The Cambridge Structural Database in molecular graphics: techniques for the rapid identification of conformational minima

Robin Taylor

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

Techniques are described for the rapid and automatic identification of the low-energy conformations of molecular residues. The techniques use data retrieved from the Cambridge Structural Database and include reduced and marginal ordering, scattergrams of principal component scores, ordinal multidimensional scaling, Andrews' plotting and single- and complete-linkage cluster analysis. When tested on a trial dataset of 110 β-1'-aminofuranoside fragments, all of the techniques proved useful. Single-linkage cluster analysis was particularly successful.

Keywords: Cambridge Structural Database, principal component analysis, multidimensional scaling, Andrews' plots, cluster analysis

received: 5 October 1985, accepted 14 October 1985

A frequent objective in molecular graphics is to identify the low-energy conformations of a biologically active molecule. The usual procedure is as follows. A starting geometry is hypothesized and then refined so as to minimize the energy of the molecule. The energy is calculated with empirical potential energy functions and the refinement involves iterative minimization techniques such as the Newton–Raphson algorithm. The method is restricted, in that it can only locate the minimum of the potential energy well that the molecule lies in at the start of the iterative procedure, i.e. it cannot take the molecule over an energy barrier into another potential well. Consequently, it is necessary to generate several different starting geometries in order to ensure that all energetically-accessible conformational minima are identified.

This can sometimes be done with the aid of the Cambridge Structural Database[1], which contains the crystal-structure atomic coordinates of over 40 000 organo-carbon compounds. For example, suppose that the molecule contains a furanoside residue. The database can be searched for crystal structures containing this residue, and the geometry of the furanoside ring can be examined in each of these structures. If it transpires that the furanoside rings are distributed over a small number of distinct conformations, then each such conformation can be used to set up a starting geometry for the molecule being studied in the molecular graphics experiment. In this way it may be possible to identify several different minima in the potential energy hypersurface of the molecule.

A common molecular residue may easily occur in several hundred of the crystal structures in the Cambridge Structural Database. In this event, it will be impracticable to examine the geometry of the residue in each structure, individually. It is therefore necessary to develop techniques which will enable a large dataset of molecular fragments to be sorted into groups, so that fragments in the same group have the same conformation as one another, while fragments in different groups have different conformations. Ideally, such techniques will be rapid and automatic, i.e. will not require the user to have a detailed understanding of the chemical system being studied. Murray-Rust and Raftery have recently investigated a number of possible techniques[2]. This paper describes further applications of their techniques and also introduces some new methods.

TRIAL DATASET

The various techniques are tested on a trial dataset of 110 β-1'-aminofuranoside fragments (see Figure 1) retrieved from the Cambridge Structural Database. This dataset is identical to that used by Murray-Rust and Motherwell in the first published application of principal component analysis to structural data[3]. A full bibliography is given in their paper.

Previous work[3] shows that there are three physical factors governing the geometry of the β-1'-aminofuranoside residue. They are:

- the degree of ring puckering, which can be measured by the Altona–Sundaralingam puckering parameter[4] τ_m;

0263-7855/86/020123-09 $03.00

Table 1. β-1′-aminofuranoside dataset. (a) Conformational descriptors. For each fragment in dataset, table gives: index number of fragment; Cambridge Structural Database reference code[1] of crystal structure from which fragment is taken; Altona–Sundaralingam puckering parameter, τ_m[4]; Altona–Sundaralingam pseudorotational phase angle, P[4]; O(1′)-C(4′)-C(5′)-O(5′) torsion angle, *T*. All angles are in degrees. (b) Classification of fragments. The fragments are divided into groups by conformation, i.e. two fragments in the same group will have similar geometries while two fragments in different groups will have appreciably different geometries. Given for each group: index numbers of fragments contained in group; orientation of exocyclic primary alcohol group (*gg*, *gt*, *tg* indicate, respectively, O(1′)-C(4′)-C(5′)-O(5′) torsion angles of approximately −60°, +60° and 180°); range of pseudorotational phase angles spanned by fragments in group; approximate description of ring conformation[4].

(a) Conformational descriptors

Fragment	Ref. code	τ_m	*P*	*T*	Fragment	Ref. code	τ_m	*P*	*T*
1	DOCYTC	37.7	3.4	−71.6	56	MRFPUR	39.7	167.5	−64.5
2	BEURID10	40.4	3.6	−73.0	57	ASTHYM10	34.7	168.0	−68.3
3	NEBULR	38.8	3.9	−76.0	58	THPYUR	40.1	168.3	−63.9
4	GUOSBH	37.3	6.8	−57.3	59	CLURID10	35.4	168.8	−66.3
5	GUANPH	34.8	8.1	−71.8	60	DXCYTD	39.0	168.8	−57.8
6	ADURPO10	39.6	8.6	−61.9	61	ADOSHC	40.2	168.8	−59.3
7	IURIDN10	36.0	8.8	−62.6	62	MCYTMS10	38.5	169.2	−57.4
8	CYTIDI10	38.7	9.0	−70.4	63	CYTIAC	40.3	169.4	−74.7
9	APAPAD10	40.8	9.1	−63.3	64	CYTIAC01	38.7	171.7	−73.7
10	ARFUAD	40.2	9.2	−62.3	65	DOCYPO01	31.4	213.1	−62.6
11	ARFUAD01	39.6	9.3	−62.3	66	DOCYPO	32.8	213.4	−63.4
12	ANIMPH01	37.3	9.5	−73.7	67	AIPCUR01	41.5	254.0	−72.6
13	APAPAD10	39.4	9.5	−57.4	68	AIPCUR10	40.9	255.8	−73.5
14	MADENS10	37.4	11.1	−73.0	69	ABHPTB	42.3	258.9	−76.3
15	TCYTDH	38.7	11.2	−63.7	70	ABHPTB	39.1	263.3	−75.1
16	VIRAZL	39.5	11.7	−62.3	71	VIRAZL01	36.5	335.8	62.4
17	DAZADN10	41.2	11.8	−58.2	72	MADENS10	36.2	349.9	59.6
18	ADPOSM	43.9	12.2	−78.0	73	ADENOS10	36.8	6.9	60.1
19	BREDIN	38.4	12.3	−62.5	74	INOSIN10	41.8	7.8	74.8
20	DTURID	38.7	12.4	−77.2	75	DHTHUR10	37.6	9.6	73.1
21	ARATUR10	39.7	13.0	−55.2	76	HDTURD10	39.4	11.1	72.2
22	DXCYTD	37.1	13.3	−61.1	77	ADPOSD	37.4	22.9	56.7
23	SALCYS	42.2	13.7	−61.8	78	ARADEN10	37.7	24.8	62.1
24	BEURID10	42.5	13.8	−77.7	79	SDGUNP01	44.3	82.5	59.1
25	UROAME	38.8	13.9	−58.8	80	SDGUNP	40.8	83.8	62.2
26	APAPAD10	40.8	14.0	−62.5	81	BROXUR10	41.2	145.5	50.2
27	MEURID	40.6	15.2	−67.7	82	CLDOUR	43.5	146.8	50.6
28	DTURID10	39.1	15.7	−75.1	83	TRFBIM	37.1	152.0	60.0
29	AGOPCD	35.7	19.2	−67.5	84	ESMINM	44.6	152.8	61.7
30	CLPURB	35.2	19.8	−62.5	85	DHURID01	42.5	154.2	52.1
31	GUOSBH	38.3	30.8	−62.8	86	TGUANS10	37.4	157.1	64.2
32	HICYTM	37.8	32.3	−60.7	87	MARAFC	38.5	168.0	74.1
33	THIRDN10	41.0	34.1	−51.7	88	ARFCYT10	35.9	169.0	68.7
34	CYTCYP20	36.3	81.8	−60.3	89	MEYRID	37.6	169.3	61.9
35	GUANSH10	44.3	139.2	−73.8	90	DMGUAN10	36.9	173.7	67.2
36	HXURID	43.9	147.2	−65.7	91	THYDIN	37.8	187.5	56.1
37	INOSND01	44.4	147.4	−73.3	92	DOXADM	36.3	194.3	68.1
38	DADPNH10	30.7	149.9	−72.0	93	CYURID	34.7	212.5	54.6
39	INOSND10	41.3	150.4	−73.4	94	AHARFU	34.8	213.3	54.6
40	SURIDP	38.0	152.8	−77.4	95	CYURID	29.3	227.0	56.1
41	IDOXUR	38.5	153.3	−67.7	96	AHARFU	29.3	227.1	55.7
42	ADURPO10	43.4	153.4	−72.9	97	CYTCYP20	2.2	251.7	42.4
43	ACADOS	37.8	154.5	−60.6	98	TEAURP10	47.5	42.2	−176.2
44	MEYRID	37.5	156.1	−68.9	99	SCGMPT10	44.2	42.7	−175.2
45	URARAF10	39.3	156.3	−63.3	100	TEAURP10	47.2	47.9	−173.0
46	XANTOS	40.5	156.5	−66.6	101	RPPYPY20	45.8	91.3	−174.7
47	URARAF01	39.0	156.6	−63.6	102	ERFIMP	40.9	156.7	172.1
48	THPRIB	40.5	159.7	−62.7	103	FDOURD	41.6	171.1	173.2
49	GUANSH10	36.2	161.4	−50.9	104	DOURID	38.6	173.0	173.5
50	ARBCYT10	37.3	162.3	−68.4	105	IURIDN10	42.2	174.8	177.1
51	INOSND01	38.2	163.2	−54.8	106	DOURID	36.5	177.9	167.6
52	INOSND10	39.1	163.6	−55.3	107	THOPAD10	39.4	181.6	162.8
53	THPRIB	43.4	164.4	−62.5	108	URIDPS10	22.5	260.7	168.3
54	BRURID10	34.0	164.8	−66.1	109	TYMCXA	36.3	164.7	−31.6
55	NAINPH10	40.2	166.9	−62.8	110	TYMCXA	39.4	171.9	−21.0

(b) Classification of fragments

Group	Fragments	—CH_2 OH orientation	*P*(°)	Approximate description
1*	1–33	*gg*	3–34	C3′ *endo*
2	34	*gg*	82	
3*	35–64	*gg*	139–172	C2′ *endo*
4	65–66	*gg*	213	
5	67–70	*gg*	254–263	
6*	71–78	*gt*	336–360, 0–25	C3′ *endo*
7	79–80	*gt*	83–84	
8*	81–92	*gt*	146–194	C2′ *endo*
9	93–94	*gt*	213	
10	95–96	*gt*	227	
11	97	*gt*	252	
12	98–100	*tg*	42–48	
13	101	*tg*	91	
14*	102–107	*tg*	157–182	C2′ *endo*
15	108	*tg*	261	
16	109–110	∼*gg*	165–172	

*Indicates favoured conformation

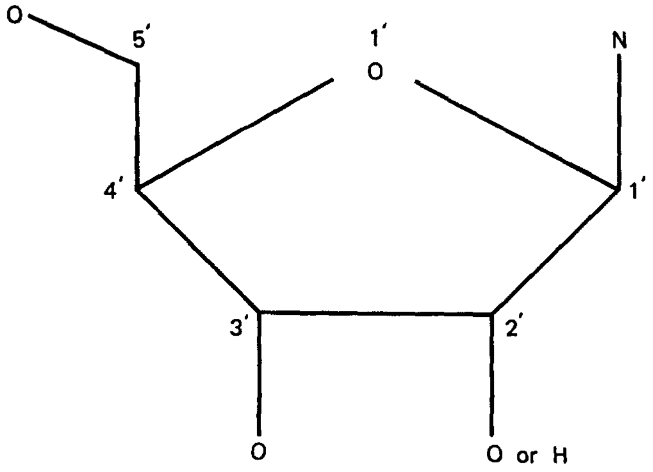


Figure 1. Structure and atom-numbering scheme of the β-1'-aminofuranoside fragment

- the direction of ring puckering, which can be measured by the Altona–Sundaralingam pseudorotational phase angle[4] P;
- the orientation of the exocyclic primary alcohol group, which can be measured by the O(1')—C(4')—C(5')—O(5') torsion angle (see Figure 1).

Using these parameters as conformational descriptors (see Table 1(a)), it is possible to sort the 110 β-1'-aminofuranoside fragments into groups according to conformation (see Table 1(b)). This manually-derived classification shows the existence of five major conformations (fragments 1–33; 35–64; 71–78; 81–92; 102–107), together with several minor ones and a number of outlying observations. The object of this paper is to establish whether the classification in Table 1(b) can be reproduced automatically (i.e. without presupposing any chemical knowledge of the β-1'-aminofuranoside residue) by various numerical and graphical techniques.

PRELIMINARY CALCULATIONS

Preamble

It is necessary to identify the regions of conformational hyperspace where the aminofuranoside fragments cluster. Methods for doing this may be classified as *graphical* or *nongraphical*. Graphical techniques aim to produce a 2D pictorial representation of the dataset (e.g. a scattergram); the natural pattern-recognition capabilities of the human eye are then used to detect clusters of observations. Since more than two parameters are needed to define the geometry of the β-1'-aminofuranoside residue (see below), graphical techniques necessitate a reduction in the dimensionality of the dataset. The first step is invariably to apply principal component analysis, which is the standard multivariate-statistical technique for reducing dimensionality[3,5,6]

Nongraphical techniques employ automatic cluster-analysis algorithms[7] to find the 'natural grouping' of a multidimensional dataset, i.e. the algorithms attempt to sort the observations into groups, so that members of a group are similar to one another but dissimilar from members of other groups. These techniques necessitate the calculation of a 'dissimilarity coefficient' for each pair of fragments, i.e. a coefficient which measures the degree of geometrical dissimilarity of the two fragments. Dissimilarity coefficients are also required by one of the graphical techniques (multidimensional scaling).

The following preliminary calculations are therefore needed. First, an initial set of parameters must be chosen to define the fragment geometries. Second, a principal component analysis must be performed. Third, a set of dissimilarity coefficients must be calculated.

Initial definition of fragment geometries[3]

The geometry of the pth fragment in the dataset can be defined initially by the thirteen torsion angles (t_{pi}, $i = 1, 2, \ldots 13$) listed in Table 2. The entire dataset is therefore described by the unstandardized data matrix $\boldsymbol{T}$, where $\boldsymbol{T}$ is of order 110 by 13 and the element in the pth row and ith column is t_{pi}. $\boldsymbol{T}$ can be converted to the standardized data matrix, $\boldsymbol{Z}$, by the transformations:

$$z_{pi} = (t_{pi} - \mu_i)/\sigma_i \tag{1}$$

where μ_i and σ_i are, respectively, the mean and standard deviation of the ith torsion angle (see Reference 3 for further details of the calculation of these quantities). Each set of values $z_{1i}, z_{2i} \ldots z_{110,i}$ has zero mean and unit variance.

Principal component analysis[3,5,6]

This section summarizes results obtained by Murray-Rust and Motherwell[3]. The objective of principal component analysis is to convert the original variables (i.e. the z_{pi}) into a new set of mutually uncorrelated variables ('principal components') which allow the dataset to be represented in the smallest possible number of dimensions. This involves the following steps. Firstly, the correlation matrix $\boldsymbol{R}$ of the z_{pi} is calculated. The element in the ith row and jth column of $\boldsymbol{R}$ is given by:

$$r_{ij} = \sum_{p=1}^{N} z_{pi} z_{pj}/(N - 1) \tag{2}$$

where N = number of observations in dataset = 110. The next step is to calculate Λ, the vector of eigenvalues of $\boldsymbol{R}$, i.e.

$$\Lambda^{\mathrm{T}} = (\lambda_1\ \lambda_2 \ldots \lambda_{13}) \tag{3}$$

Table 2. Torsion angles used in initial definition of geometry of *p*th fragment (see Figure 1).

t_{p1}	C(1')-C(2')-C(3')-C(4')
t_{p2}	C(2')-C(3')-C(4')-O(1')
t_{p3}	C(3')-C(4')-O(1')-C(1')
t_{p4}	C(4')-O(1')-C(1')-C(2')
t_{p5}	O(1')-C(1')-C(2')-C(3')
t_{p6}	N-C(1')-O(1')-C(4')
t_{p7}	N-C(1')-C(2')-C(3')
t_{p8}	O(3')-C(3')-C(2')-C(1')
t_{p9}	O(3')-C(3')-C(4')-O(1')
t_{p10}	C(5')-C(4')-C(3')-C(2')
t_{p11}	C(5')-C(4')-O(1')-C(1')
t_{p12}	O(5')-C(5')-C(4')-C(3')
t_{p13}	O(5')-C(5')-C(4')-O(1')

where λ_i is the ith largest eigenvalue. Murray-Rust and Motherwell found that[3]:

$$\lambda_1 = 7.98, \lambda_2 = 3.03, \lambda_3 = 1.95,$$
$$\lambda_4 \simeq \lambda_5 \simeq \ldots \simeq \lambda_{13} \simeq 0 \quad (4)$$

It is also necessary to calculate $\boldsymbol{E}$, the matrix of eigenvectors of $\boldsymbol{R}$. The eigenvector in the ith column of $\boldsymbol{E}$ (i.e. e_{ji}, $j = 1, 2, \ldots 13$) corresponds to the ith largest eigenvalue, λ_i. Finally, the matrix multiplication:

$$\boldsymbol{S} = \boldsymbol{ZE} \quad (5)$$

is performed. $\boldsymbol{S}$ is a matrix of order 110 by 13, the pth row of which (s_{pi}, $i = 1, 2, \ldots 13$) corresponds to the pth fragment in the dataset, the ith column (s_{pi}, $p = 1, 2, \ldots 110$) corresponding to the ith 'principal component'. Essentially, $\boldsymbol{S}$ is a new data matrix which has been derived from the original data matrix, $\boldsymbol{Z}$, by an orthogonal rotation of the coordinate axes in 13D space. The values $s_{1i}, s_{2i} \ldots s_{110,i}$ have mean = 0, variance = λ_i. Since $\lambda_4 \simeq \lambda_5 \ldots \simeq \lambda_{13} \simeq 0$, it is evident that the fourth, fifth... thirteenth principal components do not contribute to the description of the dataset, i.e. the observed variance of the dataset can be described adequately by the first three principal components alone. The geometry of the pth fragment can therefore be represented in *three* (rather than thirteen) dimensions by the coordinates s_{p1}, s_{p2}, s_{p3}. These are called the 'first, second and third principal component scores' of the pth fragment.

Calculation of dissimilarity coefficients

The geometrical dissimilarity of the pth and qth fragments can be measured by many different coefficients. Two are used in this work. They are the 'city-block dissimilarity coefficient' and the 'Euclidian dissimilarity coefficient', defined as:

$$d(t)_{pq} = \sum_{i=1}^{13} \Delta t_{pqi} \quad (6)$$

and:

$$d(t^2)_{pq} = [\sum_{i=1}^{13} \Delta t_{pqi}^2]^{1/2} \quad (7)$$

respectively[7]. Δt_{pqi} is the minimum of $|t_{pi} - t_{qi}|$ and $360 - |t_{pi} - t_{qi}|$, this definition being necessary because of the phase restriction of torsion angles to the range $-180°$ to $+180°$. Small values of $d(t)_{pq}$ and $d(t^2)_{pq}$ imply that the pth and qth fragments are similar in geometry, and *vice versa*.

GRAPHICAL TECHNIQUES FOR IDENTIFYING CONFORMATIONAL MINIMA

Preamble

Principal component analysis reduces the dimensionality of the dataset from thirteen to three, i.e. the geometry of the pth fragment can be represented by the coordinates s_{p1}, s_{p2}, s_{p3} rather than the coordinates z_{p1}, $z_{p2}, \ldots z_{p,13}$. This section examines various methods for generating a 2D pictorial representation of the 3D data. Of course, it would be possible to plot the 3D data directly on a stereo computer-graphics device. However, the generation of 2D representations is considered here in order to illustrate the various techniques that can be used in the general case when, for example, there are more than three significant principal components.

Reduced ordering

The most compact representation of the dataset is achieved by 'reduced ordering'[8,9]. If the dataset followed (exactly or approximately) a multivariate-normal probability distribution, the quantity:

$$\chi_p^2 = s_{p1}^2/\lambda_1 + s_{p2}^2/\lambda_2 + s_{p3}^2/\lambda_3 \quad (8)$$

would be chi-square distributed with three degrees of freedom. Consequently, a plot of the ordered χ_p^2 values against the expected χ^2 order statistics for a sample of size 110 from a chi-square distribution with three degrees of freedom would produce a straight line with unit slope. The observed plot is shown in Figure 2. This form of pictorial representation is too compact to be of great use, but the plot shows two interesting features. First, the marked departures from linearity — in particular, the region of the plot with very small gradient (lower left-hand corner) — imply the presence of clusters of observations. Secondly, the observations at the right-hand extremity of the plot (i.e. with large χ_p^2 values) tend to be outliers[8], i.e. fragments with atypical geometries. The ten largest χ_p^2 values correspond to the fragments numbered 69, 67, 68, 70, 108, 101, 79, 80, 100 and 98 in Table 1.

Marginal ordering

The next level of sophistication ('marginal ordering'[8]) involves plotting three normal probability plots[10]. These are of the ordered $s_{p1}/\lambda_1{}^{1/2}$, $s_{p2}/\lambda_2{}^{1/2}$ and $s_{p3}/\lambda_3{}^{1/2}$ values. For a dataset following (exactly or approximately) a multivariate-normal probability distribution, each of these plots would be a straight line with unit slope. The observed plots are shown in Figure 3. The plot of the ordered $s_{p2}/\lambda_2{}^{1/2}$ values (see Figure 3(b)) is of little use. However, the plot based on the first principal component scores (see Figure 3(a)) indicates that the distribution of these scores is strongly bimodal, and therefore allows the observations to be divided into two well defined groups. The plot based on the third component scores (see Figure 3(c)) indicates a trimodal distribution and allows the observations to be divided into three groups. By intersecting the groups found from Figures 3(a) and 3(c), it is possible to identify clearly the five important conformations found in the manually-derived classification of Table 1.

Scattergrams of principal component scores

Three 2D projections of the dataset can be obtained by plotting scattergrams of the first against the second, the first against the third, and the second against the third principal component scores. These plots, originally derived by Murray-Rust and Motherwell[3], are shown in Figure 4(a)–(c). In this Figure (and also in Figure 5; see below), different symbols are used to represent the different conformations identified in Table 1(b).

The results of the marginal ordering suggest that the most revealing projection will be obtained by plotting

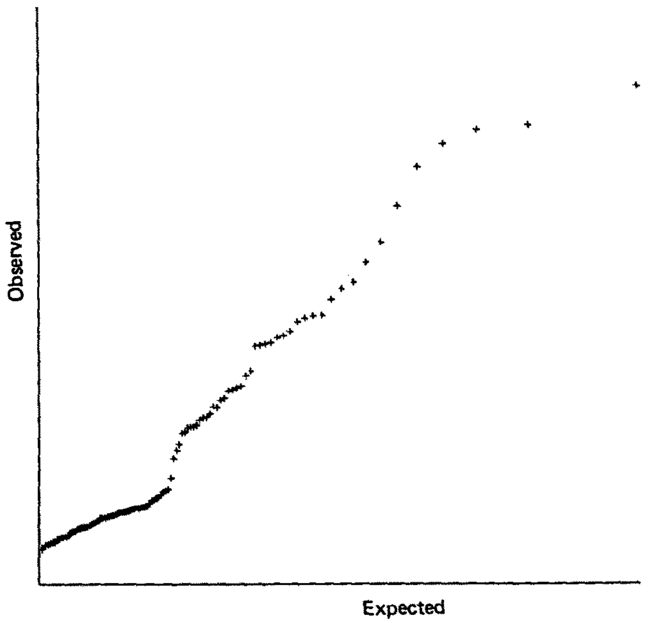


Figure 2. (above) Probability plot of ordered χ_p^2 values (see equation (8)) against expected order statistics for a sample of size 110 from a chi-square distribution with three degrees of freedom. (Scales are linear in both directions, as is the case for all other figures)

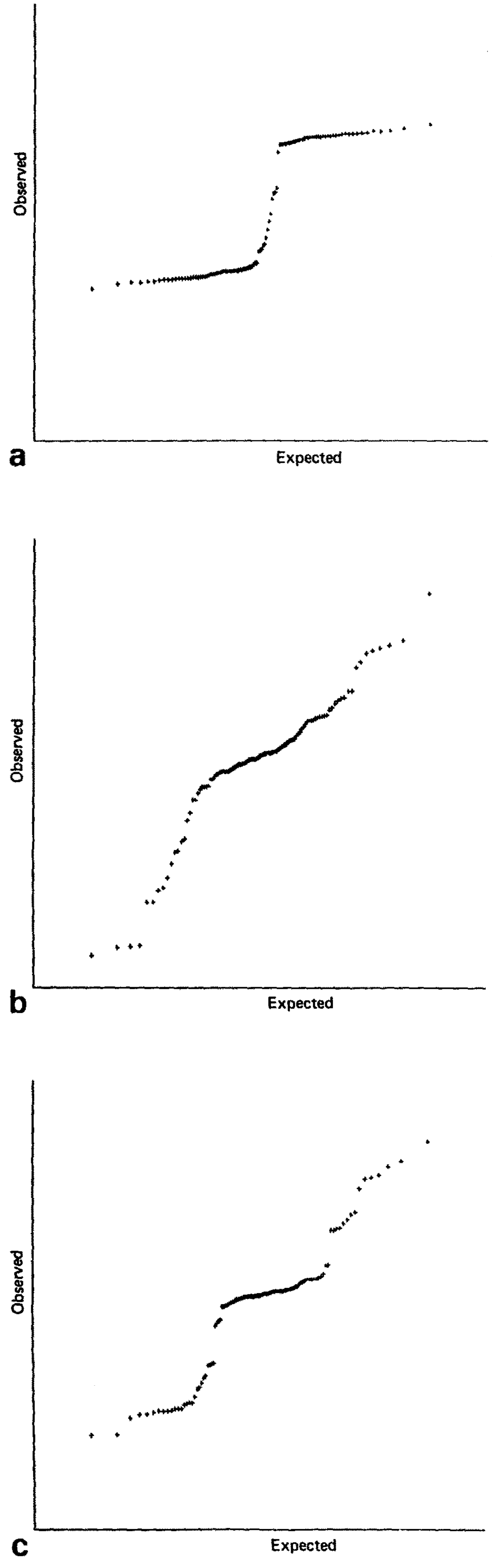


Figure 3. (right) Normal probability plots of (a) ordered $s_{p1}/\lambda_1^{1/2}$ values, (b) ordered $s_{p2}/\lambda_2^{1/2}$ values and (c) ordered $s_{p3}/\lambda_3^{1/2}$ values

the (s_{p1}, s_{p3}) values. This is confirmed by Figure 4(b), which provides an excellent pictorial representation of the dataset: fragments with the same conformation occur on the same part of the plot while fragments with different conformations are well separated and can be distinguished easily. The other two scattergrams are unsatisfactory because they do not allow the major conformations to be clearly identified (Figure 4(a) and 4(c)).

Ordinal multidimensional scaling

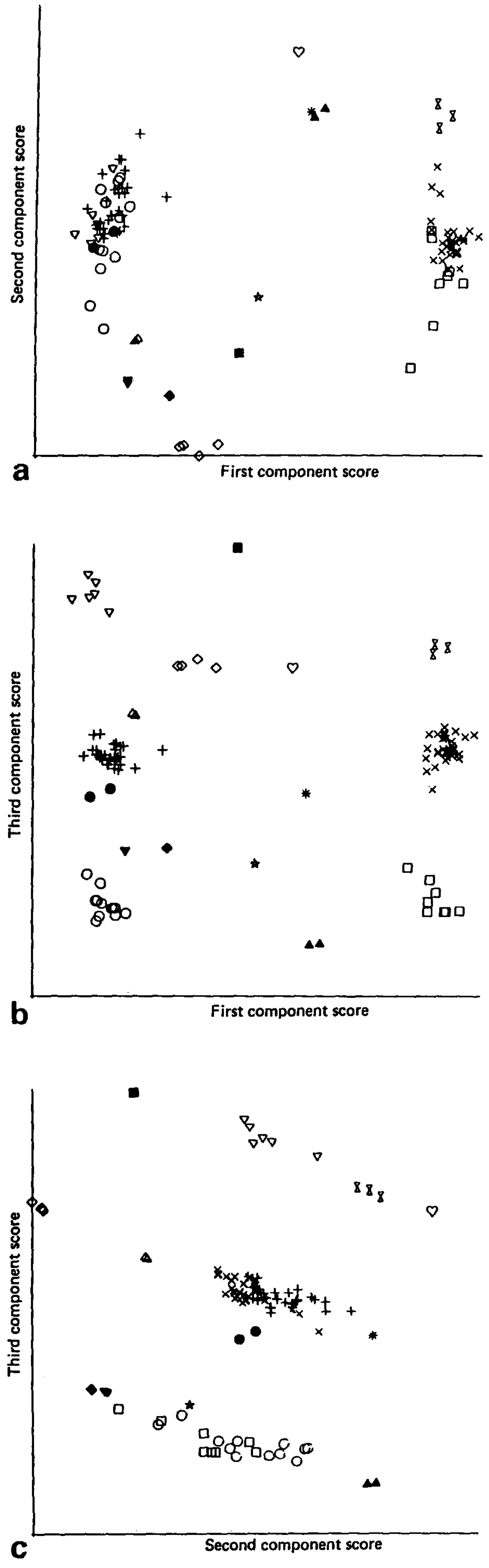
Ordinal multidimensional scaling is an iterative statistical technique for finding the 'best' n^*D representation of an nD dataset ($n^* < n$)[11]. The procedure is as follows. A starting configuration for the n^*D representation is generated (in this case, $n^* = 2$ and the scattergram shown in Figure 4(a) is a suitable starting point). This initial representation of the dataset is then improved by minimising the quantity:

$$\text{Stress} = \sum_{p<q} (D_{pq} - \delta_{pq})^2 / \geqslant_{p<q} D_{pq}^2 \qquad (9)$$

Here, D_{pq} is the distance between the pth and qth fragments in the 2D representation. The δ_{pq} are 'rank images', i.e. quantities chosen to be as close as possible to the D_{pq}, subject to the constraint that they must be monotonic with the Euclidian dissimilarity coefficients, $d(t^2)_{pq}$. Thus, if:

$$d(t^2)_{pq} < d(t^2)_{rs} \qquad (10)$$

then:



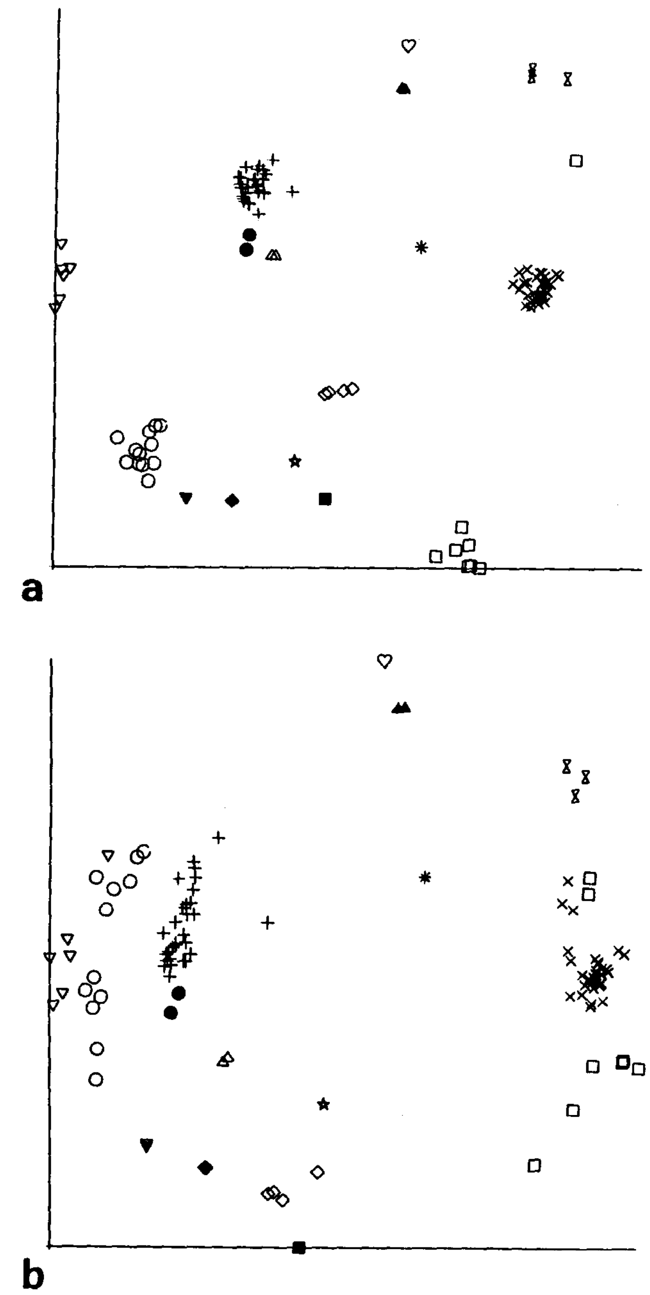


Figure 4. (left) Scattergrams of principal component scores; (a) First component score plotted against second; (b) first component score plotted against third; (c) second component score plotted against third. Key to symbols (see Table 1(b)):

×	✳	+	△	◇	□	▲	○
1	2	3	4	5	6	7	8
▼	◆	☆	⧗	♡	▽	■	●
9	10	11	12	13	14	15	16

Figure 5 (above) Scattergrams obtained by applying ordinal multidimensional scaling to the representation shown in Figure 4(a); (a) Refinement based on Euclidian dissimilarity coefficients. (b) Refinement based on city-block dissimilarity coefficients. Key to symbols (see Table 1(b)):

×	✳	+	△	◇	□	▲	○
1	2	3	4	5	6	7	8
▼	◆	☆	⧗	♡	▽	■	●
9	10	11	12	13	14	15	16

$$\delta_{pq} \leqslant \delta_{rs} \qquad (11)$$

where p, q, r, s are four fragments in the dataset.

Minimization of stress causes the starting configuration of Figure 4(a) to refine to the configuration shown in Figure 5(a). This representation of the dataset is a considerable improvement on the original, since it is now possible to identify the five important aminofuranoside conformations. However, the representation is not perfect: for example, one of the fragments belonging to group 6 of Table 1(b) has separated from the rest.

When the rank images are chosen to be monotonic with the city-block dissimilarity coefficients (i.e. the $d(t)_{pq}$) rather than the Euclidian dissimilarity coefficients, the results are not as satisfactory. Figure 4(a) now refines to the configuration shown in Figure 5(b), which is clearly unsuitable for distinguishing between fragments with different conformations. Thus, ordinal multidimensional scaling can fail in some circumstances. This is because the iterative minimization algorithm is not guaranteed to find the global (rather than a local) minimum of stress. It is theoretically possible to overcome this problem by using several different starting configurations[6].

Andrews' plots

Andrews' plotting is a simple technique for obtaining a pictorial representation of a multivariate dataset[6,9,12]. For the pth observation of an nD dataset, the Andrews' function, $f_p(\theta)$, is periodic in the range $-\pi \leqslant \theta < \pi$ and is defined as:

$$f_p(\theta) = s_{p1}/\sqrt{2} + s_{p2}\sin\theta + s_{p3}\cos\theta + s_{p4}\sin(2\theta) + \ldots \qquad (12)$$

where $s_{p1}, s_{p2}, \ldots s_{pn}$ are the coordinates of the pth observation. Only the first three terms of the series are needed for the aminofuranoside dataset. Andrews has shown[12] that for the pth and qth observations, the quantity:

$$\Delta = \int_{-\pi}^{\pi} [f_p(\theta) - f_q(\theta)]^2 d\theta \qquad (13)$$

is proportional to the squared Euclidian distance between the observations in the original multidimensional space. Consequently, aminofuranoside fragments with similar conformations should have similar Andrews' functions. If the Andrews' functions of several fragments are plotted on the same diagram, it should therefore be possible to distinguish between fragments with different conformations.

One of the problems with the technique is that only a relatively small ($< \approx 20$) number of functions can be plotted on the same diagram before it becomes too confused to be of use. In order to illustrate the method, it is therefore necessary to select an arbitrary subset of fragments from the total dataset of Table 1. This can be done with a pseudo-random number generator. The Andrews' functions of fifteen fragments chosen in this way are plotted in Figure 6(a). The plot is very revealing. One of the fragments is obviously different in geometry from the remainder (it is actually the fragment numbered 101 in Table 1). The other fourteen fragments are divided in equal proportions over two conformations (they are the fragments numbered in Table 1 as 1, 5, 7, 10, 22, 24, 29; and 38, 45, 47, 49, 55, 56, 62).

A similar set of functions, for a different randomly-chosen subset of fragments, is shown in Figure 6(b). This diagram is more difficult to interpret because the fragments in the subset are distributed over several conformations. Nevertheless, it is still possible to distinguish between the different conformations represented on the plot (referring to the numbering scheme of Table 1, the fragments plotted in Figure 6(b) are: 5, 13, 16, 17, 19, 21, 22; 44, 45, 60; 79, 80; 82, 85; 93).

NONGRAPHICAL TECHNIQUES FOR IDENTIFYING CONFORMATIONAL MINIMA

Single-linkage cluster analysis

For a multidimensional dataset containing N observations, the procedure in single-linkage cluster analysis is

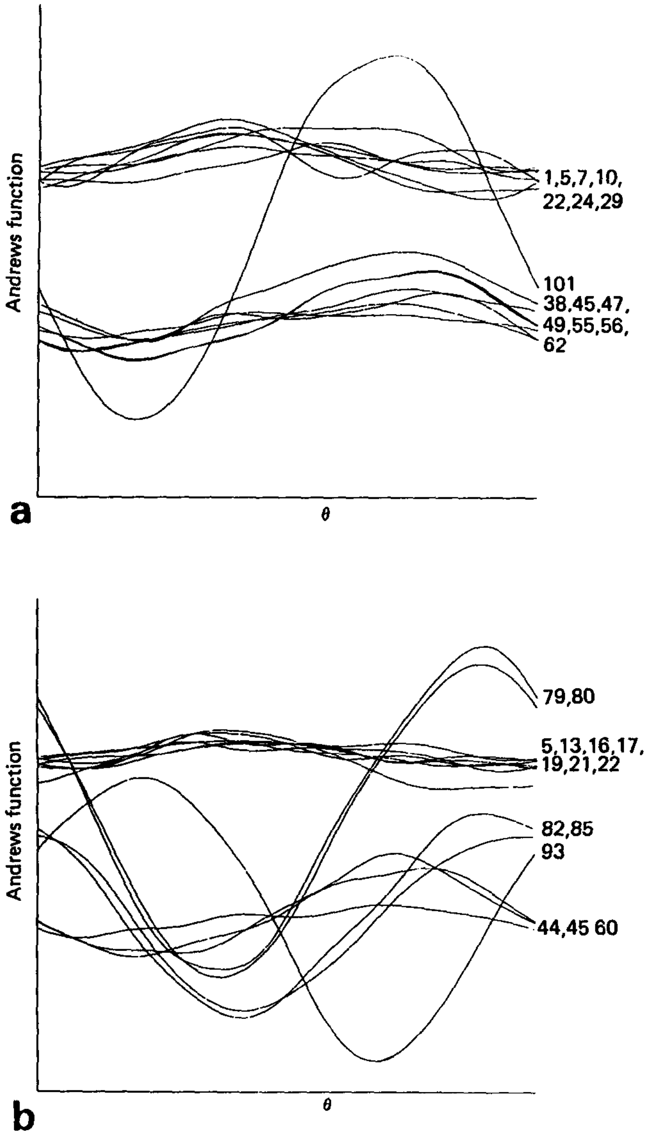


Figure 6. Andrews' functions of (a) fragments 1, 5, 7, 10, 22, 24, 29, 38, 45, 47, 49, 55, 56, 62, 101 of Table 1, and (b) fragments 5, 13, 16, 17, 19, 21, 22, 44, 45, 60, 79, 80, 82, 85, 93

as follows[7]. The dataset is initially divided into N clusters, each cluster containing one observation. The two observations with the smallest dissimilarity coefficient (i.e. the two most similar observations) are then combined into a single cluster, so that there are now $N - 1$ clusters, one of which contains two observations. Let these observations be p and q. The dissimilarity coefficient of the new two-membered cluster and any other observation, r, is now defined as the *minimum* of d_{pr} and d_{qr}, where these quantities are, respectively, the dissimilarity coefficients of observations p and r and observations q and r. In the next step, the two clusters with the smallest dissimilarity coefficient are combined; these will be two individual observations or an individual observation and the two-membered cluster formed in step 1. There are now N-2 clusters. This fusion of clusters is continued step by step until all of the observations are in the same cluster. At any stage in the analysis, the dissimilarity coefficient of two clusters is set equal to the dissimilarity coefficient of their *nearest* members. When applied to the aminofuranoside data, using the Euclidian dissimilarity coefficients $d(t^2)_{pq}$, the results are as shown in Figure 7 and Table 3. Figure 7 shows the dissimilarity coefficient of the clusters being combined against the step in the cluster analysis procedure. The 'optimum' grouping of the dataset is achieved at step 96, since the dissimilarity coefficients of the clusters being combined rise steeply after this step. The clusters found by the algorithm at step 96 are given in Table 3. Comparison with Table 1(b) shows that the automatic cluster-analysis procedure has reproduced the manually-derived classification almost exactly. Essentially the same results are obtained if the single-linkage cluster analysis is based on city-block dissimilarity coefficients instead of Euclidian dissimilarity coefficients.

Complete-linkage cluster analysis

Complete-linkage cluster analysis[7] is identical to single-linkage cluster analysis except that, at any stage in the analysis, the dissimilarity coefficient of two clusters is defined as the dissimilarity coefficient of their *most remote* pair of members. For example, if the two-membered cluster formed in step 1 contains observations p and q, then the dissimilarity coefficient of this cluster and any other observation, r, is the *maximum* of d_{pr} and d_{qr}.

When applied to the aminofuranoside dataset, using the Euclidian dissimilarity coefficients $d(t^2)_{pq}$, the results of complete-linkage cluster analysis are as shown in Figure 8 and Table 4. The Figure suggests that the optimum grouping of the dataset is achieved at about step 99; the clusters found by the algorithm at this step are given in Table 4. Comparison with Tables 1 and 3 shows that complete-linkage cluster analysis does not reproduce the manually-derived classification as well as single-linkage cluster analysis, though the results are still good. Complete-linkage cluster analysis based on city-block dissimilarity coefficients also produces satisfactory results.

Table 3. 'Optimum' set of clusters found by single-linkage cluster analysis. (Fragments are numbered as in Table 1.)

Cluster	Fragments
1	1–33
2	34
3	35–64, 109–110
4	65–66
5	67–70
6	71–78
7	79–80
8	81–92
9	93–96
10	97
11	98–100
12	101
13	102–107
14	108

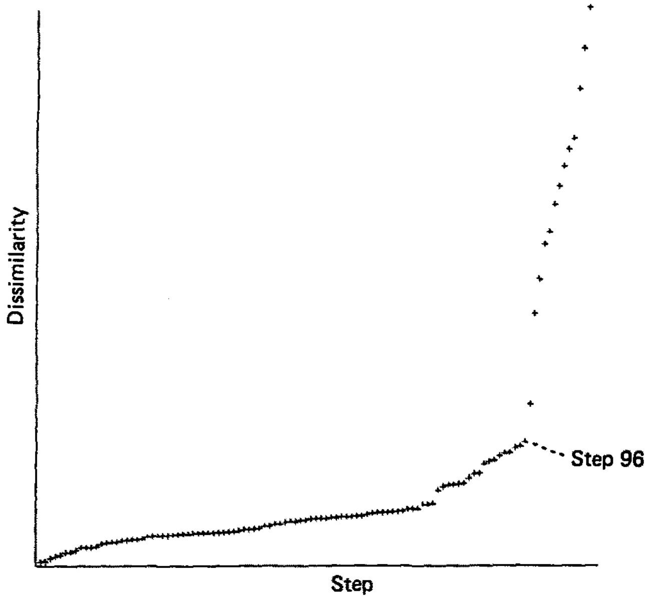


Figure 7. Plot of dissimilarity coefficient against step in cluster analysis procedure for the single-linkage cluster analysis based on Euclidian metrics

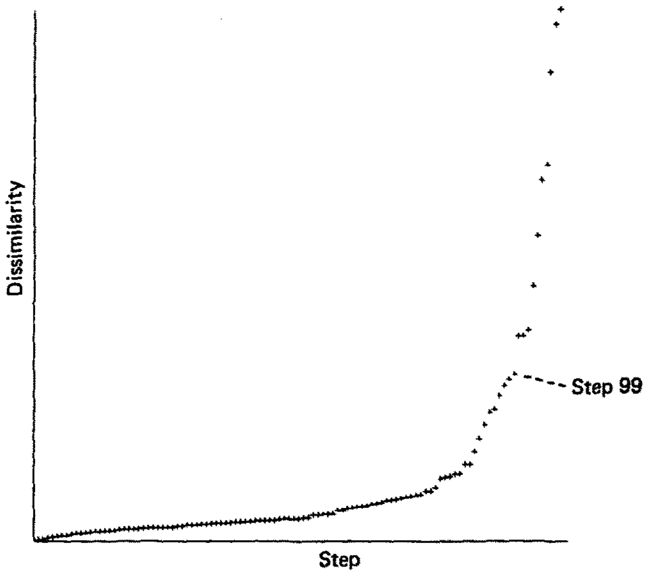


Figure 8. Plot of dissimilarity coefficient against step in cluster analysis procedure for the complete-linkage cluster analysis based on Euclidian metrics

Table 4. 'Optimum' set of clusters found by complete-linkage cluster analysis. (Fragments are numbered as in Table 1.)

Cluster	Fragments
1	1–33
2	34
3	35–64, 109–110
4	65–70
5	71–78
6	79–80
7	81–92
8	93–97
9	98–101
10	102–107
11	108

SUMMARY

Of the various graphical methods discussed in this paper, reduced and marginal ordering appear to be the least useful. However, neither technique requires much computer time and they are probably worthwhile for preliminary screening of datasets (e.g. for outliers). The Andrews' method is also valuable for preliminary screening, especially of small datasets. The plotting of principal component scores in 2D scattergrams is very successful here, as noted previously by Murray-Rust and Motherwell[3]. Those scattergrams which are unsatisfactory as pictorial representations of the dataset can, in favourable circumstances, be improved by ordinal multidimensional scaling. Of the nongraphical techniques, complete-linkage cluster analysis is moderately successful in this work and single-linkage cluster analysis produces almost perfect results. Other research groups have also obtained satisfactory results with various cluster analysis algorithms[2,13].

All of the techniques discussed in this paper attempt to find an 'optimum' low-dimensional representation of a multivariate dataset. Since this must entail some loss of information, none of the techniques is guaranteed to be successful. Nevertheless, the results obtained here are encouraging and suggest that further trials on more complicated molecular residues are worthwhile. It is noted that the techniques do not require large amounts of computer time and are therefore feasible in a real-time environment.

All calculations were performed with subroutines written by the author, most of which are available as part of the Camal subroutine library*[14].

REFERENCES

1 **Allen, F H et al.** *Acta Cryst.* Vol B35 (1979) p 2331

2 **Murray-Rust, P and Raftery, J** *J. Mol. Graph.* Vol 3 (1985) p 50

3 **Murray-Rust, P and Motherwell, S** *Acta Cryst.* Vol B34 (1978) p 2534

4 **Altona, C and Sundaralingam, M** *J. Am. Chem. Soc.* Vol 94 (1972) p 8205

5 **Murray-Rust, P and Bland, R** *Acta Cryst.* Vol B34 (1978) p 2527

6 **Chatfield, C and Collins, A J** *Introduction to multivariate analysis* Chapman and Hall, UK (1980)

7 **Everitt, B** *Cluster analysis* 2nd edn, Halsted Heinemann, London, UK (1980)

8 **Barnett, V and Lewis, T** *Outliers in statistical data* Wiley, UK (1978)

9 **Everitt, B S** *Graphical techniques for multivariate data* North-Holland, New York, USA (1978)

10 **Abrahams, S C and Keve, E T** *Acta Cryst.* Vol A27 (1971) p 157

11 **Davison, M L** *Multidimensional scaling* Wiley, New York, USA (1983)

12 **Andrews, D F** *Biometrics* Vol 28 (1972) p 125

13 **Nørskov-Lauritsen, L and Bürgi, H B** *J. Comput. Chem.* Vol 6 (1985) p 216

14 **Taylor, R** *J. Appl. Cryst.* Vol 19 (1986) p 90

*Further details of this library can be obtained from The Director, Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, UK.