use of "a" in the present context seems appropriate.
[44] We propose that augmented labels would include the ring size as a prefix, i.e. a five-membered ring annelated to two adjacent rings would be designated 5B1 or 5B2, depending on the substitution pattern.
[45] In addition, stereochemical features of atoms (configuration R, S) and of bonds (cis (Z), trans (E)) are designated by adding the steric descriptors preceded by a slash (/) at the appropriate point in the notation. The resulting linear notation is stereochemically accurate up to the present limits of the standarized Chemical Abstracts nomenclature system. Other bond labels, e.g. x for axial and q for equatorial in cyclohexane derivatives, allow the unique designation of particular conformers.
[46] G.W. Adamson and D. Bowden, J. Chem. Inf. Comput. Sci. 15(1975)215.
[47] E. Sankoff, Proc. Nat. Acad. Sci. (USA) 69(1972)4.
[48] A.K.C. Wong, T.A. Reinchert, D.N. Cohen and B.O. Aygun, Comput. Biol. Med. 4(1974)43.
[49] D. Sandoff and J.B. Kruskal, eds., *Time Warps, String Edits, and Macromolecules; The Theory and Practice of Sequence Comparison* (Addison-Wesley, Reading, MA., 1983).
[50] M.S. Waterman, Bull. Math. Biology 46(1984)473.
[51] C.L. Wilkins and M. Randić, Theor. Chim. Acta 58(1980)45.
[52] M. Randić and C.L. Wilkins, Int. J. Quant. Chem. 18(1980)1005.
[53] C.L. Wilkins, M. Randić, S.M. Schuster, R.S. Markin, S. Steiner and L. Dorgan, Anal. Chim. Acta 133(1981)637.
[54] M. Randić, Int. J. Quant. Chem. 11(1984)137.
[55] M. Randić, J. Chem. Inf. Comput. Sci. 24(1984)164.
[56] M. Randić, G.A. Kraus and B. Džomo-Jerman-Blažič, in: *Chemical Applications of Topology and Graph Theory*, ed. R.B. King (Elsevier, Amsterdam, 1983) pp. 192–205.
[57] G. Klopman, K. Namboodiri and A.N. Kalos, in: *Molecular Basis of Cancer*, Part A: *Macromolecular Structure, Carcinogens, and Oncogenes*, ed. R. Rein (Alan R. Liss, Inc., 1985) pp. 287–298.
[58] G. Klopman and O.T. Macina, J. Theor. Biol. 113(1985)637.
[59] G. Klopman and A.N. Kalos, J. Theor. Biol. 118(1986)199.
[60] G. Klopman, O.T. Macina, E.J. Simon and J.M. Hiller, J. Mol. Struct. (Theochem) 134 (1986)299.
[61] R. Carbó, L. Leyda and M. Arnau, Int. J. Quant. Chem. 17(1980)1185.
[62] P. Cleij and H.A. Van 't Klooster, Anal. Chim. Acta 150(1983)23.
[63] I. Motoc, Z. Naturforsch. 38(1983)1342.
[64] P. Broto, G. Moreau and C. Vandyeke, in: *Computer Applications in Chemistry*, ed. S.R. Heller and R. Pentenzone, Jr. (Elsevier, Amsterdam, 1983) pp. 263–284.
[65] D.H. Lafemina and P.C. Jurs, J. Chem. Inf. Comput. Sci. 25(1985)386.
[66] A.Y. Meyer, J. Chem. Soc. Perkin Trans. (1985)1161.
[67] A.Y. Meyer, J. Mol. Struct. (Theochem) 124(1985)93.
[68] B. Džomo-Jerman-Blažič, I. Fabič and M. Randić, J. Comput. Chem. 7(1986)176.
[69] W.C. Herndon, Trans. N.Y. Acad. Sci. Series II, 36(1974)200.

[70] W.C. Herndon, Int. J. Quant. Chem., Quantum Biol. Symp. 1(1974)123.
[71] J.P. Lowe and B.D. Silverman, Acc. Chem. Res. 17(1984)332.
[72] L. v. Szentpály, J. Amer. Chem. Soc. 106(1984)6021.
[73] G. Klopman, J. Amer. Chem. Soc. 106(1984)7315.
[74] B.D. Silverman, Chem.-Biol. Interactions 53(1985)313.
[75] D. Qianhuan and Z. Dunxin, Scientia Sinica, Series B28(1985)160.
[76] J.J. Kaufman, P.C. Hariharan, W.S. Koski and K. Balasubramanian, in: *Molecular Basis of Cancer*, Part B, ed. R. Rein (Alan R. Liss, Inc., 1985) pp. 263–275.
[77] A.J. Hopfinger, in: *Molecular Basis of Cancer*, Part B, ed. R. Rein (Alan R. Liss, Inc., 1985) pp. 277–286.
[78] G. Klopman, in: *Molecular Basis of Cancer*, Part B, ed. R. Rein (Alan R. Liss, Inc., 1985) pp. 287–298.
[79] L. v. Szentpály, in: *Molecular Basis of Cancer*, Part B, ed. R. Rein (Alan R. Liss, Inc., 1985) pp. 327–329.
[80] J.L. Hartwell, Survey of compounds which have been tested for carcinogenic activity, U.S. Public Health Service Publ. No. 149, Washington, D.C., 1951;
P. Shubik and J.L. Martwell, Ibid. Supplement 1, 1957; Ibid. Supplement 2, 1969;
J.I. Thompson et al., Ibid. Volumes covering 1968-69, 1970-71, 1972-73, 1974-75, 1976-77, 1978, 1979-80; Cumulative Indexes, NIH Publication No. 84-2691, Nov. 1984.
[81] J. Iball, Amer. J. Cancer 35(1939)188.
[82] The history of the quantum mechanical approach to elucidation of the mechanism of PBAH carcinogenesis is in ref. [70]. For an important early review, see C.A. Coulson, Adv. Cancer Res. 1(1953).
[83] For reviews, see W.C. Herndon, Israel J. Chem. 20(1980)276;
L. v. Szentpály and W.C. Herndon, Croat. Chem. Acta 57(1984)1621.