Acta Cryst. (1991). **B47**, 50–61

Automated Conformational Analysis from Crystallographic Data. 3.* Three-Dimensional Pattern Recognition within the Cambridge Structural Database System: Implementation and Practical Examples

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

Abstract

A unified computational procedure is described for the identification of 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. Subdivision of this multivariate data set is performed by a choice of clustering algorithms (single-linkage, complete-linkage, Jarvis–Patrick). Both asymmetric and symmetric fragments are handled routinely. The algorithms yield optimum superposition of a given conformation in a single cluster and place discrete clusters into a single asymmetric unit of conformational space. The unified procedure generates graphical and numerical indicators of clustering efficiency: (i) principal-component plots of the optimally superimposed data set, (ii) a simple statistical summary for each cluster, (iii) measures of intracluster shape and size, (iv) details of intercluster separations. Major clusters are ranked in order of decreasing population and the 'most representative fragment' (MRF: the fragment of the data set which is closest to the cluster centroid) is identified in each case. Atomic coordinates for the MRF's may be output for use as conformational alternatives in model building. The complete procedure is successfully applied to the automated conformational analysis of two very different systems, the cyclic 1-azacycloheptane moiety and the acyclic C_{17} side chain typical of cholesterol and related steroids.

1. Introduction

The two previous papers in this series (Allen, Doyle & Taylor, 1991*a*,*b*; hereafter referred to as ADT1 and ADT2) have described algorithms for the analysis of large numbers of crystallographic observations of a molecular substructure, which may be either topologically symmetric or asymmetric. The object of the work is to find clusters of observations which have closely similar conformations. If two or more well-characterized conformations are found, then each may be used as an energetically preferred alternative in model building for molecular-graphics applications.

In ADT1 we described the single-linkage (SL) algorithm (Everitt, 1980) and its modification to permit the analysis of topologically symmetric fragments. The SL algorithm is prone to the technical problem of 'chaining': an inability to distinguish two or more discrete clusters that are linked by a chain of observations of intermediate geometry. For this reason we have extended our approach (ADT2) to include symmetry-modified versions of the complete-linkage (CL; Everitt, 1980) and Jarvis–Patrick (JP; Jarvis & Patrick, 1973) algorithms, which minimize the chaining effect. All three algorithms are applicable to any multivariate data matrix where the variables have common units. For conformational analysis we use a raw data matrix of N_t torsion angles for N_f fragments. The modified algorithms take account of topological symmetry in the fragment, and the resultant clusters are drawn into their closest mutual proximity, *i.e.* a single 'asymmetric unit' in conformational space. The clustering techniques are coupled with a principal-component analysis of the symmetry-reduced data matrix, in which all N_f fragments are 'optimally' overlaid.

A trial data matrix T, containing $N_t = 6$ torsion angles for $N_f = 222$ six-membered carbocycles, (topological symmetry D_{6h}), was derived from the Cambridge Structural Database (CSD; Allen, Kennard & Taylor, 1983) and used throughout algorithm development. The search program *QUEST* (Allen &

* Part 2: Allen, Doyle & Taylor (1991*b*).

† Author for correspondence.

Davies, 1988; *CSD User Manual*, 1989) was used to locate suitable database entries. The numerical-analysis program *GSTAT* (see *e.g.* Murray-Rust & Motherwell, 1978; Murray-Rust & Raftery, 1985*a*,*b*; *CSD User Manual*, 1989) was used to generate the trial data set, whose composition is fully detailed in ADT1. Initial clustering algorithms were developed outside the *GSTAT* framework (Taylor, 1986*a*,*b*), but as the symmetry modifications progressed it became essential to incorporate them within that framework.

In this paper we describe the implementation of the generalized cluster-analysis package within *GSTAT*. In particular, we summarize the effects of various user-definable parameters, described in ADT1 and ADT2, on the actions taken by the package, both during cluster formation and in the post-processing of results. We also describe the available post-processing operations, which include (i) calculation of cluster centroids, (ii) generation of simple statistical summaries, (iii) location of the most representative fragment and output of its coordinates, (iv) indicators of intracluster shape and size, and (v) details of intercluster separations. The identification of an optimum clustering structure for a given data set is essentially a subjective judgement, based upon existing chemical knowledge and aided by some or all of the indicators described here. Finally we apply the package to two further trial data sets derived from the CSD. Both have been studied previously by less automated methods.

2. Implementation

The overall implementation scheme for the CSD cluster-analysis package is outlined in the flow chart of Fig. 1. The package has been incorporated within the framework of program *GSTAT*, which is used in its fragment-geometry mode (*CSD User Manual*, 1989) to generate the raw data matrix T of, for example, N_t torsion angles for N_f fragments. A single record, CLUST, is used to specify all variables required by the package, including whether enantiomeric fragments are to be generated. This record is followed, as required, by a set of N_s records (including the identity) specifying the topological symmetry of the fragment. A request for the generation of fragment mirror images is sufficient to invoke the symmetry-modified clustering routines, even if the fragment is topologically asymmetric.

The keys on the CLUST record permit the selection of the clustering algorithm and the integer power n to be used in calculating the dissimilarity of fragments p and q:

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

The specification of symmetry permutations and/or the conformational inversion key directs the program to generate all symmetry-allowed superpositions of the two fragments and to select as D^n_{pq} the lowest dissimilarity coefficient thus obtained. The remaining variables on the CLUST record are algorithm specific and are dealt with separately below.

Single- and complete-linkage algorithms

Two cases exist for these algorithms, depending on whether a STOP point (a step number between 1 and $N_f - 1$) is requested or not. Initial runs are usually exploratory, the output of cluster membership and graphs of dissimilarity difference *versus* step number (see ADT1) being used to select a suitable STOP value. These exploratory outputs are automatically produced at step $N_f/2$, at five further equally spaced points separated from each other by 0·1N_f steps, and at the end point (step $N_f - 1$). A user variable exists to alter the default cluster-output interval from 0·1N_f if required. The cluster-membership printouts contain simple statistics for each cluster with population

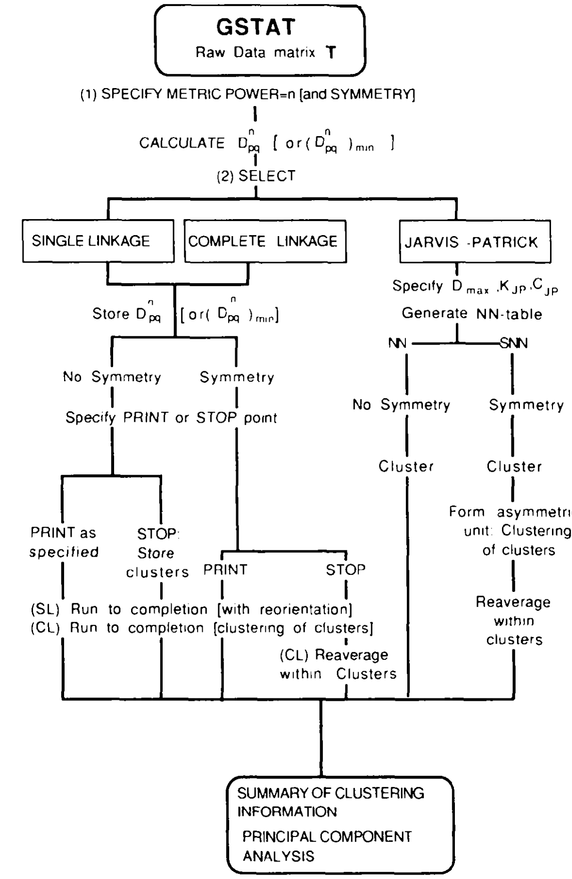

GSTAT
Raw Data matrix T
(1) SPECIFY METRIC POWER=n [and SYMMETRY]
CALCULATE D^n_{pq} [or (D^n_{pq})$_{min}$]
(2) SELECT
SINGLE LINKAGE
COMPLETE LINKAGE
JARVIS -PATRICK
Store D^n_{pq} [or (D^n_{pq})$_{min}$]
Specify D_{max} ,K_{JP} ,C_{JP}
Generate NN-table
NN
SNN
No Symmetry
Symmetry
Specify PRINT or STOP point
No Symmetry
Symmetry
Cluster
Cluster
Form asymmetric unit: Clustering of clusters
PRINT as specified
STOP: Store clusters
PRINT
STOP
(SL) Run to completion [with reorientation]
(CL) Run to completion [clustering of clusters]
(CL) Reaverage within Clusters
Reaverage within clusters
SUMMARY OF CLUSTERING INFORMATION
PRINCIPAL COMPONENT ANALYSIS

Fig. 1. Flowchart of the implementation of a conformational clustering package within the CSD software system.

$N_p \geq 3$ as (see ADT1): N_p, means, minimum and maximum values, e.s.d. of sample and e.s.d. of mean, for each of the N_t torsion angles. Since a 'final set' of clusters is indeterminate, further averaging or statistical analysis of clusters, described in § 3 and 4, is omitted. The final single cluster is then subjected to a principal-component analysis as described in § 5.

If a STOP point *is* specified then cluster membership *at that point only* is output. The membership list is retained temporarily and the algorithm allowed to go to completion either (i) naturally if no symmetry is specified, (ii) with continuous reorientation of fragments for symmetry-modified single-linkage clustering (ADT1), or (iii) by 'centroid clustering' for symmetry-modified complete-linkage clustering (ADT2). In these latter two cases a re-averaging of each cluster is carried out as described in § 3. In all three cases the statistical summary of clusters is generated and atomic coordinates may be output if required (§ 4). Principal-component analysis is then invoked (§ 5).

Jarvis–Patrick algorithm

The user-specified variables are $D_{\max}$ and $K_{\rm JP}$ (the number of nearest neighbours stored for each fragment in the nearest-neighbour table), and the Jarvis–Patrick clustering criterion, $C_{\rm JP}$, as described in ADT2. Exploratory runs require variation of these input quantities, especially the clustering criterion. The symmetry-modified algorithm (ADT2) is followed by a re-averaging of fragments (§ 3). Jarvis–Patrick results always generate a statistical summary (and coordinate output if required) as described in § 4, and are always passed to the principal-component analysis of § 5.

3. Re-averaging of clusters

This procedure is applied to the final set of clusters obtained from all of the symmetry-modified algorithms. The objective is to apply appropriate symmetry operations to the various fragments in a cluster so that, within the cluster, all fragments are optimally superimposed on one another. The initial averaging for a cluster of population N_p fragments involves reorientation of $N_p - 1$ fragments q onto a root fragment r. D^n_{rq} ($q = 2$–N_p) is minimized to give optimum torsional overlap. The (arbitrarily chosen) root fragment may, however, be on the edge of a given cluster space and, for conformations of high 3D symmetry, the initial reorientation may not yield optimum mutual overlap of fragments.

To overcome this problem further passes are made through each cluster. On the first pass the original mean torsional sequence $(\overline{\tau}_i)_m$ is used to replace the root fragment r. The symmetry permutations and inversions are used to minimize the N_p values of D^n_{mq} ($q = 1$–N_p). The cluster content and the simple statistical summary are then reprinted with a new set of overlap coefficients. This process is repeated a second time in an attempt to minimize $\sigma(\tau_i)$, the sample standard deviation, and $\sigma(\overline{\tau}_i)$, the standard deviation of the mean sequence $(\overline{\tau}_i)$. Relevant results for a cluster of 45 D_{3d} chair conformations (single-linkage, power $n = 1$) are presented in Table 1, and show significant improvements in $\sigma(\tau_i)$ and $\sigma(\overline{\tau}_i)$.

Table 1. *Effect of re-averaging procedure on* $\sigma(\tau_i)$ *and* $\sigma(\overline{\tau}_i)$

For a cluster of 45 chair-form six-membered rings obtained at step 155 with the symmetry-modified single-linkage algorithm. *A* = arbitrary root; *R*1 and *R*2 = first and second re-averaging. Torsion angles are given in °.

		τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	Σ
A	Mean	−50·8	54·3	−56·9	57·4	−54·4	50·4	
	Maximum	−44·0	60·1	−51·7	63·3	−45·0	59·3	
	Minimum	−57·6	50·6	−63·7	52·6	−65·0	40·6	
	$\sigma(\tau_i)$	3·7	2·2	2·9	2·4	4·2	4·8	20·2
	$\sigma(\overline{\tau}_i)$	0·5	0·3	0·4	0·4	0·6	0·7	2·9
*R*1	Mean	−50·2	53·3	−57·3	58·6	−54·6	50·2	
	Maximum	−40·6	57·8	−52·6	65·0	−46·9	57·2	
	Minimum	−57·6	45·0	−63·3	52·7	−61·2	42·4	
	$\sigma(\tau_i)$	4·0	2·5	2·4	2·5	2·7	3·6	17·7
	$\sigma(\overline{\tau}_i)$	0·6	0·4	0·4	0·4	0·4	0·5	2·7
*R*2	Mean	−49·9	52·9	−57·3	58·6	−54·8	50·6	
	Maximum	−40·6	57·8	−52·6	65·0	−50·5	57·6	
	Minimum	−57·2	45·0	−63·3	53·3	−61·2	42·4	
	$\sigma(\tau_i)$	3·9	2·5	2·2	2·6	2·3	3·8	17·3
	$\sigma(\overline{\tau}_i)$	0·6	0·4	0·3	0·4	0·3	0·6	2·6

4. Statistical summary and output of coordinates

The statistical summary, with optional output of coordinates, is provided for all runs except exploratory trials with the single- or complete-linkage algorithms, *i.e.* for all cases in which a final set of clusters is established. Brief summary printout of the type described below is essential for a rapid assessment of the usefulness of the cluster assignments. The cluster-membership lists, especially if they involve re-averaging (§ 3), are lengthy and the output described under the following four headings is an important diagnostic aid.

Cluster-membership summary

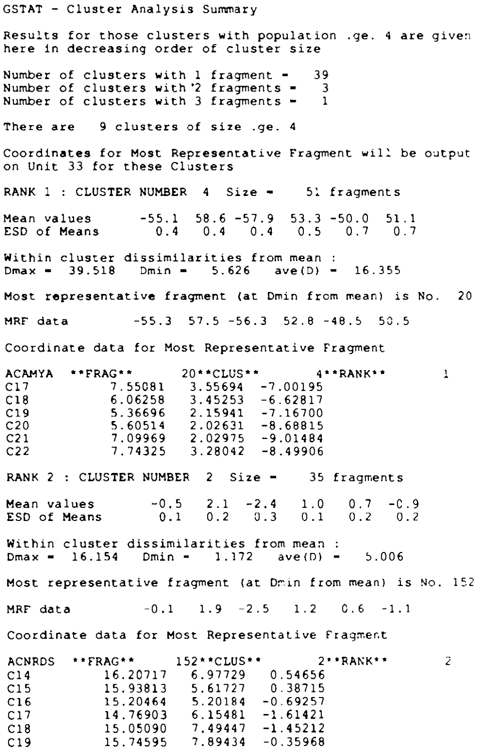
All clusters with a population $N_p \geq 2$ are assigned a sequential cluster number 1 to N_c. This list is sorted into ascending order of N_p. The number of unclustered (singleton) fragments is listed, together with the numbers of clusters having 2, 3 and ≥4 members.

Intracluster dissimilarities

The procedure described in § 3 above is repeated, with the final printed mean torsional sequence $(\overline{\tau}_i)_m$ taken as the (fixed) cluster root. Dissimilarities, minimized by toposymmetry if required, are calculated for each of the N_p members of each cluster. The D^n_{mq} ($q = 1\text{–}N_p$) are calculated in non-normalized form and expressed in degrees where n is the power factor in equation (1). Thus for $n = 1$, D^n_{mq} is the sum of absolute torsional differences from the mean sequences in degrees over all N_t torsion angles; for $n = 2$, we have the root-mean-square difference from the mean sequence. In the summary printout the average maximum and minimum values of these intracluster dissimilarities from the centroid are given. These values provide some overview of the conformational homogeneity of each cluster, especially if the $(D^n_{mq})_{\max}$ values are compared with the intercluster dissimilarities described below. Summary printout is ordered by decreasing cluster size for all clusters with $N_p \geq 4$. A small section of this output is shown in Fig. 2.

Output of atomic coordinates

One of the major objectives of this work is to provide atomic coordinates for major conformational subgroups for use in model building. Ideally these coordinates should correspond to the cluster centroid, *i.e.* to the final mean torsional sequence $(\bar{\tau}_i)_m$ for a given cluster. This would involve least-squares fitting with ring-closure constraints. In practice we have adopted a simpler approach by locating the *most representative fragment* (MRF) in each cluster. The MRF is the fragment with minimum D^n_{mq} as defined above (see Fig. 2).



```
GSTAT - Cluster Analysis Summary

Results for those clusters with population .ge. 4 are given
here in decreasing order of cluster size

Number of clusters with 1 fragment -   39
Number of clusters with '2 fragments -   3
Number of clusters with 3 fragments -   1

There are   9 clusters of size .ge. 4

Coordinates for Most Representative Fragment will be output
on Unit 33 for these Clusters

RANK 1 : CLUSTER NUMBER  4  Size -    51 fragments

Mean values        -55.1  58.6 -57.9  53.3 -50.0  51.1
ESD of Means         0.4   0.4   0.4   0.5   0.7   0.7

Within cluster dissimilarities from mean :
Dmax -  39.518   Dmin -   5.626   ave(D) -  16.355

Most representative fragment (at Dmin from mean) is No.  20

MRF data          -55.3  57.5 -56.3  52.8 -48.5  50.5

Coordinate data for Most Representative Fragment

ACAMYA  **FRAG**       20**CLUS**       4**RANK**        1
C17          7.55081   3.55694  -7.00195
C18          6.06258   3.45253  -6.62817
C19          5.36696   2.15941  -7.16700
C20          5.60514   2.02631  -8.68815
C21          7.09969   2.02975  -9.01484
C22          7.74325   3.28042  -8.49906

RANK 2 : CLUSTER NUMBER  2  Size -    35 fragments

Mean values        -0.5   2.1  -2.4   1.0   0.7  -0.9
ESD of Means        0.1   0.2   0.3   0.1   0.2   0.2

Within cluster dissimilarities from mean :
Dmax -  16.154   Dmin -   1.172   ave(D) -   5.006

Most representative fragment (at Dmin from mean) is No. 152

MRF data           -0.1   1.9  -2.5   1.2   0.6  -1.1

Coordinate data for Most Representative Fragment

ACNRDS  **FRAG**      152**CLUS**       2**RANK**        2
C14         16.20717   6.97729   0.54656
C15         15.93813   5.61727   0.38715
C16         15.20464   5.20184  -0.69257
C17         14.76903   6.15481  -1.61421
C18         15.05090   7.49447  -1.45212
C19         15.74595   7.89434  -0.35968
```

Fig. 2. Ranked summary of conformational clusters generated by the symmetry-modified single-linkage algorithm (see ADT1) with (optional) output of coordinates for the most representative fragment in each cluster. Only the two largest clusters are shown for the trial data set of six-membered carbocycles.

The program *GSTAT* already has the ability to output atomic coordinates for each fragment located in the search process (or for the complete molecule in which it occurs). This file is ordered by fragment number and is generated at the same time as the raw data matrix used in conformational analysis. A variety of coordinate-output forms are available, *e.g.* fractional or orthogonal with respect to the origin of the crystallographic unit cell, orthogonal with respect to user-defined molecular axes, *etc.* It is a simple matter to interrogate this file and retrieve the coordinates for the MRF, add some identifying information (fragment and cluster numbers and the cluster ranking, see Fig. 2) and transfer the data to a separate output unit for use in model building.

Intercluster dissimilarity tables

The calculations of intracluster dissimilarities given above provide a rapid summary of conformational variations *within* each cluster. It is also important to assess how well the clusters are separated one from another in conformational space. For this purpose we calculate the complete dissimilarity matrix based on the mean torsional sequences $(\bar{\tau}_i)_m$ for up to 20 clusters with $N_p \geq 4$ (otherwise the 20 clusters of highest rank of N_p are treated). This is a trivial calculation involving a maximum of $20 \times 19/2 = 190$ dissimilarity calculations. The D^n_{pq}, which now relate to cluster centroids for clusters p, q,..., are minimized for toposymmetry if required, and given in degrees. A typical table is shown in Fig. 3, in which the intra- and intercluster dissimilarities may be directly compared.

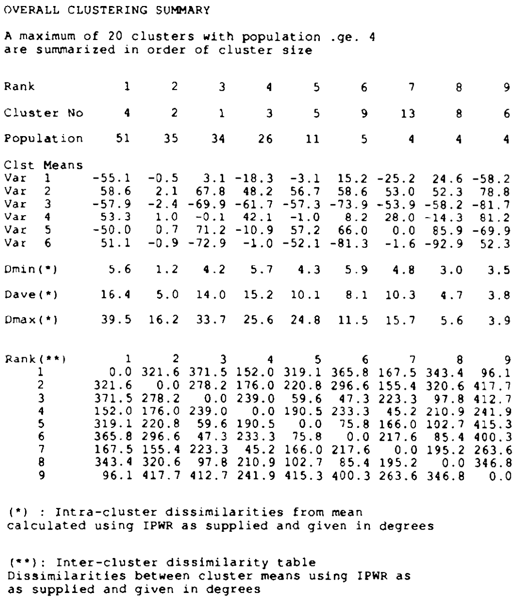
Other assessments of results

The output described above represents a rapidly computed summary of the clustering structure. Other numerical assessments of cluster shape are possible (see *e.g.* Everitt, 1980) and are being investigated. A preliminary survey of the use of principal-component analysis on individual clusters has been attempted. For the major clusters from the trial data set of six-membered carbocycles, a single principal component was dominant (*ca* 100% of variance) in each case. We think that this component can be correlated with a single parameter describing the variations of the degree of puckering within these homogeneous clusters of ring conformers. For some of the smaller clusters, representing the more flexible forms on interconversion pathways, more than one important principal component was indicated. Further work is required to establish the principal-component method as a tool for the assessment of cluster homogeneity and shape, and to provide a chemically meaningful interpretation of these results (see *e.g.* Auf der Heyde & Bürgi, 1989*a–c*). We return to this topic in a later paper (Allen & Doyle, 1991).

5. Principal-component analysis of symmetry-reduced data sets

The cluster-analysis package concludes with a call to the principal-component (PC) algorithm in *GSTAT*. The number of components to be extracted may be set by the user, otherwise a default value of 3 is used. In cases where no symmetry or inversion specification is given, then the PC analysis is applied to the raw data set. In all cases where the clustering algorithm has generated a set of final clusters (STOP specified in single- or complete-linkage, always for Jarvis–Patrick), the cluster-membership information is associated with each fragment and is passed to the PC step. This procedure allows PC scatterplots to be generated in terms of cluster membership rather than as population-density maps. By this method we were able to generate the scatterplots shown in Fig. 5 of ADT1.

The principal-component results following any of the symmetry-modified algorithms are of considerable value. They give a visual overview, as a series of 2D projections, of the clustering structure of a single asymmetric unit in *n*-dimensional conformational space. Examination of these plots provides a visual impression of the homogeneity and shape of the clusters identified by the algorithm. Examination of the intercluster spaces should indicate possible interconversion pathways (or genuine outliers). Work on an interactive-graphics system for viewing these results in 3D is planned.

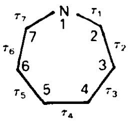


```
OVERALL CLUSTERING SUMMARY

A maximum of 20 clusters with population .ge. 4
are summarized in order of cluster size

Rank            1     2     3     4     5     6     7     8     9

Cluster No      4     2     1     3     5     9    13     8     6

Population     51    35    34    26    11     5     4     4     4

Clst Means
Var  1      -55.1  -0.5   3.1 -18.3  -3.1  15.2 -25.2  24.6 -58.2
Var  2       58.6   2.1  67.8  48.2  56.7  58.6  53.0  52.3  78.8
Var  3      -57.9  -2.4 -69.9 -61.7 -57.3 -73.9 -53.9 -58.2 -81.7
Var  4       53.3   1.0  -0.1  42.1  -1.0   8.2  28.0 -14.3  81.2
Var  5      -50.0   0.7  71.2 -10.9  57.2  66.0   0.0  85.9 -69.9
Var  6       51.1  -0.9 -72.9  -1.0 -52.1 -81.3  -1.6 -92.9  52.3

Dmin(*)       5.6   1.2   4.2   5.7   4.3   5.9   4.8   3.0   3.5

Dave(*)      16.4   5.0  14.0  15.2  10.1   8.1  10.3   4.7   3.8

Dmax(*)      39.5  16.2  33.7  25.6  24.8  11.5  15.7   5.6   3.9


Rank(**)        1     2     3     4     5     6     7     8     9
    1         0.0 321.6 371.5 152.0 319.1 365.8 167.5 343.4  96.1
    2       321.6   0.0 278.2 176.0 220.8 296.6 155.4 320.6 417.7
    3       371.5 278.2   0.0 239.0  59.6  47.3 223.3  97.8 412.7
    4       152.0 176.0 239.0   0.0 190.5 233.3  45.2 210.9 241.9
    5       319.1 220.8  59.6 190.5   0.0  75.8 166.0 102.7 415.3
    6       365.8 296.6  47.3 233.3  75.8   0.0 217.6  85.4 400.3
    7       167.5 155.4 223.3  45.2 166.0 217.6   0.0 195.2 263.6
    8       343.4 320.6  97.8 210.9 102.7  85.4 195.2   0.0 346.8
    9        96.1 417.7 412.7 241.9 415.3 400.3 263.6 346.8   0.0

(*) : Intra-cluster dissimilarities from mean
calculated using IPWR as supplied and given in degrees

(**): Inter-cluster dissimilarity table
Dissimilarities between cluster means using IPWR as
as supplied and given in degrees
```

Fig. 3. Overall-summary output for symmetry-modified single-linkage cluster analysis of the trial data set; the intra- and intercluster dissimilarity tables are shown for nine clusters with a population greater than four. IPWR is a program mnemonic for the power *n* in equation (1).

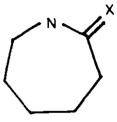
6. Practical examples

We have chosen two practical applications to illustrate the use of the package with simple symmetric fragments. The first is a cyclic fragment: 1-azacycloheptane (I), which has topological mirror symmetry in 2D, the mirror plane passing through N and the midpoint of C4—C5. The second fragment, the C17 side chain (IV) typical of cholesterol and related steroids, is acyclic. In both cases the results of manual conformational classifications (Taylor, 1989; Duax, Griffin, Rohrer & Weeks, 1980) were available for comparison.

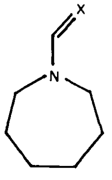


(I) Torsional permutations

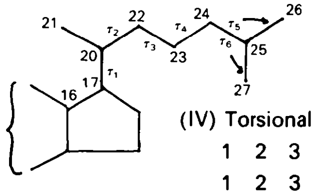
1	2	3	4	5	6	7
7	6	5	4	3	2	1



(II)

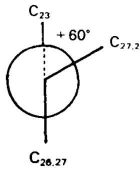


(III)

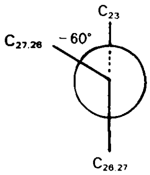


(IV) Torsional permutations

1	2	3	4	5	6
1	2	3	4	6	5



(V) (+)-synclinal



(VI) (−)-synclinal

1-*Azacycloheptane* (I)

The manual survey of Taylor (1989) was repeated. There were 36 entries in the July 1989 release of the CSD containing (I) constrained to have single intra-annular bonds (bond type = 1), with no fusion or bridging of the fragment (no cyclic routes emanating from the defined fragment), and with error-free atomic coordinates available. The CSD reference codes and short-form literature citations are given in Table 2(*a*).* The 36 entries gave rise to $N_f = 53$ independent fragments for which the $N_t = 7$ intra-annular torsion angles of (I) were generated by *GSTAT*.

All three symmetry-modified clustering algorithms were run for this torsional data set; the city-block metric [$n = 1$, equation (1)] was used throughout, together with the permutation sequences of (I) and their enantiomorphic inversions. For the single- (SL) and complete-linkage (CL) algorithms, complete clustering output was generated at steps 26, 31, 36, 41, 46, 51 and 52 ($N_f - 1$), together with the plots of D and ΔD *versus* step number. Inspection of all output led to selection of steps 41 (SL) and 45 (CL) as optimum clustering points. At this stage the SL algorithm had assigned 44 fragments to three clusters with $N_p = 30$, 11, 3 and nine singletons remained. The CL method assigned 48 fragments to four clusters with $N_p = 31$, 11, 3, 3 and only five singletons. Mean torsion angles for clusters with $N_p \geq 3$ are given in Table 3 for both methods.

The Jarvis–Patrick (JP) algorithm was run with a variety of values for $K_{\rm JP}$, $D_{\rm max}$ and $C_{\rm JP}$ (see ADT2 and *Jarvis–Patrick algorithm*). Tight constraints ($K_{\rm JP} = 10$, $D_{\rm max} = 0{\cdot}06$, $C_{\rm JP} = 6$) assigned 40 fragments to two clusters ($N_p = 30$, 10), with 13 singletons. Relaxing these constraints, particularly $C_{\rm JP}$, allows smaller clusters to form. A run with $K_{\rm JP} = 7$, $C_{\rm JP} = 2$, $D_{\rm max} = 0{\cdot}15$ generated four clusters of $N_p = 29$, 10, 4, 6 which may be readily correlated with the CL results. Results from both JP runs (JP1, JP2) are also included in Table 3, together with data for the 'most representative fragment' (MRF) for the two major clusters. The MRF data are common to all four clusterings.

All three algorithms have generated essentially identical results for the two major clusters identified in the manual analysis of a slightly smaller data set (Taylor, 1989). The comparison with gas-phase electron diffraction data (Dillen & Geise, 1979) and force-field calculations (Bocian, Pickett, Rounds & Strauss, 1975) in Table 3, shows that the major cluster corresponds closely to the chair form of the parent cycloheptane with $\tau_7(\equiv \tau_1) \rightarrow 0$. The second major cluster, 2, corresponds to a twist-chair. The smaller clusters 3 (SL, CL and JP2) and 4 (CL and

Table 2. *Short-form references to CSD entries used in the conformational analyses of* (I) *and* (IV)

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

Code	Journal	Vol.	Page	Yr
(*a*) Structure (II)				
AZBQUB	*S. Afr. J. Chem.*	34	23	81
BACMEC10	*Acta Cryst.* B	34	138	78
BAJZOB10	*Eesti NSV Tead. Akad. Toim. Keem.*	31	282	82
BILJOV	*J. Am. Chem. Soc.*	104	3929	82
BORYUC	*J. Heterocycl. Chem.*	19	481	82
BUCXAY	*Pol. J. Chem.*	55	1015	81
BUFPEX	*Chem. Ber.*	116	1547	83
BZPCHO	*Bull. Chem. Soc. Jpn*	47	1117	74
CABCOX	*Croat. Chem. Acta*	56	87	83
CABCUD	*Croat. Chem. Acta*	56	87	83
CABVIK	*Koord. Khim.*	9	306	83
CAPLAC	*Acta Cryst.* B	31	268	75
CAPLAC01	*Zh. Strukt. Khim.*	15	679	74
CAPRES	*Kristallografiya*	19	1170	74
CDHMTC	*Kristallografiya*	17	303	72
CUHMTC10	*Kristallografiya*	13	169	68
DIKVAU	*Bull. Chem. Soc. Jpn*	58	745	85
DOKMUL	*J. Med. Chem.*	29	251	86
FECYIV	*Arch. Pharm.*	319	798	86
FEFSUE	*Dokl. Akad. Nauk SSSR*	284	131	85
FENXUR	*Acta Cryst.* C	43	154	87
FENXUR01	*Acta Cryst.* C	43	154	87
FETCAI	*Inorg. Chem.*	26	822	87
FOGBEI	*J. Organomet. Chem.*	327	157	87
FOZKUA	*J. Chem. Soc. Chem. Commun.*		12	88
FULMUU	*J. Med. Chem.*	31	422	88
FULNAB	*J. Med. Chem.*	31	422	88
FULNEF	*J. Med. Chem.*	31	422	88
HEXAMC	*J. Chem. Soc. Chem. Commun.*		939	77
MECILN	*J. Antibiot.*	34	282	81
NHMDTC	*Kristallografiya*	17	111	72
TCAPLI	*Acta Chem. Scand. Ser. A*	28	175	74
XIMBZA	*Bull. Chem. Soc. Jpn*	54	964	81
XIMBZB	*Bull. Chem. Soc. Jpn*	54	962	81
XIMBZB01	*Bull. Chem. Soc. Jpn*	54	962	81
XMTCZN	*Zh. Strukt. Khim.*	13	660	72
(*b*) Structure (IV)				
ABDSCE	*Tetrahedron Lett.*		4917	79
ACNCHL	*J. Org. Chem.*	45	2264	80
AENLAN10	*Acta Cryst.* B	32	1311	76
AXSCHO	*Helv. Chim. Acta*	62	1770	79
AXCOPR	*Acta Cryst.* B	28	567	72
BABDUD	*Acta Cryst.* B	37	1793	81
BACLCH	*Acta Cryst.*	20	249	66
BADSCE	*Tetrahedron Lett.*		4917	79
BAFLID01	*Cryst. Struct. Commun.*	10	1289	81
BAGVAG	*Acta Cryst.* B	37	1881	81
BAHKEA11	*J. Lipid Res.*	24	784	83
BEXCHO	*Chem. Phys. Lipids*	18	240	77
BINKAK	*Bull. Soc. Chim. Belg.*	91	205	82
BIZZIT	*Acta Cryst.* B	38	2845	82
BODZID	*Bull. Chem. Soc. Jpn*	55	3041	82
BOGBUU	*Monatsh. Chem.*	113	439	82
BORRIJ	*Tetrahedron Lett.*	24	617	83
BSCHOL	*Helv. Chim. Acta*	59	1273	76
BUGKET	*Bull. Soc. Chim. Belg.*	92	271	83
BXCHOL10	*Acta Cryst.* B	33	3117	77
BXDCHO10	*Chem. Phys. Lipids*	26	249	80
CAZCHI	*Acta Cryst.* B	26	1362	70
CEMMAI	*J. Chem. Soc. Perkin Trans.* 1		397	84
CEMMEM	*J. Chem. Soc. Perkin Trans.* 1		397	84
CEYNAV	*J. Org. Chem.*	49	1537	84
CHENON	*Acta Cryst.* B	33	3755	77
CHLCFM	*Acta Cryst.* B	34	2872	78
CHLSOL	*Cryst. Struct. Commun.*	10	41	81
CHLSOS	*J. Org. Chem.*	33	3535	68
CHOBRH10	*Chem. Phys. Lipids*	20	43	77
CHOCAL	*J. Org. Chem.*	41	3476	76
CHOENO	*Acta Cryst.* B	32	1984	76
CHOEST20	*Acta Cryst.* B	37	1538	81
CHOESU	*Cryst. Struct. Commun.*	8	107	79
CHOLAD10	*Acta Cryst.* B	35	895	79
CHOLAU02	*Chem. Phys. Lipids*	23	179	79
CHOLAU04	*Acta Cryst.* B	36	3027	80
CHOLEU01	*Acta Cryst.* B	38	2411	82
CHOLOL	*Cryst. Struct. Commun.*	9	263	80

* Full bibliographic data for the entries in Tables 2(*a*) and 2(*b*) have been deposited with the British Library Document Supply Centre as Supplementary Publication No. SUP 53528 (13 pp.). Copies may be obtained through The Technical Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

Table 2 (*cont.*)

Code	Journal	Vol.	Page	Yr
CHOLON	*Acta Cryst.* B	33	1236	77
CHOMYS	*J. Chem. Soc. Perkin Trans.* 2		814	76
CHONON10	*J. Chem. Soc. Perkin Trans.* 2		1414	79
CHOOCT10	*Chem. Phys. Lipids*	24	157	79
CHOOLA01	*J. Lipid Res.*	27	1214	86
CHOPTS10	*Acta Cryst.*	33	2934	77
CHTYBB10	*J. Chem. Soc. Perkin Trans.* 1		805	77
CLBUST	*Acta Cryst.* B	33	3326	77
COXBST	*Acta Cryst.* B	33	3326	77
CUVJEI	*Acta Cryst.* C	41	739	85
DBRCHO	*Acta Cryst.* B	32	2730	76
DECHUO	*Khim. Fiz.*	4	329	85
DEDSOU	*Monatsh. Chem.*	115	1453	84
DHXSCH	*Tetrahedron Lett.*		4917	79
DIFTAN	*S. Afr. J. Chem.*	38	131	85
DILPIX	*Monatsh. Chem.*	116	831	85
DULCAO	*Monatsh. Chem.*	116	1221	85
EPXCHO	*Acta Cryst.* B	33	2128	77
ETCHBR	*Acta Cryst.* B	36	1460	80
EXCHOL	*Acta Cryst.* B	33	3582	77
FEKGEH	*J. Chem. Soc. Chem. Commun.*		283	87
FEMTUM	*Bull. Soc. Chim. Belg.*	96	35	87
FEMVAU	*Bull. Soc. Chim. Belg.*	96	35	87
FENGOU	*Helv. Chim. Acta*	70	37	87
FENWEA	*Kristallografiya*	31	671	86
FEYREG	*J. Org. Chem.*	51	4888	86
FIXTEL	*J. Org. Chem.*	52	2963	87
FOLSEE	*Mol. Cryst. Liq. Cryst.*	144	179	87
GASFAH	*J. Lipid Res.*	28	80	87
GAYFUH	*J. Org. Chem.*	53	2180	88
HCHLTZ10	*Acta Cryst.* C	39	297	83
IEPCHO10	*J. Chem. Soc. Perkin Trans.* 1		236	81
LUMIST	*Acta Cryst.* B	30	1695	74
SECHLS	*Helv. Chim. Acta*	64	703	81
SPINDC	*Tetrahedron*	37	1407	81
TOXCNB	*Acta Cryst.* B	32	2492	76
TSCHOL	*S. Afr. J. Chem.*	32	97	79
ZZZBID01	*Bull. Korean Chem. Soc.*	6	333	85

JP2 only) are distorted chairs in which $\tau_4 \rightarrow 0$ (cluster 3) and $\tau_6 (\equiv \tau_2) \rightarrow 0$ (cluster 4).

A survey of the chemical constitution of cluster 1 shows that these chair conformers are all derived from two chemical subgroups. The largest of these (II) has an exocyclic double bond C1=*X* (≡ C6=*X*) where *X* = O, S, N, *i.e.* ε-lactones and hetero-analogues. A smaller subgroup (III) has an unsaturated exocyclic carbon attached to the N atom. The effect of these points of unsaturation adjacent to N is to generate a degree of double-bond character in N—C1 (C6) or N—C (exocyclic) *via* conjugation with the N lone pair. The sp^2 hybridization at N results in a mean intra-annular C2—N1—C7 angle of 124·2 (5)° over the 30 fragments assigned to cluster 1 by all three algorithms. By contrast, all of the twist-chairs of cluster 2 arise from fully saturated fragments, with no exocyclic unsaturation of the N substituents. Here the mean C2—N1—C7 angle is 118·8 (8)° for the ten fragments common to cluster 2 from all three algorithms.

Table 3. *Mean intra-annular torsion angles* τ_1–τ_7 (°) *for major conformations of* 1-*azacycloheptane* (I) *derived from single-linkage* (*SL*), *complete-linkage* (*CL*) *and Jarvis–Patrick* (*JP*1 *and JP*2) *clustering*

Manual results (M: Taylor, 1989), gas-phase electron diffraction data (ED: Dillen & Geise, 1979) and force-field calculations (FF: Bocian *et al.*, 1975) are given for comparison. 'MRF' indicates the most representative fragment (which is common for all clustering methods). E.s.d.'s are in parentheses where applicable.

	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
Cluster 1, chair								
SL	30	66·6 (5)	−78·1 (5)	59·8 (5)	−61·0 (5)	80·5 (5)	−65·6 (8)	−1·0 (7)
CL	31	66·0 (8)	−77·6 (6)	59·6 (5)	−60·7 (6)	80·0 (7)	−65·4 (8)	−0·6 (8)
JP1	30	66·6 (5)	−78·1 (5)	59·8 (5)	−61·0 (5)	80·5 (5)	−65·6 (8)	−1·0 (7)
JP2	29	66·7 (5)	−78·2 (5)	59·7 (5)	−60·8 (5)	80·3 (6)	−65·5 (8)	−1·0 (7)
MRF	-	66·1	−77·9	61·4	−62·9	80·9	−65·6	−0·9
M	25	66·6	−78·6	59·6	−60·3	80·4	−66·1	−0·6
ED	–	70·7	−89·5	66·0	−66·0	89·5	−70·7	0·0
FF	–	62·7	−77·5	58·1	−58·1	77·5	−62·6	0·0
Cluster 2, twist-chair								
SL	11	85·4 (17)	−73·6 (23)	55·0 (24)	−66·6 (14)	87·3 (16)	−47·9 (26)	−32·5 (21)
CL	11	85·4 (17)	−73·6 (23)	55·0 (24)	−66·6 (14)	87·3 (16)	−47·9 (26)	−32·5 (21)
JP1	10	86·3 (16)	−72·1 (19)	53·2 (18)	−66·2 (15)	87·1 (17)	−46·6 (25)	−34·1 (21)
JP2	10	85·8 (18)	−75·2 (18)	56·5 (21)	−67·8 (9)	87·3 (17)	−47·6 (29)	−32·5 (27)
MRF		87·8	−74·9	52·2	−66·2	87·5	−46·4	−32·1
M	8	87·0	−71·4	52·4	−65·7	85·7	−44·6	−35·6
ED	-	86·5	−70·8	52·4	−70·8	86·5	−38·3	−38·3
FF	-	82·0	−66·3	50·2	−66·3	81·7	−37·6	−38·2
Cluster 3, distorted chair ($\tau_4 \rightarrow 0$)								
SL	3	62 (3)	−79 (3)	75 (3)	−17 (5)	−54 (3)	85 (1)	−67 (3)
CL	3	62 (3)	−79 (3)	75 (3)	−17 (5)	−54 (3)	85 (1)	−67 (3)
JP2	4	63 (3)	−80 (2)	78 (4)	−25 (9)	−43 (11)	78 (7)	−67 (2)
FF (chair)		58·1	−77·5	62·7	0·0	−62·6	77·5	−58·1
Cluster 4, distorted chair ($\tau_6 \rightarrow 0$)								
CL	3	−83 (2)	67 (6)	−52 (7)	61 (5)	−60 (5)	13 (4)	52 (1)
JP2	6	−80 (2)	62 (4)	−54 (8)	59 (9)	−43 (8)	−8 (11)	64 (7)
FF (chair)		−77·5	58·1	−58·1	77·5	−62·6	0·0	62·7

Force-field calculations (Hendrickson, 1967; Ermer & Lifson, 1973; Bocian *et al.*, 1975; Allinger & Chung, 1976) consistently indicate the twist-chair as the lowest-energy conformer of the parent cycloheptane. The chair is shown (again consistently) to be the next most-favoured conformation, being only 4·2–5·9 kJ mol $^{-1}$ higher in energy. Indeed, in their electron diffraction study, Dillen & Geise (1979) found that the radial distribution function was best fitted by a twist-chair/chair mixture, with a 76 (6)% abundance of the twist-chair at 310 K. Obviously the delocalization involving N in (II) and (III) is sufficient to alter the energetics in favour of the chair and, particularly in (II), require a zero intra-annular torsion angle τ_7 (or τ_1). An initial survey of an extended data set for (I), in which fusion and bridging were permitted, reveals additional clusters corresponding to boat and twist-boat conformers. These are higher in energy by *ca* 12·6 kJ mol $^{-1}$ than the twist-chair and are presumably induced by the additional steric factors present in the extended data set.

A principal-component analysis of the 53 fragments, optimally superimposed by either the SL or CL algorithms, yields four principal components which account for 98·8% of the variance (PC1 = 55·7, PC2 = 25·7, PC3 = 11·7, PC4 = 5·7%). The 2D scattergram of PC1 *versus* PC3 (Fig. 4*a*) shows the disposition of CL clusters 1–4, with cluster 2 (twist-chair) wrapping around, and in very close proximity to, cluster 1 (chair). This situation is not surprising since these two conformations interconvert *via* a pseudorotation pathway (see *e.g.* Boessenkool & Boeyens, 1980).

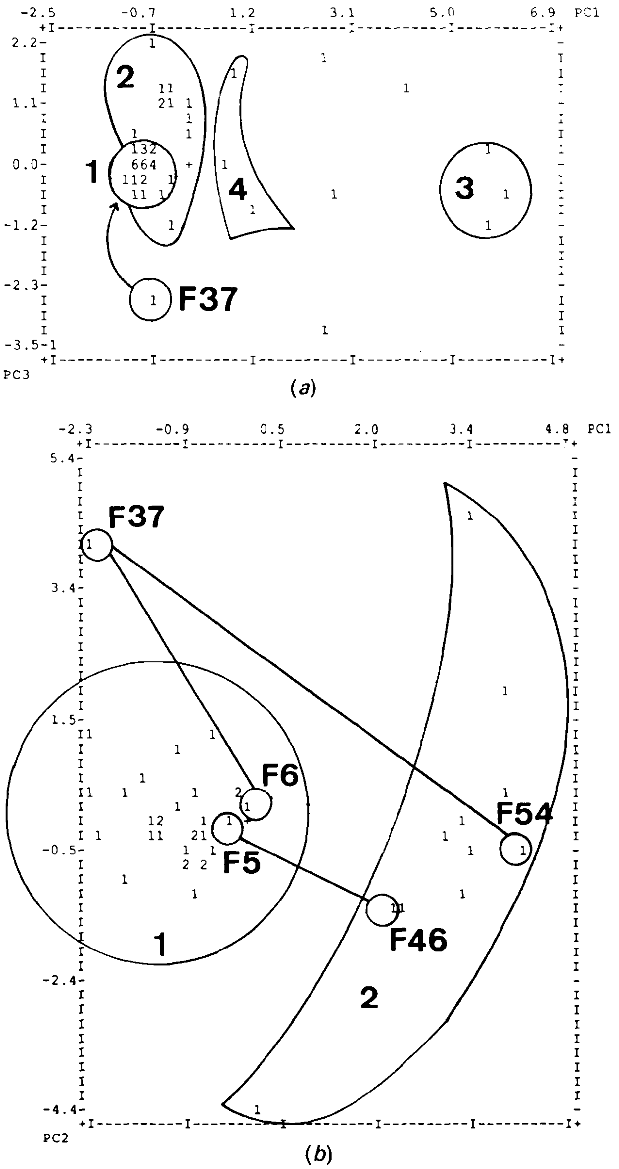
The four-dimensionality of the principal-component space, and the interconversion of seven-membered rings will be discussed more fully elsewhere (Allen & Doyle, 1991). We are more concerned here with the effectiveness of each of the algorithms, in particular with the apparently improved performance of the CL algorithm over the SL. These points are best illustrated by a PC plot (Fig. 4*b*) in which only the 42 fragments of CL clusters 1 and 2 (Table 2) are included. The omission of 11 fragments leads to a rotation of the PC axes such that PC1 *versus* PC2 affords the best resolution. The 30 chairs common to cluster 1 in both algorithms form a tight grouping in Fig. 4(*b*), with the more diffuse twist-chair cluster 2 (11 fragments in both algorithms) wrapping around it as in Fig. 4(*a*). At step 42 in the SL analysis the *minimum* available dissimilarity ($D = 0{\cdot}043^\circ$) connects fragment 5 (cluster 1) and fragment 46 (cluster 2) leading to a coalescence of clusters. This step is avoided in the CL algorithm since the *maximum* distance between any pair of entries in clusters 1 and 2 must be the next available lowest D value for coalescence to occur. The CL algorithm therefore proceeds to form cluster 4, and adds fragment 37 to cluster 1 [$D(6, 37) = 0{\cdot}075^\circ$], well before it coalesces clusters 1 and 2 at step 47 *via* $D(37, 54) = 0{\cdot}123^\circ$. Considerations such as these account for the generally higher 'stop' points given for CL clustering, and provide the basis for the chaining effect frequently observed in SL clustering and further discussed below with reference to fragment (IV).

The Jarvis–Patrick method with 'tight' constraints (JP1) identified neither of the smaller clusters 3 and 4. Fig. 4(*a*) shows these to be somewhat diffuse but well separated from clusters 1 and 2. It is obviously impossible for fragments in 3 and 4 to satisfy the JP1 clustering criterion ($C_{JP} = 6$) to enable cluster formation. Relaxation of criteria does reveal the missing clusters (JP2), whilst still preserving the overall integrity of clusters 1 and 2.

Steroid C17 *side chain* (IV)

The substructure (IV) occurs in 77 entries (CSD, July 1989) with $R \leq 0{\cdot}100$ and for which error-free coordinates are available. Short-form references are in Table 2(*b*).* This subset generates a raw data matrix of $N_t = 6$ torsion angles for $N_f = 109$ discrete fragments.

The manual classification of Duax *et al.* (1980) used 96 fragments, some of which were derived from unpublished results available in their laboratory. They classify the observed conformations into six subgroups with populations of N_p = 69, 8, 8, 5, 4 and 2 fragments. Mean torsion angles are not calculated, but complete listings were given in the paper, hence an 'idealized' summary of their results is given in Table 4(*a*). The conformers are dominated by antiplanar (*trans*, $\tau = \pm 180^\circ$) and ±synclinal (±*gauche*, $\tau = \pm 60^\circ$) torsional relationships. The vast majority of antiplanar relationships lie within the narrow range of $+165$ to -165° and are represented by a value of 180° in the idealized summary of Table 4(*a*). The ±synclinal relationships are more diffuse and are represented by a range ±(*a*–*b*) in Table 4(*a*). The



(*a*)

(*b*)

Fig. 4. Principal-component plots for the 1-azacycloheptane (I) data set. (*a*) Includes all 53 fragments and shows the SL cluster assignments at step 41, and (*b*) is derived from only those 43 fragments which occur in CL clusters 1 and 2 of Table 3. Key outliers discussed in the text are identified by fragment number in both plots.

* See deposition footnote.

Table 4. *Conformational analysis of the steroid* C17 *side chain*

τ_1–τ_6 are the torsion angles (°) as defined in (IV), N_c is a cluster identifier and N_p is the population of each cluster.

N_c	N_p	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	Range of τ_6
(*a*) Manual analysis of Duax *et al.* (1980)								
A	67	180	180	180	180	180	–	−80 to +120
A_1	41	180	180	180	180	180	–	+3 to +120
A_2	26	180	180	180	180	180	–	−12 to −105
B	7	180	180	180	+40 to +70	180	–	+42 to +85
C	6	180	180	+56 to +85	180	180	–	−52 to −89
D_1	5	180	+60 to +73	180	−67 to −99	180	–	−39 to −64
D_2	2	180	+60, 86	180	180	180	–	−3, −59
D_3	2	180	+57, +65	180	+61, +64	180	–	+56, +70
(*b*) Automated analyses using approach (i) of the text								
Single-linkage results								
A	78	180 (1)	−179 (1)	180 (1)	176 (1)	173 (1)	57 (2)	+3 to +88
B	5	−175 (2)	−172 (2)	173 (2)	74 (4)	176 (4)	63 (7)	+42 to +85
C	3	178 (3)	−170 (1)	66 (4)	−178 (3)	−179 (7)	−64 (5)	−58 to −67
D_1	3	174 (1)	63 (2)	180 (5)	−69 (1)	176 (1)	−61 (2)	−57 to −64
D_2	4	171 (2)	63 (1)	180 (2)	−178 (4)	171 (3)	−65 (4)	−59 to −76
D_3	4	176 (2)	60 (2)	176 (1)	62 (1)	180 (3)	58 (4)	+51 to +71
Complete-linkage (CL) results								
A	57	180 (1)	179 (2)	179 (1)	176 (1)	−173 (1)	62 (2)	+46 to +83
	13	−179 (1)	−169 (2)	177 (1)	−177 (2)	−176 (1)	15 (4)	−16 to +41
	10	179 (1)	−168 (1)	180 (1)	172 (3)	−164 (3)	85 (4)	+73 to +120
B	5	−175 (2)	−172 (2)	173 (2)	74 (4)	176 (4)	63 (7)	+42 to +85
C	4	177 (2)	−174 (4)	71 (5)	−178 (2)	177 (6)	−69 (5)	−58 to −84
D_1	3	174 (1)	63 (2)	180 (5)	−69 (1)	176 (1)	−61 (2)	−57 to −64
D_2	4	171 (2)	63 (1)	180 (2)	−178 (4)	171 (3)	−65 (4)	−59 to −76
D_3	4	176 (2)	60 (2)	176 (1)	62 (1)	180 (3)	58 (4)	+51 to +71
Jarvis–Patrick (JP) results								
A	61	180 (1)	180 (2)	179 (1)	175 (1)	−172 (1)	64 (1)	+48 to +85
	13	−179 (1)	−168 (2)	177 (1)	−177 (2)	−176 (1)	11 (5)	−16 to +41
B	5	−175 (2)	−172 (2)	173 (2)	74 (4)	176 (4)	63 (7)	+42 to +85
C	4	177 (2)	−174 (4)	71 (5)	−178 (2)	177 (6)	−69 (5)	−58 to −84
D_1	3	174 (1)	63 (3)	180 (5)	−69 (1)	176 (1)	−61 (2)	−57 to −64
D_2	4	171 (2)	63 (1)	180 (2)	−178 (4)	171 (3)	−65 (4)	−59 to −76
D_3	4	176 (2)	60 (2)	176 (1)	62 (1)	180 (3)	58 (4)	+51 to +71
(*c*) Automated analyses using approach (ii) of the text*								
Single-linkage (SL) results								
A_1	50	180 (1)	−171 (1)	178 (1)	177 (1)	−172 (1)	52 (4)	−16 to +88
A_2	29	180 (1)	−169 (1)	178 (1)	−175 (1)	173 (1)	−61 (1)	−48 to −82
Complete-linkage (CL) results								
A_1	43	180 (1)	−170 (1)	178 (1)	176 (1)	−170 (1)	64 (3)	+15 to +120
A_2	29	180 (1)	−169 (1)	178 (1)	−175 (1)	173 (1)	−61 (1)	−48 to −82
A_3	10	−179 (1)	−163 (10)	178 (2)	−178 (2)	−179 (1)	−7 (3)	−18 to +10
Jarvis–Patrick (JP) results								
A_1	33	180 (1)	−171 (1)	178 (1)	174 (1)	−171 (1)	68 (1)	+54 to +86
A_2	25	180 (1)	−168 (1)	178 (1)	−175 (1)	174 (1)	−61 (1)	−51 to −73
A_3	13	180 (1)	−172 (3)	177 (1)	−177 (2)	−177 (1)	9 (6)	−16 to +41

* Clusters *B*, *C*, D_1, D_2 and D_3 as for CL and JP for the automated analyses using approach (i) of the text.

data clearly define four major conformers: the fully extended (*A*), and three conformers (*B*, *C*, *D*) characterized by (+)-synclinal arrangements at τ_4, τ_3 and τ_2 respectively. The conformational group *D* was further subdivided into D_1, D_2 and D_3, for which τ_4 is (−)-synclinal, antiplanar and (+)-synclinal respectively.

The topological equivalence of the two terminal methyl groups (C26, C27) generates an ambiguity in the fragment-location process. This gives rise to the two equivalent torsion-angle sequences shown in (IV), which must be considered in our symmetry-modified cluster analysis of the raw data set. In addition to these two sequences, we must decide whether to regard mirror-image geometries of the fragment as being equivalent. Normally, we would wish to cluster mirror-image fragments together. However, this data set is special because all the compounds containing substructure (IV) are steroids with, of course, the same absolute stereochemistry. Thus, the fragment is observed in a 'constant' chiral environment which may tend to favour one mirror-image conformer (τ_1, τ_2, τ_3, τ_4, τ_5, τ_6) over its enantiomer ($-\tau_1$, $-\tau_2$, $-\tau_3$, $-\tau_4$, $-\tau_5$, $-\tau_6$). Depending on whether or not we wish to investigate this possibility, we can adopt either of two approaches:

(i) An *achiral* approach, in which we regard mirror-image geometries as equivalent. Here we must consider the permutations of (IV) together with their enantiomers. This is, in effect, the approach adopted by Duax *et al.* (1980).

(ii) A *chiral* approach, where we take account of the constant absolute stereochemistry of the steroids. Here, we consider only the permutations of (IV) and not their enantiomers.

All three algorithms were run for both methods (i) and (ii) above; clusters with population $N_p \geq 3$ were considered meaningful. Optimum clustering points for the SL algorithm were assessed as step 92 for (i) (97 fragments in six clusters), and step 93 for (ii) (100 fragments in seven clusters). The CL stop point was step 96 for both approaches with 100 [103] fragments assigned to eight [eight] clusters for (i) [(ii)]. Following our experience with the JP algorithm for 1-azacycloheptane (I), the criteria $K_{\rm JP} = 6$, $D_{\max} = 0{\cdot}10$, $C_{\rm JP} = 1$, were chosen to aid the formation of clusters with a small population. This choice assigned 94 fragments to seven clusters for (i) and eight for (ii) and, in each case, five clusters had $N_p \leq 5$. Mean torsion angles for all clusters with $N_p \geq 3$, together with the τ_6 angular range in each case, are collected in Table 4(*b*) for method (i) and in Table 4(*c*) for method (ii).

The overall success of both methods of approach can be seen clearly in Table 4. For both (i) and (ii), *all* clusters with $N_p \geq 3$ can be correlated immediately with the manual classification of Duax *et al.* (1980). Method (i) has, as expected, generated a classification identical to the published results, whilst method (ii) has yielded the expected subdivisions of conformer *A*. Groupings *B*, *C*, D_1, D_2 and D_3 are identical in all six clustering experiments (save for one 'missing' fragment in SL cluster *C* in Table 4*b*). It is only for cluster *A* that there are any gross differences between results from the three different algorithms and between results for the two approaches (i) and (ii).

A total of 82 of our 109 fragments have τ_1–τ_5 all in close proximity to 180°, conformations which may be classified as fully extended. Inspection of Table 4(*a*) shows that τ_6 in the results from method (i) is widely distributed (-16 to $+120°$) about the preferred synclinal (60°) position, as illustrated in Fig. 5(*a*). In method (ii) (Fig. 5*b*) 29 of these entries form a tight grouping (-48 to $-83°$) around the ($-$)-synclinal ($-60°$) value for τ_6; the remaining 53 are found in the same broad (-16 to $+120°$) range about the ($+$)-synclinal position noted for method (i). The CL and SL algorithms both locate all 29 ($-$)-synclinal fragments; the JP algorithm omits the three lowest τ_6 values (-48 to $-50°$) and the highest ($-83°$). All four of these entries have occasional τ_1–τ_5 values which differ by up to 20° from 180° and this is obviously sufficient for their exclusion from JP cluster A_2 (Table 4*c*).

The major differences in algorithm performance, then, lie in the ways in which the broad '($+$)-synclinal' range is subdivided in each case. These are indicated in Figs. 5(*a*) and 5(*b*). The SL algorithm, by virtue of its nearest-neighbour approach, generates clusters *A* [method (i)] and A_1 [method (ii)] with the broader τ_6 range. Essentially, the major peak at $\tau_6 \simeq 50$–$70°$ is linked with the rather smaller peak at $\tau_6 \simeq 0$–$15°$ *via* the 'chaining' effect, acting through four entries scattered in the range 15–41°. Indeed, there is a 5° discontinuity in the τ_6 distribution between 41–46° for both methods (i) and (ii). This gap is spanned by the SL algorithm, but forms a 'break point' recognized by the JP algorithm for both (i) and (ii), and by the CL algorithm in (i). Figs. 5(*a*) and 5(*b*) show that the CL algorithm is, perhaps, the least consistent for this data set. It forms an additional 'high-angle' cluster in method (i) (see Fig.

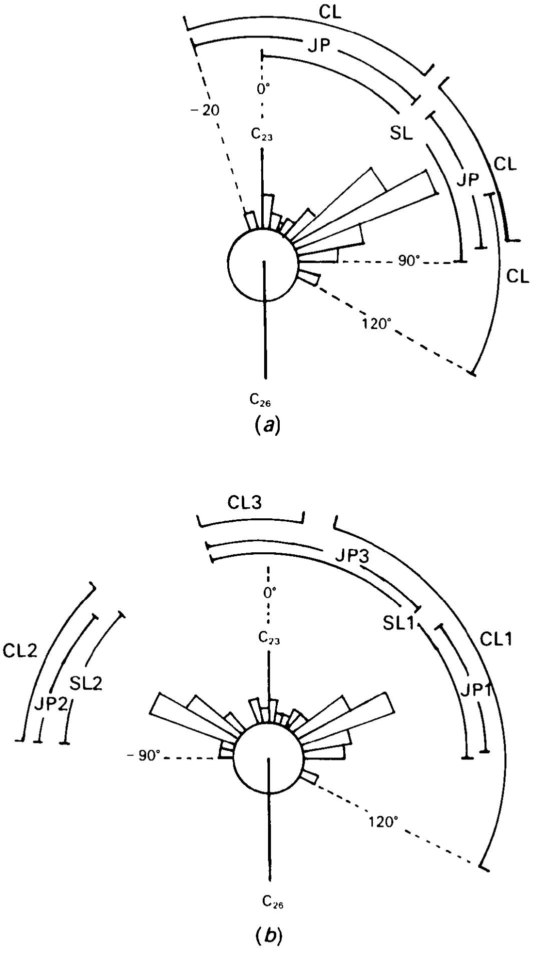


Fig. 5. Circular histograms showing the distribution of τ_6 in the fully-extended conformation (cluster *A*, Table 4) for the steroid side chain (II). (*a*) Method (i) of analysis, (*b*) method (ii). The angular composition of different subdivisions of cluster *A* generated by the three algorithms is shown in each case. This clearly illustrates the tendency of the SL algorithm to suffer from 'chaining' (see ADT2).

5*a*, Table 4*b*), which has τ_6 overlap with the main synclinal cluster. This cluster of ten fragments is associated with consistent discrepancies of τ_2 and τ_5 from 180° (Table 4*b*). It is surprising, therefore, to find a radically different CL structuring of the (+)-synclinal area in method (ii) (Fig. 5*b*, Table 4*c*). The major cluster *A* again spans the densely populated 50–70° τ_6 range, but also bridges the 41–46 and 88–116° discontinuities. The smaller A_3 cluster now spans the narrower τ_6 range of -18 to $+10°$ and is associated with a rather diffuse τ_2 distribution whose mean [$-163\,(10)°$] differs markedly from 180°.

The algorithms have generated results which are entirely sensible in chemical terms. In the achiral approach (i), all three algorithms generate a single large cluster (*A*) with a synclinal $\tau_6 \simeq 60°$. Other small subdivisions of *A* in approach (i) correspond to structures where location of C26 and C27 was complicated by high thermal motion/unresolved disorder. For the chiral approach (ii) the two major subdivisions of *A* are the (+)- and (−)-synclinal arrangements [A_1, $A_2 \equiv$ (V), (VI)]. The third subdivision detected by CL and JP clustering represent the structural anomalies noted above.

The algorithms have, therefore, succeeded in clustering those fragments which might reasonably be expected to occupy the same potential-energy well. This primary aim is achieved simply by recognizing the chemical equivalence of C26 and C27 and coding symmetry operators accordingly, as in (IV).

7. Concluding remarks

In this paper we have described the practical implementation and use of the symmetry-modified clustering algorithms of ADT1 and ADT2 in some detail. It is encouraging that all three algorithms have clearly identified the major expected conformational subdivisions for the two chosen examples. The algorithms differ only in small variations in the numbers of fragments assigned to major subdivisions, and in their ability to recognize subdivisions with low populations. The detailed analyses of each example indicate that the SL algorithm will naturally coalesce existing clusters at an earlier stage than the CL algorithm. This factor only becomes a problem if two clusters, which may be regarded as chemically discrete, are close together in conformational space, or are 'connected' in that space by a small number of mutual outliers. This SL chaining effect is recognizable by large changes in mean torsion-angle values and a rise in their e.s.d.'s. Our experiences so far indicate that, for a given data set, optimum SL clustering will occur at an earlier stage (step number) than that observed for the CL method. The furthest-neighbour approach of the CL algorithm delays the fusion of any existing cluster with either a single fragment or another existing cluster. The algorithm tends to form large numbers of clusters with low populations in its early stages, with the fusions noted above delayed to the latter stages. Chaining is usually avoided, but the delayed fusion processes can be unpredictable, as illustrated in Figs. 5(*a*) and 5(*b*) for the steroid side-chain data. Other modifications of the basic hierarchical agglomerative process are well known (Everitt, 1980). In particular, Ward's method begins in the same way as the SL method, but proceeds through a nearest-neighbour approach involving the *centroids* of clusters as they are formed. The method requires a recalculation of N_f dissimilarities after each step and is therefore more computationally intensive than either the SL or CL algorithms. We have used a gross simplification of the centroid method in our approach to the 'clustering of clusters' described in ADT2. Further experiments with modified centroid-clustering methods are in progress to provide a computationally efficient intermediate between the SL and CL algorithms.

The Jarvis–Patrick algorithm has already proved successful in the clustering of chemical compounds based on 2D chemical descriptors (Willett, Winterman & Bawden, 1986). Our work shows that the algorithm is highly effective and flexible in clustering 3D structures. Variation of the parameters $K_{\rm JP}$, $D_{\rm max}$ and $C_{\rm JP}$ provides for increased specificity of clustering which is under the control of the user. The JP results presented here and in ADT2 generally show a clustering structure which is intermediate between the hierarchical SL and CL algorithms, and which is eminently sensible in chemical terms. There remains the problem of cluster symmetrization, noted in ADT1 and ADT2. A flexible solution to this problem, applicable to all three algorithms, will be presented in a later paper (Allen & Taylor, 1991).

The algorithms studied here and described in ADT1 and ADT2 represent just three out of a myriad of approaches to the automated classification of objects on the basis of their binary or numerical attributes. The impetus for their development has come from areas as disparate as the behavioural and social sciences, biometrics, economics, *etc.*, where there is a need to extract patterns from huge volumes of multivariate data. The blind application of cluster analysis in these areas has led to some scepticism of the whole technique. Cormack (1971) introduces an excellent review of the area by stating that "The availability of computer packages of classification techniques has led to the waste of more valuable scientific time than any other 'statistical' innovation...". He points out that many data sets are homogeneous (continuous) and have no underlying cluster structure. In this case the various algorithms produce different 'dissections' of the data which, although they help to distinguish populated and empty areas of parameter space, are equivalently unsatisfactory. These strictures should not apply to the torsional data sets employed in our analyses. Here, the subgroups are most likely to be discrete, since they define conformations which are normally separated by energy barriers on the potential-energy hypersurface. The data sets have a high probability of being discontinuous in *n* dimensions and should be susceptible to clustering methods. The examples in this paper strongly support this supposition. The three algorithms have all produced 'dissections' of the data which are equivalently satisfactory in chemical terms. The only differences lie in the detailed fine structure of the results.

We have concentrated on conformational clustering for the reasons stated above, and because the results are of immediate interest, both chemically and in molecular-modelling applications. We have addressed two fundamental problems in this approach. Firstly, the well-known technical defect of chaining in the SL method (Everitt, 1980) may be avoided by use of the alternative CL and JP algorithms. Secondly, the effects of topological symmetry on the raw data set have been accounted for by major modifications to all three algorithms.

Despite this concentration on conformational aspects, we would stress that the present implementation is capable of handling data sets comprised of variables other than torsion angles, so long as all variables are expressed in the same units. Apart from gaining more extensive experience with the current implementation, further developments will be aimed at solving problems in which individual variables may have different units, *e.g.* Å and degrees. These problems require an alternative approach to the metrical basis for the calculation of dissimilarities.

We thank Dr Olga Kennard FRS for her interest in this work and the referees for suggestions which have improved its content and presentation. MJD thanks St John's College, Cambridge, and ICI Agrochemicals Division for financial assistance.

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**, 29–40.

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

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

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

Allinger, N. L. & Chung, D. Y. (1976). *J. Am. Chem. Soc.* **98**, 6798–6803.

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.

Bocian, D. F., Pickett, H. M., Rounds, T. C. & Strauss, H. L. (1975). *J. Am. Chem. Soc.* **97**, 687–695.

Boessenkool, I. K. & Boeyens, J. C. A. (1980). *J. Cryst. Mol. Struct.* **10**, 11–18.

Cormack, R. M. (1971). *J. R. Stat. Soc.* **A134**, 321–367.

CSD User Manual (1989). Version 3.4. Crystallographic Data Centre, Cambridge, England.

Dillen, J. & Geise, H. J. (1979). *J. Chem. Phys.* **70**, 425–428.

Duax, W. L., Griffin, J. F., Rohrer, D. C. & Weeks, C. M. (1980). *Lipids*, **15**, 783–792.

Ermer, O. & Lifson, S. (1973). *J. Am. Chem. Soc.* **95**, 4121–4132.

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

Hendrickson, J. B. (1967). *J. Am. Chem. Soc.* **89**, 7036–7061.

Jarvis, R. A. & Patrick, E. A. (1973). *IEEE Trans. Comput.* **22**, 1025–1034.

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.

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

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

Taylor, R. (1989). Unpublished results.

Willett, P., Winterman, V. & Bawden, D. (1986). *J. Chem. Inf. Comput. Sci.* **26**, 109–118.