# The Extent of the Relationship between the Graph-Theoretical and the Geometrical Shape Coefficients of Chemical Compounds

Peter A. Bath, Andrew R. Poirrette, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Frank H. Allen

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

Received October 5, 1994®

It has been suggested that the radius-diameter diagram provides an effective characterization of the graph-theoretical and the geometric shapes of molecules and that these two types of shape are correlated with each other. This paper reports an analysis of 25 332 organic molecules from the Cambridge Structural Database and shows that there is only a low degree of correlation between the two shape measures, even for sets of structures that have minimal conformational flexibility.

## INTRODUCTION

The geometric, or three-dimensional (3D), shape of a molecule is known to be of crucial importance in determining its biological properties. This has led to the development of a range of computational techniques for the processing of geometric shape information in applications as diverse as similarity searching,[1] structure generation,[2] the docking of small-molecule ligands into protein active sites,[3] and 3D QSAR,[4] *inter alia*. One continuing problem is the general lack of accurate 3D structural information. Specifically, the largest collection of small-molecule 3D structures, that in the Cambridge Structural Database (CSD) produced by the Cambridge Crystallographic Data Centre,[5] contains *ca.* 120 000 3D structures derived from X-ray crystallographic experiments: this is just 1% of the *ca.* 12 million molecules for which a topological, or two-dimensional (2D), structure is available in the Chemical Abstracts Service (CAS) Registry System.[6] There is hence much interest in techniques that can be used to generate 3D information from an initial 2D representation, as is evidenced by the widespread use of automatic model builders for the construction of 3D databases.[7−10]

In a recent paper,[11] Petitjean has introduced the concept of the *radius-diameter diagram* and suggested that the graph-theoretical (2D) and geometrical (3D) shapes of molecules are comparable. He states that "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". The "comparable" nature of the two types of distance is formalized when he states that "Thus, for a given family of compounds, 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[$I(G)$,$I(T)$]" (where the coefficients referred to are derived from the graph-theoretical and geometric radii and diameters as defined below). This is a strong hypothesis and one that, if correct, would be of great value since it would provide a simple tool for the investigation of shape-related phenomena without the need for accurate 3D structural information. However, the experimental work described in his paper makes use only of 2D molecules, albeit a very large number, from the CAS Registry File, and no evidence is provided that such a correlation is observed in practice. In this paper, we report an evaluation of the radius-diameter diagram with sets of molecules from the CSD for which both 2D and 3D structures are available and demonstrate that while correlations do exist between the graph-theoretical and geometric coefficients, the magnitudes of these correlations are considerably less than unity even in the case of molecules for which 3D conformational flexibility is effectively impossible.

## CALCULATION OF SHAPE COEFFICIENTS

Following Petitjean,[11] we now summarize the calculation of the graph-theoretic and geometric shape coefficients, $I_2$ and $I_3$. A 2D molecule can be represented by an undirected graph, in which the nodes of the graph represent the atoms and the edges of the graph represent the bonds. The graph-theoretical distance, $D(J,K)$, is defined as the number of bonds in the shortest path between two nodes $J$ and $K$. Let the *eccentricity* of $K$, $E(K)$, be the largest value of $D(J,K)$ for all paths involving $K$, *i.e.*

$$E(K) = \max \{D(J,K),\ 1 \le J \le N\}$$

where $N$ is the number of atoms in the molecule. The graph-theoretical *radius*, $R_2$, and *diameter*, $D_2$, are defined to be

$$R_2 = \min \{E(K),\ 1 \le K \le N\}$$

and

* Author to whom all correspondence should be addressed.
® Abstract published in *Advance ACS Abstracts*, May 15, 1995.

0095-2338/95/1635-0714$09.00/0 

$$D_2 = \max \{E(K), 1 \leq K \leq N\}$$

respectively, and the graph-theoretical shape coefficient, $I_2$, is then given by

$$I_2 = \frac{D_2 - R_2}{R_2}$$

The geometrical shape coefficient, $I_3$, is calculated in exactly the same way, except that the geometrical eccentricity is defined in terms of the longest distance, rather than the largest number of bonds in the graph-theoretical case, this change being reflected in different (Cartesian) values for the geometric radius, $R_3$, and diameter, $D_3$.

The radius-diameter diagram is defined as the bivariate *(R,D)* distribution of a population of objects and can be calculated from both a graph-theoretical and a geometrical point of view. The diagram provides a simple summary of the distribution of the chemical shapes, in either 2D or 3D, of a set of molecules.

## EXPERIMENTAL DETAILS

The graph-theoretical and geometrical shape coefficients, $I_2$ and $I_3$, were computed for 25 322 organic structures extracted from the April 1994 release of the CSD: these structures were chosen as having exactly one chemical entity per crystallographic asymmetric unit, *i.e.*, exactly one molecule per "crystal chemical unit" as defined in the CSD.[5]

The number of rotatable bonds, $N_{Rot_1}$, was calculated for each molecule in this dataset, using the following rules for the definition of a rotatable bond:

1. The bond must be single and acyclic.
2. In a bond A−B, then A can be $C_{sp^1}$, $C_{sp^2}$ [ethylenic or aromatic], or $C_{sp^3}$.
3. If atom A is $C_{sp^3}$, then B can be anything.
4. If atom A is $C_{sp^1}$ or $C_{sp^2}$, then B can be only $C_{sp^3}$, O, S, P or Si. This excludes the most common conjugated systems, $C_{sp^{1,2}}-C_{sp^{1,2}}$ and $C_{sp^2}-N$, where there are strong restrictions to free rotation.
5. Both A and B must be bonded to at least one further non-H atom so as to exclude, *e.g.*, the $O-CH_3$ bond in $C-O-CH_3$ systems, etc. Essentially, these are X−A−B−Y systems in which both X and Y are non-H.

Once the $N_{Rot_1}$ values had been calculated, the molecules were divided into subsets containing 0, 1, 2, 3−4, and ≥5 rotatable bonds.

A more restrictive enumeration of the rotatable bonds within a molecule, $N_{Rot_2}$, was obtained by using the following additional rule:

6. Take the system X−A−B−Y, where A−B is the bond of interest, *i.e.*, rotatable as defined above in calculating $N_{Rot_1}$. At least one of the atoms X (attached to A and additional to B itself) must have at least two connections to non-H atoms, and at least one of the atoms Y (attached to B and additional to A itself) also must have at least two non-H connections.

This rule serves to exclude bonds such as the C−C bond in $-C-CF_3$, *etc.*, and the $N_{Rot_2}$ counts hence represent the more "shape-defining" of the rotatable bonds in a molecule. Once the $N_{Rot_2}$ values had been calculated, the molecules were again divided into subsets containing 0, 1, 2, 3−4, and ≥5 rotatable bonds.

The molecules for which $N_{Rot_1} = 0$ were then divided into subsets containing 0, 1, 2, 3−4, and ≥5 ($N_{Rot_3}$) flexible *cyclic* bonds, which were simply defined here as cyclic $C_{sp^3}-C_{sp^3}$ or $C_{sp^3}-O_{sp^3}$ single bonds.

Petitjean's basic assumption, *viz.* that there is a strong relationship between graph-theoretical and geometrical molecular shapes, implies that one would expect a strong correlation between the $I_2$ and $I_3$ values for a set of molecules: we have hence calculated the product-moment correlation coefficient, $r$, and from this the fraction of the variance explained, $r^2$, between the sets of $I_2$ and $I_3$ values for the molecules in each of the datasets defined above. All of the correlation coefficients discussed here are significant at the 0.001 level of statistical significance. Full details of the experiments are reported by Bath.[12]

## RESULTS AND DISCUSSION

Table 1 details the values of $r^2$ between the graph-theoretical and the geometric shape coefficients for molecules with different values of $N_{Rot_1}$. As one would expect, the largest value is obtained for rigid molecules, *i.e.*, those having $N_{Rot_1} = 0$, the scatter diagram for which is shown in Figure 1. However, even for these molecules, the value ($r^2 = 0.34$) is less than might have been expected if there was a strong relationship between the two types of shape coefficient, in which case a value nearer to unity would be obtained. Moreover, the correlations drop off rapidly as one considers increasingly flexible molecules, as defined by the values of $N_{Rot_1}$. These results support the intuitive idea that the number of rotatable bonds in a molecule plays an important part in determining the 3D shape of a molecule when compared with its 2D shape.

Analogous, but even less satisfactory results are obtained when one considers the values of $r^2$ calculated for variations in $N_{Rot_2}$ (as detailed in Table 1). Although the general trend is the same, it is noticeable that the value of $r^2$ for each level of $N_{Rot_2} > 0$ is less than that for the corresponding level of $N_{Rot_1}$. This implies, again in accord with intuition, that the rotatable bonds at the edge of a molecule, *i.e.*, those that are included in the calculation of $N_{Rot_1}$ but not of $N_{Rot_2}$, are less important than the more centrally-located rotatable bonds in determining 3D molecular shape.

Thus far, we have considered only acyclic rotatable bonds. The final column of Table 1 gives the results that were obtained with the $N_{Rot_3}$ values, which take account of the limited amount of flexibility that occurs in many cyclic systems, *e.g.*, the chair/half-chair/boat/*etc.* conformations that occur for various types of six-membered rings. While the coefficients here are both higher and more constant than those obtained with the previous definitions of rotatable bonds, they are still much less than unity. The approximate constancy of the coefficients with increasing $N_{Rot_3}$ probably arises because a large proportion of the molecules considered contain six-membered rings, for which the chair form is totally dominant for energetic reasons, thus reducing the variability in shape that might otherwise have been expected. This predominance of six-membered rings is typical of any large collection of chemical structures and is thus not an artefact of the structural content of the CSD.

In view of the poor results that had been obtained thus far, an attempt was made to define a set of molecules for which one might expect the correlation between the 2D and

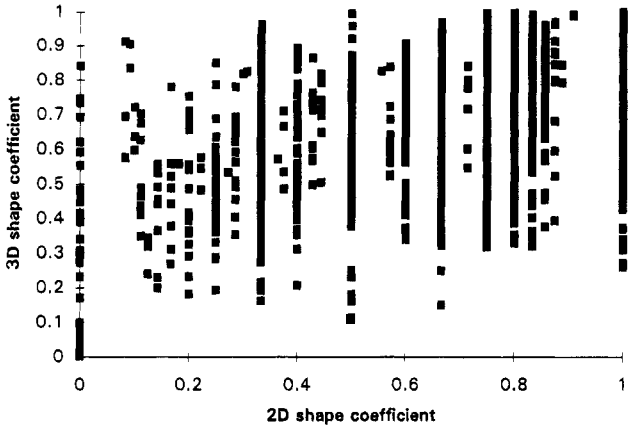


**Figure 1.** Scatter diagram to show the correlation between 2D and 3D shape coefficients for the first 4000 molecules having $N_{Rot_1} = 0$. Value of $r^2 = 0.34$, *N.B.*, only the first 4000 structures could be plotted.

**Table 1:** Values of $r^2$ between the Graph-Theoretical and Geometric Shape Coefficients for Molecules Containing Different Numbers of $N_{Rot_1}$, $N_{Rot_2}$, and $N_{Rot_3}$ bonds[a]

| no. of bonds | $N_{Rot_1}$ | $N_{Rot_2}$ | $N_{Rot_3}$ |
|---|---|---|---|
| 0 | 0.34 (7872) | 0.31 (11947) | 0.36 (3974) |
| 1 | 0.12 (3421) | 0.04 (3823) | 0.35 (392) |
| 2 | 0.08 (4188) | 0.04 (3649) | 0.42 (426) |
| 3−4 | 0.08 (5010) | 0.03 (3530) | 0.37 (656) |
| 5+ | 0.04 (4831) | 0.02 (2373) | 0.28 (2424) |

[a] The brackets contain the number of molecules in each class.

the 3D shapes to be maximized. This was done by selecting those 188 molecules that contained only a single aromatic ring assembly and that contained no flexible bonds of any of the sorts considered here, *i.e.*, both $N_{Rot_1}$ and $N_{Rot_3}$ were zero. The value of $r^2$ for this dataset was 0.28, which is little different from those obtained thus far. The 188-member subset was then further subdivided into the 132 molecules in which one or more of the atoms in the rings had single-atom non-hydrogen substituents and the 56 molecules in which hydrogen was the sole substituent: the values of $r^2$ were 0.17 and 0.53, respectively.

## CONCLUSIONS

In this paper, we have evaluated Petitjean's claim that there is a correlation between the graph-theoretical and geometric shape coefficients, $I_2$ and $I_3$. Statistically significant correlations do exist, but the correlation is far from exact in that even molecules containing unsubstituted aromatic ring systems gave an $r^2$ of only 0.53 (and the other sets of molecules gave values that were as low as 0.02). We hence believe that it is not possible to make predictions about the 3D shapes of molecules from the 2D shapes using the coefficients considered here. Although it is possible that other types of descriptor might allow such shape-based correlations to be made more precisely, we believe that effective shape similarity searching requires the use of geometric, rather than graph-theoretic, information; the development and evaluation of such descriptors is the subject of continuing study in our laboratories.[13−15]

## ACKNOWLEDGMENT

We thank the Cambridge Crystallographic Data Centre, the Science and Engineering Research Council, and Tripos Associates for funding, Helen King for helpful discussions, and the referees for their comments. This paper is a contribution from the Krebs Institute for Biomolecular Research, which is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

## REFERENCES AND NOTES

(1) Bures, M. G.; Martin, Y. C.; Willett, P. Searching Techniques for Databases of Three-Dimensional Chemical Structures. *Topics Stereochem.* **1994**, *21*, 467−511.
(2) Lewis, R. A.; Leach, A. R. Current Methods for Site-Directed Structure Generation. *J. Comput. Aid. Molec. Des.* **1994**, *8*, 467−475.
(3) Blaney, J. M.; Dixon, J. S. A Good Ligand is Hard to Find: Automated Docking Procedures. *Perspect. Drug Discov. Des.* **1993**, *1*, 301−319.
(4) Kubinyi, H., editor *3D QSAR in Drug Design. Theory, Methods and Applications;* ESCOM: Leiden, 1993.
(5) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C.F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.
(6) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111−121 (and subsequent papers in this series).
(7) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(8) Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N.; Ricketts, D. M. Comparison of Conformations of Small Molecule Structures from the Protein Data Bank with those Generated by Concord, Cobra, ChemDBS-3D and Converter and those extracted from the Cambridge Structural Database. *J. Chem. Inf. Comput Sci.* **1993**, *33*, 905−925.
(9) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(10) Pearlman, R. S. 3D Molecular Structures: Generation and Use in 3D Searching. In *3d QSAR in Drug Design. Theory, Methods and Applications;* Kubinyi, H., Ed.; ESCOM: Leiden, 1993.
(11) Petitjean, M. Applications of the Radius-Diamater Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.
(12) Bath, P. A. *Similarity Searching in the Cambridge Structural Database;* PhD Thesis; University of Sheffield, 1994.
(13) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput. Aid. Molec. Des.* **1991**, *5*, 455−474.
(14) Mitchell, E. M.; Allen, F. H.; Mitchell, G. F.; Rowland, R. S. An Integrated Approach to 2-D and 3-D Similarity Searching for the Cambridge Structural Database (CSD). In *Chemical Structures 2. The International Language of Chemistry*; Warr, W. A., Ed.; Springer Verlag: Berlin, 1993.
(15) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

CI940302I