space group $P2_1/c$. This is evidently because the CHD molecule, which possesses mirror symmetry when atom C(5) is disordered, does not lie on a mirror plane in the CHD1 structure. Consequently, when C(5) becomes ordered, no crystallographic symmetry element is destroyed. Also, as the two sites C(5*A*) and C(5*B*) are not crystallographically equivalent, the transformation proceeds in one direction and the crystal is neither twinned nor cracked.

The transfer of the enolic proton appears not to be directly connected with the ordering of C(5), but is one of the consequences which follows the ordering of C(5) in the CHD1 structure. It was suggested (Katrusiak, 1990*a*) that the transfer of H(1) can be caused by electrostatic interactions between the molecules of neighbouring chains. The shift of the chains proceeds in the direction which in the CHD1 structure would significantly decrease the distance between O(1) and O(2^{ii}) (see Fig. 4 and Table 5), the two atoms with the largest negative atomic charges in the molecule (Katrusiak, 1990*a*). A larger distance between O(1) and O(2^{ii}) can be retained if the molecule is rotated and the values of angle ρ changed from positive to negative (see Fig. 4). This rotation, however, causes strains in the hydrogen bond, which can be released by the observed change in the H-atom position. In CHD1 the carbonyl group is closer to the central line of the chain than the hydroxyl group: such a small inclination of the molecules depends on the hydrogen-bond geometry and was also observed in the crystal structure of 1,3-cyclopentanedione (Katrusiak, 1990*b*). After the rotation, the previous inclination of the CHD molecule to the central line of the chain can be restored as a consequence of the transfer of the enolic H atom to O(1) from O(2^{i}) (Table 5) in the hydrogen bond.

This study was partly supported by the Dean of the Chemistry Department, Adam Mickiewicz University, Project Adiunkt.

References

ETTER, C. E., URBAŃCZYK-LIPKOWSKA, Z., JAHN, D. A. & FRYE, J. S. (1986). *J. Am. Chem. Soc.* **108**, 5871–5876.
JOHNSON, C. K. (1965). *ORTEP*. Report ORNL-3794. Oak Ridge National Laboratory, Tennessee, USA.
KATRUSIAK, A. (1990*a*). *Acta Cryst.* B**46**, 246–256.
KATRUSIAK, A. (1990*b*). *Acta Cryst.* C**46**, 1289–1293.
KATRUSIAK, A. (1991). *Cryst. Res. Technol.* **28**(5). In the press.
SHELDRICK, G. M. (1976). *SHELX*76. Program for crystal structure determination. Univ. of Cambridge, England.

Acta Cryst. (1991). B**47**, 404–412

Automated Conformational Analysis from Crystallographic Data. 5.* Recognition of Special Positions in Conformational Space in Symmetry-Modified Clustering Algorithms

BY FRANK H. ALLEN

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

AND ROBIN TAYLOR

ICI Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY, England

(*Received* 2 *April* 1990; *accepted* 18 *December* 1990)

Abstract

This paper reports an extension to algorithms for the conformational classification of symmetrical chemical fragments on the basis of torsion-angle descriptors [Allen, Doyle & Taylor (1991). *Acta Cryst.* B**47**, 29–40, 41–49, 50–61]. The algorithms take account of 2D topological symmetry and bring all cluster centroids into a single asymmetric unit of conformational space. In some cases, however, mean conformational geometries show marginal distortions from ideal symmetric forms, *i.e.* the centroid lies close to a special symmetry position in conformational space. Here we examine a number of methods by which this proximity can be recognized. A simple, general solution is adopted based on the torsional dissimilarities, $D(C_c,C_{c'})$, between a given cluster centroid (C_c) and each of its symmetry equivalents ($C_{c'}$). Symmetry-related clusters, c', are coalesced with the original cluster, c, if $D(C_c,C_{c'}) \leq \text{MULT} \times (D_c)_{\max}$, where $(D_c)_{\max}$ is the maximum dissimilarity between any

* Part 4: Allen & Johnson (1991).

0108-7681/91/030404-09$03.00

fragment in c and the centroid C_c. MULT is normally unity, but can be altered in practical applications. This procedure yields the 'order' of the symmetrized cluster (*i.e.* the number of symmetry variants that have been coalesced) and hence the multiplicity of the special position. The implications of cluster coalescence in the generation of statistical descriptors for the torsion-angle distributions is considered. The procedures are illustrated by application to a trial data set of six-membered carbocycles.

1. Introduction

Cluster analysis (see *e.g.* Everitt, 1980) is commonly used for the classification of objects on the basis of binary or numerical descriptors. Previous papers (Allen, Doyle & Taylor, 1991*a–c*; hereafter ADT1, ADT2, ADT3) have described various clustering algorithms for the classification of 3D conformations of chemical fragments by means of torsion-angle descriptors. The algorithms have been used to indentify conformational minima that have been observed experimentally in crystal structures for which full 3D coordinate data are available in the Cambridge Structural Database (CSD; Allen, Kennard & Taylor, 1983). The impetus for this work has been the provision of well-characterized conformational geometries for use in molecular modelling.

We have examined the single-linkage (ADT1), complete-linkage and Jarvis–Patrick (1975) clustering algorithms (ADT2) as techniques for 3D pattern recognition. A particular feature of our implementations (ADT3) is their ability to take account of the topological symmetry of a fragment during the clustering process. Many fragments of interest exhibit topological symmetry and this gives rise to permutational symmetry in the $(N_f \times N_t)$ multivariate torsional data set (N_f = number of experimental observations of the fragment, N_t = number of torsion angles used to describe the fragment geometry). Further, in 3D we may have to take account of enantiomeric conformations which could be present in the data set. Full details of symmetry specifications are given in ADT1.

For symmetrical fragments, the underlying conformational space will also exhibit symmetry (see *e.g.* Dunitz, 1979; Norskov-Lauritsen & Bürgi, 1985; Auf der Heyde & Bürgi, 1989*a–c*). Our implementations of symmetry-modified clustering algorithms (ADT1, ADT2) recognize this fact, and generate a final cluster listing in which all conformations are brought into their closest mutual proximity. The cluster centroids are therefore constrained to lie within a single asymmetric unit of the appropriate multidimensional space. A complication may arise, however, if asymmetric conformational minima lie very close to special symmetry positions in that space. Examples of these positions are those that represent conformations of perfect 3D symmetry, *e.g.* the D_{3d} chair form or the C_{2v} boat form of cyclohexane. Thus some of the mean conformational geometries reported in ADT1, ADT2 and ADT3 have marginal distortions from their expected symmetries. These conformations can be 'symmetrized' by averaging the torsion angles of the original 'asymmetric' cluster together with those of its symmetry variants which fall close to it in conformational space. The problem, identified in the earlier work, is to arrive at a definition of 'close', *i.e.* to establish criteria under which symmetry-related clusters may validly be coalesced to yield a 3D conformation which is both symmetrical and chemically sensible. We report here a solution to this problem that is applicable to results generated by any of the clustering algorithms described in ADT1 and ADT2. We also consider briefly the effects of cluster coalescence on the statistical description (Allen & Johnson, 1991) of the resultant torsion-angle distributions.

2. Trial data set

We have used the trial data set of six-membered carbocycles previously employed in the development of the clustering algorithms. Full details of its derivation from the CSD, including literature citations, are given in ADT1. The data set comprises N_f = 222 rings, for which the $N_t = 6$ intra-annular torsion angles may be generated *via* the *GSTAT* program (*CSD User Manual*, 1989). The major conformations found (ADT1, ADT2) in the data set are all close to ideal symmetric forms: chairs (D_{3d}), phenyl rings (D_{6h}), boats (C_{2v}) and half chairs (C_{2h}). The previous work also reveals a number of conformations which exhibit significant asymmetric distortions.

3. Conformational space for six-membered carbocycles

The relevant conformational space is 3D and all conformations can be represented by spherical polar coordinates (Pickett & Strauss, 1970). The coordinate set Q (the total puckering amplitude), θ, φ (Fig. 1) is simply related to the three degrees of freedom, denoted as q_2, φ_2, q_3, in the Cremer & Pople (1975) description of ring pucker. The conformations of n six-membered rings can therefore be represented by a series of n concentric spheres of radii $Q_1 \rightarrow Q_n$, since each ring will have a different puckering amplitude.

Special symmetric conformations (Table 1, see *e.g.* Cremer & Pople, 1975; Boeyens, 1978) exist on these spherical surfaces as indicated in Fig. 1. The chair form occupies the north pole ($\theta = 0°$) with its enantiomer at the south pole ($\theta = 180°$). The equator

Table 1. *Canonical forms of six-membered rings*

Energy-minimized (symmetrized) torsion angles (τ) (Bucourt & Hainaut, 1965) and their vector positions (θ, φ) (Boeyens, 1978) in spherical conformational space (°). The multiplicity (M) and order (O) of each special position (see text) is indicated. Conformational descriptors are PH = phenyl, C = chair, B = boat, E = envelope, HC = half-chair, SB = screw-boat (1,3-diplanar), TB = twist-boat, as depicted in Fig. 1.

Conf.	M	O	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	φ	θ
PH	1	24	0	0	0	0	0	0	-	-
C	2	12	60	-60	60	-60	60	-60	-	0
B	6	4	0	60	-60	0	60	-60	0	90
TB	6	4	33	33	-70	33	33	-70	30	90
E	12	2	30	0	0	-30	60	-60	0	55
HC	12	2	45	-15	0	-15	45	-62	30	50
SB	12	2	40	0	-22	0	40	-60	30	68

($\theta = 90°$) is occupied by six equivalent boat forms at $\varphi = 0$, 60, ... 300°, separated one from another on a pseudorotation pathway by six twist-boats at $\varphi = 30$, 90, ... 330°. The envelope (half-boat) conformation is intermediate between the boat and chair. Six equivalent conformations exist in the northern hemisphere at $\varphi = 0$, 60, ... 300° and $\theta \simeq 55°$; enantiomers of these six conformations are in the southern hemisphere at $\theta \simeq 125°$. Similarly, six half-chairs ($\theta \simeq 50°$) and six screw-boat (1,3-diplanar) conformations ($\theta \simeq 68°$) exist in the northern hemisphere on the φ arcs ($\varphi = 30$, 90, ... 330°) connecting the chair and twist-boat conformations. Half-chair and screw-boat enantiomers are in the southern hemisphere at $\theta \simeq$ 139 and 112° respectively. The planar phenyl ring occupies the centre of the family of spheres with $Q =$ 0 and θ and φ indeterminate.

Any general conformation G, *e.g.* one which is intermediate between an envelope and a half-chair at, say, $\theta \simeq 45°$, will occur 12 times in the northern hemisphere (and at φ values of 15, 45, 75, ... 345° in this example). Its 12 enantiomers occur in the southern hemisphere at $\theta \simeq 135°$ and at the same φ values as the northern hemisphere conformers. Thus the asymmetric unit of this conformational space is 1/24th of the complete sphere. More precisely it is any one of the 12 unique 30° φ segments of the northern (or southern) hemispheres. The $\varphi = 120°$ (boat) to $\varphi = 150°$ (twist-boat) segment is illustrated in Fig. 1.

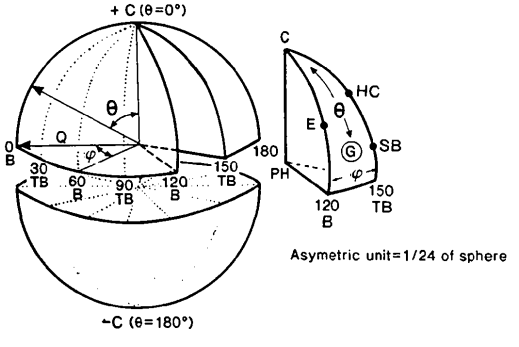

Fig. 1. Representation of conformational space for six-membered rings using the spherical polar coordinate set Q, θ, φ. Special symmetric conformations are indicated as C = chair, B = boat, E = envelope, HC = half-chair, SB = screw-boat (1,3-diplanar), TB = twist-boat. The phenyl ring (PH) is at the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

This definition of the asymmetric unit corresponds exactly to the 24 possible torsional permutation/inversion operations described in ADT1. The 12 equivalent permutations described in that paper represent the 12 possible ways in which a 2D search fragment of D_{6h} toposymmetry can be mapped onto a given target entry in the CSD. They correspond to the 12 possible (arbitrary and equivalent) atomic numbering schemes which can be imposed on each ring in the data set. For an asymmetric general-position conformation of defined geometry, each atomic enumeration generates one of 12 possible symmetry equivalents in conformational space. The remaining 12 equivalents are generated by inversion. For a six-membered carbocycle, analogous results are therefore generated by either (*a*) permuting the atomic numbering scheme against a fixed geometrical framework (as above), or (*b*) permuting the geometrical descriptors (torsion angles) with respect to a fixed atomic enumeration (as in ADT1).

If we apply torsional permutations and inversions to the symmetric conformations, then some or all of the 24 variants coalesce at one of the special positions identified above. For the D_{6h}-symmetric phenyl ring, all 24 variants coalesce at the centre of the sphere, a position of unit multiplicity and of order 24. In the case of the D_{3d}-symmetric chair, 12 variants coalesce at the north pole and 12 (enantiomeric) variants at the south pole: positions of multiplicity 2 and order 12. Each of the six boat or twist-boat conformations on the equatorial pseudorotation pathway is of order 4, whilst the 12 envelope, half-chair or screw-boat positions are of order 2. These results are summarized in Table 1. We now consider how these symmetry properties of conformational space can be detected during the clustering process. We begin by analysing the steps in this process with respect to the spherical polar construct of Fig. 1.

4. Symmetry-modified conformational clustering

The raw data set

The initial sets of N_t torsion angles generated by *GSTAT* will place a given ring in an arbitrary asymmetric unit of conformational space. This fact is an inescapable consequence of the alternative atomic numberings discussed above and in ADT1. Possible positions for six general fragments g_1–g_6, six chair fragments c_1–c_6, and six boat fragments b_1–b_6, are

illustrated in Fig. 2. We assume in this example that all fragments have similar puckering amplitudes Q, and are represented on the surface of a sphere of that radius. The individual boat and chair conformations exhibit varying deviations from ideal symmetrized forms and hence appear in general positions close to the special sites.

Dissimilarity calculations

The Minkowski metric, D (Everitt, 1980), can be used to calculate the torsional dissimilarity of two fragments (ADT1, ADT2). Thus, the dissimilarity $D(b_1,b_5)$ is a measure of the distance between fragments b_1 and b_5 in conformational space. $D(b_1,b_5)$ is obviously large for the example of Fig. 2. The symmetry-minimized dissimilarity is calculated by keeping b_1 static and moving b_5 to all of the 24 possible symmetry-equivalent positions, *i.e.* into each of the 24 asymmetric units of conformational space. This is effected by applying the user-supplied permutation/inversion list to the torsion angles for b_5. The dissimilarity $D(b_1,b_5^n)$ is calculated for the $n = 1 \rightarrow 24$ permutations and the value of n which gives rise to the minimum value of $D(b_1,b_5)$ is also recorded as the relevant overlap coefficient, $o(b_1,b_5)$. All $N_f(N_f-1)/2$ values of $D(p,q)_{\min}$ and $o(p,q)$ are calculated by this procedure.

The clustering process

The clustering process is fully decribed for all three algorithms in ADT1 and ADT2. Whilst specific details differ from algorithm to algorithm, the underlying processes are identical. A given cluster is built up around an arbitrary 'root fragment pair' selected by the algorithm on the basis of the stored minimized dissimilarities. One of these fragments is chosen as the static fragment of the pair and is called the cluster root. This root (*i.e.* g_2 in Fig. 2) may be in any of the asymmetric units of conformational space and forms a focus for the symmetry transformations (arrowed in Fig. 2) which lead to cluster formation. Roots of different clusters may be in different asymmetric units. The clustering proceeds according to the relevant algorithmic rules until a STOP point (ADT1) is reached (single or complete linkage) or the Jarvis–Patrick single pass (ADT2) is completed. If fragments g_2, b_1 and c_6 of Fig. 2 were chosen as arbitrary roots, then a possible distribution of clustered fragments in conformational space at this stage is illustrated schematically in Fig. 3(*a*). There is no guarantee that all members of a given cluster (*e.g.* that formed around c_6) will be contained within one asymmetric segment. This merely reflects the fact that fragments do not necessarily enter the cluster *via* their proximity to c_6 itself, but *via* proximity to some other fragment already assigned to that cluster.

The clustering process is concluded by bringing the largest clusters into close mutual proximity in conformational space. This is performed *via* a continuation of the single-linkage process (ADT1), or by centroid clustering in the complete-linkage and Jarvis–Patrick methods (ADT2). Assuming that the cluster formed around g_2 remains static, possible transformations of the b_1 and c_6, clusters, to new positions b_1' and c_6', are indicated in Fig. 3(*b*).

Generation of summary statistics

A wide variety of summary statistics are generated for each of the clusters detected by the algorithms. These are fully described in ADT3 and extended by Allen & Johnson (1991). Two of these descriptors are summarized here, since they are essential to the ensuing discussion of special-position location.

Mean torsion angles. For a given cluster these are derived (ADT3) by a reaveraging procedure, designed to locate the true 'centroid' of the cluster and to provide optimum overlap of fragments within

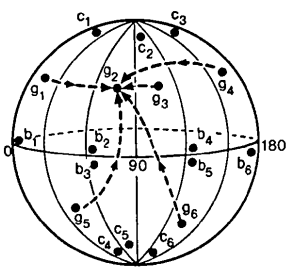

Fig. 2. Representative random conformations of six-membered rings generated by the CSD program *GSTAT* for chair (*c*), boat (*b*) and general (*g*) forms. The arrows then indicate symmetry transformations performed during cluster formation in which fragment g_2 is arbitrarily selected as the cluster root (see text).

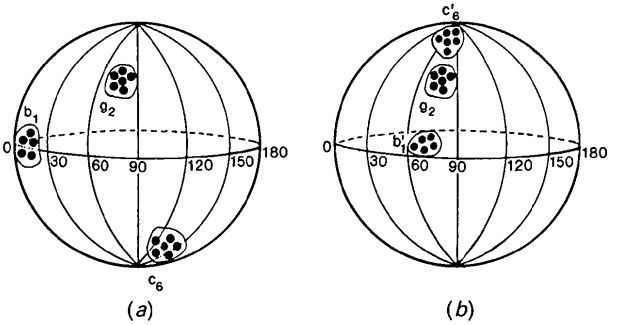

Fig. 3. Symmetry-modified clustering of six-membered ring conformations. (*a*) Clusters formed around the initial randomly chosen cluster roots g_2, b_1 and c_6'. (*b*) Symmetry transformation of clusters around b_1 and c_6 to new positions (b_1', c_6') which are in closest mutual proximity to g_2.

a cluster. The procedure is necessary since the chosen cluster root, with respect to which all incoming fragments have been oriented, may be close to the edge of the final cluster space. To allow for this situation we generate mean torsion angles for the cluster, and then reorient all fragments to provide optimum overlap with these mean values. The reorientation is performed twice to establish the final mean torsion angles reported in ADT1, ADT2 and ADT3.

Intra-cluster dissimilarities. These are calculated (ADT3) with respect to the cluster centroid established above, and using the Minkowski metric. The fragment which is closest to the centroid is termed the 'most representative fragment' of the cluster and its coordinates may be saved for use in molecular modelling. The quantity which is of most relevance to the ensuing discussion is D_{max}, the maximum dissimilarity (distance) between a cluster member and the cluster centroid. Whilst D_{max} indicates nothing about cluster shape, it does provide a good indication of cluster spread.

5. Cluster coalescence at special positions

The initial implementation (ADT3) of the symmetry-modified algorithms was primarily concerned with providing a 'dissection' of a multivariate data set, the results of which would immediately be useful in molecular-modelling applications. The mean torsion angles presented earlier for six-membered carbocycles (ADT1, ADT2) therefore reflect the 'asymmetric clustering' depicted in Fig. 3. Results for the 12 clusters of population $(N_p) \geq 4$ obtained by the Jarvis–Patrick technique (ADT2) are reproduced in Table 2(*a*). A number of these clusters have mean torsion-angle sequences that are very close to those of the ideal symmetrized forms. Chemical reasoning dictates, for example, that mean torsion angles for the phenyl cluster 1 should all be zero. Similarly the chair cluster 2 is best represented as a D_{3d}-symmetric form with mean puckering angles of $\pm 54{\cdot}6^\circ$. The problem, stated in the *Introduction* and exemplified by the discussions above, is to establish an algorithmic mechanism which permits formation of symmetrized clusters, but only in appropriate cases. The mechanism must be applicable to all three of the algorithms described in ADT1 and ADT2.

There are a number of ways of solving this problem, and each imposes a different computational overhead on the extended algorithms. Each of the solutions is briefly discussed below in decreasing order of computational complexity. Throughout the discussions we use p and q as basic-fragment designators, and p' and q' to denote any of the possible (here 24) symmetry variants of those fragments.

Use of 'fully symmetrized' dissimilarities

In this approach all of the symmetry variants of each of the original N_f fragments are included in the initial data set, *i.e.* the data set is expanded prior to analysis using the appropriate permutation/inversion operations. This results here in $24N_f$ fragments and $24N_f(24N_f - 1)/2$ unique dissimilarities of the form $D(p,q)$, $D(p,q')$ and $D(p,p')$. This approach has been used (Norskov-Lauritsen & Bürgi, 1985; Auf der Heyde & Bürgi, 1989*a–c*) in the analysis of symmetrical configuration spaces. For the present example it will result in a complete description of the spherical conformational space. Each cluster representing a 'general' conformation (*e.g.* that formed about g_2 in Fig. 3) will now appear 24 times in the resultant output summary, centred about each of the 24 occurrences of g_2. This procedure will, however, form clusters across the special positions, each cluster occurring with the correct multiplicity in the output summary: 1 phenyl, 2 chair, 6 boat clusters *etc.*, as indicated in Table 1.

This approach has the benefit that the symmetry of conformational space can be checked by inspection of the resultant cluster output. In computational terms the procedure is potentially disastrous for any data set with more than very moderate values of N_f or N_s (the number of permutational symmetry operators). Indeed, it was for this very reason that the symmetry-modified algorithms of ADT1 and ADT2 were developed, to use only the $N_f(N_f - 1)/2$ values of $D(p,q')_{min}$ as described earlier.

Use of 'partially symmetrized' dissimilarities

It would appear, in principle, that the computational overheads above can be reduced by using $D(p,q')_{min}$ (as now) together with all $N_f(N_s - 1)$ values of $D(p,p')$ as a basis for cluster formation. This modest increase in the number of stored D values has implications for the derivation of the Jarvis–Patrick nearest-neighbour table, and requires alteration of all three algorithmic bases for cluster formation.

Both of the approaches above involve modification of the minimal metrical basis for clustering derived in ADT1 and ADT2. Considerations in earlier sections of this paper indicate that effective (and simpler) solutions are possible *via* an *a posteriori* treatment of the asymmetric cluster sets, of which the data in Table 2(*a*) are an example.

Symmetrization via fragment dissimilarities

At the end of the asymmetric clustering process the program has full knowledge of cluster membership, in terms of fragment identifiers (p, q, *etc.*) and their torsion-angle sequences, transformed as

Table 2. *Mean torsional angles (°) for the ten major clusters (population $N_p > 5$) obtained with the symmetry-modified Jarvis–Patrick algorithm for the trial data set*

Column headings are: N_c = cluster number, τ_1–τ_6 are torsion angles, $(D_c)_{max}$ is the maximum centroid-fragment distance (°), O_c is the 'order' of symmetrized (coalesced) clusters (see text). Conformational descriptions are PH = phenyl, C = chair, B = boat, HC = half-chair, SB = screw-boat (1,3-diplanar), TB = twist-boat.

Class	N_c	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	$(D_c)_{max}$	O_c
(*a*) Asymmetric clusters										
PH	1	35	−1·0	0·5	1·2	−2·4	2·0	−0·3	16·1	1
C	2	46	52·1	−50·8	53·3	−57·7	58·7	−55·3	36·2	1
	3	9	37·5	−41·3	56·2	−66·1	62·3	−48·5	27·8	1
B	4	30	−72·2	70·2	0·8	−70·3	68·2	2·4	32·4	1
	5	14	−56·2	59·4	−4·4	52·6	56·7	−2·2	27·3	1
	6	9	−55·3	71·6	−6·5	−66·2	83·8	−18·8	38·0	1
	7	6	−50·8	58·2	15·2	−84·3	93·8	−27·1	53·2	1
HC	8	29	9·2	1·2	19·2	−48·7	60·6	−39·9	44·1	1
SB	9	9	−3·0	15·4	7·3	−40·3	52·7	−31·4	55·1	1
TB	10	9	−36·6	73·3	−22·2	−50·5	89·1	−36·5	50·0	1
(*b*) Values for coalesced clusters at MULT = 1·0 (default)										
PH	1	840	0·0	0·0	0·0	0·0	0·0	0·0	22·4	24
C	2	552	54·6	−54·6	54·6	−54·6	54·6	−54·6	36·2	12
	3	18	37·5	−44·9	59·2	−66·1	59·2	−44·9	37·2	2
B	4	120	−70·3	70·3	0·0	−70·3	70·3	0·0	34·0	4
	5	28	−54·4	58·0	−3·3	−54·4	58·0	−3·3	30·9	2
	6	9	−55·3	71·6	−6·5	−66·2	83·8	−18·8	38·0	1
	7	6	−50·8	58·2	15·3	−84·3	93·8	−27·1	53·2	1
HC	8	58	14·2	1·2	14·2	−44·3	60·6	−44·3	63·0	2
SB	9	18	2·2	15·4	2·2	−35·8	52·7	−35·8	62·9	2
TB	10	9	−36·6	73·3	−22·2	−50·5	89·1	−36·5	50·0	1
(*c*) Additional coalescences at MULT = 1·1										
B	5	56	−56·2	56·2	0·0	−56·2	56·2	0·0	41·1	4
	7	12	−54·5	54·5	21·2	−89·0	89·0	−21·2	81·9	2
(*d*) Additional coalescences at MULT = 1·25										
C	3	27	38·8	−42·5	55·7	−64·8	61·5	−48·7	33·0	3
TB	10	18	−29·4	73·3	−29·4	−43·5	89·1	−43·5	77·7	2
(*e*) Additional coalescences at MULT = 2·0										
B	6	18	−60·8	77·7	−12·7	−60·8	77·7	−12·7	64·7	2
HC	8	87	9·1	−2·3	22·8	−49·7	56·6	−35·9	39·7	3
SB	9	27	−3·7	11·3	11·9	−41·5	48·5	−26·3	51·4	3
TB	10	36	−36·5	81·2	−36·5	−36·5	81·2	−36·5	82·5	4

required by symmetry to place each fragment in the relevant cluster. Fragments p, q,... are now fixed in this asymmetric cluster. We may then explore conformational space around any fragment p, on a cluster-by-cluster basis, by calculating dissimilarities of the type $D(p,p')$ and $D(p,q')$. Any symmetry-related fragment, p' or q', which falls within a specified distance (D_{sym}) of p is then included in an expanding cluster; its details are added to the cluster membership arrays. For a cluster of population N_p, this approach requires the calculation of $(N_s - 1)N_p$ values of $D(p,p')$ and of $(N_s - 1)(N_p - 1)N_p/2$ values of $D(p,q')$, where N_s is the total number of symmetry operations. However, this calculation can probably be reduced to the $D(p,p')$ set alone and still produce effective symmetrization. The crux of the process is to be able to specify, or determine, a suitable value for D_{sym}.

Within the logic of the single (ADT1) and complete-linkage (ADT2) algorithms, the obvious value for D_{sym} is the maximum $D(p,q)$ value actually used by the program in cluster formation. Since the $D(p,q)$ are used hierarchically in these algorithms, D_{sym} would be equated to the $D(p,q)$ used at the STOP point specified by the user. A similarly obvious setting of D_{sym} is not, however, so readily derived for the non-hierarchical Jarvis–Patrick method. The largest $D(p,q)$ employed in constructing the nearest-neighbour (NN) table (ADT2) might appear suitable, but it can vary dramatically depending on the user choice of NN table size. We have therefore turned our attention to D_{max} (described above), the maximum distance from the cluster centroid exhibited by any member of the cluster. This quantity is calculated identically for all three algorithms from the final asymmetric cluster membership details. These considerations have led us to what appears to be the simplest solution to the problem of cluster coalescence.

Symmetrization via centroid dissimilarities

The quantity D_{max} represents our current best estimate of the spread of a cluster in conformational space. Whilst it is possible for clusters to be of any shape, we know that all members of the cluster lie within D_{max} of the centroid. Thus, in our current example, we approximate cluster shapes as spheres of radii $(D_c)_{max}$, where c is a cluster identifier, and with the cluster centroids C_c at the centres of the spheres as in Fig. 4. We are at least guaranteed that all cluster members lie within this sphere.

For a given cluster, c, we now calculate the $N_s - 1$ inter-centroid dissimilarity (distance) values $D(C_c,C_{c'})$. Symmetry-related clusters c' are then coalesced with c if $D(C_c,C_{c'}) \leq \text{MULT} \times (D_c)_{max}$, where MULT is a multiplier normally set to 1·0, but which may be altered by the user. The procedure yields the order, O_c, of the symmetrized cluster: *i.e.* the number of symmetry variants of the original asymmetric cluster, including the identity, which have been coalesced. Statistical descriptors for the symmetrized cluster, which now has a population of

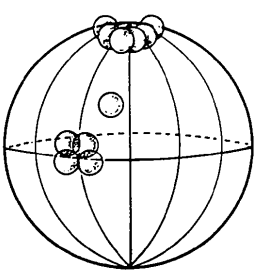

Fig. 4. Coalescence of symmetry-related variants of the asymmetric cluster set of Fig. 3. Coalescence criteria are discussed in the text. Only 6 of the 12 possible chair variants are indicated.

N_pO_c, are then derived. The statistical implications of cluster coalescence are considered later.

Cluster-coalescence results for various values of MULT are reported in Tables 2(*b–e*). For this data set, the results of Table 2(*b*) (MULT = 1·0) are in excellent accord with a visual assessment of the proximity of the asymmetric conformations (Table 2*a*) to a symmetrical form. This accord is, perhaps, improved by the additional coalescences generated with MULT = 1·1 (Table 2*c*). Thus clusters 1, 2, 4, 8, 9 are fully symmetrized [compare O_c in Table 2(*c*) with values in Table 1] at MULT = 1·0, and full symmetry for boat cluster 5 is achieved by a marginal increase in MULT. Two other clusters, chair form 3 and boat form 7, achieve only partial symmetrization (by comparison with Table 1) at the MULT = 1·10 level. This is entirely reasonable in the light of the asymmetric distortions exhibited by these conformations in Table 2(*a*). Further coalescences can be induced by raising MULT still further, and complete symmetrization of the distorted twist-boat conformation (10) is achieved at MULT $\simeq$ 2·0.

The effectiveness of the original clustering is reflected in the results of Table 2(*b*). They indicate that the symmetry variants of the cluster centroids C_c for the highly symmetric clusters 1, 2 and 4 *all* lie within the (approximate) spheres centred on the original C_c values derived from the 'asymmetric' clustering. The proximity of these original C_c values to the corresponding special position, and hence the validity of the symmetrization process, is reflected in the very small changes observed in $(D_c)_{max}$ for the asymmetric (Table 2*a*) and symmetric (Table 2*b*) forms of clusters 1, 2 and 4.

The use of $(D_c)_{max}$ as a symmetrization criterion can be criticized, since it is not a constant, but varies according to the compactness of the original asymmetric cluster. It is for this reason that boat cluster 5 is not completely symmetrized at MULT = 1·0: the original $(D_c)_{max}$ value is slightly lower (27·3°) than values for clusters 2 (36·2°) or 4 (32·4°). Conversely, the half-chair and screw-boat clusters 8 and 9 are symmetrized very early (and probably correctly) because of the broader spreads of their asymmetric forms. We feel, therefore, that the variability of the $(D_c)_{max}$ criterion is defensible, even preferable, since it effectively weights the coalescence process according to the compactness of the original asymmetric cluster. In some cases it may be appropriate to leave a very compact cluster in this form: the symmetry variants may, in fact, form an annulus about some special position, but not encompass it. We suggest that a large increase in $(D_c)_{max}$ on symmetrization may be an indicator of such behaviour. The use of the MULT operator permits coalescence properties to be examined in a systematic manner. In the present example it would be sufficient to report all results from MULT = 1·1, but we note that a fully symmetrized form of cluster 10 can be obtained at MULT = 2·0.

Whilst we have chosen $(D_c)_{max}$ as a simple, empirical criterion for determining cluster coalescence, we note that other approaches can be envisaged. For example, it seems intuitively correct to merge symmetry-related clusters if all members of the resultant single cluster lie in the same minimum of the conformational potential-energy surface. This could be revealed by statistical tests for unimodality. For example, we might choose to coalesce clusters if the distribution of fragments in the resultant merged cluster does not depart significantly from a multivariate normal distribution.

6. Statistical implications

A variety of statistics are generated (ADT3, extended by Allen & Johnson, 1991) by the cluster-analysis package implemented within the CSD program *GSTAT*. These are now amended as follows for symmetrized clusters.

The most representative fragment (*MRF*; *ADT*3)

This is now assessed as the fragment in the data set which has torsion angles which are closest to the symmetrized mean torsion angles given by the current MULT × $(D_c)_{max}$ criterion.

The overall clustering summary (*ADT*3)

This is ranked, as before, on N_p – the population of the asymmetric cluster. However, the orders O_c are now added to the summary. Intra-cluster and inter-cluster dissimilarities are calculated for the coalesced clusters of full population O_cN_p.

Statistical descriptors for individual torsion-angle distributions

These are provided (ADT3) for each individual torsion angle in each cluster. These descriptors are discussed in detail by Allen & Johnson (1991); both arithmetic and circular statistical approaches are employed. The key arithmetic values for a distribution of n torsion angles τ_i ($i = 1 \rightarrow n$) are the mean ($\bar{\tau}_a$), the sample standard deviation $\sigma(\tau_a)$ and the standard error of the mean $\sigma(\bar{\tau}_a)$, given by:

$$\bar{\tau}_a = [\textstyle\sum_{i=1}^{n} \tau_i]/n \qquad (1)$$

$$\sigma(\tau_a) = [\textstyle\sum(|\bar{\tau}_a - \tau_i|)^2/n_1]^{1/2} \qquad (2)$$

$$\sigma(\bar{\tau}_a) = \sigma(\tau_a)/(n_2)^{1/2}. \qquad (3)$$

For asymmetric clusters we have regarded the torsion-angle distributions (*e.g.* those for τ_1, τ_2,...τ_6

in Table 2) as being independent of one another, each containing N_p contributors. In these circumstances $n = n_2 = N_p$ and $n_1 = N_p - 1$. Similar considerations apply in the derivation of circular statistics (Allen & Johnson 1991).

The assumption of τ-distribution independence is clearly satisfactory for acyclic systems. It is also satisfactory for asymmetric M-membered rings. This latter statement might appear to be at variance with the $M - 3$ degrees of freedom in the puckering of M-membered rings, since treating $\tau_1 \ldots \tau_m$ as independent would imply M degrees of freedom. However, calculation of the M different torsion-angle averages tells us something about the relationships between the endocyclic valence angles and accounts for the missing three degrees of freedom. Thus, there is one degree of freedom in the puckering of cyclobutane but there is obviously more than one degree of torsional freedom: fixing one of the endocyclic torsion angles does not fix all of the others in an asymmetric ring.

For symmetrized systems the assumption of independence for *e.g.* τ_1–τ_6 in Table 2(*b*), however, is completely unacceptable. In obtaining the distributions for τ_1–τ_6 for symmetrized clusters we will have permuted each observed value into each distribution as dictated by the symmetry search. Thus each of the six phenyl-ring torsion angles is permuted four times into each of the τ_1–τ_6 distributions to obtain the $6 \times 4 = 24$ variants. For the chair form each angle permutes twice to yield the $6 \times 2 = 12$ variants. We term this the torsional frequency F_t where t ranges from $1 \rightarrow N_t$, the total number of torsion angles defining the fragment. F_t is not a constant for τ_1–τ_6, but reflects the symmetry of the special position. Thus the non-zero torsion angles of boat cluster 4 have $F_t = 1$, but $F_t = 2$ for the zero values. These F_t values are now (additionally) reported for symmetrized clusters.

It is obvious, then, that we have a fourfold excess of (equivalent) data points in each τ distribution for a phenyl ring, a twofold excess for chairs, *etc.* This excess does not affect the calculation of the mean [equation (1)] or the sample standard deviation [equation (2)]. Correct values are obtained by using $n = O_c N_p$ and $n_1 = O_c N_p - 1$ (except for one special case noted below). The principal problem concerns the value of n_2 to be used in assessing the standard error of the mean *via* equation (3). The number of different τ values included in each symmetrized τ distribution is $N_d = O_c N_p / N_t$, and it might seem that N_d is a suitable value for n_2 in equation (3).

However, the use of $n_2 = N_d = 6N_p$ in equation (3) for the six phenyl-ring τ distributions is obviously erroneous. Here symmetrization fixes zero means. Thus $\sigma(\bar{\tau}_a)$ and, indeed, its circular equivalent $\sigma(\bar{\tau}_c)$ (Allen & Johnson 1991), are also zero since the mean is known *a priori*. Further, the value of n_1 in equation (2) should properly be $O_c N_p$ (rather than $O_c N_p - 1$) for the phenyl distributions. These constraints apply to any τ distribution whose mean is fixed at a special value by symmetrization. The current software recognizes this situation for $\bar{\tau}_a = \bar{\tau}_c = 0°$ or 180°.

The use of $n_2 = N_d$ in other situations is also incorrect. N_d has values of $6N_p$ for all six τ distributions for a chair conformation, but only $4N_p$ for the non-zero torsion angles in a boat form. This would imply that mean torsion angles for a chair conformation are more precisely estimated (by a factor of $2/6^{1/2}$) than those for a boat, given that the distributions have equal sample standard deviations. This seems intuitively incorrect, and leads us to consider the problem in terms of the spherical polar coordinate set of Fig. 1. Here the chair and boat forms have fixed θ and φ values. The only degree of freedom is associated with Q, the puckering amplitude, for which the mean of the (non-zero) symmetrized torsion angles is an alternative measure. Essentially, by permuting τ_1–τ_6 we have ensured that the resultant average torsion angles correspond to perfect chair or boat forms, *i.e.* θ and φ are defined by the permutation operations. Implicitly, we have also averaged the endocyclic valence angles. There remains, then, only one degree of freedom in each individual crystallographic observation of the fragment, *viz.* that corresponding to Q. The correct value for n_2 in equation (3) is therefore N_p. We also suggest that N_p is a suitable value for n_2 for the other canonical forms of Table 1. The circular statistical treatment of Allen & Johnson (1991) has been amended to reflect these arguments.

7. Concluding remarks

The work described above represents an improvement of the symmetry-modified clustering methodology described in ADT1, ADT2 and ADT3. *A posteriori* coalescence of symmetry-related clusters is used to generate average conformations with exact 3D symmetry. The methodology is applicable to all three algorithms considered in ADT1 and ADT2, and should be applicable to any other algorithms that may be considered in the future. Because of this generality, the 'centroid dissimilarity' approach is the only method of cluster symmetrization so far encoded within the *GSTAT* program.

We note that the decision as to whether clusters should be symmetrized, and to what degree, remains effectively under user control *via* the multiplier (MULT). This parallels the user control that must also be exercized in establishing a chemically acceptable clustering structure. These user interventions are endemic in most statistical clustering systems [see

Everitt (1980) for a lengthy discussion], due to a lack of any generally accepted and robust clustering criteria. In the chemical context, where different conformations are often well separated by discrete energy barriers, we would hope to establish a decision theory to obviate user intervention. Current work is directed towards that aim, so as to generate fully automated methods for unsupervized machine learning from a large database such as the CSD.

We thank referees of earlier papers in this series for encouraging us to present the detailed discussion contained in this manuscript.

References

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*a*). *Acta Cryst.* B47, 29–40.
ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*b*). *Acta Cryst.* B47, 41–49.
ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*c*). *Acta Cryst.* B47, 50–61.
ALLEN, F. H. & JOHNSON, O. (1991). *Acta Cryst.* B47, 62–67.
ALLEN, F. H., KENNARD, O. & TAYLOR, R. (1983). *Acc. Chem. Res.* **16**, 146–153.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*a*). *Inorg. Chem.* **28**, 3960–3969.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*b*). *Inorg. Chem.* **28**, 3970–3981.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*c*). *Inorg. Chem.* **28**, 3982–3991.
BOEYENS, J. C. A. (1978). *J. Cryst. Mol. Struct.* **8**, 317–320.
BUCOURT, R. & HAINAUT, D. (1965). *Bull. Soc. Chim. Fr.* pp. 1366–1378.
CREMER, D. & POPLE, J. A. (1975). *J. Am. Chem. Soc.* **97**, 1354–1358.
CSD User Manual (1989). Version 3.4. Crystallographic Data Centre, Cambridge, England.
DUNITZ, J. D. (1979). *X-ray Analysis and the Structure of Organic Molecules*, ch. 10, pp. 447–494. Ithaca: Cornell Univ. Press.
EVERITT, B. (1980). *Cluster Analysis*, 2nd ed. London: Halstead Heinemann.
JARVIS, R. A. & PATRICK, E. A. (1975). *IEEE Trans. Comput.* **22**, 1025–1034.
NORSKOV-LAURITSEN, L. & BÜRGI, H.-B. (1985). *J. Comput. Chem.* **6**, 216–228.
PICKETT, H. M. & STRAUSS, H. L. (1970). *J. Am. Chem. Soc.* **92**, 7281–7288.

Acta Cryst. (1991). B47, 412–424

Automated Conformational Analysis from Crystallographic Data. 6.* Principal-Component Analysis for *n*-Membered Carbocyclic Rings (n = 4, 5, 6): Symmetry Considerations and Correlations with Ring-Puckering Parameters

BY FRANK H. ALLEN† AND MICHAEL J. DOYLE

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

AND THOMAS P. E. AUF DER HEYDE†

Department of Chemistry, University of the Western Cape, Bellville 7530, *South Africa*

(*Received* 28 *April* 1990; *accepted* 19 *December* 1990)

Abstract

Representative samples of four-, five- and six-membered carbocycles have been retrieved from the Cambridge Structural Database and have been used to fill, by symmetry expansion, the hyperdimensional conformation spaces spanned by the intra-annular torsion angles for these ring systems. The resulting distributions have been probed by principal-component analysis (PCA). For cyclobutane, all of the sample variance can be described in terms of a single coordinate [or principal component (PC)] which maps the degree of pucker about the ring diagonal. In the case of cyclopentane two equally important PC's fully describe the sample variance, and together they map the pseudorotation itinerary which interconverts the envelope and twist conformations of this ring. For cyclopentenes, however, a single PC (accounting for almost 80% of sample variance) maps the extent of ring pucker, whilst a second PC (accounting for the remaining 20% of variance) is found to describe minor torsional distortions away from 0° about the double bond. PCA for six-membered carbocycles (cyclohexanes and cyclohexenes) reveals three PC's: one mapping the interconversion of enantiomeric chair conformers, and two that describe the pseudorotational interchange between boat and twist-boat forms. For all three ring

* Part 5: Allen & Taylor (1991).

† Author to whom correspondence should be addressed.

0108-7681/91/030412-13$03.00