Acta Cryst. (1991). B47, 41–49

Automated Conformational Analysis from Crystallographic Data. 2.* Symmetry-Modified Jarvis–Patrick and Complete-Linkage Clustering Algorithms 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

The complete-linkage and Jarvis–Patrick clustering algorithms are 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. Both algorithms are extensively modified to handle 2D topological symmetry of the fragment and the 3D conformational enantiomers which occur in crystal structures. The modified procedures ensure that symmetry equivalents of a given conformation are optimally superimposed in a unique cluster. A modified single-linkage algorithm, based on cluster centroids, is used to place the discrete clusters in a single asymmetric unit of conformational space. Principal-component analysis provides a graphical representation of the clustering process. The complete-linkage and Jarvis–Patrick algorithms may be preferable to single-linkage cluster analysis [Allen, Doyle & Taylor (1991). *Acta Cryst.* B47, 29–40] since they minimize the effects of 'chaining' *i.e.* the linkage of major clusters through a chain of outlying observations. Both of the new algorithms have been tested using a trial data set of 222 six-membered carbocycles of known conformational complexity. The new algorithms are judged to provide a more effective conformational breakdown of the trial data set (in chemical terms) than that obtained with the single-linkage method alone.

1. Introduction

Given a large number of crystallographic observations of a molecular substructure (*e.g.* taken from the Cambridge Structural Database, CSD; Allen, Kennard & Taylor, 1983), it is of considerable interest to identify any discrete conformational groups that might be present. If two or more well-characterized conformations exist, then each can be used as an energetically preferred alternative in model building. The preceding paper in this series (Allen, Doyle & Taylor, 1991*a*; hereafter ADT1) describes how a common agglomerative clustering technique, the single-linkage algorithm (see *e.g.* Everitt, 1980), may be used for this purpose. In particular, we have shown how the algorithm can be modified to take account of the topological symmetry of the fragment; many fragments of chemical interest exhibit such symmetry, and its detrimental effects on normal principal-component and cluster analyses are summarized in ADT1. We have also shown how the results of the symmetry-modified algorithm can be passed to a principal-component analysis to obtain a visual representation of all clusters within one 'asymmetric unit' of conformational space.

The symmetry-modified single-linkage algorithm was shown to work effectively on a trial data set of 222 six-membered rings, whose conformations were defined by the six intra-annular torsion angles τ_1–τ_6. However, we cannot expect that the algorithm will always be as successful, particularly when applied to molecular fragments which can adopt a large number of poorly defined conformations, since the single-linkage method is well known to suffer from a problem called 'chaining'. This is an inability to distinguish between two clusters that are connected by a chain of observations (Fig. 1). The problem is particularly likely to occur in conformational analyses, where each cluster might represent a potential-energy well and the connecting points might lie along a valley in the potential-energy hyperspace. An example of the problem is given in the next paper in this series (Allen, Doyle & Taylor, 1991*b*).

* Part 1: Allen, Doyle & Taylor (1991*a*).

† Author for correspondence.

0108-7681/91/010041-09$03.00

© 1991 International Union of Crystallography

In the present paper we discuss two alternative clustering algorithms that are known to be less prone to chaining. They are the complete-linkage method (see *e.g.* Everitt, 1980) and the Jarvis–Patrick algorithm (Jarvis & Patrick, 1973). The latter is a nonhierarchical single-step clustering technique which has been used successfully in other areas of chemistry (Willett, Winterman & Bawden, 1986). We have tested these algorithms on the trial data set of six-membered carbocycles used in ADT1; full details of the generation of this data set from the CSD, and of its chemical and conformational composition, are given there.

2. Calculation of dissimilarities

As with the single-linkage algorithm (ADT1), both the Jarvis–Patrick and complete-linkage methods require an estimate to be made of the dissimilarity between each pair of observations in the data set. We again use the Minkowski metric for this purpose, whence the dissimilarity coefficient of fragments p and q is defined as:

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

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 minimum value of $(\Delta\tau_i)_{pq}$ is taken from (2*a*) or (2*b*) as a result of the phase restriction $0 \leq |\tau_i| \leq 180°$, where the $(\tau_i)_p$ and $(\tau_i)_q$ are the torsion angles used to describe the conformations of fragments p and q respectively. In this work, the τ_i are the intra-annular torsion angles of the six-membered carbocycles which constitute the trial data set, and the integer-power n is set to 1 (city-block metric).

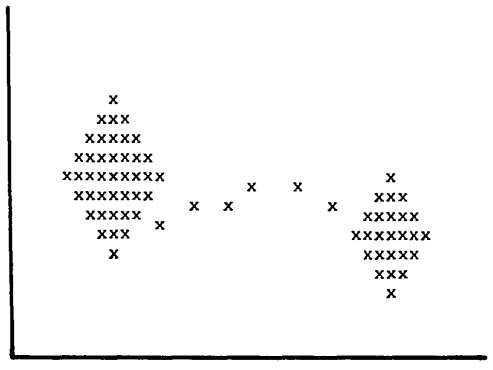

Fig. 1. A two-dimensional data set in which 'chaining' of the two major clusters may occur.

3. Unmodified Jarvis–Patrick algorithm

Nearest-neighbour table

The basic Jarvis–Patrick algorithm employs a 'nearest-neighbour table' (NN table), an array of N_f rows (N_f = number of fragments in data set) and K_{JP} columns. K_{JP} is a small, user-defined integer ($K_{JP} << N_f$). The pth row of the NN table records, in any order, the K_{JP} nearest neighbours of the pth fragment, *i.e.* the fragments deemed to be most similar to fragment p, based on the dissimilarity coefficients D^n_{pq} ($q = 1...p-1, p+1...N_f$). The NN table is an integer array, the nearest neighbours being identified by fragment number alone; *no* dissimilarity values are carried forward to the clustering process. Table 1 shows the first ten rows (fragments) of the NN table for the trial data set with $K_{JP} = 10$.

In the present application of the Jarvis–Patrick algorithm we have used the concept of an overall 'maximum dissimilarity', D_{max}. The value of D_{max} is set by the user and implies that fragment q can *never* appear in the pth row of the NN table, or *vice versa*, if $D^n_{pq} > D_{max}$. The purpose is to ensure that all fragments falling in the pth row of the NN table are genuinely similar to fragment p. In some cases, it may now be impossible to find K_{JP} nearest neighbours for some fragments (*e.g.* for outliers, which are conformationally dissimilar to all other members of the data set). In this case, the residual NN table entries are filled with zeros.

It is obvious that we need to *calculate* all of the $N_D = [N_f(N_f-1)/2]$ unique dissimilarities in assembling the NN table. However, at any stage we need only store $N_f \times K_{JP}$ of these values, where K_{JP} has been set to a maximum of 28 in the present implementation. Thus core-storage requirements are moderate. Furthermore, the summation in equation (1) can often be terminated before completion, *e.g.* as soon as the partial sum exceeds D_{max}, or is sufficiently large that neither fragment can possibly be one of the K_{JP} nearest neighbours of the other. This leads to some savings in the cpu-intensive calculation of the N_D unique dissimilarities.

Jarvis–Patrick clustering

The Jarvis–Patrick clustering technique is based on the concept of shared nearest neighbours. The fragments, p and q, are assigned to the same cluster if both of the following criteria are satisfied in the NN table:

(*a*) Fragment p is one of the K_{JP} nearest neighbours of q and fragment q is one of the K_{JP} nearest neighbours of p.

(*b*) At least C_{JP} of the K_{JP} nearest neighbours of p and q are common to both lists. The Jarvis–Patrick commonality threshold C_{JP} is specified by the user.

Table 1. *Jarvis–Patrick NN table for the first ten fragments of the trial data set of 222 six-membered rings, obtained with* $K_{JP} = 10$

Fragment (f)	The K_{JP} (= 10) nearest neighbours of (f)									
1	2	3	4	193	28	39	171	14	12	25
2	1	3	4	191	193	39	171	14	12	25
3	171	39	191	28	31	204	9	14	12	25
4	1	2	3	193	191	39	171	14	12	25
5	177	23	44	153	157	40	195	20	26	13
6	44	26	151	17	177	7	155	215	154	169
7	44	17	177	40	154	6	42	215	26	179
8	144	97	30	135	36	187	54	80	98	152
9	39	25	191	24	29	193	204	31	12	213
10	18	205	143	188	142	147	41	43	156	141

These two simple clustering rules are sufficient to enable all N_f fragments to be assigned to their respective clusters by straightforward examination of the NN table. Jarvis–Patrick clustering is a single-step procedure, in contrast to the single- and complete-linkage methods. Once the cluster membership is determined, it can be used in conjunction with the basic data matrix to generate listings of torsion angles for each conformational subgroup, together with the simple statistics detailed in ADT1.

Results from the unmodified Jarvis–Patrick method

The unmodified algorithm was applied to the trial data set, with various values of K_{JP} (the NN table width), D_{max} (the maximum dissimilarity cut-off) and C_{JP} (the commonality threshold for clustering). For $K_{JP} = 10$, $D_{max} = 0{\cdot}1$, the results for variable C_{JP}, in the range 2–9, are presented in Table 2. An increase in C_{JP} at constant K_{JP} and D_{max} obviously increases the selectivity of clustering. Thus there is a steady increase in the number of fragments in the small clusters of size ≤ 3, from 34 at $C_{JP} = 2$ to 192 at $C_{JP} = 9$, for the total data set of $N_f = 222$ fragments. Our subjective judgement (see discussion in ADT1) is that the chemical sensibility of the final clustering for $C_{JP} = 3$–6 is almost identical. Each of the 11 or 12 largest clusters represent similar expected subdivisions. It is only from $C_{JP} = 7$ onwards that these clusters become increasingly fragmented.

The clusters at $K_{JP} = 10$, $D_{max} = 0{\cdot}1$ and $C_{JP} = 5$ are summarized in Table 3. This was considered to be the optimum clustering point on chemical grounds, and by comparison with the unmodified single-linkage results (Table 4 of ADT1). In the single-linkage case, there were 38 singleton clusters with the remaining 184 fragments coalesced into 24 clusters of population ≥ 2. For the Jarvis–Patrick algorithm there are 59 singletons and the remaining 163 fragments formed only 15 clusters of size ≥ 2. The mean torsion angles for the 12 Jarvis–Patrick clusters of size ≥ 3 are given in Table 3, together with cluster numbers and sizes from the single-linkage data of ADT1.

Table 2. *Clustering ability of the unmodified Jarvis–Patrick algorithm applied to the trial data set*

The NN table length ($K_{JP} = 10$) and the overall dissimilarity cut off ($D_{max} = 0{\cdot}1$) were held constant and the commonality criterion C_{JP} allowed to vary from 2–9. N_1 N_2, and N_3 are the number of resulting clusters containing 1, 2 and 3 members. N_c is the number of clusters with ≥ 4 members, N_p^{max} is the population of the largest cluster.

C_{JP}	N_1	N_2	N_3	N_c	N_p^{max}
2	32	2	0	15	34
3	50	3	0	12	33
4	54	2	0	12	33
5	59	3	0	12	33
6	69	3	0	11	32
7	83	7	1	14	16
8	121	6	7	9	14
9	182	7	3	3	9

There are a few small but significant differences in the clustering obtained by the two unmodified algorithms. The Jarvis–Patrick method generates additional subdivisions of the phenyl and chair clusters, and also shows enhanced populations for the 'half-chair' conformations 11 and 12 (*cf.* 10 and 11 of ADT1). It fails however, to locate the very small cluster of $+-0+-0$ boat variants (cluster 6 of ADT1). Examination of the NN table showed that the three fragments forming this cluster were the *only* entries in each other's K_{JP} lists; this is a result of the $D_{max} = 0{\cdot}1$ limitation. The commonality requirement ($C = 5$) automatically excludes cluster formation in this case. Relaxation of either the C_{JP} or D_{max} settings (to $C_{JP} = 3$ or $D_{max} = 0{\cdot}25$) restores the missing cluster, but produces some undesirable side effects. Some of the symmetry variants of the half-chair/sofa conformations begin to coalesce into larger and chemically unreasonable groupings. This is undoubtedly due to the proximity in conformational space and to the low energy barriers that separate them.

4. Symmetry-modified Jarvis–Patrick clustering

Generation of the symmetry-modified nearest-neighbour table

The results from the unmodified Jarvis–Patrick algorithm suffer from the same problem as those from unmodified single-linkage clustering (ADT1), in that fragments which should be clustered together are split over several symmetry-related clusters. This arises (see ADT1) because the atom numbering of each fragment is arbitrary. Thus, in calculating D_{pq}^n from (1), the $(\tau_i)_p$ are paired with the $(\tau_i)_q$ in only one of several possible ways. Two fragments which are very similar in geometry may therefore not be recognized as such.

Our solution to the problem is identical to that used in ADT1. The dissimilarity of fragments p and q is determined by superimposing the two fragments in all possible ways (including generation of the

Table 3. *Mean torsion angles (°; e.s.d.'s in parentheses) for major clusters obtained by unmodified Jarvis–Patrick clustering of the trial data set* ($K_{JP} = 10$, $D_{max} = 0{\cdot}1$, $C_{JP} = 5$)

Here N_c^{JP} is the cluster number and N_p^{JP} is the population of the cluster obtained here, and N_c^{SL}, N_p^{SL} are the comparable values obtained by unmodified single-linkage clustering (Allen, Doyle & Taylor, 1991*a*; Table 4).

Class	N_c^{JP}	N_p^{JP}	N_c^{SL}	N_p^{SL}	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	31	1	36	0·2 (2)	−0·4 (2)	0·2 (2)	0·2 (2)	−0·4 (3)	0·2 (3)
	2	4	–	–	2·9 (11)	1·5 (6)	−4·7 (5)	3·6 (14)	0·5 (18)	−3·7 (15)
Boat	3	15	2	15	4·8 (17)	−75·5 (11)	71·4 (8)	1·1 (11)	−70·8 (10)	67·2 (16)
	4	15	3	15	0·1 (9)	65·9 (17)	−65·9 (25)	−0·4 (15)	66·7 (17)	−66·4 (22)
	5	9	4	7	65·4 (28)	0·1 (18)	−64·5 (25)	63·8 (45)	0·8 (43)	−65·5 (37)
	6	4	5	3	−69·6 (29)	74·8 (25)	−4·8 (36)	−66·6 (29)	70·0 (27)	−2·1 (32)
	7	4	7	4	−0·3 (8)	−56·8 (11)	54·5 (10)	0·5 (7)	−55·8 (10)	57·0 (8)
Chair	8	33	8	38	55·9 (7)	−54·5 (7)	53·4 (8)	−54·3 (11)	54·9 (8)	−55·5 (7)
	9	17	9	19	−55·1 (13)	53·2 (10)	−52·3 (15)	53·9 (15)	−55·3 (13)	55·4 (13)
	10	4	-		41·8 (17)	−45·1 (29)	58·6 (42)	−65·5 (34)	61·7 (45)	−51·0 (25)
Half-chair	11	9	10	7	59·5 (23)	−35·5 (38)	5·6 (29)	0·8 (27)	22·6 (46)	−52·6 (34)
	12	12	11	5	12·6 (38)	−5·5 (23)	23·2 (45)	−47·3 (37)	54·9 (25)	−37·4 (44)

mirror-image fragments if desired, see Table 3 of ADT1). A dissimilarity coefficient is calculated from (1) for each superposition and the lowest value is taken as D_{pq}^n. The dissimilarity matrix thus calculated can be used to generate an NN table, as before. Now, however, the nearest neighbours of each fragment have been determined much more reliably, since fragment overlap has been optimized.

Reorientation of fragments

It is now straightforward to determine the membership of a revised set of clusters using the symmetry-modified NN table as the new basis for Jarvis–Patrick clustering. However, in addition to determining the *membership* of each cluster, it is necessary for the symmetry-modified procedure to identify how each fragment in the cluster is to be superimposed on every other fragment. This information would be required, *e.g.* for calculating average geometries for the clusters.

In modifying the single-linkage algorithm (ADT1), we stored the symmetry operator that optimally superimposes any fragment q onto fragment p. This information was used to reorient continually fragments while clusters were grown and merged during the single-linkage process, so that at any given time the members of a cluster were all oriented correctly with respect to one another.

In principle, we could use the same approach here. However, we have chosen to implement the following simpler procedure. The Jarvis–Patrick clustering is allowed to proceed without any fragment reorientation, so that the *members* of each cluster are determined, but their optimum relative orientations are not. An arbitrary member of each cluster of size >1 is then taken as the 'cluster root'. The symmetry-optimized dissimilarity coefficients of the remaining members of the cluster are recalculated with respect to the cluster root. This calculation will superimpose each member of the cluster optimally onto the cluster root. Assuming that the cluster is reasonably homogeneous, which it will be if the cluster analysis has been successful, then each member of the cluster will be in its optimum relative orientation to every other member.

Reorientation of clusters in conformational space

The Jarvis–Patrick and single-linkage methods differ in that the latter is a stepwise, agglomerative algorithm that merges fragments or clusters one by one until all members of the data set are in the same cluster. When modified to allow for fragment symmetry, the final cluster produced by the single-linkage algorithm can be viewed as the best overlay of *all* fragments in the data set. This final cluster therefore represents a unique 'asymmetric unit' of conformational space. We have shown (ADT1) that the single-linkage method can be used to orient a set of clusters from an intermediate step of the agglomerative process so that they are in the closest mutual proximity to one another, *i.e.* intercluster dissimilarities are minimized.

The same methodology cannot be used for symmetry-modified Jarvis–Patrick analysis, since a single cluster embracing the complete data set is never formed. Whatever the values of the parameters K_{JP}, D_{max} and C_{JP}, the end point of the Jarvis–Patrick algorithm is a set of discrete clusters drawn (in general) from different asymmetric units of conformational space. It is, however, highly desirable to reorient the clusters so that they *are* in their closest mutual proximity. One possible method for this 'clustering of clusters' is to choose the largest cluster and use the reorientation method described above to overlay all other clusters onto its root. There are dangers in this approach, however, especially if the largest cluster is close to the origin of conformational space (*e.g.* phenyl rings in the trial data set). In order to be rigorous we have therefore chosen to apply the single-linkage method, with continuous reorientation, to the mean torsion angles ($\bar{\tau}_i$) of those clusters with population $N_p \geq 3$. The single-linkage

'clustering of clusters' is allowed to run to completion to generate an asymmetric unit. The cluster-overlap array (see ADT1) at the end point then indicates additional symmetry reorientations which must be applied to each of the Jarvis–Patrick clusters to bring them into a consistent asymmetric unit. The mean torsion angles are then calculated for this asymmetric unit from all clusters of population ≥ 3; clusters of population 1 or 2 are then reoriented to this mean. The final statistics, described in detail in ADT1, are generated from the fully reoriented torsional sequences of each of the Jarvis–Patrick clusters.

Numerical results from the symmetry-modified algorithm

The clustering ability of the modified Jarvis–Patrick algorithm was tested for the trial data set by varying the parameters K_{JP}, D_{max} and C_{JP}. These results are given in Table 4. The most acceptable chemical results, marked * within each subgroup, show very similar clustering structures. They vary only in the presence or absence of a number of smaller clusters ($N_p \leq 6$), and by small variations in N_p for the larger clusters. The use of symmetry-modified D''_{pq} values to assemble the NN table yields no zero-filled K_{JP} lists at $D_{max} = 0{\cdot}10$; results for $K_{JP} = 10$, $D_{max} = 0{\cdot}15$, are identical to those for $K_{JP} = 10$, $D_{max} = 0{\cdot}10$. This is not true for $D_{max} = 0{\cdot}05$ and some smaller clusters are lost at $C_{JP} \geq 5$ for the reasons noted above for the unmodified algorithm. Reduction of K_{JP} to 5 severely reduces the discriminatory power of the algorithm as C_{JP} increases. An increase in K_{JP} appears to be of no particular benefit, but optimum clustering seems to occur at higher C_{JP}/K_{JP} ratios for increasing K_{JP} and constant D_{max}. We stress that these results may only apply to this data set and are very preliminary in nature. Further results for this, and other data sets will be presented in Part 3 of this series (Allen, Doyle & Taylor, 1991*b*).

The nature of the Jarvis–Patrick algorithm requires that a subjective judgement of optimum clustering must be made from examination of a few runs using different K_{JP}, D_{max} and C_{JP} values. The results presented in Tables 2 and 4 suggest that, for reasonable settings of K_{JP} and D_{max}, C_{JP} is the most critical variable involved in cluster generation. For the trial data set the single pass with $K_{JP} = 10$, $D_{max} = 0{\cdot}10$ and $C_{JP} = 6$ was chosen as optimum for the symmetry-modified algorithm. The resultant cluster structure is summarized in Table 5, together with some comparative data from the single-linkage approach of ADT1.

The overall results of Table 5 are in excellent agreement with the single-linkage data. The major differences involve (*a*) the coalescence of the 1,2- and 1,3-diplanar conformations (single-linkage clusters 9 and 10) into Jarvis–Patrick cluster 11; (*b*) the emergence of a somewhat loose (in view of the e.s.d.'s) cluster of distorted twist-boats (cluster 12, Table 5); and (*c*) some redistribution of fragments within the boat (2–6) and within the chair (7–9) clusters.

Table 4. *Clustering ability of the symmetry-modified Jarvis–Patrick algorithm for ranges of K_{JP}, D_{max} and C_{JP}*

N_1, N_2, N_3 and N_c are the numbers of clusters containing 1, 2, 3 and ≥ 4 fragments. N_p^{max} is the size of the largest cluster. The most acceptable chemical results are marked with an asterisk.

K_{JP}	D_{max}	C_{JP}	N_1	N_2	N_3	N_c	N_p^{max}
5	0·10	1	9	5	1	12*	49
		2	16	7	4	6	19
		3	65	17	4	22	9
10	0·05	3	15	0	0	6	64
		4	27	0	0	5	57
		5	32	0	0	7	48
		6	36	1	0	9*	46
		7	52	6	1	22	14
10	0·10	3	5	1	1	5	73
		4	5	1	1	5	73
		5	6	2	0	9	48
		6	14	2	0	12*	46
		7	30	7	1	25	14
		8	72	12	5	20	9
10	0·15	5	6	2	0	9	48
15	0·10	5	3	1	0	5	73
		7	4	1	0	5	73
		9	5	1	0	6	73
		10	9	1	0	9*	56
		11	15	2	1	14	35
		12	40	5	6	19	24
20	0·10	8	1	1	0	5	74
		10	2	1	0	5	74
		12	6	1	0	6	57
		14	8	1	0	9	57
		15	10	1	0	11*	57
		16	19	4	0	15	33
		17	42	5	4	18	26

The averages of Table 5 are presented after the 'clustering of clusters' which reorients all fragments into an asymmetric unit of conformational space. The fully reoriented data matrix, with associated cluster numbers for each fragment, may be passed to the principal-component routine. Graphical output similar to that of Fig. 5 of ADT1 may then be generated.

5. Unmodified complete-linkage algorithm

The complete-linkage algorithm (see *e.g.* Everitt, 1980) is indentical to single-linkage cluster analysis (ADT1) except that, at any point in the agglomerative process, the distance between two clusters is defined as the distance between their most remote members rather than between their nearest members. The complete-linkage algorithm therefore follows exactly the path detailed in steps 0–5 of ADT1 for the single-linkage method, but with modification to steps 3 and 4. The complete process is:

Table 5. *Mean torsion angles (°; e.s.d.'s in parentheses) for major clusters obtained with the symmetry-modified Jarvis–Patrick algorithm at* $K_{JP} = 10$, $D_{max} = 0{\cdot}1$ *and* $C_{JP} = 6$ *for the trial data set*

Column headings are as for Table 3.

Class	N_c^{JP}	N_p^{JP}	N_c^{SL}	N_p^{SL}	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	35	1	35	−1·0 (2)	0·5 (2)	1·2 (2)	−2·4 (2)	2·0 (2)	−0·3 (1)
Boat	2	30	2	34	−70·9 (7)	70·5 (6)	0·1 (4)	−70·0 (6)	69·6 (6)	0·6 (5)
	3	14	3	11	−54·8 (8)	59·3 (8)	−3·7 (11)	−53·9 (13)	56·8 (9)	−2·9 (10)
	4	9	4	5	−55·3 (18)	71·6 (15)	−6·5 (12)	−66·2 (9)	83·8 (14)	−18·8 (20)
	5	6	5	4	−50·8 (13)	58·2 (21)	15·3 (39)	−84·3 (20)	93·8 (19)	−27·1 (18)
	6	4	–	–	−64·9 (9)	66·9 (13)	5·2 (15)	−76·7 (8)	79·5 (9)	−8·2 (3)
Chair	7	46	6	51	54·0 (5)	−53·2 (6)	53·7 (6)	−55·7 (8)	56·3 (7)	−54·9 (5)
	8	4	7	4	52·3 (5)	−69·9 (3)	81·2 (3)	−81·7 (7)	78·8 (4)	−58·2 (4)
	9	9	–	–	37·5 (12)	−41·1 (14)	56·5 (14)	−66·1 (19)	61·9 (17)	−48·7 (9)
Half-chair	10	29	8	26	9·2 (9)	1·2 (3)	19·2 (7)	−48·7 (7)	60·6 (8)	−39·9 (11)
Sofa and screw-boat	11	9	9	4	−3·0 (6)	15·4 (22)	7·3 (20)	−40·3 (18)	52·7 (24)	−31·4 (19)
			10	3						
Twist-boat	12	9		-	−36·6 (24)	73·3 (34)	−22·2 (26)	−50·5 (26)	89·1 (33)	−36·5 (13)

Step 0. *Calculation of dissimilarities*

The torsional dissimilarities D^n_{pq} are calculated as described above and in ADT1 [equations (1) and (2)].

Step 1. *Formation of the initial cluster*

The two most similar fragments (*a*) and (*b*), corresponding to the smallest dissimilarity coefficient, are combined to form an initial cluster of population $N_p = 2$.

Step 2. *Formation of an additional new cluster*

If D^n_{cd} is the next smallest dissimilarity and neither of the fragments (*c*) or (*d*) are members of a cluster with $N_p \geq 2$, then they are combined to form a new cluster (*c* and *d*).

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

A fragment (*c*) can only enter an existing cluster (*a*, *b*,...) if the *maximum* value of D^n_{ac}, D^n_{bc},... is still smaller than either: (i) the next available smallest D^n_{pq} value; or (ii) any available D^n_{xy} value, where one or both of *x* and *y* are clusters of population $N_p \geq 2$ (the D^n_{xy} being calculated as *maxima* as described here or at step 4 below). Fragment (*c*) enters (*a*,*b*,...) on a *furthest-neighbour* basis, which is an alternative name for the complete-linkage method.

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

Two clusters, *e.g.* (*a* and *b*) and (*c* and *d*) may merge to form a single cluster (*a*, *b*, *c* and *d*) if the *maximum* value of D^n_{ac}, D^n_{ad}, D^n_{bc} or D^n_{bd} is smaller than either (i) or (ii) at step 3 above.

Step 5. *Ending the clustering process*

Cluster formation occurs at step 1, and at every iteration of steps 2, 3 and 4 (dependent on the dissimilarity considerations described above). The process ends after $N_f - 1$ clustering steps when all fragments are in a single cluster. This exactly mirrors the single-linkage case and implementation procedures for initial cluster listings, detection and specification of a suitable stop point are exactly as described in ADT1.

The use of the furthest-neighbour criteria at steps of type 3 and 4 means that there is a tendency for these steps to be avoided in the early stages of the complete-linkage method. The algorithm therefore attempts to minimize the effects of 'chaining'. The possible functional advantages of the complete-linkage method are balanced by the fact that it is the most computationally intensive of the algorithms discussed here and in ADT1.

Results for the unmodified complete-linkage algorithm

The complete-linkage algorithm was run to completion on the trial data set and graphs of dissimilarity and dissimilarity difference *versus* step number (Figs. 2*a* and 2*b*) were used, in conjunction with a visual scan of clustering output, to select step 170 as a suitable stop point. The maximum normalized torsional dissimilarity used by the algorithm was $< 0{\cdot}079$, corresponding to a maximum mean torsion-angle difference of 2·4° for conformations assigned to the same cluster. At this point 201 fragments had been assigned to 31 clusters with $N_p \geq 2$, of which 14 had $N_p \geq 4$ and are listed in Table 6. For the single-linkage method the corresponding figures were 184 fragments in 24 clusters of $N_p \geq 2$ with 11 having $N_p \geq 4$. The single-linkage stop point (160) is ten steps below that obtained here and the number of single-linkage clusters is also lower. These differences are a direct result of the selective use of the furthest-neighbour criterion in the present algorithm.

In broad terms the subdivision of Table 6 mirrors the results of Table 4 in ADT1 and Table 3 in the present paper. There are, however, some interesting

Table 6. *Mean torsion angles (°; e.s.d.'s in parentheses) for major clusters ($N_p \geq 4$) obtained with the unmodified complete-linkage algorithm at step 170 for the trial data set*

N_c^2 is the cluster number and N_p^2 is the population; N_c^1, N_p^1 are the corresponding data for the single-linkage method (Allen, Doyle & Taylor, 1991*a*; Table 4).

Class	N_c^2	N_p^2	N_c^1	N_p^1	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	36	1	36	0·6 (3)	0·0 (3)	−0·8 (6)	1·0 (6)	−0·4 (3)	−0·4 (4)
Boat	2	19	2	15	3·7 (14)	−71·6 (20)	67·8 (18)	1·0 (9)	−67·6 (17)	65·1 (16)
	3	9	3	15	1·7 (10)	70·5 (7)	−72·4 (13)	1·4 (18)	70·6 (17)	−72·3 (8)
	4	5	4	7	70·9 (12)	−2·2 (16)	−68·9 (14)	72·8 (46)	−5·4 (45)	−64·3 (29)
	5	5	5	4	−69·0 (26)	68·3 (40)	1·3 (27)	−69·3 (4)	67·2 (30)	0·9 (20)
	6	4	6	4	64·8 (31)	−67·6 (39)	0·0 (0·9)	67·2 (36)	−64·7 (53)	−0·1 (17)
	7	7	-	-	−4·3 (23)	62·2 (36)	−57·2 (28)	−4·2 (22)	63·1 (26)	−57·5 (20)
Chair	8	34	8	38	55·7 (7)	−54·7 (6)	53·5 (8)	−54·3 (11)	54·7 (8)	−55·2 (8)
	9	17	9	19	−55·1 (13)	53·2 (10)	−52·3 (15)	53·9 (15)	−55·3 (13)	55·4 (13)
	10	4		-	41·8 (17)	−45·1 (29)	58·6 (42)	−65·5 (34)	61·7 (45)	−51·0 (25)
Half-chair	11	5	11	5	15·3 (33)	−5·6 (35)	23·9 (36)	−51·0 (31)	62·0 (19)	−43·4 (22)
	12	4	-	-	−0·9 (7)	−7·1 (48)	34·8 (62)	−53·6 (38)	45·5 (20)	−19·2 (25)
	13	4	·	·	17·1 (39)	−46·5 (32)	59·9 (9)	−40·4 (42)	9·6 (39)	2·0 (10)
Sofa	14	6	-	-	60·1 (33)	−33·4 (32)	2·1 (30)	1·1 (16)	26·9 (25)	−56·8 (32)

discrepancies: (*a*) Amongst the boat conformers there is a redistribution compared with ADT1. The smaller subgroup (4) of less puckered boats identified by the single-linkage method is now part of an enlarged complete-linkage cluster 2, whilst clusters 3 and 7 from the present algorithm are subdivisions of cluster 3 of ADT1 (Table 4). (*b*) The clustering of chair conformations here is more akin to that of the Jarvis–Patrick algorithm (Table 3 of this paper). (*c*) The distribution of the small clusters covering the more flexible intermediate forms (clusters 11–14, Table 6) is different to that obtained in either ADT1 or by the Jarvis–Patrick method.

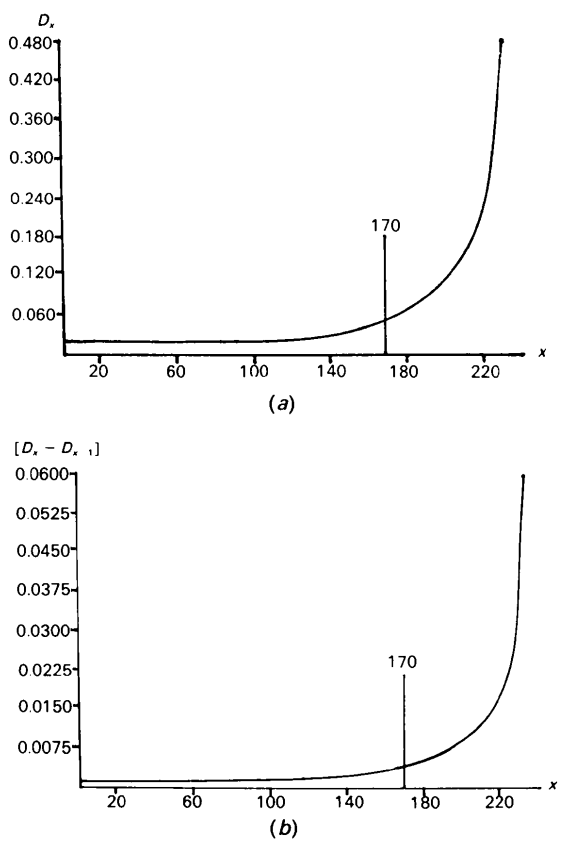

Fig. 2. (*a*) Plot of fusion dissimilarity D_x *versus* step number x, and (*b*) plot of fusion dissimilarity difference $[D_x - D_{x-1}]$ *versus* step number x for the unmodified complete-linkage results.

6. Symmetry-modified complete-linkage clustering

The complete-linkage algorithm can be modified to allow for fragment symmetry in exactly the same way as was used for the Jarvis–Patrick method, namely:

(*a*) Each pair of fragments (p and q) is superimposed in all possible ways and a dissimilarity coefficient calculated from (1) for each superposition. The lowest coefficient thus obtained is taken as D_{pq}^n.

(*b*) The new dissimilarity matrix is used to initiate complete-linkage cluster analysis. A suitable stop point in the agglomerative process is chosen, as usual.

(*c*) At the chosen stop point, the members of each cluster are oriented correctly with respect to one another by selecting an arbitrary fragment in the cluster (the 'root') and superimposing the remaining members of the cluster on this root. Difficulties may arise if the chosen root is far removed from the centroid of the cluster [these cases are dealt with in the next paper in this series (Allen, Doyle & Taylor, 1991*b*)].

(*d*) The set of clusters at the chosen stop point are placed in an asymmetric unit of conformational space by performing single-linkage analysis on the cluster means, the agglomerative process being allowed to go to completion. This determines the symmetry operations that must be performed on the various clusters to bring them into their closest mutual proximity.

Numerical results for the symmetry-modified algorithm

Plots of the dissimilarity and dissimilarity difference *versus* step number for the symmetry-modified complete-linkage run are given in Figs. 3(*a*) and 3(*b*). Consideration of these plots, and of the selected printouts of cluster membership, led to step 203 being chosen as the optimum clustering point. The maximum normalized dissimilarity was < 0·072, corresponding to a maximum mean torsion-angle difference of 2·2° for conformations assigned to the same cluster. At this stage 220 fragments had been assigned to 17 clusters with $N_p \geq 2$, of which 210 were in the 13 clusters with $N_p \geq 4$, for which data are presented in Table 7. Some other comparisons with the results from ADT1 and from this paper are given in Table 8. The complete-linkage stop point (203) is again much higher than for single-linkage (170), a further indication of the influence of the furthest-neighbour criteria at steps 3 and 4 of the complete-linkage algorithm. Table 7 should be compared with Table 5 of this paper and Table 7 of ADT1. The overall clustering structure is very similar to that generated by other algorithms, but some clusters have slightly larger populations. There are some differences in cluster assignments relating to the more flexible forms represented in Table 7 by clusters 9–13.

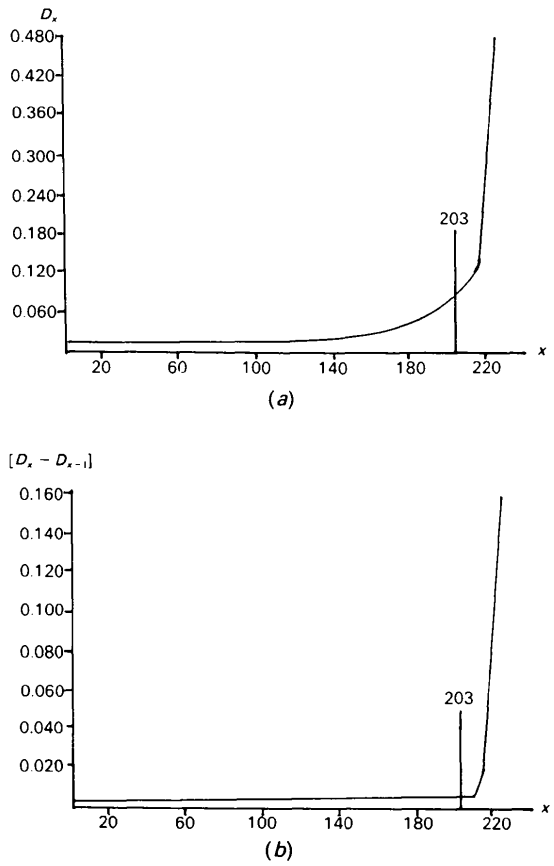

Fig. 3. (*a*) Plot of fusion dissimilarity D_x *versus* step number x, and (*b*) plot of fusion dissimilarity difference $[D_x - D_{x-1}]$ *versus* step number x for the modified complete-linkage results.

7. Discussion

Both the Jarvis–Patrick and complete-linkage methods appear to be excellent, even preferable, alternatives to the more common single-linkage technique. However, some trials with different values of the Jarvis–Patrick parameters are required to obtain optimum clustering. Applications of this algorithm to a wider variety of data sets will provide more information on suitable settings for these parameters. Detection of an appropriate STOP point remains a problem for the complete-linkage method. Graphs of the type shown in Figs. 2 and 3 can only provide a broad indication of possible settings. Other suggested indicators can be complex to program and do not have proven reliability (Everitt, 1980). We stress that 'optimal clustering' is an essentially subjective judgement, to be made primarily on the grounds of chemical sensibility. For this reason, the visual survey of cluster structures in the vicinity of an algorithmically predicted STOP point represents a vital stage in any analysis by hierarchical clustering techniques. This is equivalent to variation of clustering criteria in non-hierarchical methods, as noted above for the Jarvis–Patrick algorithm.

At this stage of development, both algorithms generate an asymmetric set of mean torsion angles for each cluster. The problem of whether to 'symmetrize' those clusters which are close to special positions in conformational space is noted in ADT1. A flexible solution, applicable to all three algorithms discussed here and in ADT1, has been developed and will be presented in a later paper (Allen & Taylor, 1991).

The trial data set of six-membered carbocyles employed here, and in ADT1, has proved a particularly interesting test of the clustering algorithms. None of the algorithms has any problem in clustering the major conformations, chair, boat and planar (phenyl), which are well separated in conformational space. The major differences, whether between different steps of the single(complete)-linkage methods, or between Jarvis–Patrick runs with different parameters, lie in the assignment of 'intermediate' conformations along the known pseudorotation and interconversion pathways (see *e.g.* Boeyens, 1978). The boat–twist-boat pathway is well represented in Table 5 of this paper by the pure boats of clusters 2 and 3, through the 'twist-boat distortion' of cluster 4, to the (albeit loose) group of

Table 7. *Mean torsion angles (°; e.s.d's in parentheses) for major clusters obtained with the symmetry-modified complete-linkage algorithm at step 203 for the trial data set*

Column headings are as for Table 6.

Class	N_c^2	N_p^2	N_c^1	N_p^1	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	1	35	1	35	1·2 (2)	0·5 (2)	−1·0 (2)	−0·3 (1)	2·0 (2)	−2·4 (3)
Boat	2	32	2	34	−71·9 (7)	71·7 (6)	−0·3 (6)	−70·4 (6)	69·6 (6)	1·2 (6)
	3	16	3	11	−54·7 (13)	57·7 (10)	−2·6 (9)	−56·0 (9)	59·7 (8)	−3·5 (9)
	4	9	4	5	−66·2 (9)	83·8 (14)	−18·8 (20)	−55·3 (18)	71·6 (15)	−6·5 (12)
	5	5	5	4	−86·3 (5)	92·0 (10)	−25·6 (11)	−52·0 (5)	56·6 (17)	17·8 (36)
Chair	6	55	6	51	51·9 (7)	−50·9 (9)	53·0 (6)	−56·3 (7)	57·5 (7)	−55·1 (5)
	7	5	7	4	50·7 (16)	−69·2 (8)	83·3 (21)	−83·3 (17)	81·5 (27)	−59·0 (9)
Half-chair	8	29	8	26	11·4 (0·6)	0·3 (5)	19·0 (7)	−48·8 (8)	61·9 (8)	−42·1 (7)
Sofa	9	4	9	4	−0·6 (21)	1·7 (17)	29·0 (16)	−59·3 (35)	61·7 (26)	−31·7 (14)
	10	4			19·3 (27)	3·8 (10)	−1·3 (6)	−24·1 (21)	47·1 (36)	−44·3 (44)
Screw-boat	11	8	10	3	5·2 (15)	17·7 (13)	−2·3 (12)	−35·2 (12)	56·4 (16)	−41·2 (16)
Twist-boat	12	4	–		−42·6 (18)	74·0 (31)	−11·3 (47)	−61·7 (48)	97·9 (35)	−34·3 (12)
	13	4	–	–	−38·0 (11)	79·6 (27)	−29·9 (18)	−44·3 (31)	86·9 (56)	−37·1 (22)

Table 8. *Comparison of clustering structures of the trial data set for the symmetry-modified single-linkage* (ADT1), *Jarvis–Patrick (this paper) and complete-linkage algorithms (this paper)*

N_1, N_2, N_3 and N_c are the number of clusters with 1, 2, 3 and ≥ 4 members. D_{max}, $\bar{\tau}_{max}$ are the maximum dissimilarity value and the corresponding maximum mean torsional difference (°) used in clustering.

Algorithm	Single-linkage	Jarvis–Patrick	Complete-linkage
Stop point	170		203
D_{max}	0·138	0·10	0·072
$\bar{\tau}_{max}$	4·1	3·0	2·2
N_1	39	14	2
N_2	3	2	2
N_3	1	0	2
N_c	9	12	13

twist-boats in clusters 12 and 13. The chair–twist-boat interconversion pathway is populated by half-chair and 1,3-diplanar (screw boat) conformations which are also spatially close to the (1,2-diplanar) sofas. Distortion of these conformations requires little energy, and it is not surprising that the algorithms have difficulty in effecting reasonable separations. It is quite pleasing that the dominant half-chair conformation for the cyclohex-1-enes of the trial data set is clearly recognized. It is doubtful, however, if any of the smaller conformational clusters, representing distortions of the dominant conformations of the subset, are either isotropic in shape or, indeed, well separated in space from each other. The assessment of intracluster and intercluster distances (dissimilarities) is obviously important in determing optimum clustering conditions. These factors are fully discussed in the next paper in this series (Allen, Doyle & Taylor, 1991*b*).

We thank Dr Olga Kennard FRS for her interest in this work and the referees for a careful reading of this paper. MJD thanks St. John's College, Cambridge, and ICI Agrochemicals Division for financial support.

References

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*a*). *Acta Cryst.* **B47**, 29–40.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*b*). *Acta Cryst.* **B47**, 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.

BOEYENS, J. C. A. (1978). *J. Cryst. Mol. Struct.* **8**, 317–320.

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

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

WILLETT, P., WINTERMAN, V. & BAWDEN, D. (1986). *J. Chem. Inf. Comput. Sci.* **26**, 109–118.