Computer analysis of molecular geometry,Part VI: Classification of differences in conformation

P Murray-Rust* and J Raftery†

* Glaxo Group Research Ltd., Greenford Road, Greenford, Middlesex UB6 0HE, UK
† Napier College, Colinton Road, Edinburgh EH10 5DT, UK

Techniques are outlined for the representation, classification and quantification of the differences of geometry in a set of molecules. For each molecule the atomic positions are defined by internal Cartesian coordinates determined either by least-squares fits to a common reference molecule or from the axes of inertia of each molecule. These Cartesian coordinates can be analysed by multivariate cluster and factor analytical techniques to classify the molecules into groups and to analyse the variation of geometry within the groups. The methods are applied to a set of 48 tripeptide fragments forming β-loops. It is shown that the type II turns form a class on their own, but that types I and III turns have very similar geometry and that their populations are not distinct. Within each group there is considerable variation in geometry.

Keywords: multivariate cluster techniques, factor analysis, CCDC datafile, tripeptides, β-turns

received 31 March 1985, accepted 15th April 1985

There is an increasing need for the automatic comparison of large numbers of conformations of a given molecule. This arises in at least two areas: when conformations are generated by a molecular modelling process such as conformational searches or molecular dynamics; and when comparing many experimental observations of molecules in crystals. The methods in this paper are described with relation to the latter, but are quite applicable to the former.

The crystal structures in the datafile of the Cambridge Crystallographic Data Centre (CCDC)[1,2] contain an enormous amount of information about molecular conformation, but the sheer number of entries <45 000) makes computer analysis essential for many problems. In this paper we show how statistical procedures can be used to provide an objective analysis of large sets of data on molecular stereochemistry.

We shall use techniques of multivariate statistics — cluster and factor analysis (principal components — to divide a set of data into groups and then to analyse the variation of geometry within the groups. In this paper we tackle the first problem; that of determining how many distinct conformations are found for a molecule or molecular fragment.

We can consider at least three common situations:

- •1 All molecules have nearly the same geometry; i.e. they can be represented by (small) deviations from a mean geometry. This case will be analysed in a later paper.
- • 2 The molecules in the set have several distinct conformations; ie they can be represented by clusters where the intercluster variance is much higher than the intracluster variance. Individual clusters can often be analysed as category 1.
- • 3 There may be a large, continuous, variation between widely different geometries. This situation can often be extremely difficult to analyse objectively, although the methods we describe here will be very useful in a preliminary survey of the problem. Where the variation involves complete rotation around one or more bonds, the continuum will be infinite and not amenable to the treatments described here. Several examples of how to treat such systems have been given[3].

In many systems divisions (1)–(3) will be blurred, particularly where there are, say, two fairly well defined conformations with some molecules of intermediate geometry. This is a very interesting chemical system since it represents a reaction pathway[4], but it is very difficult to identify by automatic methods. Nevertheless, in simple cases of this type, principal component analysis reveals these pathways (as in the conformations of aminoribofuranosides[5]).

Algorithmic methods for classifying molecular geometry are needed for reasons other than the sheer amount of data. Shapes are extremely hard to describe accurately, and, without quantitative descriptors, two serious, related errors can be made. These are: (i) that a single cluster may be incorrectly described as comprising two or more *distinct* conformations; and (ii) (conversely) that two or more distinct conformations may not be recognized as such and may be reported as having a common mean geometry. Moreover, the situation is not static; there are now ten times as many accurate structures as there were in 1970. Theories and classifications produced ten years ago may need modification because of the larger amount and accuracy of data

This work was carried out at: Department of Chemistry, University of Stirling, Stirling FK9 4LA, UK

available. In general this will lead to the detection of a greater number of conformations available to a molecule and to a better description of the pathways for their interconversion.

COORDINATE SYSTEMS

Choice of Coordinates

The geometry of a nonlinear molecule of N atoms can be completely defined by $3N - 6$ parameters. These are usually internal valence coordinates, ie bond lengths, bond angles and torsion angles. Occasionally other measures such as dihedral angles, non-bonded distances and interplanar distances are used as well. These valence parameters are simple to use and have proved very successful but they suffer from several disadvantages, which may introduce bias when the geometries of molecules are being compared.

The disadvantages are:

- In crystal structure analysis they are not the original refined parameters and the errors in them are not directly estimated. Thus an error in one coordinate of one atom in a molecule may affect four bond lengths, six bond angles and over 12 torsion angles resulting in complex (and usually unreported) effects on the variances and covariances of these valence parameters.
- There is no unique set of $3N - 6$ valence parameters. This is particularly serious for structures with rings since there are complex ring closure relationships even in as simple a system as a planar hexagon[6]. An arbitrary choice must usually be made and this can introduce bias into the analysis of the geometry.
- Quite large changes in the total geometry of a molecule may be produced by small changes in one or two valence parameters in the centre of the molecule,i.e. an amplification effect. For instance a 1° change in the C—O—C angle of the central glycosidic link of a disaccharide molecule produces movements of up to 0.1 Å at the ends of the molecule, but if only the central three atoms are used to describe this change much useful information will be discarded and the error may be quite large.
- Small, smooth variations of the *whole* geometry of a molecule such as out-of-plane deformations, twisting, bending, etc, are frequent. These involve small, correlated changes in valence parameters but may be very difficult to describe accurately using this coordinate system.
- Valence parameters carry the philosophy of localized valence fields into the analysis, placing emphasis on atom- and bond-centred properties. Whilst they are extremely successful in many areas, they may not necessarily be the best parameters for describing conformations.

Two other types of internal coordinates should be considered as alternatives

- 1Symmetry coordinates or normal coordinates
- 2 Cartesian coordinates related to internal molecular axes

Elsewhere we have described the use of (1) for analysing small molecular deformations[7,8]. For symmetrical molecules, or molecules slightly distorted from a symmetrical reference configuration, they are probably the method of choice. For a large, asymmetric, molecule whose normal coordinates are not known, there is no simple way of deriving the best set of symmetry coordinates. In this paper we shall discuss the use of internal Cartesian coordinates, (II) in describing and analysing molecular geometry. Some of their advantages are:

- They are very closely related to the parameters actually determined from crystal structure analysis. The errors (unless there is severe covariance between some of the original atomic coordinates from X-ray analysis) are easily estimated from the original e.s.d's of fractional coordinates.
- All the parameters are equivalent in type and units (scale). Thus covariances can easily be analysed whereas this is not so straightforward for angle/bond covariances, etc.
- The only subjective judgement is the method of choosing reference axes.
- All the atomic positions can be used (weighted if necessary) in the description of the conformation.
- With torsion angles there can be a problem of range (e.g. should they be expressed in the range −180° to 180° or 0° to 360°, or some other range chosen to make sure that a distribution is not split close to a mode?). With Cartesian coordinates this problem does not occur.

Against these advantages there are a few drawbacks:

- 1 The choice of reference axes is not always straightforward
- 2 $3N$ parameters are used to describe a problem with $3N - 6$ degrees of freedom
- 3 Geometries and deformations represented in Cartesian coordinates may not be as familiar as those represented by lengths and angles, and hence may be more difficult to visualize at first

Models and computer graphics to a large degree solve problem (3); we now describe how (1) and (2) can be tackled.

Choice of reference axes

The choice of the best way to superimpose molecules ('best fit') is subjective. The most commonly used method is that of least squares. Two molecules A and B are assigned a common set of Cartesian axes such that the sum of the squares (ΔAB) of the Euclidian distances between corresponding atoms, A_i and B_i, is minimized:

$$\Delta\mathrm{AB} = \sum_{i}^{N} (x_{\mathrm{A}_i} - x_{\mathrm{B}_i})^2 + (y_{\mathrm{A}_i} - y_{\mathrm{B}_i})^2 + (z_{\mathrm{A}_i} - z_{\mathrm{B}_i})^2 \quad (1)$$

When A and B are fairly similar this method is valuable particularly if there is evidence that the individual Euclidian distances between corresponding atoms should be normally distributed (which we should expect if the differences between A and B were totally due to random errors). Where molecules show large differences, least-squares fitting may obscure local regions of similar geometry in the two molecules. Thus if the only difference is, say, a fairly large rotation about a (central) bond, the least-squares fit may obscure this. Nevertheless, least-squares fitting should probably be the first method to be used in many problems and several authors have described the procedure and written algorithms or programs to compute the reference axes[9–13]. Some of the algorithms are iterative; at least one[12] is not; but all should find the same solution.

An alternative method is to use the axes of inertia of each molecule as a common reference frame. The original Cartesian coordinates of a molecule related to crystal axes, are translated so as to have the origin at the centroid of the molecule (weighted if necessary). Then the 3 × 3 matrix, **M**, of the second moments of the new coordinates is set up[14] and the eigenvectors found. These are the principal axes (axes of inertia) of the molecule. An alternative matrix, which gives different eigenvalues but the same eigenvectors, uses the inertia matrix **J**, where

$$\mathbf{J} = \sum^{N}(x^2 + y^2 + z^2)\mathbf{I} - \mathbf{M} \qquad (2)$$

which is the method used in GEOM78 (in the CCDC program package). A serious disadvantage of the use of axes of inertia is when the molecule has two equal (or nearly equal) moments of inertia (i.e. is a symmetric top) and the eigenvectors are indeterminate because of degeneracy. Because the moments of inertia are squared quantities, the directions of the eigenvectors relative to the original coordinates are not uniquely defined, and there are four possible solutions related by 180° rotations about the axes of inertia. When comparing two or more molecules each of these four orientations must be considered.

Least-squares fits and superposition of axes of inertia are exactly equivalent for molecules which differ in geometry by very small amounts (ie second order quantities can be neglected). For very large differences neither method is generally satisfactory, but least-squares fits should be used where molecules are likely to have a near-degenerate ellipsoid of inertia. The main disadvantage of least squares is that, when there are more than two molecules, and differences in geometry are large, then a least-squares fit of A to B and of A to C does not guarantee that the differences between B and C, measured by ΔBC, is a minimum. It is possible that an iterative compromise might be the best solution, or that the quantity (ΔAB + ΔAC + ΔBC) could be minimized. With axes of inertia this problem does not arise. An additional advantage of axes of inertia (which we shall not consider further here) is that molecules with different chemical connectivities, i.e. where there is not a complete 1:1 matching of atoms, can be related to a common frame. Both methods allow weighting of individual atoms, if necessary by their observed (Cartesian) variances.

Chirality

Before we can compare two molecules or molecular fragments, we must consider their relative chirality. There are many reasons why we may wish to invert the configuration of a molecule that has been retrieved from the datafile: the absolute chirality may not be known; the molecule studied may have opposite chirality to the reference molecule; the crystal may be racemic and contain equal amounts of molecules of different chiralities (the handedness of the molecule retrieved depends on the original authors' choice of chemical unit in the crystal); and in a small percentage of papers (usually fairly old) the authors have reported coordinates which give the wrong chirality for the molecule described in the publication. We may also wish to test how well a molecule fits a related one of opposite chirality. In general, therefore, we will test both the retrieved molecule and its enantiomer when superimposing molecules, although this need not be done automatically by the program.

Reference axes for molecules in a large dataset

For both methods we need to choose a reference molecule. It will be used to test chirality; for least-squares it is necessary for fitting each molecule, and for axes of inertia it is needed to determine which of the four axial orientations is the best. Ideally this molecule should be close to a 'mean geometry' but this is not usually known *a priori*. Until the distribution of geometries is known we choose a molecule at random, usually the first one in the set to be analysed. If, after the analysis, the reference molecule is found to be an outlier in the distribution, it may be necessary to choose an alternative and repeat the determination of internal coordinates.

CLASSIFICATION OF MOLECULAR GEOMETRY

The geometry of a molecule with N atoms, A, related to Cartesian axes, can be represented by a $3N$-dimensional vector, **r** (A). The difference between two molecules A and B is then represented by:

$$\mathbf{r}(\mathrm{AB}) = \mathbf{r}(\mathrm{A}) - \mathbf{r}(\mathrm{B}) \qquad (3)$$

The length of this vector is simply $(\Delta\mathrm{AB})^{1,2}$ (from equation (1)) so that least-squares fitting of A and B minimizes **r**(AB). We are fortunate in that, unlike many multivariate analyses (particularly in social sciences) which involve very heterogeneous parameters, there is no problem of scale in our space, and the Euclidean distance, $(\Delta\mathrm{AB})^{1,2}$, is an accurate measure of the difference in molecular geometry. It can even be given an estimated variance due to experimental error (simply the sum over all atoms in both molecules of the estimated variances in the Cartesian coordinates). If we could examine this $3N$-dimensional space directly we should have solved our problem, and whilst this is impossible for a molecule of any size, many methods of multivariate analysis have been developed which lead to a reduction in the effective dimensionality of a problem. Two of these are cluster analysis and factor analysis (principal components).

Cluster analysis

Cluster analysis has been developed (see, for example, Ref. 15 for a general review of the methods) to identify clusters of points (vectors) in multidimensional space. The technique is not very robust (i.e. the results may differ significantly according to the algorithms and the parameters employed in them), but because there is no scale problem it is an attractive method for describing variation in molecular geometries. There are several different algorithms commonly employed for cluster analysis, and the technique has not been used enough to compare them critically. Most attempt to link a point with the 'nearest' cluster, but which this is will depend on whether we measure to the nearest point in a cluster (in which case large extended clusters may be formed), or whether we measure to the centre of a cluster. The method we shall describe is among the simplest and uses Euclidian distances (which seem the most appropriate measure).

The reservation of non-robustness may not be too serious in practice. Certainly our example below shows that routine application of the algorithm gave a good separation into non-spherical clusters, one of which had some fine structure. The test of whether the analysis has been successful will always be in how easy the results are to describe in the languages of chemistry and molecular geometry. Cluster analysis may give poor results when a system of molecules has a very complex distribution of geometries.

A preliminary account of the use of cluster analysis has been given by one of us[16]. Recently Bürgi[17] has used the technique for mapping conformational interconversions in triphenyl phosphines.

Factor analysis (principal components)

Another commonly used method for reducing dimensionality is principal component analysis (often called factor analysis). We have described its use with valence coordinates[5,18], but it can also be used with Cartesian coordinates based on axes of inertia. There are two different stages at which it can be used: initial analysis of the data (in a qualitative manner); and quantitative analysis of a monomodal subset of the data. The latter application will be discussed elsewhere.

The aim of principal component analysis is to find a new set of orthogonal axes in the $3N$-dimensional space of parameters with (greatly) reduced dimensionality which explain most of the variance. Let us assume that a dataset consists of, say, three well-separated compact clusters of observations. When convential 2D scatter diagrams of the data are plotted, using the original parameters as measures, it is frequently found that the clusters overlap and that their identities are not clear. If, however, a plane (in $3N$ space) is constructed through the centroids of these clusters then most (often as much as 90%) of the variance can be explained by two new axes in this plane. When the data are now projected into this plane, most of the variance, especially that due to the clustering, is immediately apparent in a scatter diagram. In that case only two principal components would be necessary; if more than four are necessary then the value of the method drops rapidly. For three principal components the possibility of using stereo computer graphics is very attractive.

When internal Cartesian coordinates are used, the method of principal components requires modification from that described in Ref. 5. If the geometries of the n molecules are represented by a $3N \times n$ data matrix $\mathbf{X}$, we can form the covariance matrix as $\mathbf{X}^T\mathbf{X}$. Because all our parameters are on the same scale we do not need to convert this into a correlation matrix, but can directly find its eigenvalues and eigenvectors.

Because the molecules have been translated and rotated to give the best fits, six eigenvalues will be zero and the matrix will be singular, with maximum rank $3N - 6$. Further reductions of rank will occur if there are (as needed for a successful analysis) linear relationships between the parameters or if the number of molecules in the dataset is less than $3N - 6$. This problem of singular matrices is treated by computing a generalised inverse and this procedure is common to many statistical packages.

Combined cluster and factor analysis

For a distribution with many modes, cluster analysis should be the first technique used. In cases with compact, well-resolved clusters the resulting distance matrices and dendrograms should be easy to interpret. Where clusters are diffuse and tend to overlap it may be more useful to find the principal components of the data (or of a subset). A useful combined strategy is first of all to use cluster analysis to find a (small) number of group of molecules with widely differing conformations. The principal components for each of these groups in turn can then be found and scatter diagrams plotted. If there is any suspicion that some of the original clusters are weakly connected, then factor analysis and scatterplots can be carried out for selected aggregations of clusters. An example of this combined strategy is given below.

PROGRAM SYSTEMS

The standard programs (CONNSER, RETRIEVE, GEOM) available to users of the Cambridge Datafile* have been adopted to run with indexed sequential files (VAX/VMS + F77). In particular GEOM78 has been greately expanded to perform most of the operations described in this paper (especially factor analysis, but excluding cluster analysis). This new version GEOSTAT, is available from Cambridge to subscribers to the Datafile. It can be run, without alteration, as a standalone method of comparing molecular structures as long as they are available in the same format as data on the Cambridge Datafile.

AN EXAMPLE: β-LOOPS IN PEPTIDES

To test the methods described, a system has been chosen that is known to exhibit several conformations, some of which might be expected to be fairly flexible. In peptides with three or more amide groups it is possible (Figure 1) for the chain to adopt a conformation with an internal hydrogen bond (often designated $4 \rightarrow 1$, ie NH (residue 4) $\cdots$ O=C(residue 1)). On the basis of structures available in 1968, Venkatachalam[10] suggested that there were three distinct conformations of the chain, which he labelled I, II, and III. These three conformations play an important part in polypeptide and protein structures. The first two allow a chain to double back on itself and form a β-sheet; for an antiparallel sheet we may find an additional hydrogen bond, $1 \rightarrow 4$ (NH(1) $\cdots$ O=C(4)). This situation is also frequently found in cyclic peptides, particularly hexapeptides. Turn III is based on the model of the (infinite) 3/10 helix where hydrogen-bonds of the general formula $i + 3 \rightarrow i$ are repeated indefinitely.

* For details see Reference 2. Further information is available from: Cambridge Crystallographic Data Centre, Lensfield Road, Cambridge CB2 1EW, UK

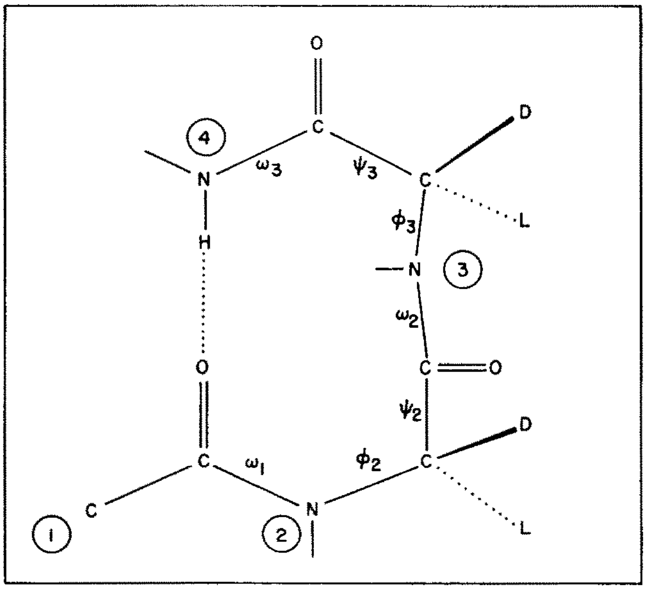

Figure 1. Parameters describing a β-loop in a tripeptide fragment. The atoms N(2) and N(3) may be substituted. Circles denote the serial numbers of the residues. The approximate values of the chain torsion angles (in degrees) for the three conformations described by Venkatachalam[10] are: $\omega_1 = \omega_2 = \omega_3 = 180$ throughout; additionally: I $\phi_2 = -60$, $\psi_2 = -30$, $\phi_3 = -80$, $\psi_3 = -10$; II $\phi_2 = -60$, $\psi_2 = 140$, $\phi_3 = 65$, $\psi_3 = 15$; III $\phi_2 = \phi_3 = 55$, $\phi_2 = \psi_3 = 35$. For a constant chirality of aminoacid sidechains, primed conformations would have opposite signs. For a constant chirality in the main chain, the unprimed conformation has L-aminoacids (i.e. substituent D is H) and the primed conformation has D-aminoacids (i.e. L = H). All twelve chain atoms (including O but not D or L) were used in the CONNSER search, and were used in GEOSTAT (see Table 1). In the analysis described in the text, a I'conformation was taken as the reference structure, so that conformations I and III have been inverted. The best fit with conformation II is found without inversion

The conformation of the chain (ie even without Cβ) lacks any symmetry in these conformations and is thus chiral. The enantiomeric conformation of the *chain* alone will be equivalent in energy, but when the aminoacid sidechains (Cβ and further atoms) are added such that the chirality at Cα is constant, the conformations are diastereoisomeric and hence unequal in energy. This is represented by unprimed and primed numbers e.g. I and I′. Equivalently if chains of constant chirality are taken and the handedness of the aminoacids is altered, the same diastereoisomerism occurs.

The most common method of describing the conformation of these loops is with the backbone torsion angles, ϕ and ψ. For a regular 3/10 helix only one value of each is needed, but for all other geometries, four angles must be considered. On the conventional Ramachandran plot this requires a pair of points and is somewhat cumbersome. Moreover there may be significant variations in ω, the torsion angle for the amide bonds, and even some small variations in some of the valence angles. Multivariate analysis of internal Cartesian coordinates may produce a simpler picture. We can also investigate the relevance of the original conformations proposed by Venkatachalam[19] which may need to be modified in the light of the larger number of peptide structures now available.

The method is exemplified by analysing crystal structures in the January 1982 version of the Cambridge Datafile.

Chirality

Many of the molecules in our dataset contain both L- and D-aminoacids, as well as the achiral glycyl and α-aminoisobutyryl fragments. We shall therefore consider backbone conformations of constant chirality (which may mean changing the chirality of some fragments retrieved by GEOSTAT). After this change of chirality we can have four possibilities: (i) both aminoacids L or one L-, one achiral; (ii) both aminoacids D- or one D-, one achiral; (iii) aminoacids of different chiralities and (iv) both aminoacids achiral. Strictly the prime notation only applies to (i) and (ii), but we shall label (i), (iii) and (iv) as unprimed and only (ii) as primed.

Retrieval of data

The program CONNSER was used to retrieve all compounds containing the fragment (Figure 1) and GEOSTAT was used to retrieve all such fragments from the Datafile. (The use of the CCDC files and programs is comprehensively described elsewhere[2]). To include prolyl residues the possibility of *N*-substitution had to be considered with the results that some other compounds, such as sarcosyl (*N*-methylglycyl) derivatives were also retrieved. All cyclic tripeptides were removed as well as three other entries: BSELGY[*] which has $R = 21\%$ and is anyway isostructural with BCPLGY; FERMAH which is essentially isostructural with ALFECB; and GPROMG which contains a UNIMOL error. The only geometrical TEST used in the FRAG routine of GEOSTAT was to limit the N(4) . . . O(1) distance to less than 3.3 Å (longer distances were arbitrarily considered not to represent a hydrogen bond). (Some cyclic tetrapeptides containing sarcosine also pass this TEST and contaminate the dataset although there is no hydrogen bond). No constraint was put on the conformation of the amide bonds at this stage. Besides genuine β-loops, therefore, there are some contaminating structures as well, particularly those with *cis*-amide linkages; reassuringly, they are clearly revealed by the cluster analysis (see below).

Reference Axes

The first fragment A retrieved by GEOSTAT was in the entry AAGAAG and this was arbitrarily taken as the reference fragment. Its axes of inertia were calculated by the procedure in GEOSTAT and coordinates for the fragment relative to these axes were taken as the reference. Only the atoms in the main chain (ie not Cβ, etc) were used in the analysis. For each subsequent fragment (B,..) the axes of inertia and Cartesian coordinates were calculated for the corresponding atoms,

[*] All compounds on the Cambridge files are identified by 6-letter codes (with 2 optional digits)

Table 1. Transformations involved in converting internal Cartesian coordinates into factor scores of the distribution in Figure 5. All these quantities are available from GEOSTAT. (a) The mean geometry of the distribution (Å) relative to axes of inertia (*X*, lowest moment; *Z*, highest moment) and its observed standard deviation (Å) are given by the 36-dimensional vectors. (b) The first vector in (a) and the following transformation matrix are used to calculate factor scores in the manner described in Table 2. (c) Components of the largest eigenvector (F_1) in Å

Atom	Coordinate (Å)	Standard deviation (Å)	Atom	Coordinate (Å)	Standard deviation (Å)	Atom	Coordinate (Å)	Standard deviation (Å)
(a)								
C(1)X	0.065	0.500	C(2')X	−0.107	0.045	N(4)X	−0.247	0.464
C(1)Y	−1.717	0.194	C(2')Y	1.344	0.135	N(4)Y	−1.458	0.091
C(1)Z	−3.216	0.255	C(2')Z	−0.070	0.078	N(4)Z	1.552	0.080
C(1')X	−0.006	0.093	N(3)X	0.784	0.086	O(1)X	−0.222	0.893
C(1')Y	−0.706	0.083	N(3)Y	0.868	0.244	O(1)Y	−1.039	0.112
C(1')Z	−2.260	0.095	N(3)Z	0.704	0.218	O(1)Z	−1.424	0.086
N(2)X	0.210	0.328	C(3)X	0.604	0.331	O(2)X	−1.101	0.251
N(2)Y	0.545	0.144	C(3)Y	0.622	0.219	O(2)Y	1.609	0.532
N(2)Z	−2.549	0.110	C(3)Z	2.103	0.180	O(2)Z	0.359	0.300
C(2)X	0.173	0.285	C(3')X	0.021	0.032	O(3)X	−0.174	0.279
C(2)Y	1.581	0.051	C(3')Y	−0.677	0.060	O(3)Y	−0.973	0.107
C(2)Z	−1.505	0.099	C(3')Z	2.478	0.080	O(3)Z	3.628	0.073
(b)								
	F_1	F_2		F_1	F_2		F_1	F_2
C(1)X	−0.621	0.189	C(2')X	−0.413	−0.090	N(4)X	−0.193	−0.044
C(1)Y	0.289	1.329	C(2')Y	0.177	0.368	N(4)Y	−0.199	0.608
C(1)Z	0.031	−1.659	C(2')Z	0.106	−0.502	N(4)Z	0.123	−0.954
C(1')X	−0.358	−0.026	N(3)X	−0.415	0.314	O(1)X	0.000	0.000
C(1')Y	0.250	0.041	N(3)Y	0.335	0.405	O(1)Y	0.194	0.624
C(1')Z	0.082	−1.172	N(3)Z	0.208	−1.015	O(1)Z	0.137	−1.068
N(2)X	−0.538	−0.365	C(3)X	−0.525	0.848	O(2)X	−1.506	−0.072
N(2)Y	0.278	1.185	C(3)Y	0.313	0.095	O(2)Y	0.000	0.000
N(2)Z	0.066	−0.926	C(3)Z	0.189	−1.039	O(2)Z	0.000	0.000
C(2)X	−0.285	−0.681	C(3')X	−0.409	−0.006	O(3)X	−0.515	−0.495
C(2)Y	0.223	0.880	C(3')Y	0.250	0.652	O(3)Y	0.243	1.032
C(2)Z	1.138	−0.499	C(3')Z	0.140	−0.925	O(3)Z	0.122	−0.856
(c)								
	F_1			F_1			F_2	
C(1)X	−0.496		C(2')X	−0.033		N(4)X	0.457	
C(1)Y	0.134		C(2')Y	−0.116		N(4)Y	−0.067	
C(1)Z	−0.181		C(2')Z	−0.124		N(4)Z	0.025	
C(1')X	0.090		N(3)X	−0.039		O(1)X	0.888	
C(1')Y	0.046		N(3)Y	0.235		O(1)Y	−0.079	
C(1')Z	−0.067		N(3)Z	0.213		O(1)Z	0.055	
N(2)X	−0.312		C(3)X	−0.284		O(2)X	−0.240	
N(2)Y	0.109		C(3)Y	0.187		O(2)Y	−0.512	
N(2)Z	−0.103		C(3)Z	0.172		O(2)Z	−0.249	
C(2)X	0.252		C(3')X	−0.025		O(3)X	−0.260	
C(2)Y	−0.137		C(3')Y	0.046		O(3)Y	0.030	
C(2)Z	0.059		C(3')Z	0.064		O(3)Z	0.024	

and ΔAB evaluated. The coordinates for B represent only one of eight possibilities (rotations and reflexion), so that the non-identity symmetry operations of point group *mmm* (*D2h*) were used to generate seven alternative sets of coordinates for each of which a ΔAB was calculated. The set which gave the lowest ΔAB was taken as being the best fit. In this way the coordinates of 63 fragments were successively transformed to fit as well as possible to A. There is a wide variation in the goodness of this fit, with ΔAB varying from 0.3–35 Å^2. For these larger differences the r.m.s. differences between corresponding atoms in superimposed structures is over 1 Å, showing that there is only a limited similarity.

Cluster analysis

The 63 sets output from GEOSTAT each contain 12 × 3 Cartesian coordinates. (Some structures, especially PROMYC which has a 15-membered ring, provided several fragments.) The cluster analysis is therefore carried out on 63 points in 36-dimensional space. We used the BMDP2M program (from the BMDP package[20] with Euclidian distance (ie ΔIJ) as the measure of proximity of two points or clusters I, J. The algorithm initially places each point in a cluster by itself. It then searches for the two clusters (which may be single points) with the shortest distance between their centroids (unweighted in this case). It then combines them into a new cluster whose new centroid is used subsequently. This process of aggregation continues until all points belong to one cluster.

A record is kept of this process which represents the structure of the clustering, and there are two common ways of presenting this information. The order of aggregation can be described by a dendrogram (tree) which shows how early or late in the process particular clusters were joined; the earlier this was, the closer the clusters are. A simplified tree is shown (Figure 2) for the geometries of the 63 peptide fragments; its particular structure is explained later. An alternative method of displaying similarity is to generate a matrix (Figure 3) representing all the Euclidian distances between pairs of points. The order of points in the matrix is derived from the aggregation process which means that points close together in hyperspace (ie in the same or neighbouring clusters) tend to form blocks in the distance matrix. The use of shading, where the smaller ΔIJ are given heavier type (Figure 3), is a quick and effective way of showing these blocks. (The distance matrix has a superficial similarity with that often used[21] to represent the conformation of a single molecule, particularly a protein. There, however, points all belong to the same molecule in 3D space and the ordering of the points is (usually) determined by the connectivity of the atoms.)

Figure 2 shows that the main difference in conformation is determined by whether the amide groups are *cis* or *trans*. The *trans*-amides form the largest group of compounds and, since the *cis* conformations are only found in *N*-substituted amides, are the only ones appropriate to the analysis of β-loops in proteins. *Cis*-Amides were thus excluded at this stage and cluster analysis repeated on just the *trans*-compounds (Figure 3). Two clusters are very apparent, the top one (AAGAAG to CLPGDH) seeming to have further substructure. The lower cluster corresponds exactly to fragments with the II conformation. (An apparent subcluster is slightly misleading since it consists of repeated observations from different molecules in the asymmetric unit of a single structure, PROMYC.) The top cluster contains three not very clearly separated subclusters which correspond to structures described by their authors as I, III, and I′.

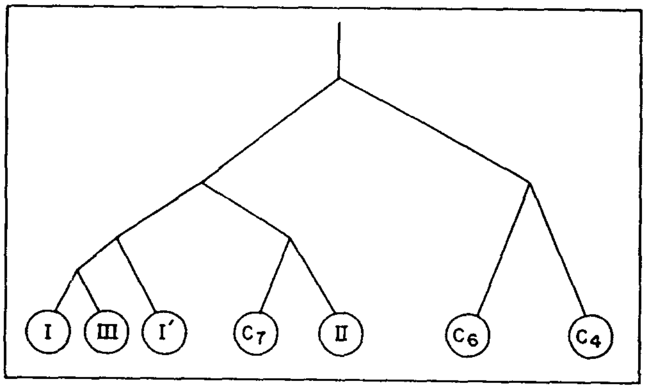

Figure 2. Dendrogram of conformations taken up by hydrogen bonded loops in tripeptides. The lower down the diagram that branching occurs, the greater the similarity in geometry. The aggregation of clusters is represented by proceeding upwards. The diagram has been simplified to include (fairly well-defined) clusters representing different conformational types. The key to the nature of the compounds in each cluster is: types of turn for trans-*amides are I, I′, II, II′, III (see text and Figure 1); types of turn for cyclic* cis-*amides (with* N-*substitution) are represented by* Cn *where* n *is the number of aminoacid or other residues in the ring*

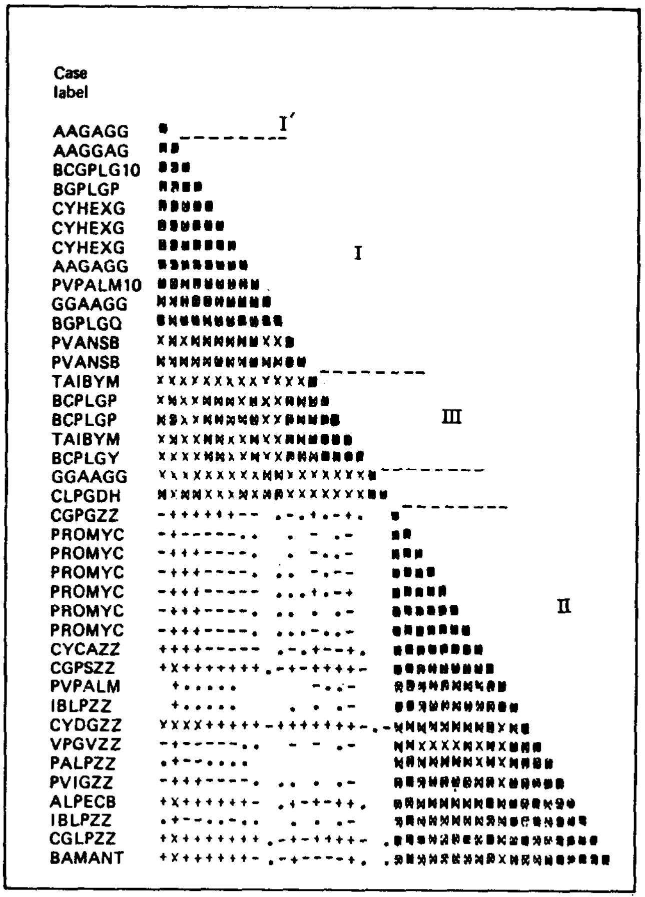

Figure 3. Distance matrix (Å) for 39 tripeptide fragments with 4 → 1 hydrogen bonds and trans-*amides. The matrix is symmetrical and only the diagonal terms and lower triangle are shown. Each cell represents the Euclidian distance in 36-dimensional space between the two structures, I and J. This distance (ΔIJ, defined in the text (equation (1)) (describes how similar the geometries of I and J are. It is represented by symbols whose denseness is roughly proportional to the degree of similarity:⊛: 0–0.5 Å; ⊗: 0.5–0.7 Å; ⋈: 0.7–0.9 Å; ×: 0.9–2.6 Å; +: 2.6–2.9 Å; –: 2.9–3.0 Å; ·: 3.90–3.2 Å; blank: > 3.2 Å. Conformations fall into two main groups: II (the lower) and I/III (upper) which includes a I′ structure and has three not very well-separated subclusters*

The classification II is very clearly supported by the analysis but the distinction between I and III is much less clear. Often the original authors made this classification not just on the geometry of the tripeptide fragment, but with knowledge of the hydrogen-bonding further along the chain (ie a 3/10 helix has a 5 → 2 hydrogen bond, an antiparallel sheet has a 1 → 4 hydrogen bond). Clearly the I and III conformations are fairly close (the r.m.s. difference per atom can be less than 0.4 Å).

Factor analysis

Figure 3 suggests that factor analysis may be useful for the *trans*-amides and this was initially done for all 39. When the 36 internal Cartesian coordinates were used for principal component analysis (program BMDP4M), two eigenvalues of the correlation matrix (88 and 7%) accounted for 95% of the variance. A scatterplot of the data as factor scores along these two axes (Figure 4) shows the two clusters very clearly. It is very obvious that there are no points intermediate between the I/III and II clusters. There is therefore no interconversion pathway of low enough energy to allow intermediates to be trapped out in the crystalline state with any frequency. The intracluster variance in the I/III cluster is high enough to warrant further factor analysis of it by itself (Figure 5). Although the subpopulations do not overlap seriously, they are not well-separated and can be seen as part of a spectrum of conformations. It is also interesting that Figure 5 describes 85% of the variance in I/III conformations and that a 2D description (of what is normally described by four torsion angles) is a very good approximation. There is no reason why 4 → 1 loops in proteins should not also be included on this diagram.

Addition of further data

A slight drawback of factor analysis is that the factor axes are linear combinations of all the Cartesian coordinates. When further structures become available their coordinates must be transformed before they can be plotted on, say Figure 5. This can be done without repeating the factor analysis by the process shown in Table 2. After the axes of inertia of the new fragment have been determined, the internal Cartesian coordinates are calculated. They are compared with the mean values of the coordinates used in the present analysis (Table 2(a)) and the appropriate rotation/reflexion applied to the new coordinates to fit them as well as possible. The new coordinates are then referred to this mean; ie the mean in Table 2(a) is subtracted. The coordinates are then represented as a 36-dimensional vector and transformed by the factor score matrix in Table 2(b), to give the values of the two factor scores in Figure 5. Only when a much larger number of structures have become available should the analysis need to be repeated. Four cluster analysis, however, it is always necessary to repeat the process afresh if new points are added to the dataset.

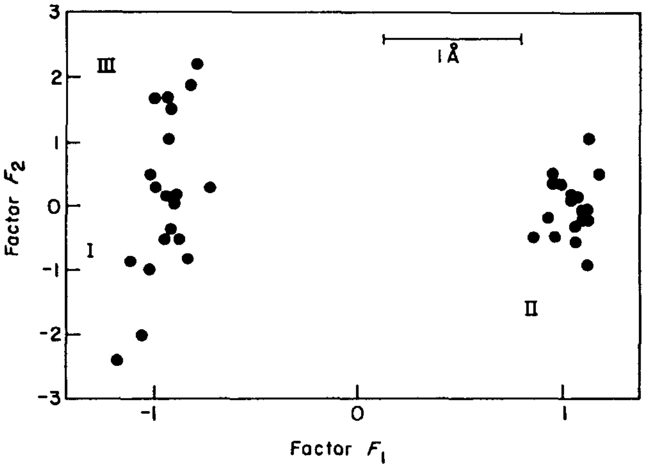

Figure 4. The conformations of 39 tripeptide loops with trans-*amides represented by the projection of the 36-dimensional Cartesian vectors onto the two largest components (factors) of the distribution,* F_1 *and* F_2. *The units are z-scores, i.e. numbers of standard deviations, but the intervals in the diagram have been scaled so that Euclidian distances are preserved, ie* F_1 *and* F_2 *as plotted, account for the correct proportions of the variance.* F_1 *accounts for a variance of 2.23 Å and this can be used to derive the Å scale marker. The structure of the clusters can be seen in Figure 3 and that of the I/III cluster is examined in greater detail in Figure 5*

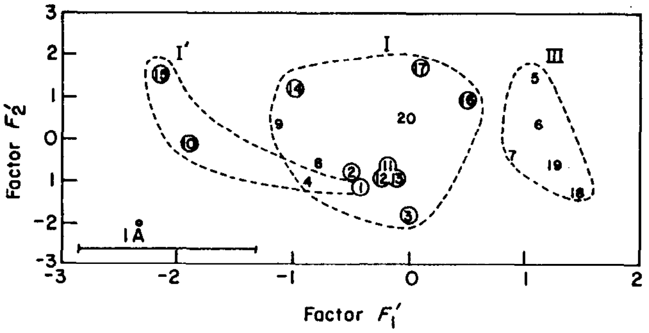

Figure 5. Factor analysis of the twenty trans-*amides in the I/III cluster. The two largest factors,* F_1' *and* F_2', *account for 88% of the variance and have been scaled to show their relative importance. (Note that* F'_1 *and* F'_2 *have no relation to* F_1 *and* F_2 *in Figure 3.) The distribution is essentially continuous, but from other considerations (see text) it can be partitioned into three groups, I, I' and III which are contiguous fragments occurring in cyclic peptides are circled and the compounds are identified by the numbers: 1, 2: AAGAGG10; 3: AAGGAG10; 4, 5: ANTAML10; 6: BAMANT10; 7: BCGPLG10; 8–10: BCPLGY20; 11: BGPLGP; 12: BGPLGQ; 13: CGLEGL; 14: CLPGDH; 15–17: CYHEXG; 18, 19: GGAAGG; 20, 21: PVANSB; 22: PVPALM10; 23, 24: TAIBYM; 25: BCPPGA*

Classification of β-Loops

The standard grouping of β-loops into widely separated I and II classes is firmly upheld by the analysis but category III is not so clear. The difference in geometry of tripeptide fragments labelled I and III is no larger than the variation of tripeptide fragments labelled I. Clearly where there is additional information (such as further hydrogen bonding) the distinction I/III can be made, but for an isolated tripeptide fragment the distinction is not clear. The large variation within the main classes I and II (often as much as an r.m.s. value per atom of 0.8 Å) implies that the β-loops are rather flexible. The amount is well beyond experimental error but how much is due to substitution patterns and how much to crystal packing forces (whose effect can be deduced from the variation in geometry of chemically identical fragments) is not clear from the somewhat limited data.

Table 2. Transformations involved in converting internal Cartesian coordinates into factor scores of the distribution in Figure 5. All these quantities are available from GEOSTAT. (a) The mean geometry of the distribution (Å relative to axes of inertia) and its observed standard deviation (Å)are given by the 36-dimensional vector. Any new structures should be referred to this geometry as origin by subtracting the first vector from the best-fit Cartesian coordinates. (b) Transformation matrix for calculating factor scores. (This should be written as a 36 row × 2 column matrix and used to postmultiply the row vector obtained after the subtraction in (a) above, giving the factor score column vector. (The components of this vector are distributed about zero with unit variance). (c) Components of the largest eigenvector (F_1') in Å

Atom	Coordinate (Å)	Standard deviation (Å)	Atom	Coordinate (Å)	Standard deviation (Å)	Atom	Coordinate (Å)	Standard deviation (Å)
C(1)X	0.534	0.069	C(2')X	−0.077	0.031	N(4)X	−0.675	0.129
C(1)Y	−1.840	0.175	C(2')Y	1.449	0.097	N(4)Y	−1.396	0.081
C(1)Z	−3.049	0.214	C(2')Z	−0.056	0.096	N(4)Z	1.527	0.072
C(1')X	−0.092	0.020	N(3)X	0.824	0.091	O(1)X	−1.063	0.070
C(1')Y	−0.745	0.093	N(3)Y	0.644	0.072	O(1)Y	−0.961	0.069
C(1')Z	−2.198	0.083	N(3)Z	0.501	0.048	O(1)Z	−1.478	0.069
N(2)X	0.505	0.100	C(3)X	0.879	0.192	O(2)X	−0.878	0.119
N(2)Y	0.446	0.135	C(3)Y	0.442	0.152	O(2)Y	2.084	0.229
N(2)Z	−2.251	0.043	C(3)Z	1.939	0.059	O(2)Z	0.601	0.201
C(2)X	−0.069	0.152	C(3')X	0.045	0.020	O(3)X	0.067	0.134
C(2)Y	1.597	0.052	C(3')X	−0.721	0.046	O(3)Y	−0.998	0.133
C(2)Z	−1.578	0.107	C(2')Z	2.416	0.036	O(3)Z	3.605	0.047

(b)

	F_1'	F_2'		F_1'	F_2'		F_1'	F_2'
C(1)X	– 0.989	3.936	C(2')X	−1.489	3.778	N(4)X	2.095	−3.139
C(1)Y	0.000	0.000	C(2')Y	0.000	0.000	N(4)Y	1.220	−2.106
C(1)Z	0.000	0.000	C(2')Z	0.000	0.000	N(4)Z	1.814	1.394
C(1')X	1.417	3.973	N(3)X	3.440	−2.663	O(1)X	3.455	2.099
C(1')Y	3.383	−1.978	N(3)Y	−2.097	−4.763	O(1)Y	0.023	−0.984
C(1')Z	2.313	0.326	N(3)Z	2.758	−0.952	O(1)Z	−0.689	0.878
N(2)X	0.000	0.000	C(3)X	0.000	0.000	O(2)X	0.000	0.000
N(2)Y	0.000	0.000	C(3)Y	−2.444	−1.217	O(2)Y	0.000	0.000
N(2)Z	2.145	2.920	C(3)Z	−1.388	−2.624	O(2)Z	0.000	0.000
C(2)X	0.000	0.000	C(3')X	6.990	2.565	O(3)X	1.422	−3.978
C(2)Y	−1.782	−0.944	C(3')Y	−2.993	−1.395	O(3)Y	0.077	−6.655
C(2)Z	0.000	0.000	C(3')Z	−2.876	−2.121	O(3)Z	2.807	1.034

(c)

	F_1'		F_1'		F_1'
C(1)X	0.011	C(2')X	−0.017	N(4)X	0.029
C(1)Y	0.168	C(2')Y	−0.086	N(4)Y	−0.023
C(1)Z	−0.200	C(2')Z	0.090	N(4)Z	−0.016
C(1')X	0.006	N(3)X	0.082	O(1)X	0.058
C(1')Y	0.090	N(3)Y	−0.056	O(1)Y	−0.029
C(1')Z	−0.075	N(3)Z	0.021	O(1)Z	0.032
N(2)X	−0.085	C(3)X	0.186	O(2)X	−0.030
N(2)Y	0.134	C(3)Y	−0.139	O(2)Y	−0.195
N(2)Z	−0.024	C(3)Z	−0.033	O(2)Z	0.193
C(2)X	−0.128	C(3')X	0.008	O(3)X	−0.119
C(2)Y	0.046	C(3')X	−0.004	O(3)Y	0.095
C(2)Z	0.100	C(2')Z	−0.003	O(3)Z	0.021

CONCLUSION

The processes described above are essentially automatic and do not require much computer time above that already required for running the CCDC programs. There is a limit to the size of the distance matrix or dendrogram that can be comfortably handled and above, say, 200 structures, the output can become somewhat unwieldy. However, all the procedures are automatic and therefore objective, so that it is possible to analyse a group of compounds without any preconceptions. They seem robust enough to suggest that they can be routinely used for initial surveys of conformational variability. In cases where they fail to give a clear picture it may be very difficult to classify the data at all. The objective measure of the similarity/difference between molecular shape may be a valuable index when analysing biological or pharmacological data. The extension to the description of the conformation of parts of macromolecules is also straightforward.

Acknowledgement

J Raftery wishes to thank the SERC for postdoctoral support.

1 *Cambridge Crystallographic Data Centre User Manual*, 2nd Edition (1978)

2 **Allen, H A, Bellard, S, Brice, M C, Cartwright, B A, Doubleday, A, Higgs, H, Hummelink, T, Hummelink-Peters, B G, Kennard, O, Motherwell, W D S, Rodgers, J R and Watson, D G** *Acta Crystallogr. B* Vol 35 (1979) p 2331

3 **Dunitz, J D** *X-Ray Analysis and the Structure of Organic Molecules* Cornell University Press, Ithaca: (1978)

4 **Bürgi, H B, Dunitz, J D and Schefter, E** *Acta Crystallogr. B* Vol 30 (1974) p 1517

5 **Murray-Rust, P and Motherwell, W D S** *Acta Crystallogr. B* Vol 34 (1978) p 2534

6 **Britton, D** *Acta Crystallogr. B* Vol 33 (1977) p 3727

7 **Murray-Rust, P, Bürgi, H B and Dunitz, J D** *Acta Crystallogr. B* Vol 34 (1978) p 1787

8 **Murray-Rust, P, Bürgi, H B and Dunitz, J D** *Acta Crystallogr. A* Vol 35 (1979) p 703

9 **Nyburg, S C** *Acta Crystallogr. B* Vol 30 (1974) p 251

10 **Diamond, R** *Acta Crystallogr. A* Vol 27 (1971) p 436

11 **Rohrer, D C and Perry, H** 'FITMOL' in **Wood, J J (Ed)** *Public Procedures: A Program Exchange for PROPHET Users* Bolt, Beranek & Newman Inc., Cambridge, MA (1978)

12 **Kabsch, W** *Acta Crystallogr. A* Vol 32 (1976) p 922

13 **Mackay, A L** *Acta Crystallogr. A* Vol 33 (1977) p 212

14 **Cruickshank, D W J (Ed)** *International Tables for X-Ray*

15 **Everitt, B S** *Cluster Analysis* Heinemann (Educational), London (1974)

16 **Murray-Rust, P** American Crystallographic Association, Spring Meeting, Gaithersburg (1982)

17 **Bürgi, H B** *Acta Crystallogr. B* (1985) submitted for publication

18 **Vaciago, A, Domenicano, A and Murray-Rust, P** *Acta Crystallogr. B* Vol 39 (1983) p 457

19 **Venkatachalam, C M** *Biopolymers* Vol 6 (1968) p 1425

20 *Biomedical Program (BMDP),* University of California (1979)

21 **Rossmann, M G and Liljas, A** *J. Mol. Biol.* Vol 85 (1974) p 177