Everitt (1980) for a lengthy discussion], due to a lack of any generally accepted and robust clustering criteria. In the chemical context, where different conformations are often well separated by discrete energy barriers, we would hope to establish a decision theory to obviate user intervention. Current work is directed towards that aim, so as to generate fully automated methods for unsupervized machine learning from a large database such as the CSD.

We thank referees of earlier papers in this series for encouraging us to present the detailed discussion contained in this manuscript.

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, 41–49.
ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*c*). *Acta Cryst.* B47, 50–61.
ALLEN, F. H. & JOHNSON, O. (1991). *Acta Cryst.* B47, 62–67.
ALLEN, F. H., KENNARD, O. & TAYLOR, R. (1983). *Acc. Chem. Res.* **16**, 146–153.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*a*). *Inorg. Chem.* **28**, 3960–3969.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*b*). *Inorg. Chem.* **28**, 3970–3981.
AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*c*). *Inorg. Chem.* **28**, 3982–3991.
BOEYENS, J. C. A. (1978). *J. Cryst. Mol. Struct.* **8**, 317–320.
BUCOURT, R. & HAINAUT, D. (1965). *Bull. Soc. Chim. Fr.* pp. 1366–1378.
CREMER, D. & POPLE, J. A. (1975). *J. Am. Chem. Soc.* **97**, 1354–1358.
CSD User Manual (1989). Version 3.4. Crystallographic Data Centre, Cambridge, England.
DUNITZ, J. D. (1979). *X-ray Analysis and the Structure of Organic Molecules*, ch. 10, pp. 447–494. Ithaca: Cornell Univ. Press.
EVERITT, B. (1980). *Cluster Analysis*, 2nd ed. London: Halstead Heinemann.
JARVIS, R. A. & PATRICK, E. A. (1975). *IEEE Trans. Comput.* **22**, 1025–1034.
NORSKOV-LAURITSEN, L. & BÜRGI, H.-B. (1985). *J. Comput. Chem.* **6**, 216–228.
PICKETT, H. M. & STRAUSS, H. L. (1970). *J. Am. Chem. Soc.* **92**, 7281–7288.

Acta Cryst. (1991). B47, 412–424

Automated Conformational Analysis from Crystallographic Data. 6.* Principal-Component Analysis for *n*-Membered Carbocyclic Rings (n = 4, 5, 6): Symmetry Considerations and Correlations with Ring-Puckering Parameters

BY FRANK H. ALLEN† AND MICHAEL J. DOYLE

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

AND THOMAS P. E. AUF DER HEYDE†

Department of Chemistry, University of the Western Cape, Bellville 7530, *South Africa*

(*Received* 28 *April* 1990; *accepted* 19 *December* 1990)

Abstract

Representative samples of four-, five- and six-membered carbocycles have been retrieved from the Cambridge Structural Database and have been used to fill, by symmetry expansion, the hyperdimensional conformation spaces spanned by the intra-annular torsion angles for these ring systems. The resulting distributions have been probed by principal-component analysis (PCA). For cyclobutane, all of the sample variance can be described in terms of a single coordinate [or principal component (PC)] which maps the degree of pucker about the ring diagonal. In the case of cyclopentane two equally important PC's fully describe the sample variance, and together they map the pseudorotation itinerary which interconverts the envelope and twist conformations of this ring. For cyclopentenes, however, a single PC (accounting for almost 80% of sample variance) maps the extent of ring pucker, whilst a second PC (accounting for the remaining 20% of variance) is found to describe minor torsional distortions away from 0° about the double bond. PCA for six-membered carbocycles (cyclohexanes and cyclohexenes) reveals three PC's: one mapping the interconversion of enantiomeric chair conformers, and two that describe the pseudorotational interchange between boat and twist-boat forms. For all three ring

* Part 5: Allen & Taylor (1991).

† Author to whom correspondence should be addressed.

0108-7681/91/030412-13$03.00

systems the reduction in dimensionality as a consequence of the PCA is compared to, and is shown to be consistent with, the dimension reduction inherent in the Cremer–Pople (CP) description of ring pucker [Cremer & Pople (1975). *J. Am. Chem. Soc.* **97**, 1354–1358]. The results for five-membered rings are also shown to be fully consistent with the alternative Altona–Sundaralingam (AS) analysis [Altona & Sundaralingam (1972). *J. Am. Chem. Soc.* **94**, 8205–8212]. The influence of outliers on the PCA is illustrated, and the results of the study are used to highlight the potentials (and pitfalls) of PCA for conformational analysis and conformational mapping.

1. Introduction

Previous papers in this series (Allen, Doyle & Taylor, 1991*a–c*, hereafter ADT1, ADT2, ADT3) have used cluster analysis to investigate the six-dimensional conformation space spanned by the ring torsion angles for six-membered carbocycles. The data were found to cluster around points representing the familiar boat, chair, twist, sofa (envelope), *etc.* conformations, and the analysis provided 'average' angular definitions for both asymmetric and symmetric (Allen & Taylor, 1991) conformers.

Whilst cluster analysis is extremely useful for identifying clouds of data points, it has two major shortcomings. Firstly, although it can reveal the distribution of the data points in conformation space, it says little about the shape of the data clouds, or about the coordinates along which they expand. This is because it does not offer, in itself, any facilities for a graphical representation of the data. Secondly, it does not help to decide whether the dimensionality of the problem may be reduced in order to make the results more comprehensible. The technique of principal-component analysis (PCA)* (Chatfield & Collins, 1980; Malinowski & Howery, 1980; Auf der Heyde, 1990) offers a means whereby these problems may be addressed.

The mathematical basis of PCA rests on an eigenanalysis of the data covariance or correlation matrix. Successive eigenvectors describe orthogonal 'axes of variance' through the data space, the direction of greatest variance being described by the vector with the largest eigenvalue, that of the second-largest variance (orthogonal to the first axis) by the vector with the second-largest eigenvalue, and so on. The eigenvectors or principal components (PC's) appear as linear combinations of the original variables. Often the variance of a sample in n-dimensional space (where n is the number of variables or parameters) may be adequately described by a considerably smaller number of PC's, which in turn might reflect some underlying physicochemical factor influencing the data distribution. Two major problems, though, commonly frustrate the interpretation of PCA results. The first is a conceptual difficulty sometimes encountered with extremely 'mathematical' techniques. The second arises from attempts to interpret the PC's in chemical terms when they may, in fact, defy any practical interpretation. Despite these problems, PCA is already well established in analytical chemistry (Malinowski & Howery, 1980; Massart & Kaufman, 1983), and is finding increasing use in conformational analysis (Murray-Rust & Motherwell, 1978; Murray-Rust & Bland, 1978; Murray-Rust & Raftery, 1985*a*; Auf der Heyde & Bürgi, 1989*c*; Hummel, Huml & Bürgi, 1988; Hummel, Roszak & Bürgi, 1988).

We report here the results of PCA on four-, five- and six-dimensional torsional data sets, containing the intra-annular angles of the appropriate carbocycle (denoted 4-C, 5-C, 6-C, respectively). These conformation spaces may be reduced to one-, two- and three-dimensional subspaces (Cremer & Pople, 1975), to provide graphical representations that can be compared to the results of the PCA. The primary purpose of this comparison is to illustrate that the PC's extracted for these carbocycles have chemical significance. We also hope to provide some pointers that might prove useful in a PCA of a higher dimensional data set.

2. Descriptions of ring conformation and conformation space

Cremer & Pople (CP; 1975) have shown how the conformation of a general N-membered ring may be described in terms of $N-3$ parameters that appear as amplitude and phase coordinates q_m and φ_m. A 4-C ring has a single puckering amplitude (the maximum atomic out-of-plane displacement), denoted q_2. A 5-C ring has an amplitude, phase pair (q_2, φ_2) that describes a pseudorotation pathway (Fig. 1, see Kilpatrick, Pitzer & Spitzer, 1947). The 6-C ring is described by a similar pair (q_2, φ_2) associated with the boat–twist-boat pseudorotation, and q_3 associated with the degree of ring pucker. Conformational space for 4-C rings is, therefore, linear, the space for 5-C rings is circular. For 6-C rings (one cyclic and two linear parameters) the space may be treated as cylindrical. It contracts to a sphere, by replacing the (q_2, φ_2, q_3) coordinates by a spherical polar set (Q, θ, φ) where $Q^2 = q_2^2 + q_3^2$ and $\tan\theta = q_2/q_3$. Fig. 2 illustrates the major characteristics of this spherical conformation space, and indicates the positions of

* PCA is often confusingly referred to as factor analysis (FA), a similar but different technique. See the discussion in Chatfield & Collins (1980, ch. 5).

familiar canonical forms (see also Allen & Taylor, 1991).

A common alternative measure of pucker in 4-C rings is provided by the two dihedral (fold) angles θ_1, θ_2 about the ring diagonals (see Dunitz, 1979; Allen, 1984). 4-C rings are close to equilateral and θ ($=\theta_1$ or θ_2) or its complement $\omega = 180 - \theta$ are used as descriptors. A sign may be attached to ω if θ is calculated as one of the trans-annular torsion angles, *e.g.* C4—C3—C1—C2. For 5-C rings an alternative method, first developed by Altona, Geise & Romers (1968) and extended by Altona & Sundaralingam (AS; 1972), is also in common use. Here the ring conformation is again described in terms of a puckering coordinate τ_m (in this case the maximum torsion angle) and a different phase angle of pseudorotation P. The CP and AS approaches differ in the definition of the origin conformer: the CP approach has an envelope (E) conformer at $\varphi = 0°$, while in the AS description it is a twist (T) conformer which has $P = 0°$.

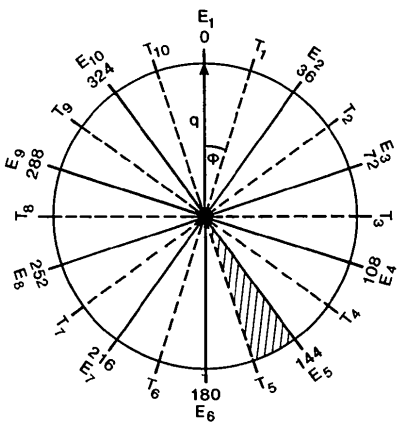


Fig. 1. Pseudorotational pathway and schematic diagram of the CP conformation space for 5-C rings. Each point on the circle represents a specific value of φ_2, while q_2 determines the radius. Points at 0, 36, 72° ... represent envelope (E) conformers, while those at 18, 54, 90° ... represent twist (T) conformers. Subscript n for E_n and T_n has no particular significance except to indicate that conformer E_{n+1} is obtained from E_n *via* pseudorotation through T_n (or *vice versa*), or that T_{n+1} is obtained from T_n *via* pseudorotation through E_{n+1} (or *vice versa*) for n taken as modulo 10. The shaded wedge represents an asymmetric unit.

Transformation of cyclic/spherical to rectangular coordinates

For this analysis, the circular and spherical CP coordinates are transformed to Cartesian equivalents. For the 5-C rings we have: CP1 $= q_2\cos\varphi_2$, CP2 $= q_2\sin\varphi_2$. The CP1 coordinate coincides with the vertical axis in Fig. 1, *i.e.* the axis mapping E conformations at $\varphi_2 = 0$, 36, 72°..., while CP2 maps the T conformations along the horizontal axis (or any other T axis displaced from it by $n \times 36°$, where $n =$ 1, 2, 3...). The AS parameters are transformed as: AS1 $= \tau_m\sin P$, AS2 $= \tau_m\cos P$, so that AS1 and CP1 both trace pure E conformers, while AS2 and CP2 coincide with T axes, as shown in Fig. 3. We note that even if AS1 and CP1 coincide exactly, AS2 and CP2 will be oriented at 180° to one another (or *vice versa*). Furthermore, the AS and CP axes can still be out of phase, though now the E axes will be $n \times 36°$ and the T axes $180 + n \times 36°$ ($n = 0$, 1, 2, 3...) out of phase (or *vice versa*). For the 6-C rings the transformations are: CP1 $= q_3$, CP2 $= q_2\cos\varphi_2$, CP3 $= q_2\sin\varphi_2$, so that the CP1 coordinate maps C conformers (the vertical axis with $\theta = 0°$ in Fig. 2), CP2 coincides with the B conformer axis (in the equator at $\varphi_2 = 0°$, Fig. 2) while CP3 traces the TB conformer axis.

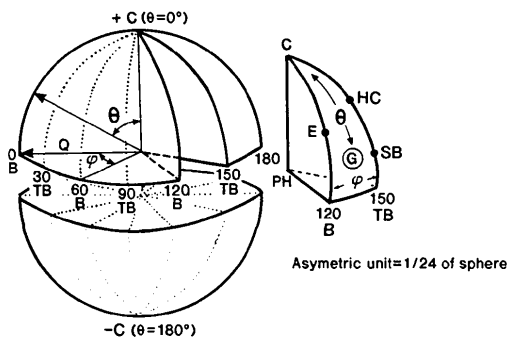


Fig. 2. Representation of conformational space for six-membered rings using the spherical polar coordinate set Q, θ, φ. Special symmetric conformations are indicated as C = chair, B = boat, E = envelope, HC = half-chair, SB = screw-boat (1,3-diplanar), TB = twist-boat. The phenyl ring (PH) is at the centre of the sphere and G is any general conformation. The isolated segment (1/24th) of the sphere is the asymmetric unit.

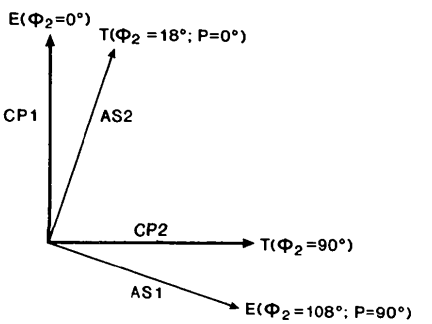


Fig. 3. Diagram illustrating the naming of the AS (light lines) and CP (heavy lines) axes, and their relative orientations. E and T signify axes corresponding to the vertical and horizontal axes in Fig. 2, respectively. Note that a counter-clockwise rotation of the AS axes by 108° superimposes AS1 and CP1, but has AS2 and CP2 at 180° to one another, while a clockwise rotation through 72° superimposes AS2, CP2 with AS1 and CP1 180° apart.

3. Data selection and retrieval

Crystallographic coordinates for 4-C, 5-C and 6-C rings were retrieved from the Cambridge Structural Database (CSD: Allen, Kennard & Taylor, 1983). Trial data sets, fully described below, were generated *via* the program *QUEST* (Allen & Davies, 1988; *CSD User Manual,* 1990) of CSD Version 4.2 dated 1 January 1990 (4-C, 5-C) and 1 January 1989 (6-C). General search constraints ensured that for all retrieved entries: (i) CSD checks had shown no numerical errors; (ii) there was no disorder in the structure; (iii) the crystallographic R value was less than 0·10; (iv) the entry was classified chemically as an organic compound; (v) no bridged rings existed in any of the entries for the 4-C or 5-C subsets (to avoid outlying, highly strained conformations). Further chemical constraints (*QUEST*) were employed to obtain: (*a*) cyclobutanes, all 4-C atoms sp^3 hybridized; (*b*) cyclopentanes, all 5-C atoms sp^3 hybridized (these rings are referred to as 5anes); (*c*) cyclopent-1-enes, 3 C atoms sp^3, 2 C atoms sp^2 and connected by an endocyclic double bond (these rings are referred to as 5enes). The first 200, 300 and 170 listed entries of types (*a*), (*b*) and (*c*) in the randomly ordered CSD file yielded 247 independent 4-C, 353 5ane and 198 5ene fragments respectively. In addition, a mixed data set (5mixs) composed of approximately equal numbers of cyclopentanes and cyclopentenes, with a total of 347 independent fragments, was also formed.

The 5-C rings were processed in these three ways for the following reasons. First, in order to map the conformation space for cyclopentane (Fig. 1), all cyclopentenes were excluded, since the double bond locks the fragment into an E conformation. Second, the 5ene sample was used to examine the conformational options open to these fragments. Finally, the 5mixs sample attempts to mimic the conformational complexity that a chemically non-specific search for 5-C rings might uncover. This approach to data selection is particularly important for 6-C rings, which can exhibit greater conformational complexity. A subset of 71 6-C rings [designated 6-C(1)] was selected from the trial data set described in ADT1, so as to best illustrate the conformational mapping process. The subset comprised: normal chairs (16), distorted (highly-puckered) chairs (4), normal boats (8), highly-puckered boats (from norbornane) (14), phenyl rings (10), half-chairs (10), twist boats (3), screw-boats (1,3-diplanar) (4), distorted envelopes (2). An additional larger subset [6-C(2)] was retrieved from CSD (1 January 1990) using criteria (i)–(iv) and constraining (*via QUEST*) five contiguous bonds of the ring to be either single, double or triple; the remaining bond was fixed as a single bond. The number of atoms present in the retrieved entries was restricted to be ≤ 24. A total of 582 entries yielded 952 independent ring fragments.

4. Computational methodology

The CSD program *GSTAT* (*CSD User Manual,* 1990) will locate substructural fragments in a molecular connectivity representation established using distance criteria. A wide variety of geometrical parameters may then be calculated systematically for each located fragment. Algorithms for PCA were introduced into *GSTAT* by Murray-Rust & Raftery (1985*a*,*b*). The algorithms perform the necessary eigenanalysis of the correlation matrix, $B(N_p,N_p)$, obtained from a data matrix, $A(N_f,N_p)$, containing N_p parameters for each of N_f fragments. In this work the N_p parameters are intra-annular torsion angles. The PC 'scores', *i.e.* the coordinates of each fragment in n-dimensional PC space, are scaled by the variance accounted for, and plotted as scattergrams for each unique pair of the mutually orthogonal PC axes. A number of extensions to the functionality of *GSTAT* were required for the work described in this paper.

Symmetry expansion of the raw data sets

The effects of the topological symmetry of a 2D representation of a chemical fragment on the relative ordering of the N_p geometrical parameters has been discussed (ADT1; Murray-Rust, 1982; Allen & Taylor, 1991). The origin of these effects is the number of possible equivalent enumerations of the atoms of the fragment in 2D, each one giving rise (at random) to a different order in which the geometrical parameters are calculated from the corresponding 3D coordinates. The earlier work (ADT1, ADT2, ADT3) described a scheme in which the torsion angles were permuted according to this topological symmetry, together with an inversion operator to generate conformational enantiomers. In essence (Allen & Taylor, 1991) this process successively places a given fragment into each of the possible asymmetric units of the relevant conformational space. Alternatively, the topological symmetry can be used implicitly to transform each representative point into its symmetry-related siblings, so that all topological isomers are now included in the basic data set. This technique has been used successfully (Norskov-Lauritsen & Bürgl, 1985; Auf der Heyde & Bürgi, 1989*a*–*c*) as a precursor to cluster analysis; it is an essential feature of the PCA and correlation analyses reported here.

Two expansions are performed. For the PCA, based on torsion angles, there are $4N$ permutations/inversions (see ADT1) of the basic angle sequence for an N-membered ring of symmetry D_{Nh}. The $4N$ variants fill the conformational space spanned by the

torsion angles for each fragment. For the 5enes symmetry is reduced to C_s and there are only four torsional variants: two permutations and two inversions. A CP expansion must also be performed to generate an isomer in each of the asymmetric units of the CP space. The 20 (24) asymmetric units for 5-C (6-C) rings are illustrated in Figs. 1 and 2. For the 5enes the CP dimensionality reduces from two (q_2, φ_2) to one (q_2); the double bond effectively prohibits pseudorotation, and q_2 simply measures the out-of-plane displacement of the ring atom opposite the double bond.

Previously (ADT1, ADT2, ADT3) topological isomers were generated by direct permutation of the torsion angles. However, we cannot permute the CP parameters directly: the phase information, for example, is related by a phase shift between isomers rather than by a simple permutational mechanism. Thus, *GSTAT* now applies permutation and inversion operations to the atomic coordinate set for each located fragment. In a general case of N_s permutations and N_s inversions we obtain a symmetry-expanded data matrix $A_s(2N_sN_f,N_p)$. The 4-C (D_{4h}), 5anes (D_{5h}), 5mixs (D_{5h}), 5enes (C_s), and 6-C(1) (D_{6h}) samples were expanded as indicated. For the larger 6-C(2) sample of 952 fragments, we assumed that the randomness of the atomic enumerations would ensure a distribution that would begin to reflect the inherent symmetry of the conformation space. This allows us to compare PCA results for a random distribution with those obtained from a fully symmetrized one.

Calculation of puckering parameters

Code for the systematic calculation of CP puckering parameters was kindly supplied by Professor Dieter Cremer (program *RING*88; Univ. of Köln, Germany). Within *GSTAT* the user must indicate which N atoms of the coded fragment, in cyclic order, form the ring. Suitable names must also be supplied for the $n-3$ puckering parameters, *e.g.* Q2, PHI2, Q3 for a 6-C ring. These names may be used in later *GSTAT* commands, *e.g.* (*a*) to select those rings having specified ranges of pucker, or (*b*) to plot histograms or scattergrams of specified parameter(s) for the complete data set. For five-membered rings the alternative AS treatment was also introduced to *GSTAT*. Again the user must specify a (cyclically ordered) list of torsion angles, and provide two parameter names (*e.g.* TAUMX, PHI) for outward use by the program. The basic development of AS is used with the first-specified torsion angle taken as τ_0 in their equation.

Inter-parameter correlations

Various routines from the *CAMAL* library (Taylor, 1986) have been added to *GSTAT* to give the following functionality: (*a*) Generation of a correlation–covariance matrix $C(N_p,N_p)$ in which elements of the lower triangle contain inter-parameter correlation coefficients, elements of the upper triangle contain inter-parameter covariances, and the diagonal elements contain the N_p individual parameter variances. (*b*) The ability to perform linear-regression analyses of one parameter on another. Relevant mathematical treatments are given in most statistical texts (*e.g.* Snedecor & Cochran, 1980) and in the *CSD User Manual* (1990).

One of the problems of PCA (see above) lies in relating the PC axes to chemically meaningful variations in the input data. *GSTAT* will now add the PC scores (coordinates) for each fragment to the calculated table of user-defined geometry. An extended data matrix $A'(N_f,N_p+m)$, where m is the number of PC's required to account for $>99\%$ of the variance in the original A matrix, is now available to the correlation and regression routines. A variety of individual geometrical parameters (or linear combinations thereof) may then be tried, in a search for chemically meaningful correlations with the PC scores. This facility is used extensively below, in assessing correlations between the various puckering parameters and the PC results.

Elimination or retention of individual fragments

For some data sets one or more conformations are identified as outliers and the PCA results and correlations were examined both with and without these fragments. In other cases, it was instructive to study PC results based on one, or a very few, representatives. These elimination/retention requirements are effected by two new, mutually exclusive, commands in *GSTAT*: KILL/KEEP, which operate on individual fragment numbers.

Cartesian transformations of circular or spherical polar coordinates

These were effected by the 'TRAnsform' command in *GSTAT* (Murray-Rust & Raftery, 1985*a*,*b*) which permits linear combination of pairs of existing parameters *via* FORTRAN-like arithmetic, trigonometric and other operators.

Limitations on the size of the data set

At present, generation of the correlation–covariance matrix $C(N_p,N_p)$ is limited to 1950 fragments, even though no such restrictions apply to PCA. Thus, the maximum number of independent fragments which can be fully treated is limited to $1950/2N_s$, *i.e.* a maximum of 121 4-C, 97 5-C and 81 6-C rings, so that a number of smaller samples were analyzed in each case. The 4-C rings were split into four samples with 61, 67, 52 and 67 independent

fragments respectively. The 5anes were processed as seven samples (5anes1–7) containing 42, 48, 45, 54, 48, 60, 56 independent fragments, the 5enes as four samples (5enes1–4) of 52, 44, 51, 51 fragments and the 5mixs as four samples (5mixs1–4) containing 85, 87, 83, 92 independent fragments.

5. Principal-component analyses

4-C *rings*

Only one PC, accounting for a minimum 99·98% of sample variance, was obtained for each symmetrized sample. The corresponding eigenvector always had the (vertical) form [0·5, −0·5, 0·5, −0·5]. These components describe the linear relationship between τ_1–τ_4 in the four-dimensional torsion-angle space, and correspond to the one mode of pucker (the fold about the ring diagonal) which cyclobutane can exhibit. They indicate that, in an equilateral 4-C ring, any change in τ_1 implies an equal change in τ_3, and equal but opposite changes in τ_2 and τ_4. The PC exhibited exact correlations (r = 1·000) with both the puckering angle, ω, and the CP puckering parameter, q_2. These results prove that the PC accounts completely for the puckering observed in our samples, and that it traces the CP coordinate in the four-dimensional torsional space. Representative histograms showing the structural identity of the three techniques are shown in Fig. 4 for one of the samples.

5-C *rings*

In all cases PCA revealed just two PC's which together account for a minimum 99·98% of the sample variance. For the 5anes and the 5mixs the two PC's had identical eigenvalues, each describing *ca* 50% of sample variance, while for the 5enes the first PC accounted for *ca* 79% and the second PC for the remaining 21%.

Table 1 lists the results of PCA for each sample, giving the loadings (or coefficients) for τ_1–τ_5 in each PC, the amount of variance accounted for, as well as the 'symmetry' of each PC – E or T. The latter is determined from an investigation of the loadings, which represent the relative contributions that each torsion angle makes to the distortion of a given ring along the PC coordinate; hence they indicate the conformation mapped by that PC. For example, PC2 for 5anes2 has the form

$$\mathrm{PC2} = 0{\cdot}9\tau_1 - 18{\cdot}8\tau_2 + 29{\cdot}4\tau_3 - 28{\cdot}8\tau_4 + 17{\cdot}2\tau_5.$$

Here, the coefficients of τ_1–τ_5 trace the pattern one might expect for an imperfect E conformer (Fig. 5). A perfect E conformer of C_s symmetry and with torsion-angle numbering as shown in Fig. 5 would, of course, have a τ_1–τ_5 pattern of (0, $-x$, y, $-y$, x). Similarly, PC1 for 5anes2 has the torsion-angle pattern of an (imperfect) T conformer with approximate C_2 symmetry. The PC's almost always exhibit identifiable and distinct symmetry, although it is exact only for the 5enes. In this case PC1 maps perfect E(C_s) conformers, while PC2 maps very small distortions maintaining a T(C_2) conformation away from

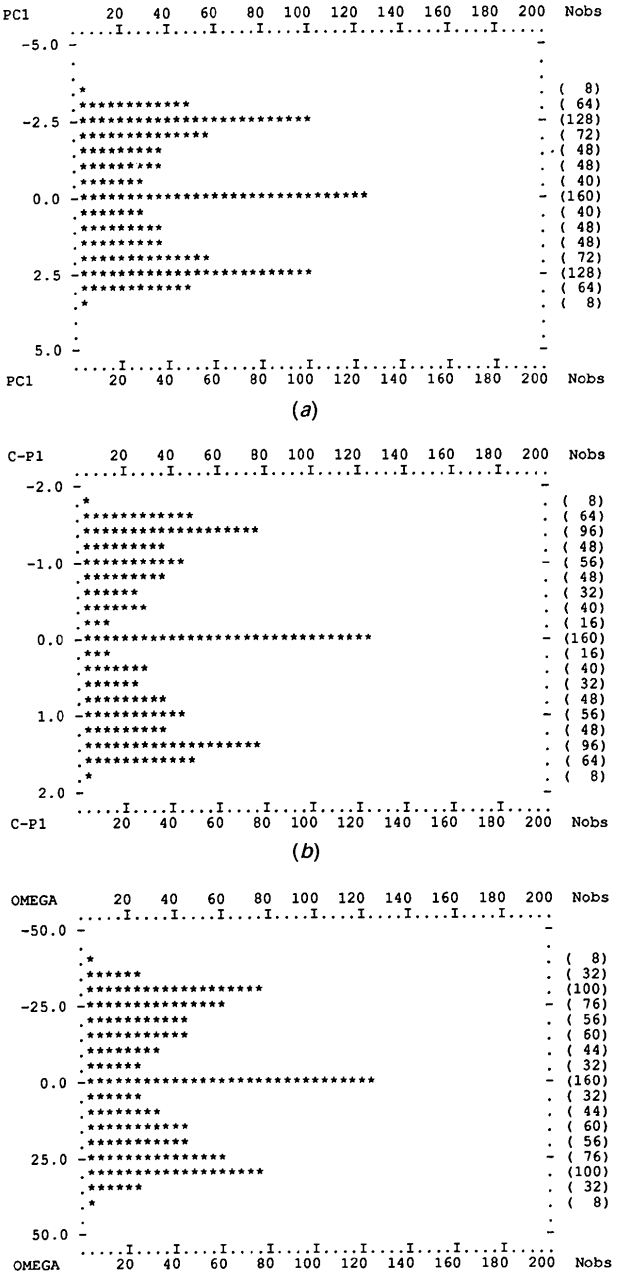


Fig. 4. Histograms of (*a*) PC1 scores, (*b*) CP values (multiplied by factor 5) and (*c*) puckering angles for sample 1 of the 4-C rings.

Table 1. *Loadings for* τ_1–τ_5 *in* PC1 *and* PC2 *for* 5-C *ring samples*

% = percentage of sample variance accounted for by PC; S = symmetry of PC (see text). Loadings in parentheses are those obtained prior to exclusion of outliers.

Sample	PC	τ_1	τ_2	τ_3	τ_4	τ_5	%	S
5anes1	PC1	1·0 (−14·8)	−19·5 (−4·3)	30·6 (21·8)	−30·0 (−31·0)	17·9 (28·3)	49·99	E
	PC2	−31·8 (−27·8)	25·1 (31·2)	−8·8 (−22·7)	−10·8 (5·5)	26·3 (13·7)	49·99	T
5anes2	PC1	−30·6	24·2	−8·5	−10·4	25·3	49·99	T
	PC2	0·9	−18·8	29·4	−28·8	17·2	49·99	E
5anes3	PC1	29·9	−23·6	8·3	10·1	−24·8	49·99	T
	PC2	0·9	−18·4	28·8	−28·2	16·8	49·99	E
5anes4	PC1	−29·0 (9·1)	22·9 (−23·4)	−8·0 (28·8)	−9·8 (23·1)	24·0 (8·7)	49·99	T
	PC2	0·9 (−27·3)	−17·8 (16·7)	27·9 (0·2)	−27·3 (−17·1)	16·3 (27·4)	49·99	E
5anes5	PC1	−30·3	23·9	−8·4	−10·3	25·1	49·99	T
	PC2	0·9	−18·6	29·1	−28·5	17·0	49·99	E
5anes6	PC1	−29·7	26·0	−12·3	−5·9	22·0	49·99	TE
	PC2	−3·3	−14·7	27·2	−29·3	20·1	49·99	ET
5anes7	PC1	1·1	−21·4	33·5	−32·8	19·5	49·99	E
	PC2	−34·8	27·5	−9·7	−11·8	28·8	49·99	T
5mixs1	PC1	−25·1	19·8	−7·0	−8·5	20·8	49·98	T
	PC2	0·8	−15·4	24·1	−23·6	14·1	49·98	E
5mixs2	PC1	0·8	−15·4	24·1	−23·6	14·0	49·98	E
	PC2	25·1	−19·8	7·0	8·5	−20·8	49·98	T
5mixs3	PC1	−25·5 (18·4)	20·1 (−25·1)	−7·1 (22·2)	8·6 (10·7)	21·1 (−4·7)	49·98	T
	PC2	0·8 (−17·3)	−15·6 (3·1)	24·5 (12·2)	−24·0 (−22·9)	14·3 (24·8)	49·98	E
5mixs4	PC1	23·7	−18·7	6·6	8·0	−19·6	49·97	T
	PC2	0·7	−14·5	22·8	22·3	13·3	49·97	E
5enes1	PC1	0	−14·0	21·7	−21·7	14·0	79·01	E
	PC2	2·8	−2·1	0·7	0·7	−2·1	20·98	T
5enes2	PC1	0	13·6	−21·1	21·1	−13·6	79·14	E
	PC2	2·5	−1·9	0·6	0·6	−1·9	20·85	T
5enes3	PC1	0	−16·0	24·9	−24·9	16·0	79·37	E
	PC2	2·6	−1·9	0·6	0·6	−1·9	20·63	T
5enes4	PC1	0	−17·2	26·6	−26·6	17·2	29·50	E
	PC2	2·3	−1·8	0·6	0·6	−1·8	20·48	T

almost planar 5-C rings. The major portion (almost 80%) of sample variance is bound up with PC1 of E symmetry, due to the limitation imposed by the double bond. For both 5mixs and 5anes the distinction between E and T coordinates is always clear, except for the sample 5anes6. This sample exhibits the minimum difference between the smallest loadings for a torsion angle in the two PC's, suggesting that both PC's map intermediates between ideal E and T conformations. We (subjectively) label the PC with the smallest loading as ET (= E, distorted towards T), and the other PC as TE (= T, distorted towards E).

Table 2 presents the correlations between the PC and the CP coordinates. In all 5anes and 5mixs the CP parameters are near perfectly correlated with the PC of corresponding symmetry, but in three samples (5anes1, 5anes4 and 5mixs3) this was only achieved after rejection of outliers. In 5anes1 the outlier comprised a cyclopentane ring whose pucker (q_2) was

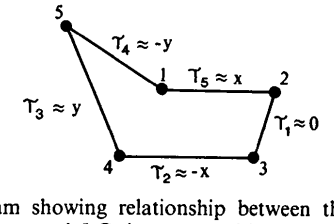


Fig. 5. Diagram showing relationship between the five torsion angles of a general 5-C ring approaching E conformation.

smaller by $> 4\sigma$ than the mean of the sample. An investigation revealed that the structure (Varughese & Chacko, 1978) exhibited disorder which was not flagged in the CSD. For 5anes4, the outlier (Ferguson, Parvez, McKervey, Ratananukul & Vibuljan, 1982) contained completely flat cyclopentene rings in addition to normal cyclopentanes, an accidental oversight in our data processing. Exclusion of these outliers had rather drastic effects: originally the correlations were between coordinates of different symmetry [*e.g.* PC1(E) with CP2(T), r = 0·881 for 5anes1], while after exclusion the correlations involved coordinates of the same symmetry (*i.e.* E with E, T with T). For 5mixs3, the cyclopentene fragment was joined to cyclohexene to form a conjugated system which led to severe distortion of the molecule (Lindeman, Timofeeva, Chernov, Reshetova & Struchkov, 1988), hence explaining its anomalous behaviour.

The PC/CP correlation coefficients for the 5anes and 5mixs are 0·999 in all cases save that of 5anes6. This implies [arccos(0·999) = 2·6°] that the PC coordinates are, on average, just under 3° out of phase with the CP coordinates. On purely numeric grounds we cannot exclude the possibility that this phase shift is due to round-off error by the single-precision algorithm. On the other hand, the PC/AS correlations for the 5ane samples (Table 2) all lie in the region of r = 0·82, clearly beyond the possibility of round-off error. This implies that the AS axes are on

Table 2. *List of* PC/CP *and* PC/AS *correlation coefficients* (× 1000)

Symmetry of the coordinates (E or T) is given in parentheses. Correlations given in parentheses are those obtained prior to exclusion of outliers; only correlations ≥0·720 are listed.

Sample		CP1(E)	CP2(T)	AS1(E)	AS2(T)
5anes1	PC1(E)	999	(881)	−827	(899)
	PC2(T)	(881)	999	(900)	827
5anes2	PC1(T)		999		813
	PC2(E)	999		−825	
5anes3	PC1(T)		−999		−823
	PC2(E)	999		−826	
5anes4	PC1(T)	(949)	999	(−953)	827
	PC2(E)	999	(949)	−827	(952)
5anes5	PC1(T)		999		827
	PC2(E)	999		−827	
5anes6	PC1(T)		994		740
	PC2(E)	994		−740	
5anes7	PC1(E)	999		−825	
	PC2(T)		999		826
5mixs1	PC1(T)		999		822
	PC2(E)	999		−822	
5mixs2	PC1(E)	999		−825	
	PC2(T)		−999		−815
5mixs3	PC1(T)		999 (−730)	(−978)	823
	PC2(E)	999 (−730)		−825	(968)
5mixs4	PC1(T)		−999		−826
	PC2(E)	999		−826	
5enes1	PC1(E)	−997	990	1000	
	PC2(T)				1000
5enes2	PC1(E)	998	−991	−1000	
	PC2(T)				1000
5enes3	PC1(E)	−998	993	1000	
	PC2(T)				1000
5enes4	PC1(E)	−999	995	1000	
	PC2(T)				1000

average about 35° out of phase with both the PC and the CP axes (assuming perfect PC/CP correlation), instead of the exact phase shift of $n \times 36°$ ($n = 0, 1, 2, 3...$) described in §2. This incremental difference may be ascribed to our current AS algorithm, which is not independent of the ordering of ring torsion angles. Thus, Jeffrey & Taylor (1980) have shown that the difference between CP and AS phases (φ_2 and P) will always involve a discrepancy ε, which can be of the order of 4° dependent on the ring being analysed.* Here, the discrepancy results in additional random scatter about the AS coordinates, and hence leads to reduced collinearity with the PC and CP axes. Even for the apparent outlier sample 5anes6, for which we could not justify the exclusion of any fragment, this CP/AS phase shift is maintained: while the CP coordinates are 6·3° out of phase with the PC's, the AS coordinates are shifted by 42·3°, *i.e.* 36° out of phase with the CP axes. One of the four PC/CP and PC/AS correlations always has an opposite sign to the remaining three, since one or other of the AS coordinates will always be anti-parallel to one of the CP coordinates, or *vice versa* (see §2).

Fig. 6 depicts PC, CP and AS scatterplots for representative samples of 5-C rings. In Fig. 6(*a*) the pseudorotation itinerary of cyclopentane (Fig. 1) is clearly mapped by the combination of PC1 and PC2; the similarity of this mapping to those of the CP and AS parameters is striking. Fig. 6(*b*) reveals the picture which would conceivably emerge from permutational expansion following a chemically non-specific search for 5-C rings. Here the perimeter of the circle is formed by cyclopentanes, while the 'spokes' result from cyclopentenes that are frozen into an E conformation by the double bond.

The pattern of PC/CP and PC/AS correlation coefficients ($|r| = 0·999$ between PC/CP coordinates of equivalent symmetry) which characterizes the 5anes and 5mixs is not reproduced by the 5enes. Here PC1(E) correlates almost equally to both CP1(E) and CP2(T) with non-identical $|r|$'s in the range 0·990 to 0·999, while PC2(T) no longer correlates with either ($|r|$ is always less than 0·140). However, the PC's now correlate perfectly with the AS coordinates of equivalent symmetry.

These changes are due to the imposition of C_s symmetry on the cyclopentene rings, eliminating the possibility of pseudorotation to a T conformer. The resulting one-dimensional nature of CP and AS space is clearly illustrated in Fig. 6(*c*). The data are not distributed along the CP (0 and 90°) coordinates, but along an E coordinate which is displaced from 0° by $n \times 36°$ ($n = 1, 2, 3...$); in this instance the E coordinate at 144° (324°) is a result of the chosen atomic enumeration. In the AS distribution, though, the data are distributed along the E coordinate at 90°. Again, this is a consequence of numbering the torsion angles (or atoms) so that τ_1 is the torsion angle about the double bond (see Fig. 5). The effect of applying the C_s atomic permutations (1 2 3 4 5; 4 3 2 1 5) is akin to measuring τ_1 twice: once clockwise and once anticlockwise. Hence, for each representative point displaced from an ideal AS coordinate by an increment x (perhaps as the result of experimental error) there will be a partner with an equal but opposite displacement of $-x$. The data are therefore scattered symmetrically about and along the AS1 coordinate, which in this case is coincident with PC1. The perfect PC2/AS2 correlation necessarily results from the mutual orthogonality of PC1/PC2 and AS1/AS2; it does not imply that the variance along the AS1/AS2 coordinates is in the same proportion as that along PC1/PC2. Indeed, the average variance for the 5enes along AS1 and AS2 is 1·963 and 0·021, respectively, while that along PC1 and PC2 is 3·963 and 1·036. Almost 99% of the variance along the AS coordinates therefore lies along AS1 – as it should, since the problem in this space is one-dimensional.

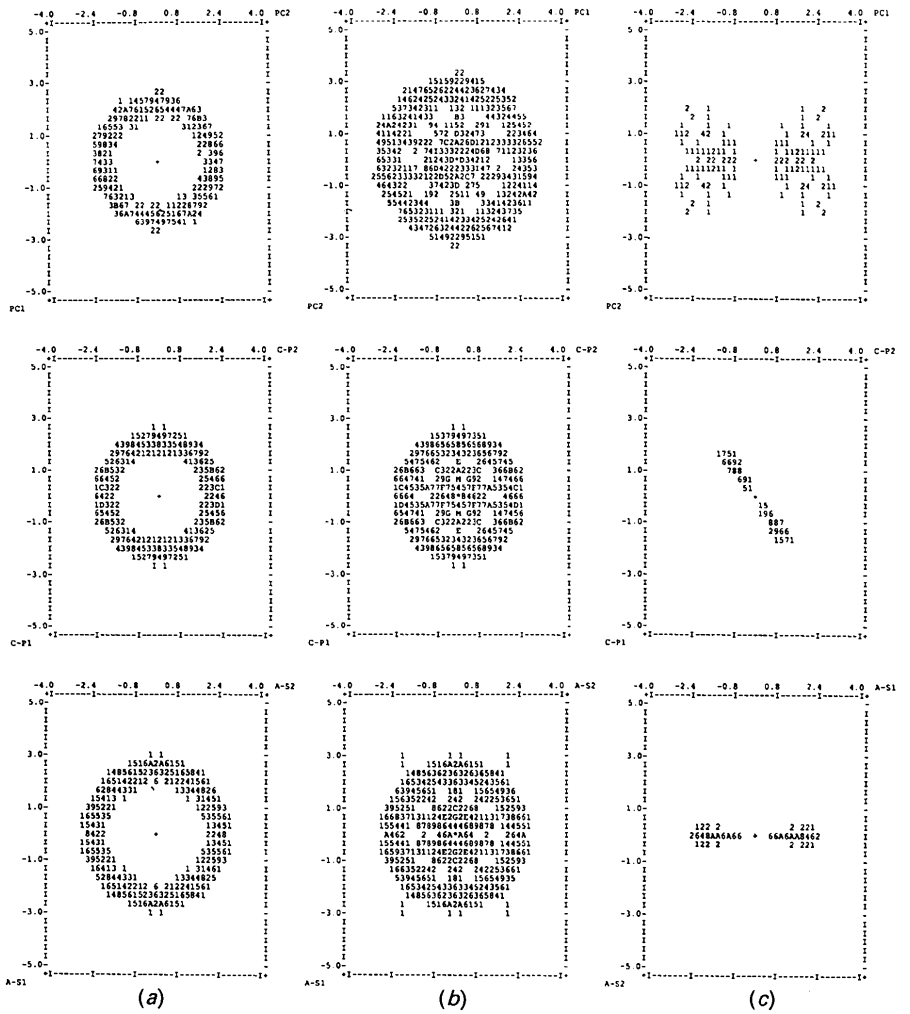
Even though the representative points are symmetrically distributed about the AS coordinates, each individual AS phase (P) will still differ from the equivalent CP phase (φ_2) by a whole number of

* A modification to the AS algorithm, in which the phase P is independent of numbering, has been published (Rao, Westhof & Sundaralingam, 1981). It will be implemented in *GSTAT* in due course.

phase differences, plus the random incremental value ε (Jeffrey & Taylor, 1980). In contrast to the 5anes and 5mixs, this effect will be manifested for the 5enes in inexact and irregular PC/CP correlations, due to the exact PC/AS ones. Moreover, since the (one-dimensional) data are now not coincident with either CP1 or CP2 (Fig. 6*c*), the PC describing the distribution will be correlated to both CP axes; this is observed in the correlation coefficients of Table 2. Since the data lie on a line inclined at 144° (and 324°) to the CP1(E) axis at 0°, it follows that the distribution is unequally angled towards the two CP axes, giving rise to slightly different correlation coefficients with them.

PCA for the 5enes has revealed two PC's which account for about 79 and 21%, respectively, of sample variance. In AS space these PC's coincide with the AS axes, but the variance along them is not in the same proportion as that along the PC's. In CP space, meanwhile, the first PC coincides with a coordinate running between the two CP axes, and the second PC coincides with nothing. In both cases, therefore, PC2 describes variance which cannot be accounted for in terms of the CP or AS methods. An examination of the correlation matrix for all 5ene samples reveals that the only other parameter correlated with PC2 (apart from AS2) is τ_1, the torsion angle about the double bond. This indicates that PC1 maps the extent of pucker at the envelope tip, the major source of variance in the sample, while PC2 describes additional variance arising from minor torsional distortions about τ_1. Expressed chemically, this means that while the double bond prohibits pseudorotation, it does not prevent τ_1 from being distorted away from 0°.

The PC plot of Fig. 6(*c*) would suggest an appreciable amount of variance in the second dimension, certainly much more than is the case for either the AS or the CP plot. The cut-off point, evident in the PC plot, results from the exclusion of flattened frag-



(*a*) (*b*) (*c*)

Fig. 6. Scattergrams of PC, CP and AS scores (top, middle and bottom, respectively) for samples (*a*) 5anes1, (*b*) 5mixs1 and (*c*) 5enes4. The CP scores have been scaled by a factor of 5, the AS scores by 0·055.

Table 3. *Listing of the coefficients for the torsion angles in the* PC's *for* 6-C *ring samples*

% = percentage variance accounted for by each PC, and S = symmetry of PC (see text).

Sample	PC	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	%	S
6-C(1)	PC1	−33·7	33·7	−33·7	33·7	−33·7	33·7	49·38	C
	PC2	−21·5	−12·1	33·6	−21·5	−12·1	33·6	25·28	TB
	PC3	26·4	−31·8	5·4	26·4	−31·8	5·4	25·28	B
6-C(2)	PC1	38·7	−38·5	38·8	−39·8	40·2	−39·4	75·36	C
	PC2	2·0	20·2	−22·2	1·7	20·3	−22·2	14·47	B
	PC3	−19·7	11·4	8·7	−20·2	11·3	8·4	10·11	TB

ments with $q_2 < 0{\cdot}1$. An initial data survey had revealed a number of almost planar rings, usually with conjugated bonds, where φ_2 was far from 144° (or 324°). These were inconsistent with the rest of the sample, since the φ_2 values were random and did not correspond to any recognizable position on the pseudorotation itinerary: as the rings flatten, the E/T descriptions become more and more meaningless. In practice, these exclusions scarcely influenced the results of PCA.

Table 4. *Correlations* (×1000) *between* PC *and* CP *coordinates for* 6-C *samples*

The symmetry of each coordinate is given in parentheses.

Sample		CP1(C)	CP2(B)	CP3(TB)
6-C(1)	PC1(C)	1000		
	PC2(TB)		−774	
	PC3(B)			774
6-C(2)	PC1(C)	1000		
	PC2(B)		996	
	PC3(TB)			−992

6-C *rings*

In both the permutationally expanded data set 6-C(1) and in the untreated 6-C(2), as well as in numerous other samples which we tested, PCA extracts three PC's which together account for a minimum 99·94% of variance. Table 3 lists their torsion-angle loadings, the variance accounted for by each PC and its symmetry (established as for the 5-C rings). Table 4 lists the PC/CP correlations.

The PC's for the symmetrized 6-C(1) data set manifest exact symmetry: PC1 maps the C coordinate along which the relationships between τ_1–τ_6 maintain D_{3d} symmetry, while both PC2 and PC3 trace coordinates with C_2 symmetry. The loadings of PC2 suggest a TB conformation, while those of PC3 indicate a coordinate maintaining a B conformation. However, the data in Table 4 suggest that PC2 (TB symmetry) is badly correlated with the CP axis of B symmetry, *i.e.* the PC correlates with a CP axis of different symmetry. In fact, the correlation places PC2(TB) at 39·3° to the CP2(B) axis, *i.e.* within 9·3° of the next CP(TB) axis, which lies at 30° to CP2(B). It can be seen, therefore, that PC2(TB) lies close to a CP axis of similar symmetry, even though this may not be immediately obvious. A similar argument can be advanced that PC3(B) correlates closely with a CP B coordinate, rather than the TB coordinate which the correlation matrix would suggest.

For the unexpanded 6-C(2) data set the PC coordinates approach C, B and TB conformations, though (naturally) they do not exhibit ideal symmetry. There is a straightforward correlation between PC and CP coordinates of similar symmetry, with the axes being misaligned by about 6°. In both 6-C samples, therefore, we observe a slight rotation of the PC axes away from the CP coordinates within the plane of the B⟷TB pseudorotation itinerary. Moreover, the maximum variance in both samples is described by the C axis, suggesting that variation in chair conformers is most important in accounting for conformational variability amongst 6-C rings.

The results from the symmetrized data set are similar but not identical to those of the random sample: both groups yield just three PC's, two of which are of near equal importance. Furthermore, in both cases PC1 (mapping the chair conformers) is a unique axis in terms of both the percentage sample variance that it accounts for, and of its exact correlation with CP1 or (q_3). For sample 6-C(2) the former may simply result from a preponderance of chair conformers, but for 6-C(1) this is not so, since this sample contains fewer chairs (20) than it does boat conformers (22) or others (29). It appears that PC1 and, by implication, CP1 (or q_3), represent inherently unique coordinates. Fig. 7 depicts scattergrams of the PC scores and the CP parameters for both 6-C data sets. The diagrams represent two-dimensional sections through the conformational sphere of Fig. 2, with all data projected onto the equatorial plane (Figs. 7*a*,*c*) or onto a plane perpendicular to it (Figs. 7*b*,*d*). As with the 5-C rings the structural similarity between the conformational mapping emerging from PCA, and that traced by the CP parameters is strikingly obvious.

6. Discussion

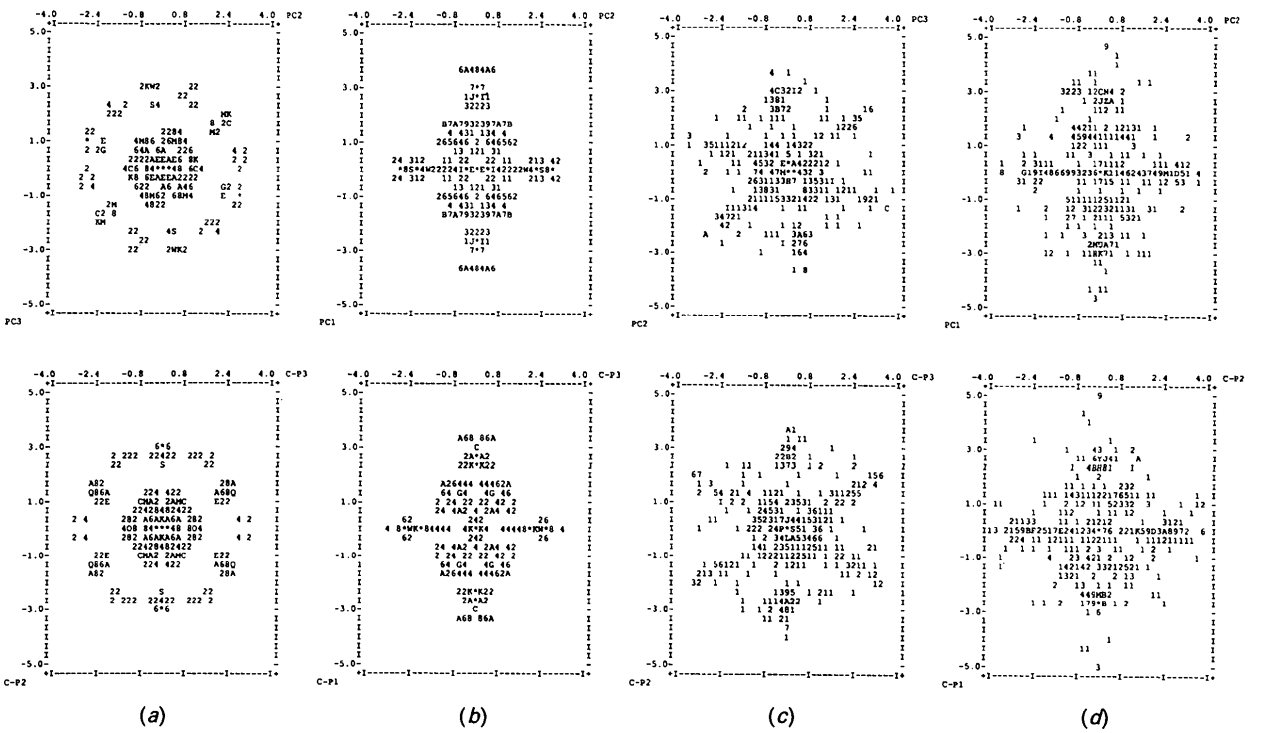
PCA of the torsional conformation spaces examined here has provided a dimension reduction which is in complete accordance with that afforded by the CP method. Whilst this reduction is a natural (geome-

tric) consequence of linear relationships between torsion angles under the constraint of ring closure, these relationships could, in general, only have been derived with considerable effort. In any event, our purpose was not to deduce these geometric relationships, but to show how PCA can be used to analyse (and to reduce) hyperdimensional spaces. The carbocyclic systems were chosen deliberately, since their hyperdimensional conformation spaces can be visualized in lower dimensions (*via* the CP and AS methods), permitting a better comparative interpretation of the PCA results. Our approach was foreshadowed by Murray-Rust & Motherwell (1978), who examined the molecular geometry of β-nucleosides using PCA. They did not consider the topological symmetry of the five-membered ring fragments as separate units, but they did discover that two PC's account for the ring geometry, and that the PC scatterplot bore a strong resemblance to an AS pseudorotational plot.

The present study also shows that PCA reproduces correctly the 'characteristics' of the CP conformation spaces. Thus, for 6-C rings the CP puckering coordinates q_2, φ_2, q_3 may be deemed to "span the conformational space of a six-membered ring, which contains one pseudorotational subspace of dimension two (q_2,φ_2) and one inversional subspace of dimension one (q_3)" (Cremer, 1980). This characteristic subdivision is mirrored by the PCA on the symmetrized 6-C sample. For 5-C rings only a pseudorotational subspace of dimension two (q_2,φ_2) exists, and again this is reproduced by PCA of the 5anes and 5mixs samples. Moreover, the energetic equivalence (Lifson & Warshel, 1968; Kilpatrick, Pitzer & Spitzer, 1947; Dunitz, 1979) of the E,T and B,TB coordinates in the 5-C and 6-C CP conformational spaces is reproduced by the equivalence in variance accounted for by the corresponding PC's.

In most other studies a correlation coefficient of $r = 0{\cdot}999$ would imply an exact linear relationship between the two corresponding parameters. Here, due to the single-precision arithmetic, it was initially uncertain whether 0·999 was significantly different from 1·000, and consequently it was not certain whether or not the PC's coincided exactly with the CP coordinates. At the start of this study, though, we were expecting to find this exact correlation, an expectation which we now believe somewhat oversimplified the problem.

When a PCA yields k identical eigenvalues of a correlation or covariance matrix, then this means (Chatfield & Collins, 1980, p. 65) that: (i) the corresponding k eigenvectors (and hence the PC's) will have the same variance; (ii) any other orthonormal set of k eigenvectors could have been chosen in the appropriate subspace of k dimensions; (iii) a



(*a*) (*b*) (*c*) (*d*)

Fig. 7. Scattergrams of PC and CP scores for symmetrized sample 6-C(1) [(*a*) and (*b*)], and random sample 6-C(2) [(*c*) and (*d*)].

spherical variance in the corresponding k dimensions is implied. The PCA on the 5anes, 5mixs and on the symmetrized 6-C(1) data revealed two equivalent PC's with identical variance in each case [implication (i), above]. However, by implication (ii) the resultant PC's are not unique and any other orthonormal set could have been chosen. Thus, the PC's extracted for the 5-C rings are never of exact E or T symmetry (except for the 5enes), but are always slightly displaced. Had they been of exact symmetry, then they would have been equivalent to the E and T coordinates of CP (or AS) space. Since they are not, this suggests that the PC/CP correlation coefficients are meaningful, and do indicate a slight misalignment of coordinates. This conclusion is supported by the results for 5anes6, where the PC's are the most displaced of all 5-C samples, both in terms of the r value (0·994) and the torsion-angle loadings (see §5). These displacements are even more pronounced for sample 6-C(1), where the smallest loading in PC3 (B symmetry) is larger than the equivalent loading of PC2 for 5anes6 (−5·1 as opposed to −3·3), and coordinate misalignment is correspondingly larger (9·3° compared to 6·3°). Thus, any almost exact correlations between the PC and the CP coordinates were somewhat fortuitous, and might look quite different for other samples.

Implication (ii), above, is actually a consequence of (iii), since a spherical variance implies that no two orthogonal axes can account for the variance in a unique way. The PC plots of Figs. 6 and 7 reveal circular (or spherical) distributions which obviously reflect similar distributions in the hyperdimensional torsion spaces. These result from the equipotential nature of the corresponding pseudorotation coordinates (see Lifson & Warshel, 1968; Kilpatrick, Pitzer & Spitzer, 1947; Dunitz, 1979, p. 435). Hence there are no *a priori* reasons why the PCA should yield axes that are coincident with CP (or AS) coordinates of exact symmetry, *i.e.* coordinates lying along special directions in these conformation spaces.

Finally, we have observed the possible dramatic effect of outliers on the PCA. Results can change from being chemically unintelligible to being chemically informative on the exclusion of just one entry. This fact is well recognized in standard texts on PCA which stress the importance of investigating (and taking into account) any 'erroneous' data, whether this results from experimental error or from chemical effects (Malinowski & Howery, 1980, ch. 4 and ch. 9, pp. 130–132). In this study PCA was perhaps particularly sensitive to the effects of outliers, due to the circular/spherical nature of the data distributions. However, we would urge all prospective users to examine their data very carefully for outliers, using standard deviation analysis, histograms and (especially) the initial PC plots.

Despite these strictures, we conclude that, when the PCA technique is applied to fully symmetrized data sets, the PC plots are invaluable (*a*) for providing chemically sensible mappings of conformation space, and (*b*) for the visualization of any conformational interconversion pathways that may be present. The unambiguous mapping of the pseudorotational pathways in 5-C and 6-C systems (Figs. 6 and 7) is clear evidence of the visual utility of the PC approach to conformational problems.

References

ALLEN, F. H. (1984). *Acta Cryst.* **B40**, 64–72.

ALLEN, F. H. & DAVIES, J. E. (1988). *Crystallographic Computing*, Vol. 4, edited by N. W. ISAACS & M. R. TAYLOR, pp. 271–289. Oxford Univ. Press.

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

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*b*). *Acta Cryst.* **B47**, 41–49.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*c*). *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**, 404–412.

ALTONA, C., GEISE, H. J. & ROMERS, C. (1968). *Tetrahedron*, **24**, 13–32.

ALTONA, C. & SUNDARALINGAM, M. (1972). *J. Am. Chem. Soc.* **94**, 8205–8212.

AUF DER HEYDE, T. P. E. (1990). *J. Chem. Educ.* **67**, 461–469.

AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*a*). *Inorg. Chem.* **28**, 3960–3969.

AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*b*). *Inorg. Chem.* **28**, 3970–3981.

AUF DER HEYDE, T. P. E. & BÜRGI, H.-B. (1989*c*). *Inorg. Chem.* **28**, 3982–3991.

CHATFIELD, C. & COLLINS, A. J. (1980). *Introduction to Multivariate Analysis*. London: Chapman & Hall.

CREMER, D. (1980). *Isr. J. Chem.* **20**, 12–19.

CREMER, D. & POPLE, J. A. (1975). *J. Am. Chem. Soc.* **97**, 1354–1358.

CSD User Manual (1990). Version 4.2. Crystallographic Data Centre, Cambridge, England.

DUNITZ, J. D. (1979). *X-ray Analysis and the Structure of Organic Molecules*. Ithaca: Cornell Univ. Press.

FERGUSON, G., PARVEZ, M., MCKERVEY, M. A., RATANANUKUL, P. & VIBULJAN, P. (1982). *Acta Cryst.* **B38**, 2316–2318.

HUMMEL, W., HUML, K. & BÜRGI, H.-B. (1988). *Helv. Chim. Acta*, **71**, 1291–1302.

HUMMEL, W., ROSZAK, A. & BÜRGI, H.-B. (1988). *Helv. Chim. Acta*, **71**, 1281–1290.

JEFFREY, G. A. & TAYLOR, R. (1980). *Carbohydr. Res.* **81**, 182–183.

KILPATRICK, J. E., PITZER, K. S. & SPITZER, R. (1947). *J. Am. Chem. Soc.* **69**, 2483–2488.

LIFSON, S. & WARSHEL, A. (1968). *J. Chem. Phys.* **49**, 5116–5129.

LINDEMAN, S. V., TIMOFEEVA, T. V., CHERNOV, S. V., RESHETOVA, I. G. & STRUCHKOV, YU. T. (1988). *Bioorg. Chem.* **14**, 397–405.

MALINOWSKI, E. R. & HOWERY, D. G. (1980). *Factor Analysis in Chemistry*. New York: John Wiley.

MASSART, D. L. & KAUFMAN, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley.

MURRAY-RUST, P. (1982). *Acta Cryst.* **B38**, 2765–2771.
MURRAY-RUST, P. & BLAND, R. (1978). *Acta Cryst.* **B34**, 2527–2533.
MURRAY-RUST, P. & MOTHERWELL, W. D. S. (1978). *Acta Cryst.* **B34**, 2534–2546.
MURRAY-RUST, P. & RAFTERY, J. (1985*a*). *J. Mol. Graphics*, **3**, 50–59.
MURRAY-RUST, P. & RAFTERY, J. (1985*b*). *J. Mol. Graphics*, **3**, 60–68.
NORSKOV-LAURITSEN, L. & BÜRGI, H.-B. (1985). *J. Comput. Chem.* **6**, 216–228.
RAO, S. T., WESTHOF, E. & SUNDARALINGAM, M. (1981). *Acta Cryst.* A**37**, 421–425.
SNEDECOR, G. W. & COCHRAN, W. G. (1980). *Statistical Methods*, 7th ed. Ames: Iowa State Univ. Press.
TAYLOR, R. (1986). *J. Appl. Cryst.* **19**, 90–91.
VARUGHESE, K. I. & CHACKO, K. K. (1978). *Cryst. Struct. Commun.* **7**, 149–152.