Angular Spectroscopy: Rapid Visualization of Three-Dimensional Substructure Dissimilarity Using Valence Angle or Torsional Descriptors

Frank H. Allen*

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

Peter A. Bath and Peter Willett

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received September 8, 1994®

A simple method, termed "angular spectroscopy", is developed for the rapid visual assessment of 3D shape diversity (conformations, metal coordination geometries) that is exhibited by a specific chemical substructure as observed in a number of different crystal structures. If there are $q = 1 \rightarrow N_s$ instances of the substructure in 3D and each conformation is defined by $i = 1 \rightarrow N_t$ torsion angles, then we can calculate a set of dissimilarity coefficients D^n_{pq} that relate each of the q instances to some fixed reference conformation p. The Minkowski metric, adapted to take account of permutational isomerism and enantiomorphic inversions, is used to calculate city-block ($n = 1$) or Euclidean ($n = 2$) dissimilarities. The N_s values of D^n_{pq} provide a unidimensional representation of the multivariate parameter space and can be plotted as a simple histogram. Multiple peaks in the histogram, or torsional spectrum, indicate the presence of multiple conformations in the dataset. Dissimilarity calculations based on valence angle descriptors can be used to assess the different coordination geometries that may exist around a metal of fixed ligancy. The reduction in dimensionality of the representation, i.e., from N_t to unity, can lead to information loss and to the accidental overlap of peaks due to different conformations. To combat this problem, two simple modifications of the Minkowski metric have been investigated which generate multiplicative (M^n_{pq}) and cumulative (C^n_{pq}) dissimilarities, respectively. When all three types of coefficient are applied to a variety of example substructures, then the known conformational or configurational diversity in these datasets is clearly revealed. It is found that the multiplicative coefficient, M^n_{pq}, is generally effective in minimizing peak overlap.

INTRODUCTION

The concept of structural similarity, and its relationship to structural properties and biological activity, is of considerable importance in rational molecular design.[1] During the 1980s, algorithms were developed[1a−3] that permit nearest-neighbor or similarity searching and structure clustering to be implemented efficiently for large databases of two-dimensional (2D) structures. Algorithms operate through a direct comparison of the fixed-length bit-encoded representations of structure that also act as screens in 2D substructure search procedures. However, it is three-dimensional (3D) shape similarity that is likely to be crucial in imbuing a subset of molecules with similar biological and molecular recognition properties. The fact that subsets of database molecules have closely similar chemical characteristics is important but does not mean that their 3D shapes are necessarily closely similar as well, particularly as the number of degrees of conformational freedom increases.

The availability of large databases of experimentally determined or computed 3D structures has provoked increased interest in the quantification of 3D similarity. What is needed, obviously, are representations of 3D molecular shape that can be stored in these databases and which can be rapidly compared with the 3D shape representation of a query molecule. In two recent papers[4,5] we have discussed a variety of geometry-based representations for 3D similarity searching applied, primarily, to experimental 3D data stored in the Cambridge Structural Database (CSD)[6] and the Protein Data Bank (PDB).[7] This work is continuing, and further results will be reported in due course.

3D similarity searching, then, is a browsing mechanism by which we hope to scan those database molecules that have overall shape characteristics that are closely similar to those of an input query molecule. Hits from such a procedure may be expected to differ in both chemical constitution and in the connectivity of their atoms. In essence, it is the structural diversity that can be contained within the fixed shape (or conformation) of the query molecule that is of primary interest to the browser.

In this paper, however, we apply similarity calculations to the equally important inverse problem: to provide a rapid assessment of the conformational (shape) diversity exhibited by a 3D database molecule or substructure of some predefined (fixed) 2D connectivity. Such procedures are particularly important for a database of experimental 3D structures such as the CSD. Here, a 2D substructure search may yield tens, hundreds, or even thousands of examples of a flexible fragment for which 3D conformations have been determined in many different chemical environments. Mappings and classifications of these energetically-accessible conformations represent valuable structural knowledge, since they provide qualitative information about the potential energy hypersurface and give information that is vital in model building procedures.

The conformation of a flexible substructure is appropriately described by N_t torsion angles for each of the N_s instances of the substructure located in the CSD. This yields a

* To whom all correspondence should be addressed.

® Abstract published in *Advance ACS Abstracts,* January 15, 1995.

0095-2338/95/1635-0261$09.00/0

multivariate data matrix $\mathbf{T}(N_s,N_t)$, and, in recent years, a variety of numerical methods have been used to analyze such matrices.[8] Principal component analysis (PCA), a variable-directed technique, has been used to reduce the dimensionality of the problem (from N_t) so as to generate graphical mappings of conformational space.[9-11] These maps can reveal conformational interconversion pathways or indicate discrete clusters of substructures that have closely similar conformations. In these cases, a numerical classification of conformations can usually be achieved by cluster analyses[11-15] based on torsional dissimilarities between the substructures. The algorithms developed so far[12-14] will also select a "most representative substructure" from each subgroup that can be used in model building operations. Further, given a dataset of reasonable size, the predominance of one or a very few conformations can generally be taken to indicate that they are energetically preferred (see, e.g., ref 11) although it is not possible to obtain quantitative energy estimates from the relative populations of the subgroups.[16]

Numerical analyses of crystallographic observations are a valuable adjunct and can often be an effective alternative, to a variety of computational procedures (e.g., *ab initio*, semiempirical, or force-field calculations) that are commonly used to locate local and global conformational energy minima. These procedures are sometimes of limited applicability for computational reasons or for lack of suitable required input parameters. Even where they are applicable, it is still valuable to compare the computational results and, indeed, new experimental crystallographic results, with conformational classifications obtained from the ever-increasing pool of existing knowledge contained, for example, in databases such as the CSD.

Whilst PCA and cluster analyses are proving to be valuable aids to knowledge acquisition from the CSD, they do have some drawbacks: (a) they have significant cpu overheads, (b) they are only semiautomatic and require manual intervention and visual interpretation to obtain optimal results, and (c) they do not compare an input conformation with existing database information, as envisaged in the previous paragraph.

This paper, then, describes procedures that are designed to provide a rapid visual overview of the conformational complexity of a given chemical substructure. Torsional dissimilarity calculations that underpin the clustering algorithms[12-14] are modified and extended so as to compare the conformations of existing database substructures with that of an input "reference" or "query" conformation. The result is a simple histogram, or "angular spectrum", in which the density of peaks gives a satisfactory indication of conformational diversity. The technique is also used to reveal the different coordination geometries that may exist about a metal center of fixed ligancy, by use of dissimilarity calculations based on ligand−metal−ligand valence angles.

TORSIONAL DISSIMILARITY CALCULATIONS

The Minkowski Metric. All clustering algorithms employ some measure of the "dissimilarity" or "distance" of all pairs of objects in the input dataset. Thus, for a torsional data matrix $\mathbf{T}(N_s,N_t)$, i.e., N_s instances of the chosen substructure with the conformation of each instance defined by N_t torsional descriptors (in degrees), we may calculate a square dissimilarity matrix $\mathbf{D}(N_s,N_s)$ which is symmetric about a zero diagonal. As in the earlier clustering experiments,[12-14] we calculate the torsional dissimilarity coefficient for objects p and q, D^n_{pq}, by using the Minkowski metric[17] applied to the torsion angles τ_i ($i = 1 \rightarrow N_t$)[18]

$$D^n_{pq} = [\sum_{i=1}^{N_t} (\Delta\tau_i)^n_{pq}]^{1/n} \quad (1)$$

where

$$(\Delta\tau_i)_{pq} = |(\tau_i)_p - (\tau_i)_q|/180N_t \quad (2a)$$

or

$$(\Delta\tau_i)_{pq} = [360 - |(\tau_i)_p - (\tau_i)_q|]/180N_t \quad (2b)$$

and the values of $(\Delta\tau_i)_{pq}$ used in (1) are the *minimum* values generated by (2a) and (2b) for each $(\tau_i)_{p,q}$; this arises due to the phase restriction $-180 < \tau_i \leq 180°$. The power factor n in (1) is an integer variable usually taken as $n = 1$ (city-block metric) or $n = 2$ (Euclidean metric). The D^n_{pq} values are normalized to lie in the range 0−1 by use of the denominator $180N_t$ in (2a and 2b). For N_s substructures, there are $N_s(N_s - 1)/2$ unique D^n_{pq} values that correspond to the upper (or lower) triangle of the full dissimilarity matrix $\mathbf{D}(N_s,N_s)$.

Table 1. Atomic (and Torsional) Permutations[12] for the 3D Substructure (II) Arising from the D_{6h} Topological Symmetry of the 2D Substructure I

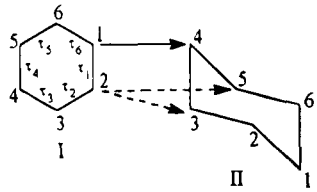
permutation no.	atomic (torsional) permutations					
1	1	2	3	4	5	6
2	2	3	4	5	6	1
3	3	4	5	6	1	2
4	4	5	6	1	2	3
5	5	6	1	2	3	4
6	6	1	2	3	4	5
7	6	5	4	3	2	1
8	5	4	3	2	1	6
9	4	3	2	1	6	5
10	3	2	1	6	5	4
11	2	1	6	5	4	3
12	1	6	5	4	3	2

Adaptations Due to Permutational Isomerism. Many substructures of conformational interest, e.g., simple rings and ring systems, acyclic side chains, etc., are small and exhibit topological symmetry. Thus, the 2D representation of cyclohexane I has plane symmetry D_{6h}. In locating subgraph isomorphisms of (I) as a substructural query with target molecules in any 2D database II then atom 1 of the query may be mapped equivalently onto any of the atoms of II in a given target. Further, if atom 1 (query) maps to atom 4 (target), then atom 2 (query) may map either to atom 3 or to atom 5 of the target. This gives rise to 12 independent, but completely equivalent, mappings of the query I to a cyclohexane ring II in a given target molecule. The substructure search routines in the CSD system[6] will arbitrarily choose just one of these mappings.

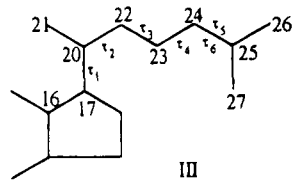
Essentially, there are 12 possible and topologically equivalent atomic enumerations for substructure (II) as located in CSD target molecules (Table 1). Any one of these enumerations will satisfy a 2D substructure search. However, in the CSD each atom is further described by 3D crystallographic coordinates which form the basis for geometrical calculations. Apart from the special case of a (D_{6h}) planar ring, the 3D symmetry of a general cyclohexane ring is lower than the 2D topological symmetry; e.g., chairs (ideally D_{3d}), boats (ideally C_{2v}), etc. Hence, in 3D the alternative atomic enumerations give rise to alternative orderings of the cyclic sequence of torsion angles that characterizes (or approximates) these symmetries: an illustrative table is given in ref 12.

Adaptations Due to Enantiomorphic Inversion. Even in the absence of permutational isomerism, each 3D substructure will have an enantiomorph of equal interest and both enantiomorphs are likely to occur in different target molecules in the CSD. Since torsion angles are enantiomorph sensitive (i.e. a change in enantiomorph reverses the signs of all torsion angles), then we must also consider all 12 enantiomorphic torsional sequences for substructure I. Thus, a total of 24 permuted/inverted torsional sequences are possible for I, any one of which may be chosen at random by the software system for inclusion in the basic torsional matrix $\mathbf{T}(N_s,N_t)$.

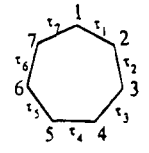
Computational Procedures. In our current code for dissimilarity calculations,[12,13] topologically symmetric substructures are handled: (a) by specifying torsional permutations explicitly as instructions to the program, (b) requiring the application (or not) of an inversion operator to each permuted sequence, and (c) calculating D^n_{pq} from eq 1 by keeping the $(\tau_i)_p$ constant and allowing the $(\tau_i)_q$ to adopt all possible permutations/inversions specified in the instruction set. For cyclohexane (I) this procedure gives 24 discrete D^n_{pq} values for a pair of general (asymmetric) 3D rings p and q. The lowest of these values is taken as the value of D^n_{pq} for inclusion in the full dissimilarity matrix, i.e., the shortest distance between rings p and q in the symmetrical conformational space.



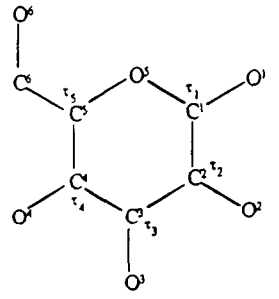
I II



III



IV



V

TORSIONAL SPECTRA BASED ON THE MINKOWSKI METRIC

Methodology. The conformational clustering algorithms of Allen et al.[12,13] make use of all of the $N_s(N_s - 1)/2$ unique torsional dissimilarities that connect the unique pairs of instances of a given substructure. However, it occurs to us that eqs 1 and 2, modified for permutation/inversion if required, can also be used to calculate dissimilarities or distances of each of these N_s instances (q: eq 1) from a fixed input conformation (p: eq 1). Further, it is to be hoped that a histogram of these N_s dissimilarity values will give a clear indication of the conformational diversity of the dataset.

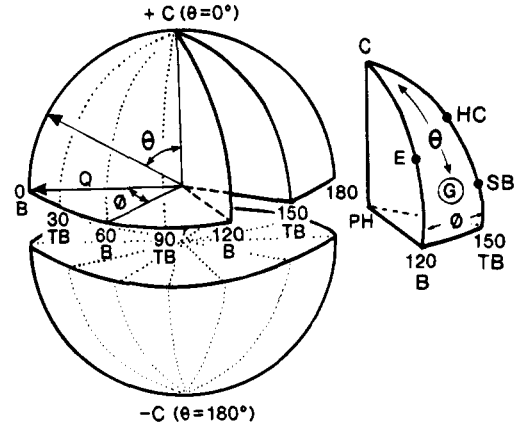


Figure 1. Representation of conformational space for six-membered rings using the spherical polar coordinate set Q,θ,ϕ.[19,20] The special symmetric conformations indicated are chair (C), boat (B), envelope (E), half-chair (HC), screw-boat (or 1,3-diplanar) (SB), twist-boat (TB), and phenyl (PH). G is any general conformation, and the isolated segment (1/24th of the sphere) is the asymmetric unit.

This methodology is illustrated in Figure 1 by reference to the three-dimensional conformational space for six-membered rings.[19,20] Here each conformation, normally represented by a sequence of $N_t = 6$ torsion angles, may also be represented[19] in a space of minimum dimensionality (3) by use of spherical polar coordinates Q (the total puckering amplitude), θ and ϕ. These coordinates are simply related to the three degrees of freedom generated by the Cremer–Pople[20] analysis of ring pucker. The conformation of N_s instances of a six-membered ring can be represented by N_s concentric spheres of radii $Q_1 \rightarrow Q_{Ns}$ since each ring will have a different puckering amplitude.

Special symmetric conformations[21] exist on these spheres as indicated in Figure 1. The chair form occupies the north pole ($\theta = 0°$) with its enantiomer at the south pole ($\theta = 180°$). The equator ($\theta = 90°$) is occupied by six permutationally equivalent boat forms at $\phi = 0, 60,...300°$, separated from each other by six twist-boats at $\phi = 30, 90,...330°$. The envelope (or half boat) conformation is intermediate between the boat and the chair: six equivalent conformations exist in the northern hemisphere at $\phi = 0, 60,...300°$ and at $\theta \approx 55°$, their enantiomers exist in the southern hemisphere at $\theta \approx 125°$. Similarly, six half-chair and six screw-boat (1,3-diplanar) conformations exist in the northern hemisphere, at $\theta \approx 50°$ and $\theta \approx 68°$, respectively, on the ϕ-arcs connecting chairs and twist-boats ($\phi = 30, 90,...330°$). The enantiomers are in the southern hemisphere at $\theta \approx 130°$ (half-chairs) and $\theta \approx 112°$ (screw-boats). The planar phenyl ring occupies the center of the family of spheres with $Q = 0$ and θ,ϕ indeterminate.

Any general conformation, G in Figure 1, for example a conformation which is intermediate between an envelope (E) and a half-chair (HC) and which has $\theta \approx 45°$, will have 12 permutational equivalents (at $\phi = 15, 45, 75,...345°$) in the northern hemisphere and its 12 enantiomers at $\theta \approx 135°$ (and at the same ϕ values) in the southern hemisphere. Thus, an

asymmetric unit of this conformational space is 1/24th of the complete sphere. Specifically (Figure 1), it is any one of the 12 unique 30° ϕ segments of the northern or southern hemispheres. This definition of an asymmetric unit corresponds exactly to the 24 possible torsional permutation/inversion operations described earlier.

The torsional dissimilarity coefficient D^n_{pq} (eq 1) for a pair of six-membered rings represents the normalized distance between their conformations. However, if we hold the conformation of ring p static, for example, at the torsional sequence +60, −60, +60, −60, +60, −60° that corresponds to an idealized chair, then the N_s unique values of D^n_{pq} will represent the closest normalized distance of each of the $q(= 1 \rightarrow N_s)$ observed conformations from the north pole $(+c)$ of Figure 1. Thus, a histogram of the D^n_{pq} values should show a number of peaks ranging from zero (chair conformations) to a maximum value that corresponds to boat and twist-boat conformers. Peaks corresponding to phenyl rings and the intermediate half-chair, envelope, and screw-boat forms should fall between these two limiting values. While it may not be possible to separate all conformations that may be present in a given dataset, it is to be expected that the histogram will give a valuable insight into the conformational diversity that is contained within that dataset. We refer to these histograms as "torsional spectra" (and analogously so for the "valence angle spectra" discussed later in this paper).

Results for Six-Membered Rings. The trial dataset of 222 six-membered rings used previously[12] in the development of cluster analysis algorithms was used here to generate initial torsional spectra. Cyclohexane itself occurs predominantly in the chair form, but the original dataset was chosen so as to embody conformational diversity through the deliberate inclusion of phenyl rings, cyclohexene rings, and cyclohexane rings in constrained (norbornane) environments. Symmetry-modified[13] Jarvis−Patrick clustering[22] of this dataset revealed a mixture of chairs (59), boats (63), phenyl rings (35), half-chairs (29), screw boats (9), and twist-boats (9). The remaining 18 rings were highly distorted variants, primarily of the chair and boat forms.

Torsional spectra, calculated using the ideal (D_{3d}) chair form as the invariant "origin" conformation and torsional permutational sequences that are identical to the atomic permutations of Table 1, are shown in Figure 2a (D^1_{pq} = city-block metric) and Figure 2b (D^2_{pq}: Euclidean metric). Both show a large peak corresponding to chair conformations close to $D^n_{pq} = 0$. The D^1_{pq} plot of Figure 2a then shows three additional major peaks: a "half-chair" peak with a small shoulder corresponding to screw-boats and a large doublet due to the boat and phenyl rings. In Figure 2b, however, six major conformational variants are clearly observed in the D^2_{pq} spectrum. Not only the small population of screw-boats is now differentiated from the half-chairs, but also the boat peak is split into a doublet in agreement with the earlier detailed clustering experiments.[12,13] These initial results, based on the simple Minkowski metric of eq 1, thus provide a clear indication of the conformational diversity of this particular dataset. Obviously, ideal torsion angles for any other known conformation, e.g., a boat form, may be used to define the invariant origin and generate a similar display of conformational diversity.

Results for Steroidal C_{17} Side Chains. The side chain (III) is typical of cholesterol and related steroids. Recently,[13] a symmetry-modified single linkage clustering was performed for 109 instances of III derived from 77 structures retrieved from the CSD. The six torsion angles, $\tau_1-\tau_6$, illustrated in III were used as conformational descriptors, and the torsional permutations (1 2 3 4 5 6;1 2 3 4 6 5) must be considered in the analysis. For this work, we have repeated the clustering experiment on an enlarged dataset of 151 observations obtained from 111 CSD entries. Idealized torsion angles for the seven major clusters (those with a population, N_p, greater than two observations) are given in Table 2.

Table 2. Idealized Torsion Angles (in deg) for the Six Major Conformational Clusters (Populations, N_p, Greater Than Two Observations) for the Steroidal Side Chain III

cluster	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
A	112	180	180	180	180	180	60
B	7	180	−60	180	180	180	60
C	6	180	180	180	60	180	60
D	5	180	180	60	180	180	60
E	5	180	60	180	180	180	60
F	4	180	60	180	60	180	60
G	3	180	-60	180	60	180	60

As expected, the major conformation for this acyclic chain is the fully extended form A (Table 1). Other variants involve either one (B, C, D, and E) or two (F and G) synclinal or anticlinal (±60°) relationships in the $\tau_1-\tau_4$ sequence. If we use the idealized torsion angles for conformer A as the invariant origin (p) for D^n_{pq} calculations via eq 1, then it is obvious, by inspection, that there will be three dominant peaks in the D^1_{pq} or D^2_{pq} histograms. These encompass conformers A, B, C, D, and E, and F and G, respectively, as shown in Figure 3a,b, where the peaks have populations that closely mirror the population sums from Table 2. Since this conformational overlap problem is likely to be typical in acyclic (and some cyclic) systems, we have sought simple modifications to the basic Minkowski metric (eq 1) that are designed to resolve these overlap situations.

MODIFIED MINKOWSKI METRICS

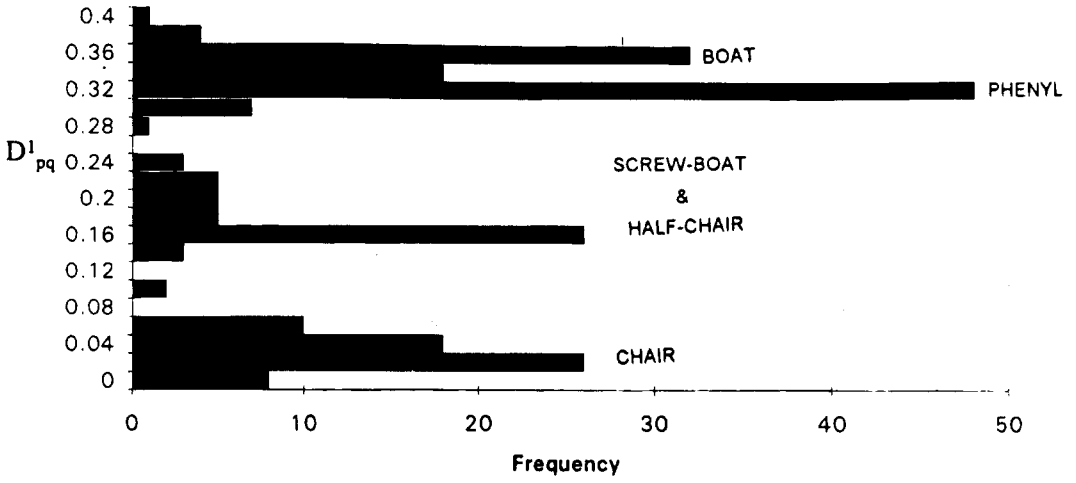
Multiplicative and Cumulative Dissimilarity Calculations. Each of the conformers B−E (Table 2) differs from the chosen origin conformation A by having a single clinal (±60°) relationship along the chain defined by $\tau_1-\tau_4$, although these relationships occur at different bond locations along the side chain (III). Thus, under eq 1, all values of D^n_{Aq} (q = B, C, D, and E) will be approximately equal. Similar equalities occur for conformers F and G (Table 2) which both exhibit two separate and positionally distinct clinal relationships. To resolve these problems, we have sought simple modifications to the Minkowski metric of eq 1 that take account of the *position* of a given torsion angle in an ordered sequence. If i is the position of a given torsional descriptor in the sequence $i = 1 \rightarrow N_t$, then we may modify eq 1 by multiplying each individual difference $(\Delta\tau_i)_{pq}$ by its position indicator, thus

$$M^n_{pq} = [\sum_{i=1}^{N_t} \{i(\Delta\tau_i)_{pq}\}^n]^{1/n} \quad (3)$$

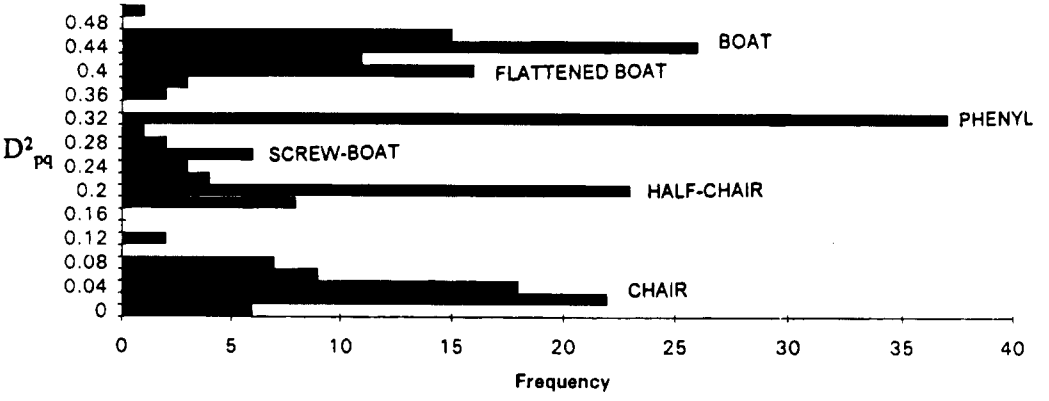
where M denotes the multiplicative nature of the metric. Normalized $\Delta\tau_i$ values are now given by modification of, e.g., eq 2a to

$$(\Delta\tau_i)_{pq} = |(\tau_i)_p - (\tau_i)_q|/180\sum_{i=1}^{N_t} i \quad (4)$$

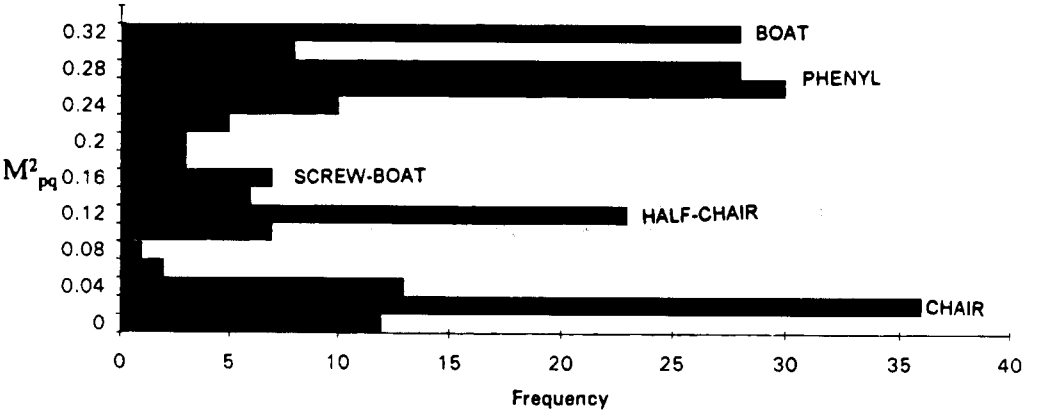
with analogous treatment of eq 2b.



(a)



(b)



(c)

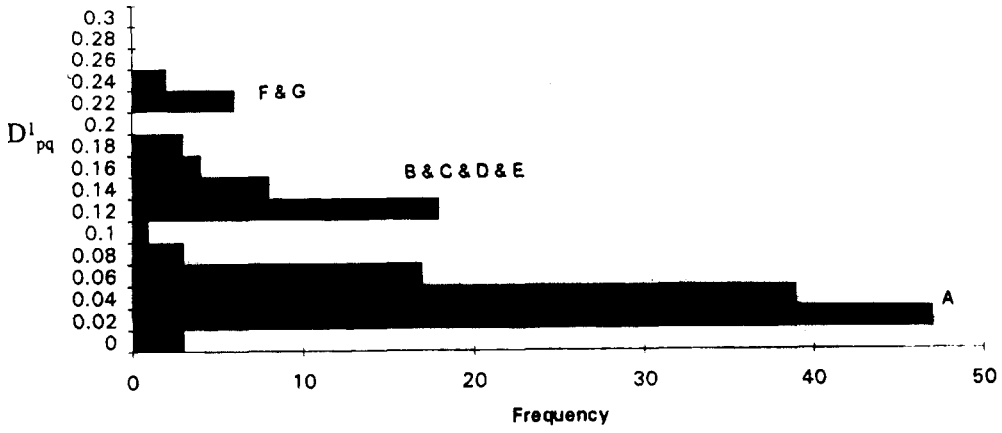
Figure 2. Torsional spectra for a dataset of 222 six-membered rings.[12] The dissimilarities plotted are (a) D^1_{pq}, (b) D^2_{pq}, and (c) M^2_{pq}.

A second metrical modification involves alteration of the basic torsional sequences $(\tau_i)_p,(\tau_i)_q$ so that each individual τ_i value is replaced by the cumulative sum (c_i) of the absolute values $|\tau_i|$ up to and including that position in the sequence, thus

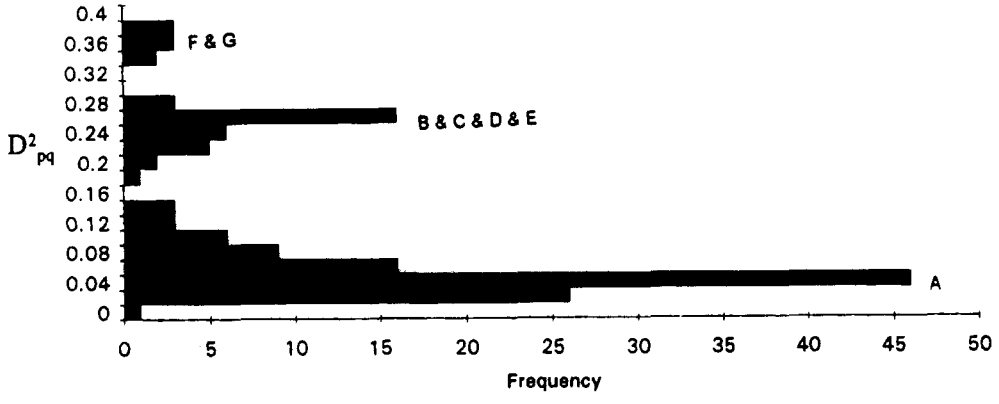
$$(c_i)_p = \sum^{i} |(\tau_i)_p| \qquad (5)$$

i.e., if the $(\tau_i)_p$ are 180, 180, 180, 180, 180, 60°, then the $(c_i)_p$ become 180, 360, 540, 720, 900, 960°. The cumulative dissimilarity coefficient, C^n_{pq}, is then calculated via eq 1 applied to the $(c_i)_{p,q}$ sequences and normalization is achieved using eq 4a applied to the $(c_i)_{p,q}$ sequences.

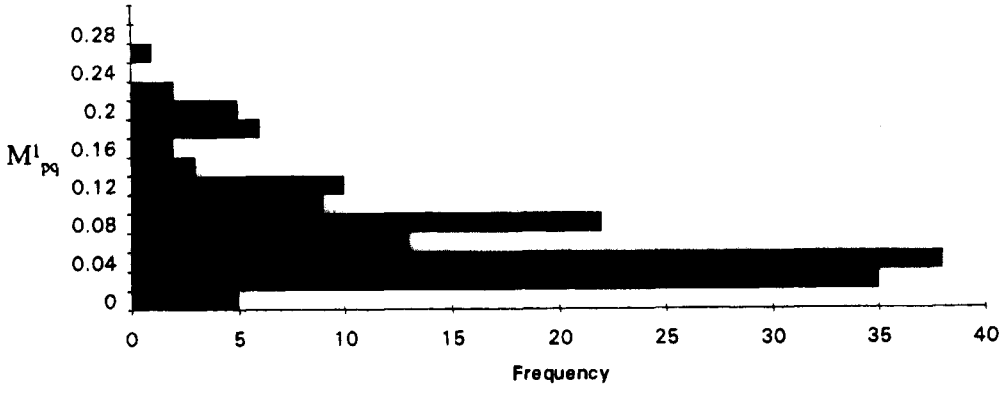
Multiplicative and Cumulative Dissimilarities for the Steroid C_{17} Side Chain. Histograms of the multiplicative dissimilarity coefficients M^1_{pq} (city block) and M^2_{pq} (Euclidean) for the steroid dataset are shown in Figure 3c,d. Analogous plots of the cumulative coefficients C^1_{pq} and C^2_{pq} are shown in Figure 3e,f. Idealized torsion angles for conformer A were used as the invariant origin in all calculations. In both cases, the city-block metrics (Figure 3c,e) are less well resolved than their Euclidean counterparts (Figure 3d,f). The M^2_{pq} plot shows six distinct peaks, that are due to conformers (Table 2): A (fully extended), B + E ($\tau_2 = \pm 60°$), C ($\tau_4 = 60°$), D ($\tau_3 = 60°$), and F + G ($\tau_2 =$


D^1_{pq}
0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
F & G
B & C & D & E
A
0
10
20
30
40
50
Frequency

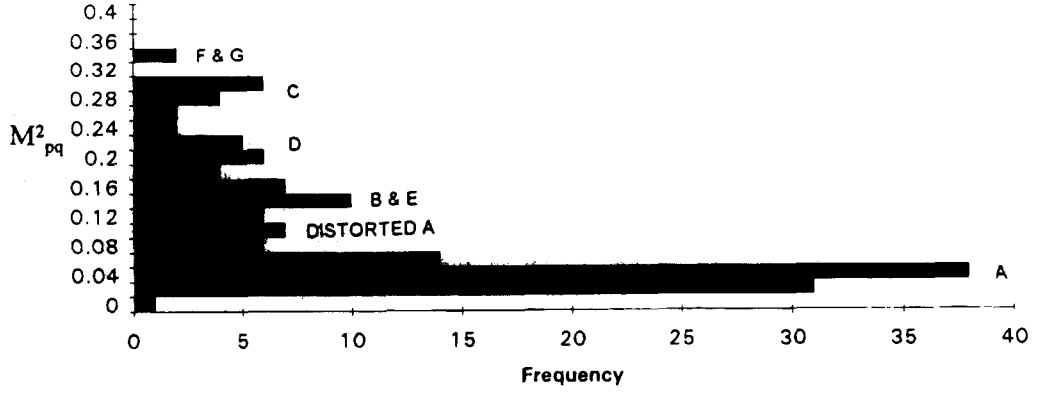
(a)


D^2_{pq}
0.4
0.36
0.32
0.28
0.24
0.2
0.16
0.12
0.08
0.04
0
F & G
B & C & D & E
A
0
5
10
15
20
25
30
35
40
45
50
Frequency

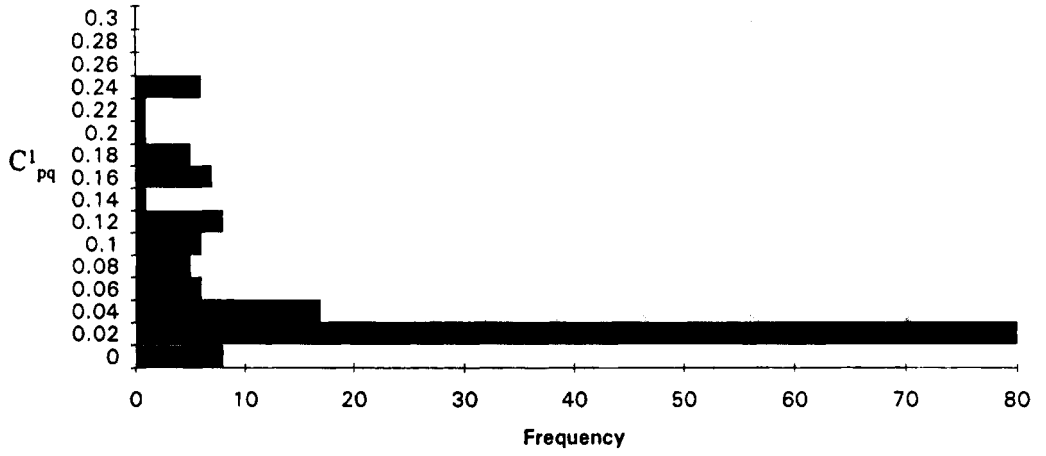
(b)


M^1_{pq}
0.28
0.24
0.2
0.16
0.12
0.08
0.04
0
0
5
10
15
20
25
30
35
40
Frequency

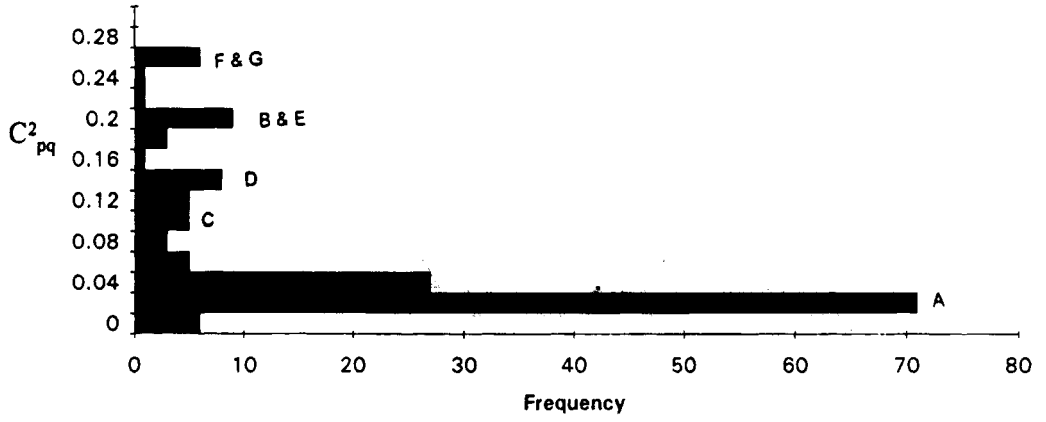
(c)


M^2_{pq}
0.4
0.36
0.32
0.28
0.24
0.2
0.16
0.12
0.08
0.04
0
F & G
C
D
B & E
DISTORTED A
A
0
5
10
15
20
25
30
35
40
Frequency

(d)



(e)



(f)

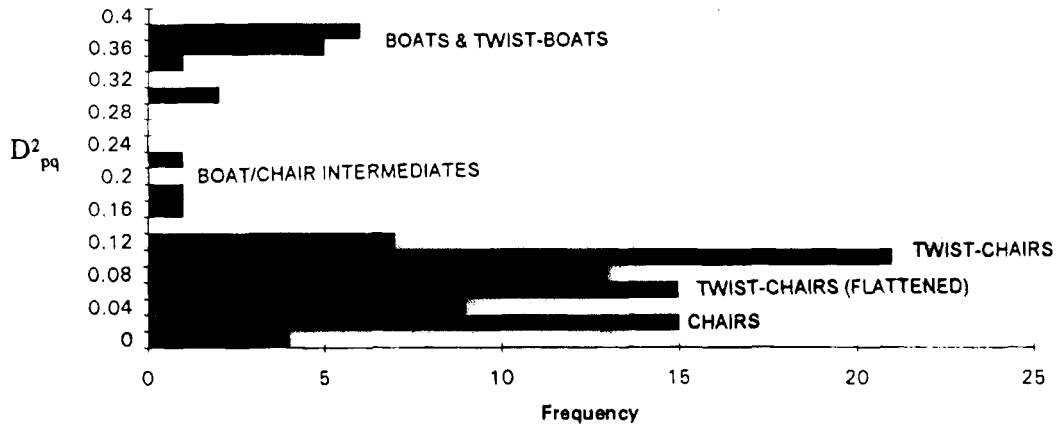
Figure 3. Torsional spectra for 151 instances of the steroidal C_{17} side chain (III). The dissimilarities plotted are (a) D^1_{pq}, (b) D^2_{pq}, (c) M^1_{pq}, (d) M^2_{pq}, (e) C^1_{pq}, and (f) C^2_{pq}.

$\tau_4 = \pm 60°$). The sixth peak, at $M^2_{pq} \approx 0.08$, is due to resolution of some distorted variants of conformer A in which there is rotation of the terminal isopropyl group. This effect is not observed in the C^2_{pq} plot (Figure 3f), and here conformers C and D form a broad singlet. Nevertheless, both the multiplicative and cumulative Euclidean coefficients provide an improved unidimensional representation of the steroid data set. Further modifications to the calculation of dissimilarities to take account of the signs of the torsion angles are being considered.

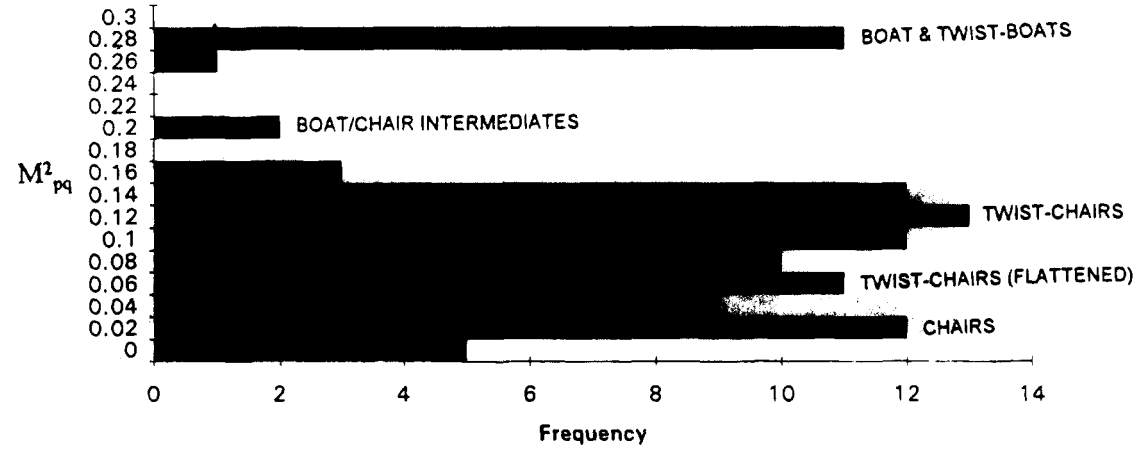
Multiplicative and Cumulative Dissimilarities for the Six-Membered Rings. Figure 2c shows the torsional spectrum based on M^2_{pq} values computed for the 222 substructures in the six-membered ring dataset.[12] The fixed origin, a perfect (D_{3d}) chair with ±60° torsion angles, is identical to that used to compute the D^1_{pq} and D^2_{pq} values for Figure 2a,b. The cumulative C_{pq} dissimilarity plots (not shown) are disappointing and give only three discrete peaks by comparison with the six peaks of the D^2_{pq} plot (Figure 2b). However, the M^2_{pq} plot of Figure 2c exhibits five distinct peaks. There is a clear separation of the half-chair and screw-boat conformers in this linear representation, but the boat doublet of Figure 2b is not observed at the chosen resolution.

Torsional Spectra for Cycloheptane Rings. A recent cluster analysis[11] of the 101 nonbridged cycloheptane rings (IV) located in the January 1992 Version of the CSD revealed the presence of chairs (30), twist-chairs (38), flattened twist-chairs (10), and distorted boats (10). The remaining, unclustered rings were twist-boats, boat-chair intermediates, and a small number of highly distorted twist-chairs. In Figure 4a,b we show torsional spectra based (a) on D^2_{pq} values (eq 1) and (b) on M^2_{pq} values (eq 3). Torsion angles for the normal chair conformation were used as the fixed origin sequence, $(\tau_i)_p$, and the 14 torsional permutations needed here are analogous to those for six-membered rings (Table 1). Both plots show the principal chair/twist-chair/boat conformers as clearly resolved peaks. Further, both plots are able to discriminate further the flattened twist-chairs from their more highly puckered counterparts and mimic very clearly the results of the full cluster analysis.

Stereochemical Partitioning of Hexopyranose Sugars. As a final example of torsional spectroscopy, we illustrate its use in the stereochemical partitioning of a dataset of 249 instances of the hexopyranose substructure (V). The problem here, of course, is that the 2D search fragment (V) has five stereocenters, at each of the ring C atoms, and can have $2^5 = 32$ possible stereoisomers in three dimensions. Not all of these are likely to be present in the CSD and it is of interest to identify rapidly the major stereoisomers that are present in the available crystallographic data. In this case we have used improper torsion angles (projected valence angles)[23] to define the configuration at each stereocenter:[15] if we look along the bond from each asymmetric carbon toward its exocyclic O or C substituent, then the improper torsions $(\tau_1 - \tau_5)$ are the projections of the intraannular



(a)



(b)

Figure 4. Torsional spectra for 101 seven-membered rings:[11] (a) D^2_{pq} values and (b) M^2_{pq} values.

valence angles at C_1-C_5 onto the plane that is perpendicular to the C_n-O,C substituent vector. Individual angles will be ca. ±120°, and, for the five stereocenters, there are 32 possible sequences of $\tau_1-\tau_5$, i.e., 32 possible sign permutations, that identify the complete set of stereoisomers.[15]

These sequences of improper torsions have recently[15] formed the basis of a complete Jarvis–Patrick[22] clustering designed to identify the stereochemical diversity in the 249 instances of V retrieved from the CSD. A total of 14 stereochemical partitions were identified by the clustering procedure, and these results are summarized in Table 3.

In calculating dissimilarities, we note that substructure (V) is asymmetric. A unique atomic numbering scheme can be assigned, and no permutation of the $(\tau_i)_q$ is required. Further, the 3D coordinates should report the correct absolute stereochemistry (but see ref 15), hence inversions should not be considered in this case.

Torsional spectra based on D^n_{pq} and C^n_{pq} values (eqs 1 and 5, $(\tau_i)_p$ = 120, 120, 120, 120, 120) are disappointing, showing a maximum of four peaks. However, the spectra of M^n_{pq} values (eq 3) are again well resolved, showing nine peaks for $n = 2$ and 11 peaks for $n = 1$. The M^1_{pq} spectrum is shown in Figure 5 with peaks identified by their cluster numbers from Table 3. Only three peaks represent overlaps of stereoisomers, with both major and minor partitions of the dataset clearly demarcated.

Table 3. The 14 Stereochemical Partitions of the 249 Hexopyranose Substructures (V) Obtained by Jarvis–Patrick Clusterings of Improper Torsion Angles[15] [a].

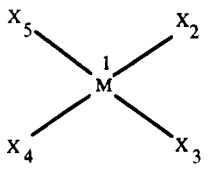
cluster	N_p	signs of $\tau_1-\tau_5$	chemical name
1	104	+ − + − +	β-D-glucose
2	74	− − + − +	α-D-glucose
3	25	+ − + + +	β-D-galactose
4	19	− − + + +	α-D-galactose
5	8	− + + − +	α-D-mannose
6	3	− + + + +	α-D-talose
7	3	− + − − +	α-D-altrose
8	3	+ − − + +	β-D-gulose
9	3	+ + + − +	β-D-mannose
10	2	− − − + +	δ-D-gulose
11	2	− − − − +	α-D-allose
12	1	+ − − − +	β-D-allose
13	1	− + − + +	α-D-idose
14	1	+ − + − −	α-L-idose

[a] Only the signs of these angles are given (see text): their absolute numerical values are close to 120° in all cases. N_p is the population of each cluster.

VALENCE ANGLE SPECTRA

Just as torsion angles are the natural geometrical descriptors of conformation, so the $n(n-1)/2$ valence angles subtended at a metal atom of ligancy n are the natural description of the shape of the metal coordination sphere. For a given coordination number n, a small number of geometries is usually possible, and this number increases for the higher coordination numbers. It is obviously useful to

Table 4. Valence Angle Permutations for Tetracoordinate Metal Centers



end	2	2	2	3	3	4
apex	1	1	1	1	1	1
end	3	4	5	4	5	5
θ_i, i =	1	1	3	4	5	6
perm	i_1	i_2	i_3	i_4	i_5	i_6
1	1	2	3	4	5	6
2	1	3	2	5	4	6
3	2	3	1	6	4	5
4	2	1	3	4	6	5
5	3	1	2	5	6	4
6	3	2	1	6	5	4
7	1	4	5	2	3	6
8	1	5	4	3	2	6
9	4	5	1	6	2	3
10	4	1	5	2	6	3
11	5	1	4	3	6	2
12	5	4	1	6	3	2
13	2	4	6	1	3	5
14	2	6	4	3	1	5
15	4	2	6	1	5	3
16	4	6	2	5	1	3
17	6	2	4	3	5	1
18	6	4	2	5	3	1
19	3	5	6	1	2	4
20	3	6	5	2	1	4
21	5	3	6	1	4	2
22	5	6	3	4	1	2
23	6	3	5	2	4	1
24	6	5	3	4	2	1

obtain a rapid overview of coordination sphere diversity prior to some in-depth analysis, e.g., of interconversion pathways[24] or pseudorotation mechanisms.[25]

We may calculate angular dissimilarities D^n_{pq}, M^n_{pq}, and C^n_{pq} by use of eqs 1, 3, and 5, applied to sequences of valence angles $(\theta_i)_p$,$(\theta_i)_q$. The fixed sequence $(\theta_i)_p$ should correspond to the values for one of the common archetypal geometries for the relevant ligancy, e.g., 120, 120, and 120° for trigonal planar three-coordination. The phase problem that gives rise to the alternative eqs 2a,b does not exist for valence angles, and eq 2a only need be used with the θ_i. Further, while we need not consider enantiomorphic coordination spheres, the problems caused by permutational isomerism will occur in many cases, since definition of ligand atoms will frequently be generalized in the 2D search query.

As an example of valence angle spectroscopy, we have retrieved 435 examples of tetracoordinate Cu atoms from the CSD, by locating instances where the four ligand atoms (X in Table 4) are any non-H atom and, hence, are topologically equivalent. There are 24 possible permutational equivalents of the atom labels 2, 3, 4, and 5 assigned to the ligand atoms and 24 possible permutations of the six valence angles subtended by pairs of ligand atoms at the central metal Cu(1). These permutations are shown in Table 4 and must be applied to the observed angle sequences $(\theta_i)_q$ so as to locate the minimum value of D^n_{pq}, M^n_{pq}, or C^n_{pq} with respect to the fixed origin sequence $(\theta_i)_p$. As with the torsional dissimilarities discussed above, it is this minimum value for each crystallographically observed coordination sphere (q) that is retained for inclusion in the angular spectrum.

The valence angle spectra given by the D^2_{pq} and M^2_{pq} values obtained for the 435 instances of the tetracoordinate copper substructure are illustrated in Figure 6a,b. The fixed origin (p) was defined by $(\theta_i)_p$ = 90.0, 180.0, 90.0, 90.0, 180.0, and 90.0 (square planar coordination) for the valence angle sequence 213, 214, 215, 314, 315, and 415 in the basic atomic enumeration at the head of Table 4. Both spectra show two major peaks, as expected, corresponding to the square planar and tetrahedral arrangements of atoms 2−4. However, both spectra show significant density between the two major peaks, indicative of an interconversion pathway: square planar ↔ tetrahedral. The small peak on this pathway corresponds to an intermediate form best described as a trigonal pyramid with a Y-shaped basal plane for which angles of ca. 135, 135, and 90° are observed.

DISCUSSION

The angular spectra illustrated in this paper can be simply and rapidly computed from crystallographic or other obser-

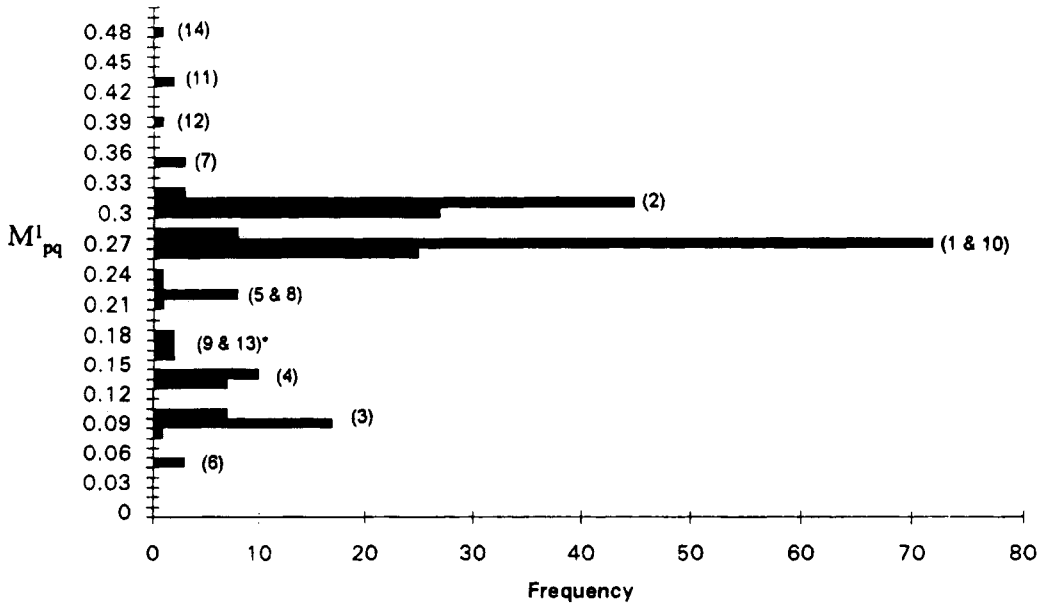
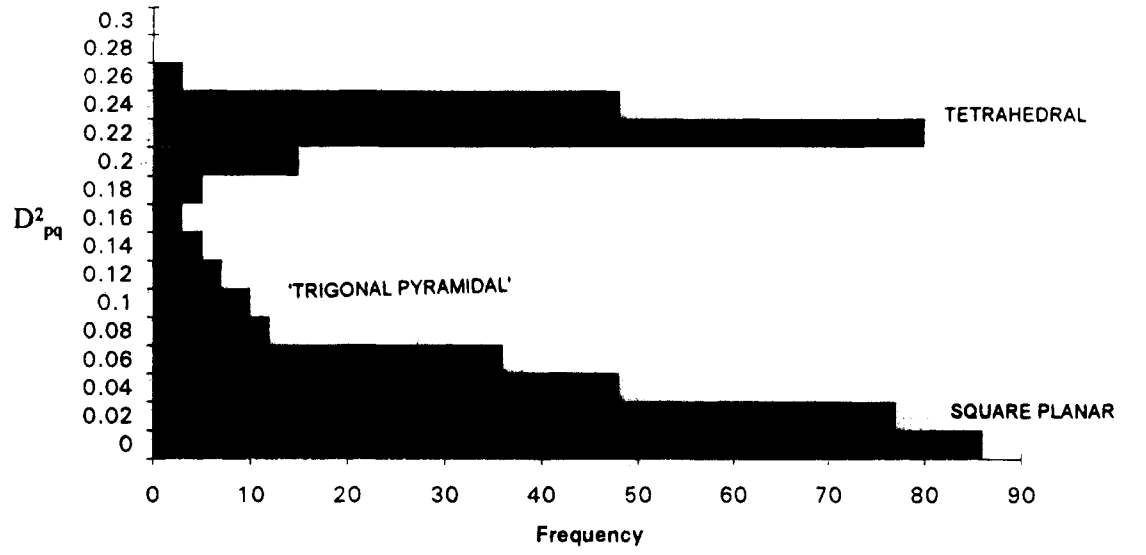
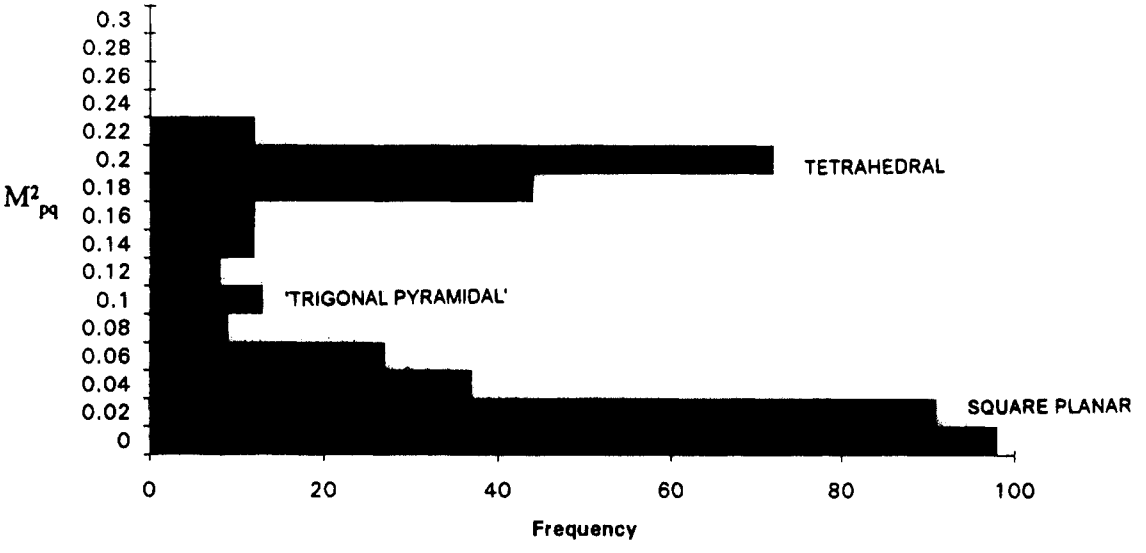


Figure 5. Torsional spectrum (M^1_{pq} values) for 249 hexopyranose sugars.[15]



(a)



(b)

Figure 6. Valence angle spectra for 435 instances of the CuX_4 coordination sphere: (a) D^2_{pq} values and (b) M^2_{pq} values.

vations of 3D molecular structure. These dissimilarity plots provide an effective visualization of the diversity of 3D shapes exhibited by a specific chemical substructure taken from a wide variety of molecular environments. We note, in any case, that dissimilarity calculations using eq 1 are an essential precursor to more extensive multivariate analyses, such as the clustering experiments reported elsewhere.[12,13] Thus, in practical applications in 3D data analysis, we envisage that angular spectra might be used routinely to assess conformational or configurational diversity before invoking the more cpu-intensive multivariate procedures. One advantage of this two-step approach to shape analysis is that *all* substructural objects are represented in the spectrum, and their linear relationships to other objects can be visualized. In a full cluster analysis, some objects may be assigned to low-occupancy classes whose relationship to each other, and to the more dominant high-occupancy classes, may not be immediately obvious in the results of the full numerical analysis.

We recognize, however, that the dissimilarities D^n_{pq}, M^n_{pq}, and C^n_{pq} computed with object p as a fixed origin are a small subset of the full dissimilarity matrix used in a cluster analysis, since the full matrix is built by allowing every object in turn to act as the origin. Thus, the dissimilarities used to generate angular spectra are linear representations of multidimensional data, and it is not surprising that this dimension reduction results in occasional coalescences of different conformations at nearly identical dissimilarity values. Nevertheless, given the original dimensionalities (5–7) of the original angular datasets, it is encouraging that the unidimensional spectra obtained for the chemically diverse examples are so well resolved in Figures 2–6. This good resolution is consistently obtained (at least for the datasets examined so far) by use of the multiplicative modification (eq 3) of the Minkowski metric. This result encourages us to try M^n_{pq} values, rather than the current[12,13] D^n_{pq} values, as the dissimilarity basis for subsequent clustering operations.

There are a number of ways in which an origin conformation or configuration can be chosen. For well-studied substructures, such as those used in this paper, it is possible to use the standard angular parameters of any of the well-known archetypal forms. These are commonly available in

textbooks or in the literature and are well accepted values derived from experimental and/or calculational methods. In other cases, we may have *de novo* experimental or calculational results, e.g., from a new crystal structure or from an original force-field or *ab initio* study. It would then be appropriate to use these novel results to provide the fixed origin, so as to relate the new data to existing results stored in a database such as the CSD. This procedure is even more appropriate if we note that conformations or configurations that are closely similar to that of the origin will *always* generate a peak that is close to zero in the spectrum. This peak can never be obscured by the accidental peak coalescences noted above, which can only occur when two or more different 3D shapes are at very similar and nonzero distances from the chosen origin.

In its present implementation, the technique has primarily been used for analysis of angular datasets retrieved for crystallographic observations stored in the CSD. Here it is usually a simple matter to scan a listing of these parameters and choose a representative angular sequence as the spectral origin. Usually the listing will show that the dataset is comprised of just a few common and different conformations or configurations. It is then appropriate and simple to select the representative of one of these common archetypes as the origin. Given that peak coalescence can occur, as noted above, then it may be advisable to generate spectra using different origins (since computational overheads are small), so as to improve the visualization of the conformational or configurational diversity present in a given dataset.

The one problem with the routine application of the methods described in this paper is that permutational isomerism must be systematically treated. In our present implementation of eqs 1, 3, and 5, within a developmental version of the CSD program GSTAT,[6] the permutation sequences of the torsion or valence angles must be entered manually by the investigator. Care is required to generate geometrical permutation lists, such as that illustrated in Table 4, from the alternative atomic enumerations of the substructure that are possible due to topological symmetry. Current developments in the CSD software system are aimed at automating the determination of atomic and geometrical permutational symmetry for a variety of numerical analysis applications.

ACKNOWLEDGMENT

We thank the Science and Engineering Research Council, U.K. for a Research Studentship (CASE award) to P.A.B.

REFERENCES AND NOTES

(1) (a) Willett, P. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press: Letchworth, UK, 1987. (b) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest-neighbour searching in an on-line chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *36*, 36−41.
(3) Willett, P.; Winterman, V.; Bawden, D. Implementation of non-hierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *36*, 109−118.
(4) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirrette, A. R.; Rice, D. W.; Thorner, D. A.; Wild, D. J.; Willett, P.; Allen, F. H.; Taylor, R. Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 617−630.
(5) 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.
(6) 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.
(7) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. The Protein Data Bank. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography, Chester, UK, 1987.
(8) Taylor, R.; Allen, F. H. Statistical and numerical methods of data analysis. In *Structure Correlation*; Bürgi, H.-B., Dunitz, J. D., Eds.; VCH Publishers: Weinheim, Germany, 1994.
(9) Murray-Rust, P.; Motherwell, W. D. S. Computer retrieval and analysis of molecular geometry: 3. Geometry of the β-1′-aminoribofuranoside fragment. *Acta Crystallogr.* **1978**, *B34*, 2534−2546.
(10) Allen, F. H.; Doyle, M. J.; Auf der Heyde, T. P. E. Automated conformational analysis from crystallographic data. 6. Principal component analysis for n-membered carbocyclic rings (n = 4−6): symmetry considerations and correlations with ring-puckering parameters. *Acta Crystallogr.* **1991**, *B47*, 412−424.
(11) Allen, F. H.; Howard, J. A. K.; Pitchford, N. A. Symmetry-modified conformational mapping of the medium rings. 1. Cycloheptane. *Acta Crystallogr.* **1993**, *B49*, 910−928.
(12) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated conformational analysis from crystallographic data. 1. A symmetry-modified single-linkage clustering algorithm for three-dimensional pattern recognition. *Acta Crystallogr.* **1991**, *B47*, 29−40.
(13) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated conformational analysis from crystallographic data. 2. Symmetry-modified Jarvis−Patrick and complete-linkage clustering algorithms for three-dimensional pattern recognition. *Acta Crystallogr.* **1991**, *B47*, 41−49.
(14) Allen, F. H.; Doyle, M. J.; Taylor, R. Automated conformational analysis from crystallographic data. 3. Three-dimensional pattern recognition within the Cambridge Structural Database System: Implementation and practical examples. *Acta Crystallogr.* **1991**, *B47*, 50−61.
(15) Allen, F. H.; Fortier, S. Stereochemical and conformational classification of the hexopyranose sugars using numerical clustering methods. *Acta Crystallogr.* **1993**, *B49*, 1021−1031.
(16) Bürgi, H.-B.; Dunitz, J. D. Can statistical analysis of structural parameters from different structural environments lead to quantitative energy relationships? *Acta Crystallogr.* **1988**, *B44*, 445−448.
(17) See, e.g.: Everitt, B. *Cluster Analysis*, 2nd ed.; Halstead Heinemann: London, 1980.
(18) Taylor, R. The Cambridge Structural Database in molecular graphics: techniques for the rapid identification of conformational minima. *J. Mol. Graphics* **1986**, *4*, 123−131.
(19) Pickett, H. M.; Strauss, H. L. Conformational structure, energy and inversion rates of cyclohexane and some related oxanes. *J. Am. Chem. Soc.* **1970**, *92*, 7281−7290.
(20) Cremer, D.; Pople, J. A. A general definition of ring puckering coordinates. *J. Am. Chem. Soc.* **1975**, *97*, 1354−1358.
(21) See, e.g.: Allen, F. H.; Taylor, R. Automated conformational analysis from crystallographic data. 5. Recognition of special positions in conformational space in symmetry-modified clustering algorithms. *Acta Crystallogr.* **1991**, *B47*, 404−412.
(22) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Computing* **1973**, *C22*, 1025−1034.
(23) Allen, F. H.; Rogers, D. The use of a connectivity or bonding array in molecular geometry calculations. *Acta Crystallogr.* **1969**, *B25*, 1326−1330.
(24) Klebe, G.; Weber, F. Description of coordination geometry in tetrahedral metal complexes by symmetry-deformation coordinations. *Acta Crystallogr.* **1994**, *B49*, 50−59.
(25) Auf der Heyde, T. P. E.; Bürgi, H.-B. Molecular geometry of d^8 five-coordination. 3. Factor analysis, static deformations and reaction coordinates. *Inorg. Chem.* **1989**, *28*, 3982−3989.

CI9401027