Applications of the Radius–Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds

Michel Petitjean

Institut de Topologie et de Dynamique des Systèmes, associé au CNRS, Université de Paris VII,
1 rue Guy de la Brosse, 75005 Paris, France

Received December 27, 1991

The shape coefficient (I) of a chemical compound is defined as the ratio $(D-R)/R$, where R is the generalized radius and D is the generalized diameter. Chemical compounds have both a graph-theoretical and a geometrical-shape coefficient, and the properties of these coefficients are examined in this paper. The graph-theoretical bivariate repartition of the (R,D) pairs, named here the "radius–diameter diagram", has been computed for members of a large file of compounds derived from the Chemical Abstracts Services Registry File. This analysis shows many shapes and structural formulas to be absent and suggests that, prior to 1978, organic chemistry had evolved in only a few specialized directions.

INTRODUCTION

Many structural indices may be found in the chemical literature.[1-4] Most of them have been used in QSAR studies, and in such cases, they must meet two different requirements. They must be simple, i.e., easy to understand and compute, and they must be sensitive to small structural perturbations, which reflect slight variations in physical or chemical properties. Such indices can be used for any property with which it is observed to be correlated, but a different structural index may be more suitable for a different property. The correlations that have been reported are usually derived from well-defined structural families which do not contain structurally diverse compounds. Thus, correlations dealing with fatty acids or with sugars have been published, but no attempt has been made to compare such distinct compound types. No universal indices have been discovered; the ability of structural indices to explain or predict chemical or physical properties is usually limited to specific areas of chemistry.

Chemical structures have two shape coefficients. The first is based upon graph theory and is calculated from the compound's atom-bond graph. The other is the geometrical-shape coefficient, which is determined by its geometrical shape. Shape coefficients are not intended to support accurate QSAR calculations. Rather, they are a means for the classification of large collections of compounds and offer a compromise between QSAR indices, which are accurate but applicable to uninterestingly restricted groups of chemicals, and complex and intractable molecular descriptors obtained from the analysis of very large databases. The definitions of these indices and the coefficients may be applied simultaneously to structural formulas and to three-dimensional solids. This unusual property provides a method of comparing the topology and the geometry of chemical compounds.

DEFINITIONS AND GENERAL PROPERTIES OF RADIUS AND DIAMETER

The concepts of "radius" and "diameter" that are introduced here are normal generalizations based upon the ordinary radius and diameter, and their properties are examined below. The following definitions apply to any metric space. They apply to both graphs and Euclidian solids and, thus, structural formulas and molecular solids. Aside from the "object", the "shape coefficient" and the "radius–diameter diagram", all the definitions are unremarkable.

Let $d(x,y)$ be the distance between two points, x and y, in a metric space.

Let us define an object as a closed bounded subset.

Let x be a point in a given object. The eccentricity $E(x)$ of the x point is the upper bound of $d(x,y)$ taken from the set of all y points in the object.

The radius R of an object is the lower bound of the eccentricity $E(x)$, taken from the set of all the x points. A point for which $E(x) = R$ is called the center of the object, but this center may be nonunique. A point lying at a distance R from a center is an R-extremal point.

The diameter D of an object is the upper bound of the eccentricity $E(x)$, taken from the set of all the x points. A point with $E(x) = D$ is a D-extremal point.

Let us define $I = (D-R)/R$ as the shape coefficient of the object, and let us define the radius-diameter diagram as the bivariate (R,D) distribution of a population of objects.

For a compact object, the following properties may be simply inferred:

1. $E(x)$ varies from R to D. It assumes only finite values and reaches both bounds.
2. An object containing only a single point (x_0) is such that $E(x_0) = R = D = 0$. (I is undefined in this particular situation).
3. An object containing at least two points has at least two D-extremal points, e_1 and e_2. It also has at least one center and one R-extremal point.
4. For any object, $|E(x) - E(y)| \leq d(x,y) \leq \mathrm{Inf}(E(x), E(y)) \leq D$.
5. The triangle inequality applied to the center c and to the D-extremal points e_1 and e_2 shows that D varies from R to $2R$, depending on the shape of the object. Consequently, I varies from 0 to 1.
6. When $D = R$ ($I = 0$), then $E(x) = R = D$ for all x, and all the points are centers, R-extremal, and D-extremal points.
7. When $D = 2R$ ($I = 1$), any D-extremal point is also R-extremal, but all R-extremal points need not be D-extremal and the center need not be unique.

No particular assumption is made concerning the distance needed to determine these properties. Additional properties can be deduced when distance is specified in the set of nodes in a nonedge-valued graph or in Euclidian n-dimensional space.

0095-2338/92/1632-0331$03.00/0

GRAPH-THEORETICAL PROPERTIES AND RELATIONSHIP TO CHEMICAL NOMENCLATURE

The structural formula of a compound is usually represented by an undirected graph whose nodes stand for the atoms and edges for the bonds.[5] The graph-theoretical distance between two nodes is the smallest number of edges between them.[6] When there is no path between two nodes, the distance is infinite. It is possible to represent multiple bonds either with multiple edges or with single edges that are colored.[5] The multiplicity of the edges does not affect the values of the distances, and thus these two representations are equivalent. The hydrogen-suppressed graph and the complete graph are not equivalent, but the properties of the graph-theoretical distance apply to both.

Any connected partial subgraph—a substructure or a fragment—is an object, but the most natural chemical unit is the component, which is a maximum cardinality compact object. As an example, sodium trifluoroacetate has two components, the trifluoroacetate anion and the sodium cation. In this molecule, the carbonyl and trifluoromethyl groups are objects, but the sodium trifluoroacetate is not an object because the distances d(C,Na), d(O,Na) and d(F,Na) are infinite. Thus, the trifluoroacetate component is a natural chemical unit. The following properties[6] are present:

1. Distances, eccentricities, radii, and diameters all have integer values.
2. An acyclic object (i.e., one containing no cyclic bonds) has either an even D value with $D = 2R$ and a unique center or an odd D value with $D = 2R - 1$ and two centers c_1 and c_2 with $d(c_1,c_2) = 1$. There may be non-acyclic objects with $D = 2R$ or $D = 2R - 1$; an example would be n-octylbenzene.
3. An object for which $D = R$ is strictly cyclic, and all its bonds are cyclic. There may, however, be strictly cyclic objects, naphthalene, for example, for which $D > R$. Objects which have both cyclic and acyclic bonds are nonstrictly cyclic.

The graph-theoretical shape coefficient of a molecule may be interpreted as a measure of the balance between its cyclic part and its acyclic part. If $I = 0$, the structure must be strictly cyclic; no acyclic bonds are permitted. An acyclic structure with an even diameter and no cyclic bonds implies $I = 1$. Some examples of I values for common single-component compounds are given in Figure 1, and a detailed calculation of I is given in Figure 2.

The center of a component is a focus chosen so as to minimize the number of concentric layers surrounding it. A center need not be unique. Each D-extremal atom is a focus which maximizes the number of concentric layers surrounding it. The first of the concentric layers contains the focus itself, the second layer contains the neighbors of the focus, the third the next neighbors, and so on. These concentric layers are often explored by processing algorithms such as the Cahn–Ingold–Prelog convention or substructure searching programs.[7] In such cases, concentric numbering of the atoms around the center of a component is a useful way to optimize time-critical or space-critical algorithms.

For most hydrocarbons and their derivatives, chemical nomenclature is closely related to the value of D. Thus methanes ($D = 0$), ethanes ($D = 1$), and higher families are obtained. Names are derived from the name of the hydrocarbon corresponding to the longest carbon chain in the structure. The major deviations from this canon occur when the longest chain contains heteroatoms or when any of the many specific cyclic systems are present.

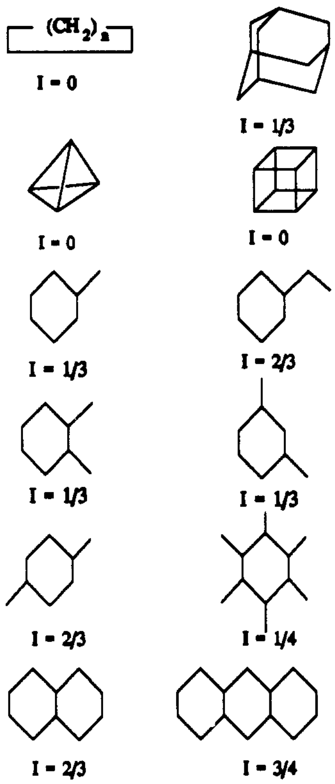
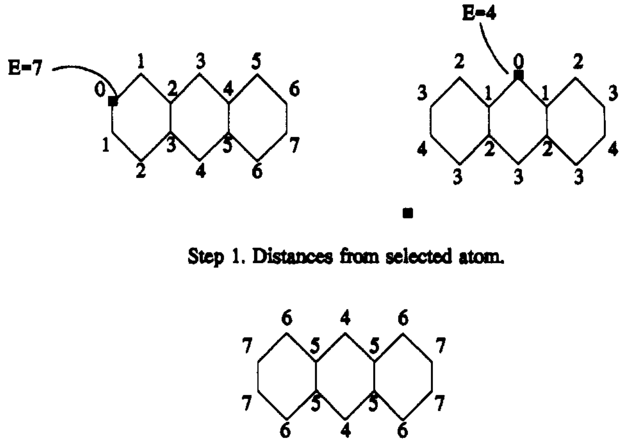


Figure 1. Examples of values of the graph-theoretical shape coefficient, $I = (D - R)/R$, for common structures. Unsaturated and saturated compounds have the same I value, because the distances, radii, and diameters do not depend upon bond multiplicity.



Step 1. Distances from selected atom.

Step 2. Atom eccentricities.

R = 4 (lowest eccentricity) and D = 7 (highest eccentricity).

Thus I = (D - R)/R = 3/4

Figure 2. Detailed calculation of the graph-theoretical shape coefficient of anthracene. In the first step, an atom is selected and all other atoms are labeled with their distance from the selected atom. Only two atoms are shown in this example. The highest value label is the eccentricity of the selected atom. In the second step, every atom is labeled with its eccentricity: the lowest label is the radius of the structure; the highest, the diameter.

GEOMETRICAL PROPERTIES AND THEIR RELATIONSHIP TO THE NOTION OF CONVEXITY

The natural distance in a Euclidian space is the normal Euclidian distance. In order to distinguish local shapes and concavities in a solid, we define the distance $d(x,y)$ between two points inside this solid as the smallest path between x and y. All the continuum of points in the path are included in the object.

Let R^n be the multidimensional real space over the real field, with a finite dimension, n. Let l be the distance induced by a scalar product defined over R^n.

Let S be a closed bounded subset in R^n. S is thus compact. A path between two points, x and y, in S is an ordered sequence of points $(x, x_1, x_2, ..., y)$, beginning with x and ending with y, such that each pair of successive points defines a segment included in S. The length of the path $(x, x_1, x_2, ..., y)$ is the sum of the lengths of all successive segments, each length of a segment being the distance l between its end points. When the set of all paths between x and y is not empty, $d(x,y)$ is the lower bound of the length of these paths, and when no path exists between x and y, $d(x,y)$ is infinite.

Thus d is a distance on S (see Appendix), such that $d(x,y) \geq l(x,y)$, and equality occurs when the smallest path is reduced to the segment $[x,y]$ itself.

The shape of an object is closely related to the concept of *convexity*. In a convex object, the smallest path between x and y is always the segment (x,y) itself: $d(x,y) = l(x,y)$.

In order to facilitate an understanding of the relationship between R-extremal, D-extremal, and extremal points (in a convex sense), the definitions and properties of convex-extremal points are given below. They are general and apply to spaces more general than R^n.[8–10]

The convex hull H of a set S (convex or not) is the intersection of all convex sets containing S. H is also the set of all convex linear combinations of points pertaining to S. H is unique and contains S. The convex hull of a convex set S is S itself.

A point x of the convex hull H of a set S is called convex-extremal if no open segment included in H contains x. An alternative definition of convex-extremal points is the set of x points in the convex hull H such that H-$\{x\}$ is convex. The set of convex-extremal points of S is unique and included in S.

Any point in a closed bounded set S is a convex linear combination of the convex-extremal points of S. In general, this combination is not unique. A well-known exception is the nondegenerate n-simplex: the coefficients of the unique n-tuple are called the barycentric coordinates and are invariant under affine transformation. A closed bounded convex set is the set of convex linear combinations of its convex-extremal points.

Thus the d-distance has all the properties of the l-distance for convex objects. The most remarkable of these properties[8] are as follows:

1. The radius R is the radius of the smallest n-sphere containing the object. This n-sphere is unique, and its center is the unique center of the object. Note that the smallest n-sphere (referred to the l-distance) containing a given object, convex or otherwise, is always unique, but its center may not pertain to the object. This smallest n-sphere is sometimes called the circumscribed n-sphere, but such terminology is ambiguous. The three vertices of a triangle are usually thought to lie on the boundary of its circumscribed circle; when an angle in the triangle approaches 180°, the radius tends to infinity while R is equal to the half-length of the longest side.

2. All R-extremal points are convex-extremal points, and all D-extremal points are convex-extremal, but there may be convex-extremal points which are neither R-extremal or D-extremal.

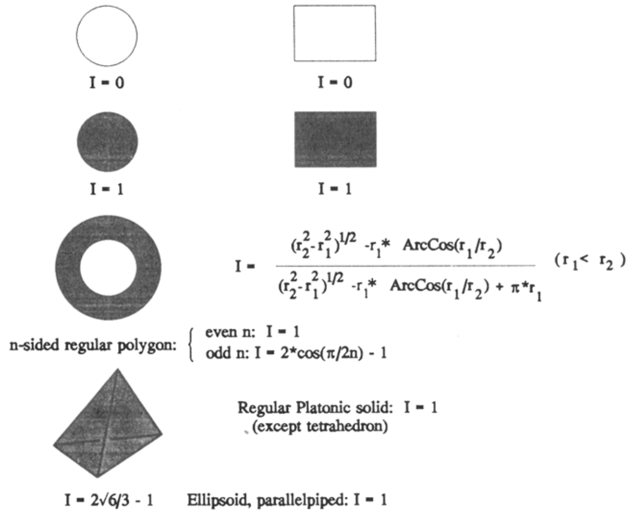

I = 0
I = 0
I = 1
I = 1
I = (r_2^2 - r_1^2)^{1/2} - r_1* ArcCos(r_1/r_2) / (r_2^2 - r_1^2)^{1/2} - r_1* ArcCos(r_1/r_2) + π*r_1
(r_1 < r_2)
n-sided regular polygon:
even n: I = 1
odd n: I = 2*cos(π/2n) - 1
Regular Platonic solid: I = 1
(except tetrahedron)
I = 2√6/3 - 1
Ellipsoid, parallelpiped: I = 1

Figure 3. Examples of values of the geometrical-shape coefficient, $I = (D - R)/R$ for common solids.

3. D has a minimum value for a given radius R:

$$D^2 \geq R^2(2 + 2/n) \qquad \text{(a)}$$

4. If e_1 and e_2 are two D-extremal points which satisfy $d(e_1,e_2) = D$, the cosine of the $(e_1$-c-$e_2)$ angle has a maximum value:

$$\cos\,(e_1\text{-}c\text{-}e_2) \leq -1/n \qquad \text{(b)}$$

5. If m denotes the midpoint of the segment e_1-e_2, $d(c,m)$ has a maximum value:

$$d^2(c,m) \leq R^2(n-1)/2n \qquad \text{(c)}$$

When one of the inequalities (b) or (c) is satisfied, then all three equalities (a), (b), and (c) will be satisfied, and there are exactly $n+1$ R-extremal points, which are all D-extremal and which are the vertices of a regular n-simplex. The regular n-simplex itself satisfies the three equalities.

A convex object always has a shape coefficient:

$$I \geq \sqrt{(2 + 2/n)} - 1$$

If I is smaller than this minimum value, the object cannot be convex, but there may be nonconvex objects with any I value in [0;1].

None of the previous properties assumes a specific expression of the scalar product. Obviously, the radius and diameter concepts have the usual meaning for a full circular or a spherical object, for which $I = 1$. Examples of I values for some common solids are given in Figure 3.

APPLICATIONS TO MOLECULAR SOLIDS

For chemical applications, the dimensionality of n is 3 and the ordinary scalar product should be selected. It should be noted that all "planar" compounds such as polycyclic aromatics are three-dimensional objects whose shape depends not only on their atomic coordinates but also on the shape of their orbitals. All the geometric properties listed above apply to chemical objects. Thus, for example, the minimal I value for a convex object is reached with the regular tetrahedron, the smallest convex chemical object with a given radius. No convex chemical object has an $(e_1$-c-$e_2)$ angle smaller than the angle in a regular tetrahedron, whose value is arccos (–1/3) or 109° 28′ 16″.

The (R,D) pair and the shape coefficient do not uniquely describe the geometries of three-dimensional molecular shapes,

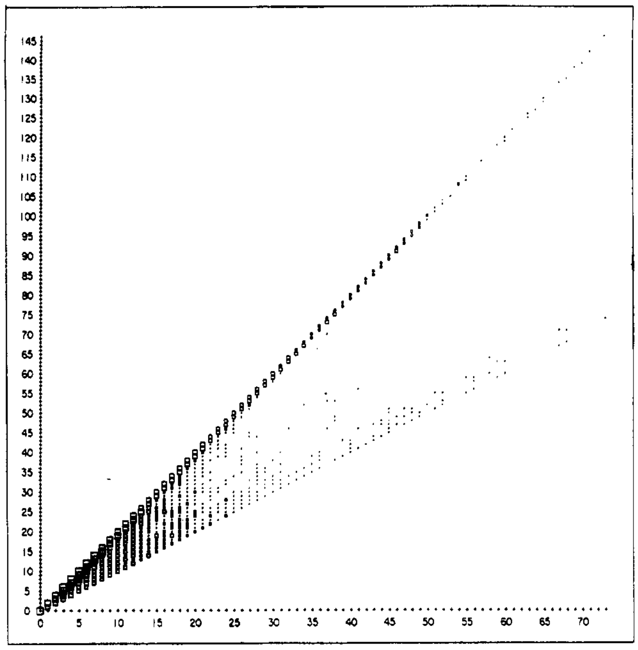


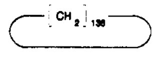
Figure 4. Bivariate distribution of the 4 019 514 components in a graph-theoretical radius–diameter diagram. This is based upon the data in Table I. The radius, R, is plotted along the x-axis, and the diameter, D, on the y-axis; thus $I = (y/x) - 1$. There are 3 424 428 compounds in the database, but many of these (e.g., salts) have more than one component. There are a total of 4 019 514 components in the database. The sides of the squares in the graph are proportional to the logarithm of the number of components they represent. All the observations fall between the lines $D = R$ ($I = 0$) and $D = 2R$ ($I = 1$). Among the 4 019 514 (R,D) pairs, only 509 are unique; as an example, there are 403 027 occurrences of structures for which $R = D = 0$.

but no geometric index or pair of indices could be expected to do so. The geometrical radius–diameter diagram of a set of chemical objects is a simple summary of the distribution of their chemical shapes. It avoids the excessive complexity of descriptions involving numerous coordinates as well as the oversimplification of an index which may be unique but which contains none of the shape information.

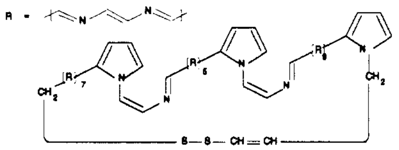
APPLICATION TO THE LARGE CHEMICAL ABSTRACTS FILE

The same observations are relevant to the graph-theoretical radius–diameter diagram. The main difference between the graph-theoretical and the geometrical radius–diameter diagrams stems from the difference between a structural formula and the corresponding molecular shape. The structural formula carries incomplete and idealized geometrical parameters. Hybridizations, for example, may be computed to provide bond angle information, but overall the structural formula, at best, offers a pale reflection of the molecular properties traditionally associated with the compound. No single topological index can provide more information than the structure, but a pair of topological indices represents an efficient tool with which to examine large amounts of data.

The graph-theoretical radius–diameter diagram has been computed for 3 424 428 structures in a Chemical Abstracts Service (CA) Registry File obtained from CA in 1978.[5] The computation algorithm is easily built by using concentric layers, as described previously.[5] It shows, surprisingly, that the repartition of the components leads to clusters around the limit areas $I = 0$ and $I = 1$. Only 509 (R,D) pairs are observed and, except in cases where R or D has a low value, intermediate I values are absent. This result is shown in both Figure 4 and Table I. It is also seen in the cumulative I distribution in Figure 5. Here, most I values are close to either 0 or 1, and only 263 different I values are observed. This means that, prior to 1978, high molecular weight compounds and their skeletons appeared mostly in two areas. Compounds typical of those found close to the $I = 0$ area have regular or quasiregular vertex graphs such as are found in macrocycles such as the cycloalkane **1** (RN 63217-83-4) or cyclopolyimines such as **2** (RN 65000-54-6). Compounds found in the $I = 1$ area are typically polypeptides such as the proinsulin connecting peptide, $C_{138}H_{220}N_{36}O_{50}$ (RN 57327-90-9), polymers, or biopolymers such as $C_{144}H_{206}N_{36}O_{37}$ (RN 54442-02-3), and even when they contain cyclic units, large polymers may have a value of I that is close to 1. It should be noted that the calculation of I is independent of atom and bond type; cyclohexane and pyridine are considered to have the same skeleton and both have $I = 0$. Some of these topologically extremal compounds are identified in our earlier paper.[5]

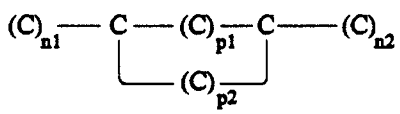


1



2

The existence of these two opposite areas may be interpreted in terms of a cyclic–acyclic balance. Compounds with higher molecular weights have a highly dominant part which is either cyclic (I close to 0) or acyclic (trees in which I is close to 1). Some very simple skeletons corresponding to missing I values are shown below.



n_1	n_2	p_1	p_2	R	D	I
6	0	19	19	20	26	0.300
3	3	19	19	20	26	0.300
3	0	20	20	21	24	0.143
2	1	20	20	21	24	0.143

Many other unobserved skeletons may be built using the data in Table I. The large size of the file ensures the statistical significance of the radius–diameter diagram. In comparison with known organic compounds, however, most of the unobserved structures should not be synthetically unapproachable and neither a technical or a historical explanation for this distribution is proposed.

The graph-theoretical and geometrical-shape indexes are two different descriptions of the same molecule. If a rough topology–geometry correlation is assumed to be possible (e.g., a structure containing a single large ring, for example, cyclotetradecane, has graph-theoretical I value of 0, and a possible associated closed filiform shape also yields a geometrical I value of 0), then the unobserved intermediate I values imply a lack of range of geometrical shapes in the file. Thus, for a given family of compounds, the topology–geometry correlation, sometimes called the topology–topography correlation, may be thought of as an ordinary correlation (in the regression sense) between the graph-theoretical coefficient $I(T)$ and the geometrical coefficient $I(G)$ and computed as the ordinary correlation coefficient $r[I(G),I(T)]$.

Table I. Distribution of Components in Radius–Diameter Diagram

R	*D*	comp	*R*	*D*	comp	*R*	*D*	comp	*R*	*D*	comp	*R*	*D*	comp
0	0	403027	12	19	289	26	27	2	29	36	2	34	52	1
1	1	9483	13	19	44	27	27	6	34	36	2	50	52	1
1	2	90059	14	19	6	14	28	5778	36	36	6	52	52	1
2	2	2260	15	19	219	15	28	63	19	37	2130	27	53	243
2	3	41436	16	19	21	16	28	10	20	37	4	37	53	1
3	3	7488	17	19	241	17	28	7	21	37	1	38	53	1
2	4	57738	18	19	13	18	28	3	26	37	1	52	53	1
3	4	33622	19	19	16	19	28	6	28	37	1	27	54	171
4	4	1072	10	20	34592	20	28	6	31	37	4	28	54	2
3	5	162485	11	20	243	22	28	2	35	37	1	28	55	278
4	5	5903	12	20	250	24	28	10	19	38	1607	51	55	1
5	5	420	13	20	32	25	28	1	20	38	22	52	55	1
3	6	200376	14	20	15	26	28	2	21	38	3	55	55	1
4	6	64112	15	20	5	27	28	3	22	38	4	28	56	196
5	6	932	16	20	40	28	28	4	26	38	2	41	56	1
6	6	495	17	20	7	15	29	6059	30	38	1	56	56	1
4	7	274721	18	20	26	16	29	8	36	38	2	29	57	157
5	7	4024	19	20	7	17	29	4	38	38	1	29	58	142
6	7	1058	20	20	23	18	29	17	20	39	2053	30	58	1
7	7	373	11	21	26544	19	29	3	21	39	7	56	58	2
4	8	325601	12	21	191	20	29	5	23	39	4	30	59	187
5	8	18374	13	21	44	22	29	8	24	39	1	55	59	1
6	8	1595	14	21	16	25	29	1	26	39	1	56	59	1
7	8	652	15	21	9	26	29	5	35	39	2	59	59	1
8	8	335	16	21	9	27	29	1	36	39	2	30	60	162
5	9	356157	17	21	18	28	29	1	39	39	1	31	60	1
6	9	1661	18	21	22	29	29	1	20	40	1580	58	60	1
7	9	811	19	21	12	15	30	4378	21	40	13	60	60	1
8	9	504	20	21	2	16	30	31	22	40	1	31	61	160
9	9	374	21	21	11	17	30	5	24	40	2	31	62	92
5	10	357682	11	22	20534	18	30	4	28	40	2	32	63	102
6	10	14451	12	22	197	19	30	2	31	40	2	59	63	1
7	10	757	13	22	126	20	30	68	36	40	2	60	63	1
8	10	481	14	22	33	21	30	4	40	40	9	32	64	170
9	10	286	15	22	8	22	30	4	21	41	1203	58	64	4
10	10	250	16	22	10	23	30	3	22	41	4	33	65	102
6	11	314533	17	22	10	25	30	1	23	41	1	33	66	97
7	11	1140	18	22	80	26	30	3	24	41	1	34	67	137
8	11	269	19	22	24	27	30	3	26	41	1	67	67	1
9	11	427	20	22	6	28	30	4	40	41	1	34	68	88
10	11	134	21	22	5	30	30	5	41	41	1	68	68	1
11	11	155	22	22	11	16	31	4552	21	42	1039	35	69	60
6	12	270911	12	23	16864	17	31	12	22	42	6	35	70	95
7	12	4873	13	23	120	18	31	5	23	42	1	37	70	1
8	12	417	14	23	10	19	31	1	24	42	1	36	71	35
9	12	254	15	23	13	20	31	7	40	42	3	67	71	1
10	12	262	16	23	9	21	31	5	42	42	2	68	71	1
11	12	80	17	23	6	22	31	3	22	43	999	36	72	68
12	12	200	18	23	12	23	31	3	23	43	1	37	73	108
7	13	202923	19	23	17	27	31	1	39	43	1	37	74	76
8	13	325	20	23	3	29	31	1	40	43	1	73	74	1
9	13	740	21	23	4	31	31	1	43	43	1	38	75	152
10	13	102	22	23	3	16	32	3308	22	44	722	38	76	49
11	13	119	23	23	1	17	32	21	23	44	7	39	77	82
12	13	85	12	24	12467	18	32	12	27	44	3	39	78	46
13	13	98	13	24	138	19	32	2	28	44	8	40	79	73
7	14	165761	14	24	14	20	32	1	43	44	1	40	80	49
8	14	2711	15	24	11	21	32	1	44	44	4	41	81	22
9	14	490	16	24	5	24	32	1	23	45	629	41	82	32
10	14	234	17	24	35	28	32	5	24	45	1	42	83	28
11	14	66	18	24	4	29	32	1	25	45	2	42	84	45
12	14	138	19	24	5	30	32	2	27	45	3	43	85	28
13	14	61	20	24	20	31	32	1	45	45	2	43	86	28
14	14	103	22	24	10	32	32	2	23	46	552	44	87	32
8	15	126591	24	24	24	17	33	3052	24	46	15	44	88	40
9	15	850	13	25	12234	18	33	15	25	46	1	45	89	20
10	15	123	14	25	140	19	33	1	32	46	1	45	90	36
11	15	117	15	25	17	20	33	5	37	46	1	46	91	126
12	15	149	16	25	137	22	33	2	41	46	1	46	92	26
13	15	86	17	25	19	24	33	1	45	46	1	47	93	11
14	15	33	18	25	47	27	33	1	24	47	438	47	94	37
15	15	65	19	25	5	28	33	1	25	47	1	48	95	14
8	16	96357	20	25	12	29	33	6	43	47	1	48	96	20
9	16	1291	21	25	1	30	33	1	44	47	1	49	97	20
10	16	322	22	25	2	31	33	1	45	47	1	49	98	26
11	16	61	23	25	2	33	33	1	46	47	1	50	99	9
12	16	127	24	25	2	17	34	2338	47	47	1	50	100	21

Table I (Continued)

R	*D*	comp	*R*	*D*	comp	*R*	*D*	comp	*R*	*D*	comp	*R*	*D*	comp
13	16	42	25	25	9	18	34	19	24	48	391	51	101	1
14	16	57	13	26	8403	19	34	3	25	48	2	51	102	6
15	16	38	14	26	187	20	34	1	38	48	1	52	103	5
16	16	48	15	26	12	21	34	2	44	48	1	52	104	1
9	17	80508	16	26	17	25	34	1	45	48	1	53	105	2
10	17	232	17	26	2	28	34	4	48	48	1	54	108	11
11	17	125	18	26	10	30	34	3	25	49	398	55	109	4
12	17	125	19	26	5	32	34	2	26	49	1	55	110	4
13	17	65	21	26	2	34	34	2	38	49	1	57	114	1
14	17	15	22	26	5	18	35	2891	46	49	4	59	118	1
15	17	74	23	26	3	19	35	2	47	49	1	60	119	1
16	17	13	24	26	6	20	35	3	49	49	1	60	120	1
17	17	31	25	26	2	21	35	5	25	50	322	61	122	1
9	18	59047	26	26	4	22	35	2	26	50	4	63	125	1
10	18	537	14	27	8065	26	35	2	45	50	1	63	126	5
11	18	146	15	27	8	28	35	1	48	50	1	64	127	2
12	18	203	16	27	16	29	35	5	49	50	1	65	129	1
13	18	14	17	27	6	31	35	1	26	51	264	65	130	1
14	18	48	18	27	6	33	35	1	27	51	1	67	134	1
15	18	18	19	27	20	35	35	3	45	51	1	68	135	1
16	18	31	20	27	3	18	36	2563	47	51	1	69	138	1
17	18	7	21	27	4	19	36	19	48	51	1	70	139	1
18	18	50	23	27	4	20	36	2	51	51	1	71	142	1
10	19	45409	24	27	2	21	36	4	26	52	231	73	146	2
11	19	241	25	27	4	28	36	6	27	52	11			

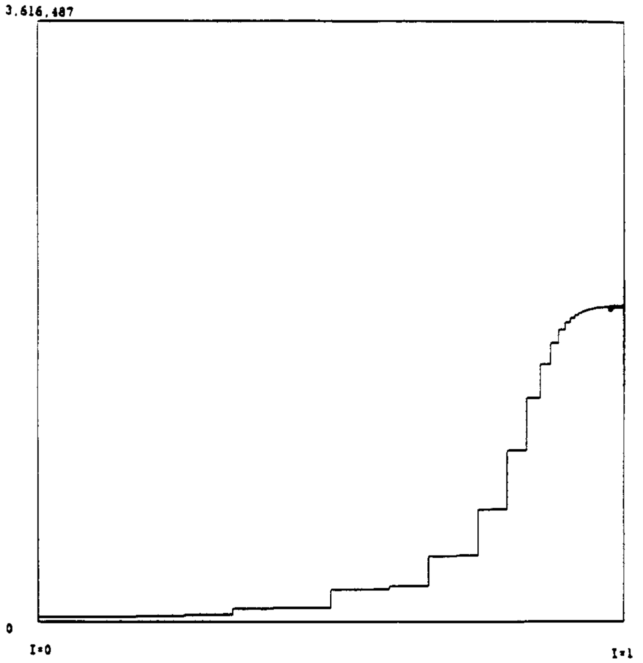


Figure 5. Cumulative distribution of the graph-theoretical shape coefficient of 4 019 514 − 403 027 = 3 616 487 non-monoatomic components. The data were computed from Table I. The 403 027 monoatomic components have no *I* value because their radius is zero (see the beginning of Table I). Among the 3 616 487 *I* values, only 263 are unique. The frequencies of the *I* = 0 and *I* = 1 values are 23 462 and 1 726 186, respectively.

CONCLUSION

The classification of chemical compounds using either their graph-theoretical or their geometrical shapes implies two opposing requirements. One must avoid complex, impractical descriptions which lead to as many chemical classes as there are compounds, in which case the aim of the exercise—grouping of the compounds—is lost. On the other hand, too simple a classification scheme loses touch with the shape characteristics. The existence of these two opposed requirements have led us to build the bivariate index described here rather than use any of the numerous univariate indices that are in the literature.

The radius–diameter diagram allows classification of the shapes of compounds and has remarkable properties for both graph-theoretical and geometrical shapes. Comparisons between the two diagrams are possible because the graph-theoretical distance in the structural representation and the usual geometrical distance are comparable and because the radius and diameter concepts are uniquely defined for any given distance.

Aside from our earlier paper,[5] no computation appears in the literature of topological indices for a large number of compounds. Application of the graph-theoretical radius–diameter diagram to the large file from the CA Registry has revealed an unexpected partitioning of the compounds in the file and suggests that very few of the possible graph-theoretical shapes have been observed. Such an observation would not be possible if univariate indices had been used.

ACKNOWLEDGMENT

I thank Roger Attias and Jacques-Emile Dubois for helpful discussions, and I also thank the reviewers, who suggested substantial improvements to this paper.

APPENDIX

Let us establish that d, defined above, is a distance. Let S be any subset in R^n, and l any distance on R^n. Let x, y, and z be three points in S.

That $d(x,y) = d(y,x)$ and $x = y \Rightarrow d(x,y) = 0$ is obvious from the definition of d. Assume that $d(x,y) = 0$. There is at least one path from x to y included in S because $d(x,y)$ has a finite value. The lower bound of the length of the paths from x to y included in S is greater or equal to the lower bound of all the lengths of the paths from x to y in the whole space, i.e., $l(x,y)$. Then $l(x,y) = 0$ and $x = y$.

Triangle inequality: Assume there are points x, y, and z such that $d(x,z)$ and $d(z,y)$ both exist. There is at least one path included in S from x to y, going through z. Then $d(x,y)$ exists, and the lower bound of all the lengths of the paths from x to y included in S cannot be greater than the length of this particular path, which is $d(x,z) + d(z,y)$. Thus, the triangle inequality stands.

REFERENCES AND NOTES

(1) Diudea, M. C.; Minailiuc, O.; Balaban, A. T. Molecular Topology. IV. Regressive Vertex Degrees (New Graph Invariants) and Derived Topological Indices. *J. Comput. Chem.* **1991**, *12*, 527–535.

(2) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Center Concept and Derived Topological Centric Indexes. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 106–113.

(3) Randić, M.; Hansen, P. J.; Jurs, P. C. Search for Useful Graph Theoretical Invariants of Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 60–68.

(4) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(5) Petitjean, M.; Dubois, J.-E. Topological Statistics on a Large Structural File. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 332–343.

(6) Berge, C. Graphes et Hypergraphes. Dunod Université, Paris, 1973 (ISBN 2-04-009755-4).

(7) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.

(8) Eggleston, H. G. Convexity. In *Cambridge Tracts in Mathematics and Mathematical Physics*, No. 47; Smithies, F., Todd, J. A., Eds.; Cambridge University Press: Cambridge, 1966.

(9) Berberian, S. K. Lectures in Functional Analysis and Operator Theory. In *Graduate Texts in Mathematics*; Springer-Verlag: New York, 1974; Vol. 15 (ISBN 0-387-90080-2).

(10) Grunbaum, B. Convex Polytopes. In *Pure and Applied Mathematics*. Courant, R., Bers, L., Stoken, J. J., Eds.; Interscience-John Wiley: New York, 1967; Vol. XVI.