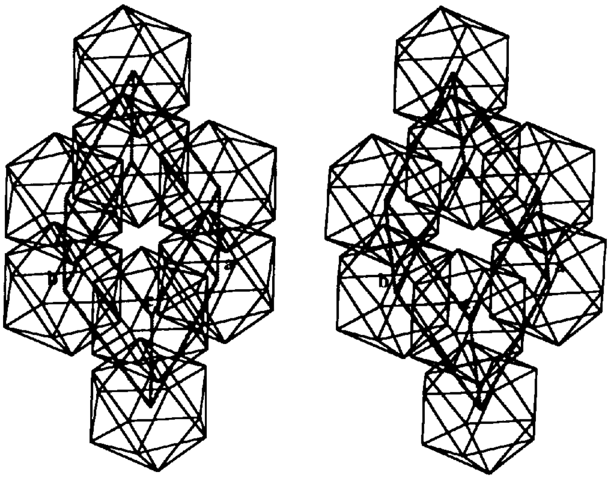


Fig. 6. A packing diagram showing the unit cell defined in line 1 of Table 1 and the particles oriented according to the rotation function in Fig. 5.

Diffraction data for this study were collected at Brookhaven National Laboratory in the Biology Department single-crystal diffraction facility at beamline X12-C in the National Synchrotron Light Source. This facility is supported by the US Department of Energy, Office of Health and Environmental Research.

References

Blum, M., Metcalf, P., Harrison, S. C. & Wiley, D. C. (1987). *J. Appl. Cryst.* **20**, 235–242.

Durbin, R. M., Burns, R., Moulai, J., Metcalf, P., Freymann, D., Blum, M., Anderson, J. E., Harrison, S. C. & Wiley, D. C. (1986). *Science*, **232**, 1127–1132.

Finch, J. T., Crowther, R. A., Hendry, D. A. & Struthers, J. K. (1974). *J. Gen. Virol.* **24**, 191–200.

Hendry, D., Hodgson, V., Clark, R. & Newman, J. (1985). *J. Gen. Virol.* **66**, 627–632.

Howard, A. J., Gilliland, G. L., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H. & Salemme, F. R. (1987). *J. Appl. Cryst.* **20**, 383–387.

Kim, S. (1989). *J. Appl. Cryst.* **22**, 53–60.

McPherson, A. (1982). *Preparation and Analysis of Protein Crystals.* New York: John Wiley.

Minor, W. & Bolin, J. T. (1990). In preparation.

Olson, N. H., Baker, T. S., Bomu, W., Johnson, J. E. & Hendry, D. A. (1987). *Proceedings of the 45th Annual Meeting of the Electron Microscopy Society of America*, edited by G. W. Bailey, pp. 650–651. Baltimore, MD: Electron Microscopy Society of America.

Rossmann, M. G. (1979). *J. Appl. Cryst.* **12**, 225–238.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.

Acta Cryst. (1991). **B47**, 29–40

Automated Conformational Analysis from Crystallographic Data. 1. A Symmetry-Modified Single-Linkage Clustering Algorithm for Three-Dimensional Pattern Recognition

By Frank H. Allen* and Michael J. Doyle

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 *November* 1989; *accepted* 12 *September* 1990)

Abstract

Single-linkage cluster analysis is used to identify discrete conformational subgroups for a chemical fragment from crystal structure data. Fragment conformations are defined by N_t torsion angles for N_f occurrences of the fragment in the Cambridge Structural Database. Conformational analysis is complicated by (*a*) the 2D topological symmetry of the fragment, giving rise to permutations of torsion-angle sequences, and (*b*) by the presence of 3D enantiomers in the original crystal structures. All steps in the single-linkage algorithm are modified to use fragment symmetry to obtain the optimum torsional overlap of all fragments. Thus, all symmetry equivalents of a given conformation are grouped into the same cluster and the final set of clusters represents an asymmetric unit of conformational space. Principal-component analysis is used to provide a visual mapping of the clustering process. The complete procedure is shown to be effective when applied

* Author for correspondence.

to a test data set of 222 six-membered carbocycles of known conformational complexity.

1. Introduction

Investigation of the low-energy conformations of molecules or substructures is fundamental to the process of molecular modelling. A number of computational procedures exist for this purpose but suffer from some limitations, *e.g.* (*a*) the size of the molecule or fragment that can be processed may be limited for computational reasons; (*b*) reliable force-field parameters may not be available; (*c*) there may be a need to postulate, *a priori*, a number of starting geometries in order to scan the whole of conformational space. It is reasonable, however, to assume that conformations observed in crystal structures are close to one or more minima on the potential-energy hypersurface. This series of papers describes methods by which the crystal structure data may be analysed to reveal details of the conformational surface for any specified substructure.

The Cambridge Structural Database (CSD; Allen, Kennard & Taylor, 1983) now contains 3D structural information from 73 893 X-ray and neutron diffraction analyses of organocarbon compounds (as of 1 July 1989). The CSD represents an ever-growing compendium of conformational data for a very broad spectrum of chemical compounds, and a common chemical substructure may occur in several hundred of these crystal structures. It is therefore necessary to develop rapid automatic techniques by which such large datasets may be sorted into conformational subgroups. The subgroups may then be ranked in order of population: if two or more well-populated subgroups exist, then a representative or averaged conformation from each may be used as an (energetically preferred) alternative in model building.

For a given substructure, the experimentally observed conformations can be expressed in terms of N_t torsion angles (*e.g.* the six intra-annular torsion angles in cyclohexane), which can readily be derived for the N_f occurrences of the fragment in the CSD. Two main techniques have been applied to the analysis of such multivariate data sets. Principal-component analysis (Murray-Rust & Bland, 1978; Murray-Rust & Motherwell, 1978; Murray-Rust & Raftery, 1985*a*,*b*) can be used to construct the M_t mutually orthogonal principal components (linear combinations of the original N_t torsion angles) which account for most of the variance in the multivariate data set. In many cases $M_t << N_t$ and the dimensionality of the problem is thus reduced. Pairs of the M_t principal-component 'scores' for the N_t fragments may then be plotted as 2D scattergrams so that conformational subgroups may be identified by visual inspection. Occasionally the directions of the principal-component axes can be interpreted in chemical terms, especially for cyclic systems. This topic will be discussed in a separate paper (Allen & Doyle, 1991).

A variety of agglomerative clustering algorithms have also been applied to multivariate torsional data sets (Norskov-Lauritsen & Bürgi, 1985; Murray-Rust & Raftery, 1985*a*,*b*; Taylor, 1986*a*). The first step in these techniques is to calculate the conformational dissimilarity of each pair of observations in the data set. This information is then used to break down the observations into 'clusters', each cluster containing fragments of similar conformation. The results may be presented as numerical tabulations of torsion angles for each discrete cluster, from which an average conformation is readily derived.

Principal-component analysis and cluster analysis have been shown to work well when applied to chemical fragments which are asymmetric. In these cases a unique and unambiguous atomic numbering can be applied to the fragment. However, many fragments exhibit symmetry in their 2D and 3D (sub)structures, *e.g.* carbocyclic and heterocyclic ring systems, metal environments in ligand complexes, *etc.* In these common cases there can be severe difficulties in interpreting the results from multivariate analyses in terms of the underlying conformational minima. These difficulties are exemplified by a principal-component analysis of phosphate groups (Murray-Rust, 1982) and by a cluster analysis of bis(triphenylphosphine)metal complexes (Norskov-Lauritsen & Bürgi, 1985).

In this paper we first illustrate the problems caused by fragment symmetry with reference to a trial data set of six-membered carbocycles. We then describe how the standard single-linkage (nearest-neighbour) clustering algorithm (Everitt, 1980) can be modified to take account of the topological symmetry of the fragment (here D_{6h}) to generate a unique and asymmetric set of conformational clusters. These clusters can then be tabulated numerically, or displayed graphically using principal-component scatterplots based on a symmetry-modified data set generated by the new clustering algorithm.

2. Trial data set

A trial data set comprising six-membered carbocycles was chosen because its conformational variants (chair, boat, half-chair, twist-boat, *etc.*) are well known. The data set was retrieved using program *QUEST* (Allen & Davies, 1988) of the CSD System Version 3.4 as released on 1 January 1989. Six-membered rings were located within the general restrictions that hits should (*a*) be organic molecules, (*b*) have no reported disorder in their crystal structures, (*c*) contain error-free atomic coordinate data, and (*d*) have reported *R* factors $\leq 0{\cdot}100$. To ensure conformational variety within a data set of manageable size, three chemically discrete substructure searches were performed to locate:

9*a*) The first 40 entries in the CSD containing a fully saturated ring comprising six Csp^3 atoms; this ensured the presence of chair conformations in the trial data set.

9*b*) The first 40 entries containing a fully saturated norbornane system (seven Csp^3 atoms) to ensure the presence of boat conformations.

(*c*) The first 50 entries containing the cyclohex-1-ene system to ensure the presence of half-chair and other intermediate conformations.

The total subset (*a*) + (*b*) + (*c*) also contained a large number of planar phenyl rings. In the event, the initial data set contained nine duplicated entries, retrieved in response to two of the substructure searches. Duplicates were eliminated to yield 121 unique entries.

All numerical calculations, including development of the symmetry-modified clustering algorithm, were carried out within the framework of the program *GSTAT*. This is the CSD System Version 3 successor to *GEOM*78 (Murray-Rust & Motherwell, 1978) and *GEOSTAT* (Murray-Rust & Raftery, 1985*a*,*b*). A basic function of the program is the preparation of systematic tabulations of user-defined geometry for a specified chemical fragment. Here the 2D framework connectivity (Fig. 1, left) of a six-membered carbocycle was matched against the molecular connectivity of each crystal structure by *GSTAT*. A total of 318 six-membered rings were located in the 121 unique entries retrieved by *QUEST*. This number exceeded array dimensions (250 rings) in the development version of the algorithm and was reduced to 222 rings from the 81 entries having $R \leq 0{\cdot}080$; short-form literature references to these 81 entries are given in Table 1.* The trial multivariate data set generated *via* *GSTAT* is, therefore, a matrix

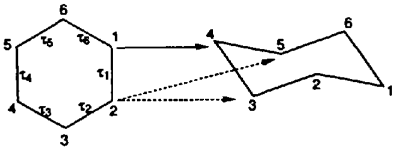


Fig. 1. One of the six possible mappings of atom 1 of a 2D search fragment (left) onto a 3D 'target' in the CSD (right). The two possible mappings for atom 2 are shown with broken lines.

Table 1. *Short-form references to the* 81 *CSD entries from which the trial data set of six-membered carbocycles was generated*

The table is ordered alphabetically by CSD reference code. Full bibliographic details have been deposited (see deposition footnote).

Code	Journal	Vol.	Page	Yr
ADDMPY10	*J. Med. Chem.*	25	427	82
ABAXES	*J. Chem. Soc. Perkin Trans.* 1		808	78
ABBUMO10	*Acta Cryst.* B	37	188	81
ABCOCX	*Acta Cryst.* B	34	147	78
ABDSCE	*Tetrahedron Lett.*		4917	79
ABIPIM	*Acta Cryst.* B	32	2683	76
ABSINT10	*Kristallografiya*	30	682	85
ACAMYA	*J. Cryst. Mol. Struct.*	9	199	79
ACANOB	*Acta Cryst.* B	32	2852	76
ACBNFP	*Acta Cryst.* B	35	1512	79
ACCITR10	*J. Chem. Soc. Perkin Trans.* 1		2393	77
ACESTA	*S. Afr. J. Chem.*	33	45	80
ACESTC	*S. Afr. J. Chem.*	33	45	80
ACFPCH	*Acta Cryst.* B	37	1119	81
ACHNAP10	*Acta Cryst.* B	32	1918	76
ACINST	*Carbohydr. Res.*	82	303	80
ACLOLN	*Tetrahedron Lett.*	22	2823	81
ACLYCA10	*Tetrahedron*	40	1783	84
ACMBPN	*Acta Cryst.* B	36	3128	80
ACNCHL	*J. Org. Chem.*	45	2264	80
ACNCHO	*Acta Cryst.* B	35	983	79
ACNODC	*Cryst. Struct. Commun.*	9	1039	80
ACNRDS	*Aust. J. Chem.*	33	2737	80
ACOHKT	*Helv. Chim. Acta*	63	2230	80
ACOLET10	*J. Chem. Soc. Perkin Trans.* 1		1494	79
ACONTN10	*Acta Cryst.* B	37	379	81
ACSCLR	*Aust. J. Chem.*	33	1783	80
ACSESO10	*Cryst. Struct. Commun.*	5	99	76
AFDECO	*Acta Cryst.* B	32	2886	76
AHDITX	*Acta Cryst.* B	26	207	70
AHMUNO10	*Acta Cryst.* B	32	1269	76
ALDRIN	*J. Chem. Soc. Perkin Trans.* 2		2153	72
AMAPRG	*Acta Cryst.* B	33	2392	77
AMBELL	*Acta Cryst.* B	32	1394	76
AMHPEN10	*Bull. Chem. Soc. Jpn*	46	1021	73
AMTBTZ	*Acta Cryst.* B	37	177	81
AMTTCD	*Tetrahedron Lett.*		1547	79
ANDBRB10	*Cryst. Struct. Commun.*	6	291	77
ANDEDP10	*J. Am. Chem. Soc.*	100	4282	78
ANDIDO	*Acta Cryst.* B	29	2247	73
ANDRAN	*Acta Cryst.* B	35	666	79
ANONAL10	*J. Chem. Res.*	15	429	78
ANOTDC10	*J. Org. Chem.*	46	5264	81
ANYCLA	*Cryst. Struct. Commun.*	5	775	76
AOIAND	*Helv. Chim. Acta*	55	375	72
AOTETC	*J. Am. Chem. Soc.*	93	7290	71
AUSTIN	*J. Am. Chem. Soc.*	98	6748	76
AXATRA	*Heterocycles*	6	1805	77
AXBCHX	*Acta Cryst.* B	34	1195	78
AXCMHN	*Cryst. Struct. Commun.*	2	391	73
AXMCHD10	*Acta Chem. Scand. Ser. B*	29	1059	75
AZCHLN	*Acta Cryst.* B	36	2337	80
AZNAND	*Cryst. Struct. Commun.*	2	33	73
AZPNHX	*Chem. Ber.*	114	423	81
BABBIP	*Acta Cryst.* B	37	1762	81
BABXUX	*Zh. Strukt. Khim.*	22	100-3	81
BAHZEP10	*J. Am. Chem. Soc.*	106	2200	84
BANJEF	*Chem. Ber.*	114	3533	81
BAPOCM10	*J. Am. Chem. Soc.*	90	74	68
BARFAB	*Cryst. Struct. Commun.*	10	1539	81
BAZCHP	*Acta Cryst.* B	28	2754	72
BCYLON10	*Acta Cryst.* B	28	3228	72
BEGCUL	*Acta Cryst.* B	38	1043	82
BEHDUN	*J. Chem. Soc. Perkin Trans.* 2		361	82
BEHFAV	*J. Chem. Soc. Perkin Trans.* 2		111	82
BEJXIX	*J. Org. Chem.*	46	4021	81
BEJXOD	*J. Org. Chem.*	46	4021	81
BEJXUJ	*J. Org. Chem.*	46	4021	81
BENBCL	*Acta Cryst.* B	30	828	74
BEPPOB	*Cryst. Struct. Commun.*	11	211	82
BERLIT	*Cryst. Struct. Commun.*	11	207	82
BEVZOR	*J. Org. Chem.*	47	265	82
BEWNOG	*Chem. Ber.*	115	1875	82
BEXGUG	*Can. J. Chem.*	60	501	82
BIBXEP	*Cryst. Struct. Commun.*	11	721	82
BIDLIJ	*Kristallografiya*	27	273	82
BIGDUQ	*J. Am. Chem. Soc.*	104	3131	82
BILVAT	*J. Org. Chem.*	47	2761	82
BIWDUG	*Aust. J. Chem.*	35	989	82

* Full bibliographic data for the entries in Table 1 have been deposited with the British Library Document Supply Centre as Supplementary Publication No. SUP 53526 (11 pp.). Copies may be obtained through The Technical Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

Table 1 (*cont.*)

Code	Journal	Vol.	Page	Yr
BLONGA10	*Acta Cryst.* B	28	3234	72
BMCLMH	*Recl J. R. Neth. Chem. Soc.*	99	118	80

$T(N_f, N_t)$ containing $N_t = 6$ torsion angles (τ_1–τ_6 in Fig. 1) for the $N_f = 222$ fragments.

3. Effects of fragment symmetry

The effects of fragment symmetry can be observed directly in a listing of the basic data matrix *T*. Two representative sections are shown in Table 2: (*a*) for boat conformations and (*b*) for chairs. Misalignment of torsional sequences is obvious. This arises from the atom-by-atom, bond-by-bond mapping of the specified 2D fragment of symmetry D_{6h} (Fig. 1) onto each 'target' in the CSD. There are six possible ways in which atom 1 of the fragment can be mapped to a given target, leaving, in each case, two alternatives for mapping atom 2. This gives rise to 12 possible mappings of the 2D fragment onto a given target and the fragment-mapping routine in *GSTAT* will arbitrarily choose one of these 12 alternatives.

If the target itself has D_{6h} symmetry (a planar phenyl ring with all τ values equal to zero), then all mappings are equivalent and no problems arise. However, in the general case, the targets are 3D objects with symmetry lower than D_{6h}, *e.g.* boats (C_{2v}), chairs (D_{3d}), *etc.* It is only the 2D connectivity representations of these 3D objects which have D_{6h} symmetry. The use of this D_{6h} 2D representation in the mapping process gives rise to the misalignments of Table 2. Remembering also that each 3D target has an enantiomorph of equal interest, there is a total of 24 ways in which the 2D 'search' representation can be mapped onto the target (Table 3). We now show how these ambiguities in the mapping process manifest themselves in (*a*) principal-component analysis, and (*b*) normal single-linkage cluster analysis.

4. Principal-component analysis

A principal-component analysis* of the trial data set was performed using routines introduced into *GSTAT* by Murray-Rust & Raftery (1985*a*,*b*). The results showed that three mutually orthogonal principal components PC1, PC2 and PC3, account for 47·7, 32·2 and 20·0% ($\Sigma = 99·9\%$) of the total variance. Principal-component scores for the 222 fragments are plotted as scattergrams in Fig. 2: (*a*) PC1 *versus* PC2, and (*b*) PC2 *versus* PC3. A full analysis of this principal-component space in chemical terms will be given later (Allen & Doyle, 1991); only the general features are of interest here.

Scattergram (*a*) of Fig. 2 shows three large peaks, 1, 2 and 2′, at PC1 = 0, +2·8 and −3·2 respectively and all with PC2≃0. Correlation of the PC scores with the original data set shows that 1, the central peak close to the origin, arises from phenyl rings of approximately D_{6h} symmetry. The slight offset from the origin is due to small deviations from ideal symmetry in the observed experimental data. Peaks 2 and 2′ at PC2≃0 both correspond to chair conformations. The larger peak, 2, at PC1≃2·8 corre-

Table 2. *Representative sections of the basic data matrix for the trial data set*

f is a fragment number and τ_1–τ_6 (°) are the intra-annular torsion angles.

(*a*) Boat conformations

f	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
63	1·7	70·7	−73·6	2·2	70·2	−72·0
69	−1·2	−71·7	71·3	1·7	−70·8	70·4
114	−78·1	72·9	0·2	−69·8	64·4	5·2
121	70·1	−1·0	−72·5	72·8	−4·7	−65·8
131	−70·5	−0·6	71·1	−68·7	−2·5	72·7
134	68·2	−73·2	1·8	70·9	−75·6	4·7

(*b*) Chair conformations

f	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
1	−60·5	59·7	−61·1	63·4	−60·4	58·9
9	56·1	−56·0	56·4	−53·9	54·7	−57·2
12	56·4	55·9	54·4	−59·1	59·9	−56·2
28	−50·9	54·0	−59·3	65·0	−61·2	52·8

Table 3. *The* 24 *possible torsion-angle sequences which can be generated due to the mapping of a* 2*D fragment of topological symmetry* D_{6h} *onto* 3*D targets in the* CSD

The sequences are given relative to τ_1–τ_6 of Fig. 1, negative signs represent enantiomers. The columns INV and *s* are fully explained in the text.

Torsional sequence (τ)						INV	*s*
1	2	3	4	5	6	1	1
2	3	4	5	6	1	1	2
3	4	5	6	1	2	1	3
4	5	6	1	2	3	1	4
5	6	1	2	3	4	1	5
6	1	2	3	4	5	1	6
6	5	4	3	2	1	1	7
5	4	3	2	1	6	1	8
4	3	2	1	6	5	1	9
3	2	1	6	5	4	1	10
2	1	6	5	4	3	1	11
1	6	5	4	3	2	1	12
−1	−2	−3	−4	−5	−6	−1	1
−2	−3	−4	−5	−6	−1	−1	2
−3	−4	−5	−6	−1	−2	−1	3
−4	−5	−6	−1	−2	−3	−1	4
−5	−6	−1	−2	−3	−4	−1	5
−6	−1	−2	−3	−4	−5	−1	6
−6	−5	−4	−3	−2	−1	−1	7
−5	−4	−3	−2	−1	−6	−1	8
−4	−3	−2	−1	−6	−5	−1	9
−3	−2	−1	−6	−5	−4	−1	10
−2	−1	−6	−5	−4	−3	−1	11
−1	−6	−5	−4	−3	−2	−1	12

* This technique is frequently described as 'factor analysis' [see *e.g.* Murray-Rust & Bland (1978)] and is so designated in the printed commentary from *GSTAT*. Recent statistical terminology (Chatfield & Collins, 1980) draws a clear distinction between principal-component analysis (as programmed in *GSTAT*) and factor analysis, a technique with similar aims but using a different underlying mathematical model.

sponds to rings with the torsion-angle sign sequence of fragments 9 and 12 in Table 2(*b*). The smaller peak, 2′, at PC1 ≃ −3·2 corresponds to the enantiomeric sequence exemplified by fragments 1 and 28 (Table 2*b*). The difference in population (peak height) of peaks 2 and 2′ is a direct consequence of the random mapping of the search fragment (Fig. 1) onto the targets in the CSD.

The other main feature of scattergram (*a*) of Fig. 2 is a line of population density at PC1 = 0. Again this population density is asymmetric about the origin. The orthogonal view along PC1 [Fig. 2, scattergram (*b*)] shows a large central peak, which is a superposition of peaks 1, 2 and 2′ of scattergram (*a*). The line of density in (*a*) at PC1 = 0 is resolved into six small peaks, surrounding the origin peak, in the orthogonal view of scattergram (*b*). Three of these peaks arise from boat conformations with torsion-angle sequences corresponding to fragments 63, 114 and 121, respectively in Table 2(*a*); a further three enantiomeric peaks are exemplified by the (enantiomeric) fragments 69, 134, and 131 of Table 2(*a*). Again there is asymmetry in the peak heights caused by random mapping of the substructural fragment.

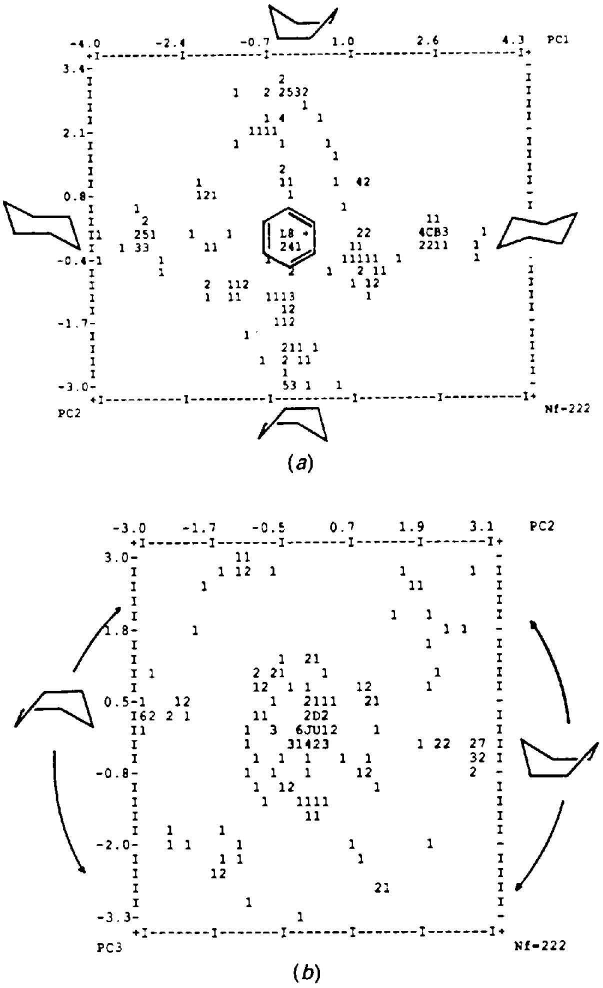


Fig. 2. Principal-component plots derived from the raw data set of 222 six-membered carbocycles: (*a*) PC1 *versus* PC2, (*b*) PC2 *versus* PC3. The frequency of incidence of fragments is recorded on a scale of 1–35, where *A* = 10, *B* = 11, ... *Z* = 35.

The principal-component analysis has, in fact, produced a very clear visual classification of the trial data set. However, considerable manual work or additional programming is required to provide listings of torsion angles, means, e.s.d.'s of means *etc*., for individual conformational minima from the principal-component results. Further, in cases of topological symmetry, it is essential to place the small subgroups 2 and 2′ and the six peaks, grouped about the origin of scattergram (*b*) and which represent boat conformations, into two larger subgroups representing chairs and boats respectively.

The principal-components method is excellent for asymmetric fragments, but we have shown that the random mapping of symmetric fragments onto CSD entries causes two major problems of interpretation: (*a*) a given conformational subgroup will be spread over a number of much smaller symmetry-related groups; (*b*) the symmetry of the principal-component scattergams is partial rather than exact, the smaller symmetry-related groupings are of unequal (even zero) population and can be difficult to locate in the scattergrams.

5. Single-linkage cluster analysis

Single-linkage and complete-linkage algorithms are two of the most commonly used methods of agglomerative clustering (Everitt, 1980). They have recently been applied to crystallographic data (Norskov-Lauritsen & Bürgi, 1985; Taylor, 1986*a*,*b*). In this section we describe the normal single-linkage algorithm in some detail. This description is included because each step of the single-linkage algorithm must be modified to take account of symmetric fragments. We also describe the application of the normal (unmodified) algorithm to the trial data set. Single-linkage clustering comprises the steps described below.

Step 0. *Calculation of dissimilarity coefficients*

All clustering algorithms employ some measure of the dissimilarity between pairs of objects (fragments) p and q. Here we calculate the conformational dissimilarity coefficient D^n_{pq} by calculating the Minkowski metric (Everitt, 1980) from the torsion-angle sets τ_i ($i = 1 \rightarrow N_t$):

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

where

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

or

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

The power factor n in (1) is an integer variable usually taken as $n = 1$ (city-block metric) or $n = 2$ (Euclidean metric). The city-block metric was used for all examples quoted in this paper. The values of the $(\Delta\tau_i)_{pq}$ in (1) are taken as the *minimum* value generated by (2*a*) and (2*b*), which arise from the phase restriction $-180 < \tau_i \leq 180°$. The D^n_{pq} values are normalized to lie in the range 0–1 by the denominator of $180N_t$ in (2*a*) and (2*b*). For N_f fragments we hence construct a square dissimilarity matrix of order N_f which is symmetrical about the (zero) diagonal. Thus only the upper or lower triangle $[N_f(N_f - 1)/2$ coefficients] need to be calculated and stored.

Step 1. *Formation of the initial cluster*

The N_f fragments are initially assigned to N_f clusters of occupancy one. Larger groupings are then formed on the basis of the D^n_{pq} values. The process is initiated by combining the two most similar fragments [*i.e.* the two fragments (*a* and *b*, say) which have the smallest dissimilarity coefficient] into a single cluster. There are now $N_f - 1$ clusters, one of which has occupancy 2. In the single-linkage algorithm the dissimilarity of the new cluster (*a* and *b*) and any other fragment (*c*) is taken as the *minimum* value of D^n_{ac} and D^n_{bc}. We now examine the dissimilarity matrix for the next smallest value. Two alternatives occur: a coefficient of the type D^n_{cd} is next smallest and we proceed to step 2, or a coefficient of the type D^n_{ac} *or* D^n_{bc} is next smallest, whence we proceed to step 3.

Step 2. *Formation of an additional new cluster*

If D^n_{cd} is next smallest, neither (*c*) nor (*d*) having been clustered with any other observation, then (*c* and *d*) form a new cluster of occupancy 2 and the number of clusters is reduced by one. The dissimilarity of (*c* and *d*) to any fragment (*e*) is assessed as described at step 1. However, we must now take into account the dissimilarity of the existing cluster (*a* and *b*) and the new cluster (*c* and *d*). For the single-linkage algorithm this is taken as the *minimum* value of D^n_{ac}, D^n_{ad}, D^n_{bc} and D^n_{bd}. We now find the next smallest dissimilarity D^n_{pq} and proceed on the basis of whether it is: (i) fragment–fragment (*p* and *q* not in any current cluster) in which case step 2 is reiterated; (ii) cluster–fragment (*p* in a cluster of size >1, *q* is not) in which case proceed to step 3; (iii) cluster–cluster (*p* and *q* already in *different* clusters of size >1) in which case proceed to step 4; (iv) cluster–cluster (*p* and *q* in same cluster already) in which case ignore, choose next lowest D^n_{pq} and assess *via* (i)–(iv).

Step 3. *Addition of a fragment to an existing cluster*

If D^n_{ac} or D^n_{bc} is the next smallest dissimilarity and (*a* and *b*) already form a cluster, then (*c*) is added to form (*a*, *b* and *c*) of occupancy 3. The total number of clusters is reduced by one. Fragment (*c*) enters the cluster by virtue of its proximity to *either* (*a*) or (*b*). The dissimilarity of (*a*, *b* and *c*) to any fragment (*d*) is the *minimum* value of D^n_{ad}, D^n_{bd} and D^n_{cd} in the single-linkage method. We now locate the next smallest value D^n_{pq} and proceed as described at (i)–(iv) of step 2.

Step 4. *Addition of a cluster to a cluster*

If D^n_{bc} is the next smallest dissimilarity, and (*b*) and (*c*) are in different clusters (*a* and *b*) and (*c* and *d*) then they merge to form (*a*, *b*, *c* and *d*) of occupancy 4. The total number of clusters is reduced by one. The clusters merge by virtue of the proximity of fragment (*c*) to fragment (*b*). We now locate the next smallest D^n_{pq} and proceed as described at (i)–(iv) of step 2.

Step 5. *Ending the clustering process*

Cluster formation occurs at step 1, and at every iteration of steps 2, 3 and 4. In each case the number of clusters is reduced by one from the original N_f singletons. The process ends naturally at step $N_f - 1$ when all fragments are in a single cluster. There is an extensive literature (Everitt, 1980) describing methods for detecting the optimum clustering point between steps 1 and $N_f - 1$. In our implementation, we initially generate a listing of all clusters at step $N_f/2$ (N_f even) or $(N_f - 1)/2$ (N_f odd), and then at five equally spaced steps up to the final step $(N_f - 1)$. Thus, for $N_f = 222$ in the trial data set, listings occur at steps 111, 133, 155, 177, 199 and 221. The listing for a given cluster at any step contains the fragment number and torsion angles of its members, together with their mean, the e.s.d.'s of mean and sample, and the population of the cluster. After step $N_f - 1$ we generate two graphs which summarize the clustering process:

(i) A plot of step number (*X*) *versus* the 'fusion dissimilarity' D^n_f, *i.e.* the value of D^n_{pq} which gave rise to cluster formation at that step (Fig. 3*a*).

(ii) A plot of step number (*X*) *versus* a ΔD^n_f value calculated as the positive difference between D^n_f values at steps X and $X - 1$ (Fig. 3*b*).

A sharp rise in either of these plots indicates that higher D^n_{pq} values are entering the clustering process, *i.e.* very dissimilar fragments or clusters are being merged. An optimum step number is selected by a visual scan of the initial cluster listings and the plots, and will be guided by chemical expectations. This step number is input as a program parameter to generate final output *only* at the selected optimum step.

Results from the unmodified single-linkage algorithm

Cluster step 160 (of 221) was selected as a stop point for the trial data set by use of the plots of Figs. 3(*a*) and 3(*b*) and a visual scan of the cluster listings noted above. At this stage 184 (of 222) fragments had been assigned to 24 clusters of population $N_p \geq 2$, leaving 38 fragments as singletons. Mean torsion angles (with e.s.d.'s) are shown in Table 4 for the top 11 clusters with $N_p \geq 4$. The single-linkage results represent a numerical expression of the principal-component scattergrams of Figs. 2(*a*) and 2(*b*). A single large cluster ($N_c = 1$) contains all phenyl rings. The boat conformers from norbornanes with puckering angle ~70° are spread over five clusters representing five of the six possible symmetry sets of Table 2(*a*). The single-linkage algorithm has also formed a sixth cluster of normal boats with puckering angle ~56°. This separation of cluster 7 from its more puckered equivalent, cluster 2, is quite justifiable in view of the e.s.d.'s obtained. There are two well populated, enantiomorphic clusters of chair conformations, and a number of smaller clusters representing half-chair, twist-boat and sofa conformations, of which clusters 10 and 11 are the largest.

The symmetry problems identified earlier are clearly apparent in the single-linkage results. The asymmetry of cluster populations is obvious in Table 4 where N_p is 19 and 38 for the two enantiomeric chairs, and N_p ranges from 1 to 15 for the individual boat variants of Table 2(*a*).

6. Symmetry-modified single-linkage clustering

We have chosen to modify the single-linkage algorithm to take account of 2D fragment symmetry.

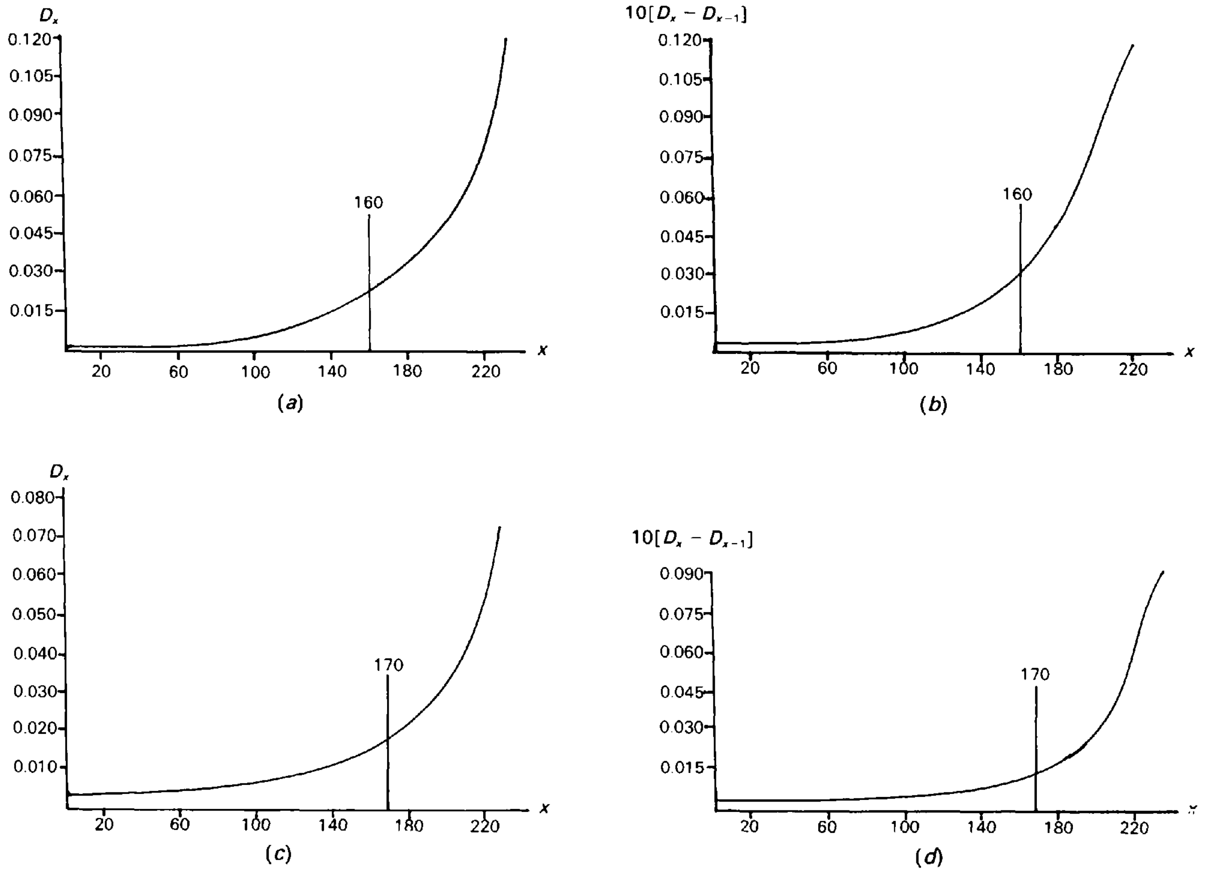


Fig. 3. Graphs of fusion dissimilarity, D_x, *versus* step number x (*a* and *c*) and of fusion-dissimilarity difference $[D_x - D_{x-1}]$ *versus* step number x (*b* and *d*). The graphs (*a* and *b*) are for the unmodified algorithm and show higher D_x and $[D_x - D_{x-1}]$ values than in those (*c* and *d*) derived from the modified algorithm. The stop step chosen is indicated on all four graphs.

Table 4. *Mean torsion angles (°; e.s.d.'s in parentheses) for major clusters obtained by the unmodified single-linkage algorithm at step* 160 *for the trial data set*

N_c = cluster number, N_p = population of cluster.

Class	N_c	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	36	0·6 (3)	0·0 (3)	−0·8 (6)	1·0 (6)	−0·4 (3)	−0·4 (4)
Boat	2	15	4·8 (17)	−75·5 (11)	71·4 (8)	1·1 (11)	−70·8 (10)	67·2 (16)
	3	15	0·1 (9)	65·9 (17)	−65·9 (25)	−0·4 (15)	66·7 (17)	−66·6 (22)
	4	7	63·6 (33)	0·5 (17)	−63·5 (31)	62·2 (35)	0·5 (18)	−63·4 (19)
	5	4	−70·6 (26)	72·2 (4)	−1·3 (5)	−69·7 (3)	69·6 (24)	−0·4 (21)
	6	4	67·9 (4)	−71·4 (10)	0·8 (6)	70·7 (16)	−69·3 (35)	0·7 (21)
	7	4	−0·3 (8)	−56·8 (11)	54·5 (10)	0·5 (7)	−55·8 (10)	57·0 (8)
Chair	8	38	54·3 (9)	−53·7 (8)	54·1 (9)	−55·4 (12)	55·4 (9)	−54·7 (7)
	9	19	−53·9 (15)	53·7 (10)	−53·4 (15)	53·9 (14)	−53·6 (17)	53·1 (20)
Half-chair	10	7	59·8 (20)	−38·8 (34)	8·3 (29)	2·5 (20)	17·6 (39)	−48·6 (28)
	11	5	12·9 (27)	−2·2 (9)	21·2 (29)	−50·4 (37)	62·7 (22)	−42·6 (25)

This is the most commonly used algorithm (Everitt, 1980) and is conceptually simple in its consistent selection of *minima* from the dissimilarity matrix.

Symmetry specification

The modified algorithm requires knowledge of the torsional sequences (s, Table 3) which are equivalent under the 2D topological symmetry of the fragment. The 12 non-enantiomeric sequences for the D_{6h} symmetry of the trial data set have INV = 1 in Table 3. In order to be general, our implementation requires the user to specify the equivalent torsional sequences individually. Each complete symmetry specification consists of N_s records, each containing N_t numerical values. The symmetry matrix $S(N_s, N_t)$ should form a group. Enantiomeric sequences can be specified in two ways. (a) A simple flag is set by the user to 1 or 0, instructing the algorithm to consider automatically (1) or not (0) the all-sign inverted torsional sequences, *e.g.* the 12 additional sequences with INV = −1 in Table 3. (b) The positive sequences are repeated with explicit negative signs specified to generate the complete explicit set of 24 permutations for the current example. Mode (a) is appropriate for all rigid fragments, but mode (b) may be necessary for flexible systems. This point is further discussed in Part 3 of this series (Allen, Doyle & Taylor, 1991b).

It is possible to detect the fragment symmetry automatically, but we have not yet implemented this option. Instead, some shorthand notations for specific symmetries are provided from which the symmetry matrix $S(N_s, N_t)$ can be derived. For D_{6h}-symmetric fragments (see Table 3) the matrices which give rise to the forward (N_s = 1–6) and reverse (N_s = 7–12) rotations of the original sequence form Abelian groups of order 6. For the general case of D_{nh} symmetry we use a keyword CROT to generate the $2n$ cyclic rotations and fill $S(N_s, N_t)$ automatically. Individual symmetry specifications may be given in any order and are used to modify steps 0–4 of the single-linkage algorithm as follows.

Step 0. *Calculation of symmetry-modified dissimilarity coefficients*

The D^n_{pq} are calculated from equations (1) and (2) by keeping the torsional sequence $(\tau_i)_p$ static and allowing the $(\tau_i)_q$ to adopt all of the N_s variants (and enantiomers if required). The *minimum* D^n_{pq} value obtained in this loop is stored in the dissimilarity matrix. The symmetry operator ($\pm s$), which relates the permutation of the $(\tau_i)_q$ to their sequence in the original data matrix $T(N_f, N_t)$, is stored in a corresponding overlap matrix O^n_{pq}. The basic multivariate data matrix T is never altered: a given fragment q may adopt any or all of its N_s possible sequence variants in the generation of minimized dissimilarities with different static fragments p. The matrix of minimized dissimilarities is used in all subsequent steps in the modified algorithm. This exhaustive calculation results in optimum overlap of the $(\tau_i)_q$ onto the $(\tau_i)_p$, despite the fact that dissimilarities within the loop on N_s show a very wide range, owing to the 3D symmetry of the particular conformations represented by fragments p and q.

Step 1. *Formation of the initial cluster*

If fragments (a and b) form the initial cluster according to the minimum symmetry-modified D^n_{pq}, then (a), the lowest-numbered fragment, is taken as the root of the cluster. This information is stored in a cluster-overlap array of dimension N_f, by setting C_a = 1, *i.e.* the root has torsion angles identical to those in the original data matrix. The symmetry relationship of the $(\tau_i)_b$ to the static $(\tau_i)_a$ is stored in the cluster-overlap array by setting $C_b = O^n_{ab}$.

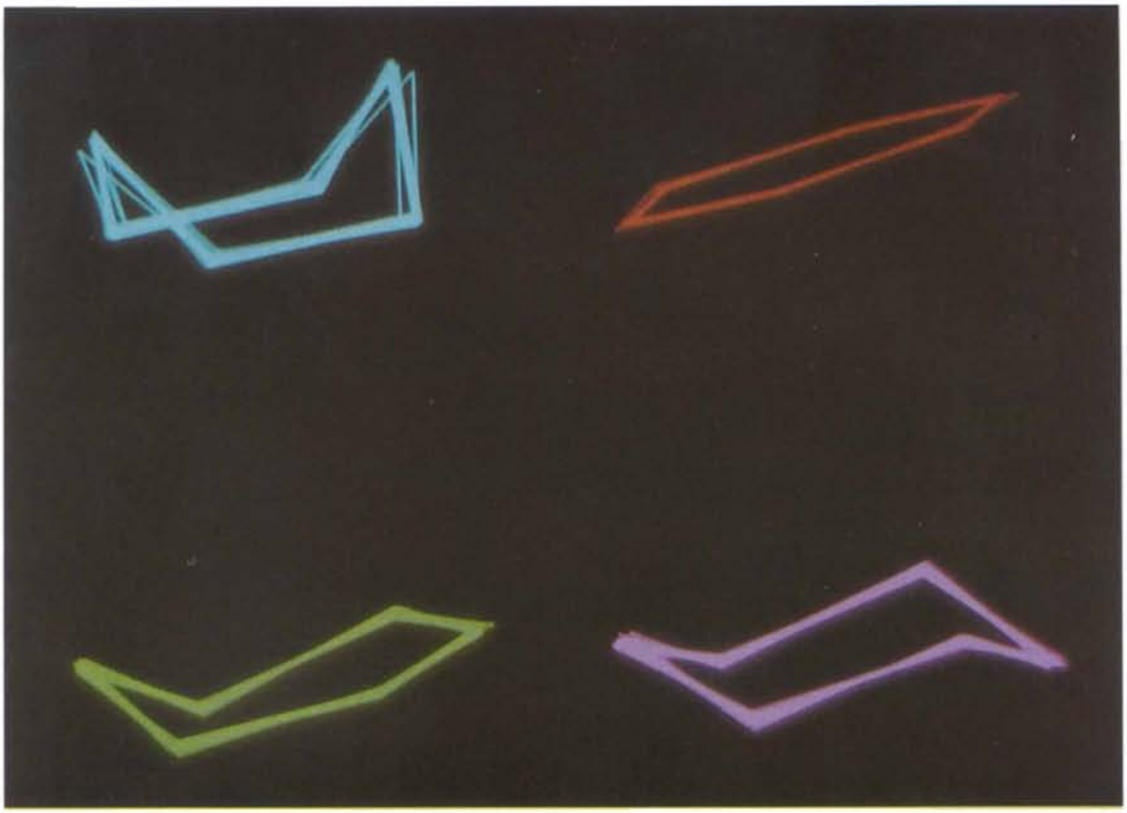
Step 2. *Formation of an additional new cluster*

If fragments (c and d) form a new cluster of occupancy 2, then (c) is regarded as the root (C_c = 1) and the symmetry relationship of (d) to (c) is stored by setting $C_d = O^n_{cd}$.

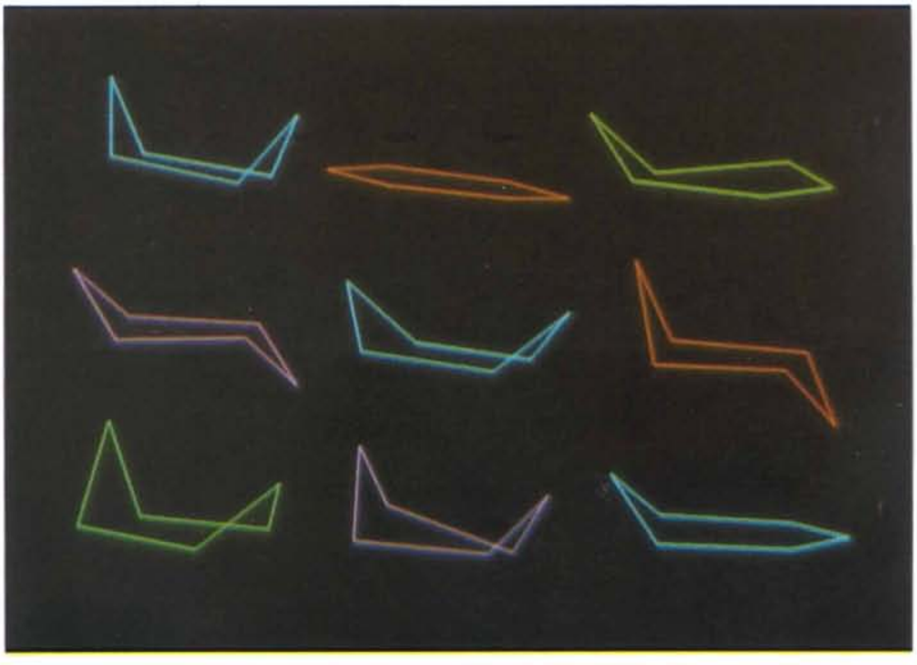
Step 3. *Addition of a fragment to an existing cluster*

Fragment (c) may enter cluster (a, b, ...) where (a) is the root, in two different ways:

(i) D^n_{ac} is a minimum, (c) enters (a and b) *via* its proximity to (a), the root of the cluster for which C_a = 1 (from steps 1 or 2 above). Since we only store the upper triangle of the dissimilarity matrix, there are two possible values for C_c. If (c) > (a) then $C_c = O^n_{ac}$. If (c) < (a) then the stored overlap coefficient is O^n_{ca}, *i.e.* the symmetry operation (S_a) which gives optimum overlap of the $(\tau_i)_a$ onto a static $(\tau_i)_c$. To locate the operator (S_c) which reverses this situation

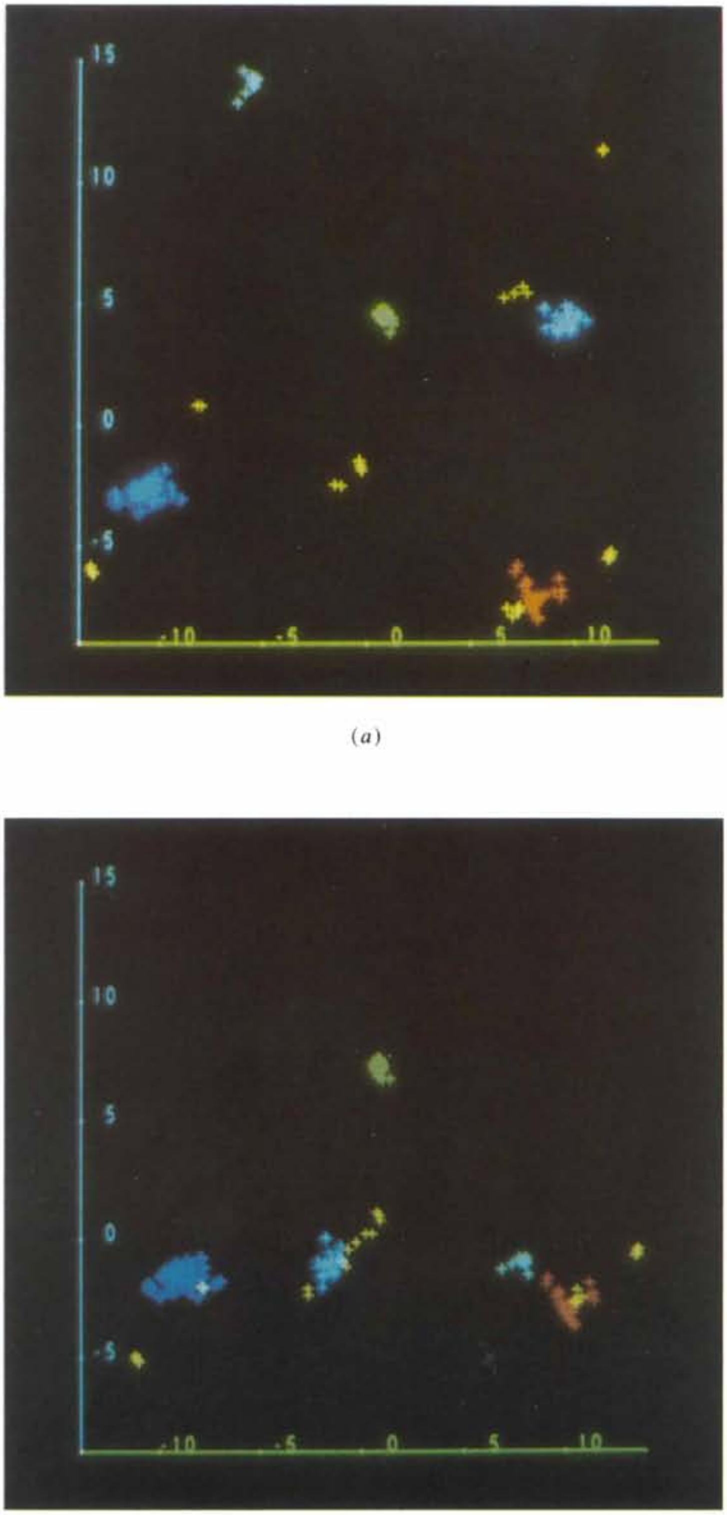


(*a*)



(*b*)

Fig. 4. (*a*) Overlay plots of all members of the four major clusters of Table 6 (after symmetry-modified single-linkage analysis). Top: norbornane boats (cluster 2, $N_f = 34$) and phenyl rings (cluster 1, $N_f = 35$); bottom: half-chairs (cluster 8, $N_f = 26$) and chairs (cluster 6, $N_f = 51$). (*b*) The 'most representative fragment' from each of the top nine clusters of Table 6 having a population $N_p \gtrsim 4$. Top row (left to right): norbornane boats ($N_c = 2$), phenyl rings ($N_c = 1$) and half-chairs ($N_c = 8$). Middle row (left to right): chairs ($N_c = 6$), 'normal' boats ($N_c = 3$) and distorted chairs ($N_c = 7$). Bottom (left to right): distorted boats ($N_c = 5$), distorted boats ($N_c = 4$) and sofas ($N_c = 9$).



(*a*)

(*b*)

Fig. 5. Principal-component scattergrams PC1 (vertical) *versus* PC2 (horizontal) for the 183 fragments assigned to clusters of population $N_p \geq 2$ by symmetry-modified single-linkage algorithm at step 170. The largest clusters are colour coded as follows (N_c with respect to Tables 6 and 7): green (phenyl rings, $N_c = 1$); red (norbornane boats, $N_c = 2$); turquoise ('normal' boats, $N_c = 3$); dark blue (chairs, $N_c = 6$); light blue (half-chairs, $N_c = 8$). All other fragments are coloured yellow. (*a*) Incomplete re-orientation (see text - clusters in random 'asymmetric units'). (*b*) Complete re-orientation (see text - clusters now in closest proximity). (*a*) and (*b*) are drawn with the same (arbitrary) scale for direct comparison.

we first apply (S_a) to the original torsional sequence $(\tau_i)_a$, so that it now has optimum overlap with $(\tau_i)_c$, and store the result in a subsidiary test array. We then apply all of the N_s symmetry operations in turn to the test array, until the resulting torsional sequence is identical to the $(\tau_i)_a$ held in the original data matrix T. The symmetry operation which gives this result is then (S_c) and this value is stored in the cluster-overlap array as C_c. There is no need to recalculate any dissimilarities in this operation.

(ii) D^n_{bc} is a minimum, (c) enters (a and b) *via* its proximity to (b), which is *not* the root of the cluster. Again there are two derivations of the overlap coefficient C_c *with respect to the cluster root* (a). If $(c) > (b)$ we first apply the operator stored in O^n_{bc} to torsional sequence (c) and store the result in a subsidiary array. We then apply the symmetry operator C_b, which relates (b) to the root (a). This results in a new subsidiary array containing a set of $(\tau_i)_c$, which has been re-sequenced twice, and represents the best overlap of $(\tau_i)_c$ onto the root sequence represented by $(\tau_i)_a$. We now apply all N_s operators to the original sequence of $(\tau_i)_c$ from the basic data matrix T. This process is stopped as soon as the sequence $(\tau_i)_c$ is generated. The operator which gives this result is then stored in the cluster-overlap array as C_c. In cases where $(b) > (c)$ we must first perform the overlap reversal described at (i) above before proceeding as for $(c) > (b)$. For many puckered rings the C_c value resulting from the above rigorous approach is simply O^n_{ac} or its inverse. However this cannot always be guaranteed.

Table 5. *Results of the symmetry-modified single-linkage clustering algorithm*

Cluster 2 at step 170 contains 34 boat-form rings from the norbornanes of the trial data set. Torsion angles are given in °.

Cluster number 2

Frag.	INV.*s*	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
49	8	0·0	71·4	−71·5	0·0	71·5	−71·4
59	−4	3·8	67·7	−72·3	1·9	70·0	−73·0
63	4	2·2	70·2	−72·0	1·7	70·7	−73·6
65	−12	1·2	72·6	−73·4	0·2	72·3	−73·0
66	1	1·2	72·6	−73·4	0·2	72·3	−73·0
67	4	0·5	70·1	−70·2	−0·3	70·2	−70·4
68	−12	1·6	69·0	−70·2	0·0	70·1	−71·4
69	9	1·7	71·3	−71·7	−1·2	70·4	−70·8
74	4	1·7	69·6	−71·7	−0·1	72·6	−72·3
79	−9	8·3	65·7	−67·1	−6·4	77·3	−80·1
86	12	8·1	66·0	−71·8	4·9	68·2	−75·5
87	12	5·9	64·9	−70·1	2·4	72·6	−76·7
91	12	9·0	63·7	−68·8	−1·1	74·5	−80·6
95	12	2·8	67·8	−70·7	1·1	72·0	−74·9
99	12	2·6	68·1	−70·3	1·0	72·2	−74·5
104	9	7·6	66·9	−68·6	−5·3	78·1	−80·6
114	7	5·2	64·4	−69·8	0·2	72·9	−78·1
121	−5	4·7	65·8	−70·1	1·0	72·5	−72·8
122	9	0·2	74·8	−74·3	−0·6	74·5	−74·6
123	−4	0·4	74·6	−73·6	−0·1	74·5	−74·9
128	4	3·0	69·1	−71·7	1·9	70·0	−73·7
130	−1	1·0	70·0	−68·1	0·6	67·1	−71·6
131	−8	2·5	68·7	−71·1	0·6	70·5	−72·7
132	12	7·9	63·2	−63·2	−7·9	76·9	−76·9
134	6	4·7	68·2	−73·2	1·8	70·9	−75·6
136	−6	4·2	68·7	−72·8	1·5	69·2	−73·3
137	−6	2·9	69·2	−71·9	1·6	70·4	−74·0
139	−3	2·4	69·4	−66·6	−0·4	66·3	−71·3
167	−10	0·4	71·0	−68·3	0·4	68·9	−73·4
168	3	0·9	68·0	−63·5	−2·2	67·2	−69·8
219	−12	0·2	62·0	−60·2	−3·0	65·0	−63·0
220	−2	0·6	66·6	−63·1	−2·4	66·7	−69·1
221	4	−0·5	63·3	−62·4	−1·6	64·4	−63·2
222	11	−0·6	68·8	−63·8	−1·7	66·8	−68·3
N_{obs}		34	34	34	34	34	34
Mean		2·9	68·3	−69·5	−0·3	70·9	−73·2
Maximum		9·0	74·8	−60·2	4·9	78·1	−63·0
Minimum		−0·6	62·0	−74·3	−7·9	64·4	−80·6
E.s.d. sample		2·8	3·1	3·7	2·5	3·3	3·9
E.s.d. mean		0·5	0·5	0·6	0·4	0·6	0·7

Step 4. *Addition of a cluster to an existing cluster*

Clusters (a, b, ...) and (c, d, ...) are fused if *e.g.* D^n_{bc} is the current smallest dissimilarity. Here we arbitrarily denote one of the clusters, say (a and b), as the 'primary cluster' and the cluster-overlap coefficients for C_a, C_b, ... remain unaltered. The coefficients C_c, C_d, ... must, however, be changed to reflect the fact that (a) is now the root of a composite cluster (a, b, c, d, ...). We first obtain C_c as described at (3) above and then apply this symmetry operation to all other members of the secondary cluster (c, d, ...).

Step 5. *Ending the clustering process*

The single-linkage algorithm is allowed to run to completion at step $N_f - 1$ with all fragments in the same cluster. A suitable end-point is assessed from a visual scan of the clustering process and examination of the X *versus* D and X *versus* ΔD plots, as described above for the normal algorithm.

Presentation of numerical results

In printing the torsion-angle tables for any cluster at any step (including the last), the current cluster-overlap coefficients are applied to the (τ_i) of the basic data matrix T. Table 5 shows a typical cluster generated by the symmetry-modified algorithm and contains the highly puckered (norbornane) boat conformers taken from step number 170. All of the fragments of Table 2(a) are now to be found in this cluster of 34 entries, which has fragment 66 as its root.

For each cluster we present a simple statistical summary (see Table 5) comprising (i) the number of observations N_p in the cluster, (ii) the maximum and (iii) the minimum values of each (τ_i), (iv) the means $(\bar{\tau}_i)$, (v) the sample standard deviations, $\sigma(\tau_i)$ of the τ_i, and (vi) the standard errors, $\sigma(\bar{\tau}_i)$ of the means:

$$\bar{\tau}_i = \sum_{i=1}^{N_p} \tau_i / N_p \qquad (3)$$

$$\sigma(\tau_i) = \left\{ \left[\sum_{i=1}^{N_p} (\bar{\tau}_i - \tau_i)^2 \right] / (N_p - 1) \right\}^{1/2} \qquad (4)$$

$$\sigma(\bar{\tau}_i) = \sigma(\tau_i)/(N_p)^{1/2}. \qquad (5)$$

The unweighted mean ($\bar{\tau}_i$, equation 3) is used here, rather than any form of weighted mean, since the unweighted mean has been shown (Taylor & Kennard, 1986) to be preferable for soft parameters such as torsion angles. The statistical summary is generated for all clusters with $N_p \geq 3$.

Further analysis of each cluster is possible, in terms of cluster shape, best overlay of the symmetry variants, identification of the 'most representative' member, *etc*. A full discussion of these topics is presented in Part 3 of this series (Allen, Doyle & Taylor, 1991*b*).

Numerical results for the trial data set

The graph of dissimilarity against step number for the symmetry-modified algorithm is shown in Fig. 3(*c*); the corresponding graph of dissimilarity differences is shown in Fig. 3(*d*). Consideration of these graphs, in conjunction with the complete cluster listings at steps 111, 133, 155, 177, 199 and 221, led to the selection of step 170 as representing optimum clustering. At this stage 183 fragments had been assigned to 13 clusters of size $N_p \geq 2$ leaving 39 singletons. These data should be compared with the 184 fragments in 24 clusters at step 160 of the unmodified algorithm.

Table 6 lists the mean torsion angles for the 10 clusters with $N_p \geq 3$ and these results should be compared with those of Table 4 from the unmodified algorithm. Apart from the D_{6h} phenyl rings, the major clusters of Table 4 (2–6 = norbornane boat, 8–9 = chair, 10–11 = half-chair) are now replaced by a single cluster in each case in Table 6 (2 = norbornane boat, 6 = chair, 8 = half-chair). These latter clusters (2, 6, 8 of Table 6) now represent compact clusters, as evidenced by their low $\sigma(\bar{\tau}_i)$ values, but their populations (N_p) do not reflect the sums of their 'contributors' from Table 4. Thus the symmetry-modified algorithm has generated two clusters of chair-form rings: the normal chairs (cluster 6, 51 fragments) and a smaller cluster of highly puckered chairs (cluster 7, 4 fragments). Their N_p sum (55) is close to the 57 chairs + enantiomorphic chairs of clusters 8 and 9 of Table 4, but the division is now much more chemically useful. There is no doubt that the high e.s.d.'s for cluster 9 of Table 4 are due to the presence of a number of highly puckered rings. Similar comments apply to a comparison of clusters 2–6 of Table 4 and the normal and highly puckered norbornane boats of clusters 2, 4 and 5 of Table 6. Clusters 4 and 5 are separated from cluster 2, and also from each other, since cluster 4 contains one torsion angle >80° whilst cluster 5 exhibits two such angles.

The symmetry-modified algorithm also enhances a number of the smaller conformational subgroups of Table 4. The 12 half-chairs of clusters 10 and 11 (Table 4) are now augmented by a number of smaller clusters $N_p \leq 3$ and individuals to give cluster 8 (Table 6) with 26 members. Similarly the small cluster 7 ($N_p = 4$, Table 4) of normal boat conformations, flattened with respect to the dominant norbornane fragments, is now enhanced to the eleven-membered cluster 3 of Table 6. Finally two conformations of cyclohexane [the sofa (1,2-diplanar) and screw-boat (1,3-diplanar)] are revealed for the first time by the symmetry-modified algorithm as clusters 9 and 10 of Table 6. These few fragments were spread over too many symmetry variants to be revealed as clusters in the unmodified analysis of Table 4.

Table 6. *Mean torsion angles (°; e.s.d.'s in parentheses) for major clusters obtained with the symmetry-modified single-linkage algorithm at step* 170 *for the trial data set*

N_c = cluster number, N_p = population of cluster.

Class	N_c	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	35	1·1 (1)	0·6 (2)	−1·0 (2)	−0·4 (1)	2·1 (2)	−2·4 (3)
Boat	2	34	2·9 (5)	68·3 (5)	−69·5 (6)	−0·3 (4)	70·9 (6)	−73·2 (7)
	3	11	1·0 (4)	−57·5 (6)	52·4 (12)	3·0 (10)	−56·4 (6)	57·0 (3)
	4	5	−15·2 (6)	81·3 (5)	−66·0 (9)	−8·2 (10)	73·9 (11)	−58·6 (7)
	5	4	24·6 (6)	52·3 (5)	−58·2 (4)	−14·3 (5)	85·9 (5)	−92·9 (5)
Chair	6	51	−50·8 (6)	54·7 (3)	−57·9 (4)	58·2 (4)	−54·3 (6)	50·1 (7)
	7	4	−58·2 (4)	78·8 (4)	−81·7 (7)	81·2 (3)	−69·9 (3)	52·3 (5)
Half-chair	8	26	18·3 (6)	1·0 (4)	10·9 (6)	−42·1 (7)	61·7 (7)	−48·2 (7)
Sofa	9	4	52·5 (12)	−25·3 (12)	1·4 (7)	0·1 (9)	27·9 (12)	−54·4 (19)
Screw-boat	10	3	−36·1 (14)	51·5 (9)	−32·9 (8)	−2·4 (13)	17·7 (8)	2·4 (8)

Presentation of results in graphical form

The most obvious graphical illustrations of the effectiveness of the symmetry-modified algorithm are (*a*) superimposed conformational plots for each discrete cluster, and (*b*) plots of a typical fragment from each cluster. A number of illustrations of this type are shown in Fig. 4 for the trial data set. Fig. 4(*a*) gives a visual impression of variance within a cluster, while Fig. 4(*b*) confirms that the automatic assignments are chemically sensible. They do not, however, show how individual clusters are related to *each other* in conformational space.

The most important visual representation of the symmetry-modified results is given by a principal-component analysis of the results at step $N_f - 1$ *i.e.* when the entire data set has been assigned to a single cluster. For the *unmodified* algorithm the single cluster is identical to the original data matrix *T*; the principal-component plots before and after clustering are therefore identical and are illustrated in Fig. 2. For the *modified* algorithm the principal component plots of the final, single cluster now have some interesting properties, as a result of the continuous re-orientation of fragments by repetition of steps 1–4 of the symmetry-modified algorithm. The final single cluster now represents the best overlay of *all* fragments. The final cluster (step $N_f - 1$) from the symmetry-modified algorithm therefore represents a unique 'asymmetric unit' of conformational space.

The formation of this unique asymmetric unit is clearly illustrated by the coloured scatterplots of Fig. 5. In our initial implementation the clustering process was simply terminated at the chosen stop point, here at step 170, and the final (step $N_f - 1$) cluster was never created. It is data from this implementation which are collected in Table 6. It is obvious that the two boat clusters 2 and 3, for example, are not yet in their closest mutual proximity in conformational space. A principal-component analysis, based on the 183 re-oriented fragments in clusters with $N_p \geq 2$, had three components (PC1 = 62·6, PC2 = 34·7, PC3 = 2·6%) accounting for 99·9% of total variance. The scattergram of PC1 *versus* PC2 scores in Fig. 5(*a*) shows the distribution of major clusters. In particular the normal boats (cluster 3, Table 6) are well separated from the norbornane boats of cluster 2. Termination at the chosen stop point yields chemically sensible clusters, but drawn from random asymmetric units of conformational space.

This problem is avoided by recording the cluster-membership details at the chosen stop point (here 170), and then allowing the process to go to completion at step $N_f - 1$ (here 221). We then re-assign the stored cluster numbers to each fragment before generating statistics and entering the principal-component analysis. Results from this additional process are given in Table 7 and in Fig. 5(*b*). The torsion-angle means for clusters 2 and 3 (Table 7) are now correctly sequenced so as to bring them into closest mutual proximity. The principal-component plot (Fig. 5*b*) is now based on the 183 fully re-oriented fragments (PC1 = 60·6, PC2 = 34·5, PC3 = 4·7%, $\sum = 99{\cdot}8\%$ of total variance) and is drawn with the same axial scales as Fig. 5(*a*). The (turquoise) normal boat cluster 3 is now in close proximity to the (red) norbornane boat cluster 2 in Fig. 5(*b*); re-location of other clusters can also be observed by comparing Figs. 5(*a*) and 5(*b*).

Asymmetric versus symmetric clusters

The procedures so far described result in the formation of discrete clusters *within* an asymmetric unit of conformational space. For this reason the mean torsion angles (Table 7) defining the cluster centroids all correspond to asymmetric conformations. It is obvious, however, that many of the reported asymmetries are very minor: the cluster centroid lies close to a position of special symmetry in conformational space. Thus, our chemical sense would predict mean torsion angles of zero for a phenyl ring, and might expect a D_{3d} chair-form ring with a mean puckering angle of 54·3° to be reported. Similar 'symmetrizations' of the data reported in Table 7 might be deemed appropriate for other conformations.

Table 7. *Mean torsion angles* (°; *e.s.d.'s in parentheses*) *for major clusters obtained with the symmetry-modified single-linkage algorithm at step* 170 *for the trial data set*

Here (*cf.* Table 6) the averaging has taken place using the overlap coefficients after the final step, but with the cluster membership codes of step 170. N_c = cluster number, N_p = population of cluster.

Class	N_c	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	35	−0·4 (1)	2·1 (2)	−2·4 (3)	1·1 (1)	0·6 (2)	−1·0 (2)
Boat	2	34	2·9 (5)	68·3 (5)	−69·5 (6)	−0·3 (4)	70·9 (6)	−73·2 (7)
	3	11	−3·0 (10)	56·4 (6)	−57·0 (3)	−1·0 (4)	57·5 (6)	−52·4 (12)
	4	5	15·2 (6)	58·6 (7)	−73·9 (11)	8·2 (10)	66·0 (9)	−81·3 (5)
	5	4	24·6 (6)	52·3 (5)	−58·2 (4)	−14·3 (5)	85·9 (5)	−92·9 (5)
Chair	6	51	−54·7 (3)	57·9 (4)	−58·2 (4)	54·3 (6)	−50·1 (7)	50·8 (6)
	7	4	−58·2 (4)	78·8 (4)	−81·7 (7)	81·2 (5)	−69·9 (6)	52·3 (5)
Half-chair	8	26	−18·3 (6)	48·2 (7)	−61·7 (7)	42·1 (7)	−10·9 (6)	−1·0 (4)
Sofa	9	4	−25·3 (12)	52·5 (12)	−54·4 (19)	27·9 (12)	0·1 (9)	−1·4 (7)
Screw-boat	10	3	2·4 (13)	32·9 (8)	−51·5 (9)	36·1 (14)	−2·4 (8)	−17·7 (8)

This process of cluster symmetrization, *i.e.* the coalescence of two or more symmetry variants of a given cluster which are close together in conformational space, is not straightforward. A variety of approaches are possible, which extend the present algorithms, and which may produce a variety of results. The crux of the problem is to arrive at an acceptable definition of 'closeness' within which the coalescence of symmetry-related variants of a given cluster is allowed to occur. This is an essentially subjective judgement, as a perusal of the results of Table 7 will indicate, hence it must remain under user control if at all possible. A number of solutions to the problem have been examined and one, which is both flexible for the user and applicable to many different clustering algorithms, is currently being fully tested. Full details will be presented in a later paper in this series (Allen & Taylor, 1991).

7. Discussion

The symmetry-modified single-linkage algorithm described in this paper has performed well in identifying the conformational minima contained in a trial data set of six-membered carbocycles. The procedure does, however, suffer from two possible technical disadvantages.

Firstly, the location of a suitable stop point to represent optimum clustering is a practical problem common to all agglomerative clustering algorithms. An extensive literature on the subject exists (Everitt, 1980, pp. 64–67). We have chosen two very simple graphical indicators, *viz* the plot of dissimilarity *versus* step number (commonly employed in many systems) and the plot of dissimilarity *differences versus* step number (which we believe to be novel). However, we would stress the necessity of examining complete listings of all clusters at various points before arriving at a decision as to the optimum step. A clear example is provided by the trial data set, where step 170 is the last step that preserves a distinction between the highly-puckered norbornane boats (cluster 2, Tables 6 and 7) and the more normal boats (cluster 3, Tables 6 and 7). The choice of the optimum-clustering step is really a choice as to the number of clusters (N_c) into which the basic data set is to be subdivided. In the case of conformational clustering, whether symmetry modified or not, this choice is subjective, and must be made in the light of what is chemically sensible for the fragment under study. The cluster listings, dissimilarity graphs and principal-component plots provide an impression of the multivariate data set, but the final decision is always in the hands of the user.

Secondly, it is well known (Everitt, 1980) that the single-linkage algorithm is prone to problems caused by the 'chaining' effect. If two well-populated and discrete clusters (*a*) and (*b*) are connected by a chain of outlying observations, then the single-linkage algorithm has a tendency to coalesce (*a*) and (*b*) and the outliers into a single large cluster. Occasionally this can happen quite early in the process, with a resultant distortion of the overall results.

A large number of alternative clustering algorithms exist which attempt to address these two fundamental problems. In the next paper (Allen, Doyle & Taylor, 1991*a*) we examine two of these procedures, and show how they can also be modified to take account of topological symmetry, to provide realistic alternatives to the single-linkage method.

We thank Dr Olga Kennard FRS for her interest in this work and the referees of this paper for suggesting improvements which are incorporated in the manuscript. MJD thanks St. John's College, Cambridge, and ICI Agrochemicals Division for financial support.

References

Allen, F. H. & Davies, J. E. (1988). *Crystallographic Computing*, Vol. 4, edited by N. W. Isaacs & M. R. Taylor, pp. 271–289. Oxford Univ. Press.

Allen, F. H. & Doyle, M. J. (1991). *Acta Cryst.* In preparation.

Allen, F. H., Doyle, M. J. & Taylor, R. (1991*a*). *Acta Cryst.* **B47**, 41–49.

Allen, F. H., Doyle, M. J. & Taylor, R. (1991*b*). *Acta Cryst.* **B47**, 50–61.

Allen, F. H., Kennard, O. & Taylor, R. (1983). *Acc. Chem. Res.* **16**, 146–153.

Allen, F. H. & Taylor, R. (1991). *Acta Cryst.* **B47**. Submitted.

Chatfield, C. & Collins, A. J. (1980). *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Everitt, B. (1980). *Cluster Analysis*, 2nd ed. London: Halstead Heinemann.

Murray-Rust, P. (1982). *Acta Cryst.* **B38**, 2765–2771.

Murray-Rust, P. & Bland, R. (1978). *Acta Cryst.* **B34**, 2527–2533.

Murray-Rust, P. & Motherwell, W. D. S. (1978). *Acta Cryst.* **B34**, 2534–2546.

Murray-Rust, P. & Raftery, J. (1985*a*). *J. Mol. Graphics*, **3**, 50–60.

Murray-Rust, P. & Raftery, J. (1985*b*). *J. Mol. Graphics*, **3**, 60–69.

Norskov-Lauritsen, L. & Bürgi, H.-B. (1985). *J. Comput. Chem.* **6**, 216–228.

Taylor, R. (1986*a*). *J. Mol. Graphics*, **4**, 123–131.

Taylor, R. (1986*b*). *J. Appl. Cryst.* **19**, 90–91.

Taylor, R. & Kennard, O. (1986). *J. Chem. Inf. Comput. Sci.* **26**, 28–35.