Acta Cryst. (1991). B47, 62–67

Automated Conformational Analysis from Crystallographic Data. 4.* Statistical Descriptors for a Distribution of Torsion Angles

BY FRANK H. ALLEN† AND OWEN JOHNSON

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

(*Received* 13 *March* 1990; *accepted* 12 *September* 1990)

Abstract

The methods of linear and circular statistics are used to derive summary descriptors for a unimodal distribution of torsion angles (τ_i). The arithmetic methods of normal linear statistics are complicated by the phase restriction $-180 < \tau_i \leq 180°$. Phase adjustments must be made to generate a correct arithmetic mean ($\bar{\tau}_a$) which spans $\pm 180°$ in a Newman projection of the τ_i. A general single-pass algorithm is described for the calculation of this mean, its standard error $\sigma(\bar{\tau}_a)$ and the sample standard deviation $\sigma(\tau_a)$. The single-pass technique is restricted to distributions in the range $r \leq 180°$; a preliminary pass is required for broader distributions. The τ_i are, however, properly represented as a circular distribution and the established formalisms of circular statistics should be applied. Here a *circular mean* or *preferred direction* ($\bar{\tau}_c$) may be derived in a single pass for a distribution of any range. The variance of the distribution may be assessed in terms of the *concentration*, $\bar{R}$, of data points around the mean $\bar{\tau}_c$. A circular standard error $\sigma(\bar{\tau}_c)$ and a circular sample standard deviation $\sigma(\tau_c)$ may then be derived. It is shown that the arithmetic and circular descriptors are numerically similar, except for broad distributions. The circular method has computational advantages in minimizing phase-shift operations and the results are more realistic and reliable when used in further statistical tests.

1. Introduction

Previous papers in this series (Allen, Doyle & Taylor, 1991*a–c*; hereafter ADT1, ADT2, ADT3) have described symmetry-modified cluster-analysis algorithms for the identification of conformational minima. The techniques were applied to a number of substructural fragments located *via* searches of the Cambridge Structural Database (CSD; Allen, Kennard & Taylor, 1983; Allen & Davies, 1988). Fragment conformations were defined by N_t torsion angles for each of the N_f occurrences of the fragment. The algorithms (ADT1, ADT2) then attempt to divide this multivariate data matrix into homogenous conformational subgroups, taking account of fragment toposymmetry and enantiomorphic inversions where appropriate. For each conformational subgroup, of population N_p, we obtain N_t distributions of torsion angles, each distribution containing N_p values.

An integral part of this work involves the derivation of simple summary statistics for each of the N_t distributions, assumed to be unimodal if the clustering process has been successful. In our first implementation (ADT3) we used the standard formulae of linear statistics to calculate the arithmetic mean $\bar{\tau}_a$, its standard error $\sigma(\bar{\tau}_a)$, and the sample standard deviation $\sigma(\tau_a)$, for each distribution. These quantities are important (ADT1, ADT3): (*a*) in assessing the most representative fragment in any cluster, *i.e.* that fragment whose torsion-angle sequence $\tau_i (i = 1 \rightarrow N_t)$ most closely matches the sequence of means $(\bar{\tau}_a)_i$; (*b*) in giving some indication, through the $\sigma(\bar{\tau}_a)_i$ and $\sigma(\tau_a)_i$, of the conformational homogeneity of any cluster; and (*c*) in assessing intercluster separations in conformational space, in terms of dissimilarities between the mean torsional sequences for each unique pair of clusters.

The accepted definition of a torsion angle τ (Klyne & Prelog, 1960) generates the phase restriction $-180 < \tau \leq 180°$, *i.e.* τ is a circular function for which $\pm 180°$ are equivalent limiting values. This restriction causes obvious problems in the systematic generation of arithmetic means, $\bar{\tau}_a$, for distributions of τ; these difficulties are briefly summarized in the first section of this paper. We also show how some of these problems can be overcome and assess the computational requirements of the solutions proposed.

The methods of linear statistics are, of course, formally correct only when applied to distributions of linear functions [*e.g.*, bond lengths (Allen, Kennard, Watson, Orpen, Brammer & Taylor, 1987)]. For torsion angles, and for other angular data, the methodology of circular statistics is more appro-

* Part 3: Allen, Doyle & Taylor (1991*c*).

† Author for correspondence.

0108-7681/91/010062-06$03.00

priate (see, *e.g.*, Mardia, 1972; Upton & Fingleton, 1989). In the second part of this paper we apply this formalism to the calculation of the circular mean direction, $\bar{\tau}_c$, and its associated error estimates. We also examine the circumstances in which the arithmetic mean, $\bar{\tau}_a$, may be regarded as a satisfactory approximation to $\bar{\tau}_c$.

2. The arithmetic mean

The arithmetic mean of a distribution of torsion angles $\tau_i (i = 1 \rightarrow n)$ is:

$$\bar{\tau}_a = \left(\sum_{i=1}^{n} \tau_i \right) / n \qquad (1)$$

and the sample standard deviation is given by:

$$\sigma(\tau_a) = \left[\sum_{i=1}^{n} (\bar{\tau}_a - \tau_i)^2 / (n-1) \right]^{1/2}, \qquad (2a)$$

or by:

$$\sigma(\tau_a) = \left\{ \left[n \sum_{i=1}^{n} (\tau_i)^2 - \left(\sum_{i=1}^{n} \tau_i \right)^2 \right] / [n(n-1)] \right\}^{1/2}. \qquad (2b)$$

Equation (2*b*) is computationally preferable, since the necessary summations may be accumulated during a single pass through the distribution. The standard error of the mean is then:

$$\sigma(\bar{\tau}_a) = \sigma(\tau_a) / n^{1/2}. \qquad (3)$$

Equations (1)–(3) are applicable so long as the range, r, of the distribution (however broad) does not span the $+180°$ limit (*e.g.*, r_a in Fig. 1). Where the range does span $+180°$ (r_b in Fig. 1) the simple mean $b_1 = -45°$ (Fig. 1) is generated for $\tau_i = -110, -175, +150°$, rather than the correct value $b_2 = -165°$ (Fig. 1) which is obtained if $+150°$ is expressed as $-210°$. For computational efficiency we require a treatment which is independent of convention and by which (1)–(3) may be applied in a single pass through the distribution.

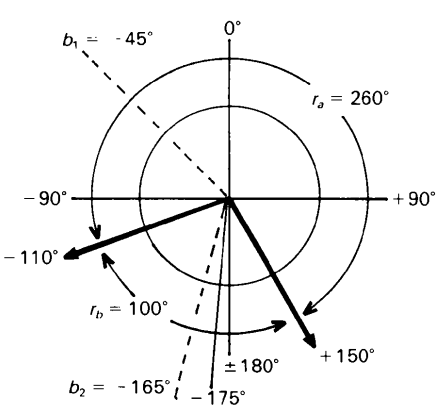

Fig. 1. Two complementary τ distributions. Distribution (*a*) spans the clockwise range r_a from -110 to $+150°$. Distribution (*b*) spans the clockwise range r_b from $+150$ to $-110°$ and includes $-175°$ as a data point; b_1 is the straightforward arithmetic mean of $\tau_i = -110, -175, +150°$, b_2 is the correct arithmetic mean after allowing for the phase change at $\pm 180°$.

One way to accomplish this is to redefine the (arbitrary) torsion-angle origin of 0° to lie *within* the observed distribution. Since we have no *a priori* knowledge of that distribution, it is convenient to redefine the value of the first angle in the list as $\tau'_1 = 0°$. The original value of τ_1 now becomes a constant offset (denoted τ_o) to be applied to the other $\tau_i (i = 2 \rightarrow n)$ to effect an origin shift to a new distribution τ'_i. If we further define for each torsion angle in turn:

$$X = \tau_i - \tau_o (= 0° \text{ for } \tau_1) \qquad (4)$$

then for $\tau_2 \rightarrow \tau_n$ we have:

$$\tau'_i = X \qquad \text{for } -180 < X \leq 180° \qquad (5a)$$

$$\tau'_i = |X| - 360 \qquad \text{for } X > 180° \qquad (5b)$$

$$\tau'_i = 360 - |X| \qquad \text{for } X < -180°. \qquad (5c)$$

Equations (5*b*) and (5*c*) can be verified by reference to distribution *b* of Fig. 1. If $\tau_o = -110°$, then $X = 260°$ for $\tau_i = 150°$ and τ'_i is then $-100°$ by equation (5*b*). Conversely, if $\tau_o = 150°$, then $X = -260°$ for $\tau_i = -110°$ and τ'_i is then $+100°$ by equation (5*c*).

We may now apply equations (4), (5) and (1), (2*b*), (3) in a single pass through the original distribution to obtain $\bar{\tau}'_a$, $\sigma(\bar{\tau}'_a)$ and $\sigma(\tau'_a)$. The σ's will be numerically equivalent to the values for the unprimed distribution, since they depend on squares of angular differences. The true arithmetic mean is then obtained by setting:

$$X = \bar{\tau}'_a + \tau_o \qquad (6)$$

and re-applying (5*a*)–(5*c*) as appropriate.

The use of the phase-shifted τ'_i distribution to solve the $\pm 180°$ problem does, however, have limitations for *any* distribution of τ_i wider than 180°, irrespective of whether or not it includes the $\pm 180°$ point. Thus distribution (*a*) of Fig. 1, previously amenable to equations (1)–(3) in its original 'unshifted' form, will only be correctly transformed to its τ_i equivalent [*via* equations (4), (5)] if the chosen τ_o lies within the range r_o^a indicated on Fig. 2. With τ_o chosen in this way, all values of $|\tau_i - \tau_o|$ will be $\leq 180°$. The τ_o validity range (r_o) for any range (r) of τ_i is given by:

$$r_o = [r - 2(r - 180)]° = (360 - r)°. \qquad (7)$$

Thus r_o decreases to 0° as r approaches 360°, indicating (correctly!) that the mean of a circular distribution is increasingly meaningless, and almost certainly not unimodal, as its range increases to 360°. Equation (7) also indicates that, as r *decreases* from 180°,

then r_o increases above 180°, *i.e.* τ_o need not necessarily be chosen from amongst the τ_i themselves. Distribution (*b*) (+150 to −110°) of Fig. 1 is the circular complement ($r_b = 100°$) of distribution (*a*); here r_o^b (Fig. 2) is now 260° or $360 - r_o^a$. It is the choice of τ_o from within the distribution which preserves the single-pass application of equations (4)–(6), a choice which is generally valid for distributions with $r \leq 180°$.

If we can be convinced that calculation of the mean value of a τ distribution with $r > 180°$ has some validity, then equations (4)–(6) may be used provided τ_o is correctly placed. To do this we must sacrifice the convenience of a single-pass procedure to gain initial knowledge of the τ_i distribution so that τ_o may be chosen correctly. Specifically, we need to know the bounds of the distribution (τ_x, τ_y, Fig. 2) and which arc (*a* or *b*) carries that distribution. This can be done by sorting the τ_i, to obtain the ordered distribution τ_k, and then locating the largest gap between neighbouring points k and $k+1$ in the sorted list. Gaps may be assessed using the city-block metric (see ADT1) employed in dissimilarity calculations, recast as:

$$G_k = \min[(|\tau_{k+1} - \tau_k|), (360 - |\tau_{k+1} - \tau_k|)] \quad (8)$$

to account for gaps spanning the +180° position. The adjacent pair, τ_k, τ_{k+1}, which gives rise to the largest G_k, then represents τ_x, τ_y in Fig. 2. Any other τ_k will serve to indicate the relevant arc (*a* or *b*) for the distribution. In this case τ_o should not be chosen directly from the τ_i values, but calculated as the bisector of the arc containing the distribution (+20° for *a* in Fig. 2). This ensures that τ_o lies within the r_o for the distribution. Equations (4), (5), (1)–(3) and (6) may now be applied, but with (4) now covering all $i = 1 \rightarrow n$ torsion angles. We have tested this two-pass generation of statistics from a sorted τ_k distribution and find it to be completely general. This complexity can be avoided, however, by the procedures detailed in the next section.

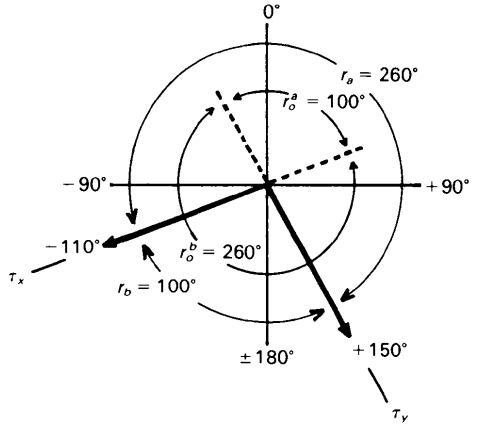

Fig. 2. The r_o ranges [see text and equation (7)] for distributions (*a*) and (*b*) of Fig. 1.

3. The circular mean

The Newman (1955) projection of a torsion angle τ (*ABCD*) may be regarded (Fig. 3) as a unit vector *CD* with direction $\tau°$ [consistent with the Klyne & Prelog (1960) convention] measured from the vector *BA* chosen to coincide with the (vertical) *y* axis. It then has components $\sin\tau$ and $\cos\tau$ parallel to *x* and *y* respectively. The resultant vector, *R*, of a distribution of torsion angles taken in any order is then:

$$R^2 = \left(\sum_{i=1}^{n} \sin\tau_i\right)^2 + \left(\sum_{i=1}^{n} \cos\tau_i\right)^2 \quad (9)$$

whence the *circular mean* or *preferred direction* (Upton & Fingleton, 1989) is given by:

$$\bar{\tau}_c = \tan^{-1}\left[\left(\sum_{i=1}^{n} \sin\tau_i\right)\left(\sum_{i=1}^{n} \cos\tau_i\right)\right]$$

$$= \tan^{-1}(C_x/C_y) \quad (10)$$

where C_x and C_y are the sums of the vector components along *x* and *y* respectively. Because of the torsional phase change at ±180°, a more complete definition is needed to ensure the correct radial value of $\bar{\tau}_c$:

$$\bar{\tau}_c = \begin{cases} \tan^{-1}(C_x/C_y) & \text{for } y > 0 \\ \tan^{-1}(C_x/C_y) + 180° & \text{for } x > 0,\ y < 0 \\ \tan^{-1}(C_x/C_y) - 180° & \text{for } x < 0,\ y < 0. \end{cases} \quad (11)$$

Equation (11) differs from that given by Upton & Fingleton (1989), who use the 0–360° convention for angular measurement often employed in this area of

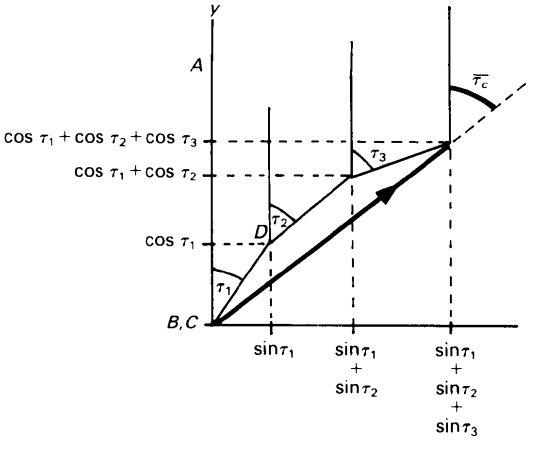

Fig. 3. The resultant vector, *R*, of a distribution of torsion angles, τ_1–τ_3, represented as unit vectors whose direction is measured clockwise from the *y* axis.

statistics. Obviously $\bar{\tau}_c$ may be evaluated for any range r in a single pass through the torsional distribution.

The length of the resultant vector R [equation (9)] provides some measure of the variance of a circular distribution. If all unit vectors are collinear, *i.e.* all τ_i are equal, then $R = n$, the number of angles in the distribution. If the data have mutually opposed directions which cancel completely (*e.g.*, $\tau_1 = +30$, $\tau_2 = -150°$), then $R = 0$ and there is no preferred direction. This result should be compared with a finite arithmetic mean, $\bar{\tau}_a$, of $-60°$. The mean vector length:

$$\bar{R} = R/n \tag{12}$$

then represents the *concentration* (Upton & Fingleton, 1989) of the τ_i about the mean direction $\bar{\tau}_c$. Values of R approaching unity correspond to increasingly narrow unimodal distributions, *i.e.* to reducing variances in the samples.

Various methods have been proposed (Mardia, 1972; Batschelet, 1981) for relating R or $\bar{R}$ to some estimate of circular variance. Here we use the treatment of Fisher & Lewis (1983), who approximate V, the *circular variance* of $\bar{\tau}_c$ for a well-populated unimodal distribution as:

$$V = n(1 - \varphi)/4R^2 \tag{13}$$

where:

$$\varphi = \left[\cos(2\bar{\tau}_c) \sum_{i=1}^{n} \cos(2\tau_i) + \sin(2\bar{\tau}_c) \sum_{i=1}^{n} \sin(2\tau_i) \right] / n. \tag{14}$$

The $(1 - \alpha)$ confidence interval for the mean direction is then given by the central limit theorem as:

$$\bar{\tau}_c \pm \sin^{-1}[u_\alpha (2V)^{1/2}] = \bar{\tau}_c \pm \sin^{-1}(u_\alpha \hat{\sigma}_c) \tag{15}$$

where u_α is the upper $0{\cdot}5\alpha$ point of a unit normal distribution and $\hat{\sigma}_c$ is termed (Fisher & Lewis, 1983) as the *estimated circular standard error* of $\bar{\tau}_c$, *i.e.* the circular analogue of the standard error of the sample mean.

The quantity $\hat{\sigma}_c$ is dimensionless, only attaining physical reality (in degrees) in combination with u_α *via* the arcsin function. For a normal distribution of a linear variable x, the standard error of the mean, $\sigma(\bar{x})$ [equation (3)], represents a 68·3% confidence interval (since $u_\alpha = 1{\cdot}0$). Hence we may estimate a $\sigma(\bar{\tau}_c)$ in degrees as:

$$\sigma(\bar{\tau}_c) = \sin^{-1}[(2V)^{1/2}] = \sin^{-1}(\hat{\sigma}_c) \tag{16}$$

whence the standard deviation of the sample may be estimated by analogy with equation (3) as:

$$\sigma(\tau_c) = n^{1/2} \sigma(\bar{\tau}_c). \tag{17}$$

We perform these transformations to provide a direct comparison between the arithmetic and circular results, thus providing a formalism for the assessment of τ_i distributions which is common to that for linear distributions. Comparison of equations (15) and (16) shows that this common formalism is acceptable only when:

$$\sin^{-1}(u_\alpha \hat{\sigma}_c) \simeq u_\alpha \sin^{-1}(\hat{\sigma}_c) \tag{18a}$$

or equivalently when:

$$\sin[u_\alpha \sigma(\bar{\tau}_c)] \simeq u_\alpha \sin[\sigma(\bar{\tau}_c)]. \tag{18b}$$

Equations (18) are obviously true for small values of $\hat{\sigma}_c$ and $\sigma(\bar{\tau}_c)$. Indeed, for the 99·75% confidence limit ($u_\alpha = 3{\cdot}0$) usually employed in crystallography (Jeffrey & Cruickshank, 1953), the approximation of equation (18) becomes untenable beyond a $\sigma(\bar{\tau}_c)$ value of *ca* 15° corresponding to a $\hat{\sigma}_c$ value of *ca* 0·26.

A useful property of equation (15) is that if $u_\alpha (2V)^{1/2}$ is greater than unity, then the confidence interval is undefined. In such cases the mean direction, $\bar{\tau}_c$, is so poorly determined that it may not be worth reporting. For $u_\alpha = 1{\cdot}0$, the maximum possible value of $\sigma(\bar{\tau}_c)$ from equation (16) is 90°. From a computational viewpoint, equations (13)–(17) are efficient, since they depend upon summations in (14) which can be accumulated, along with those of equation (10), in a single pass through the τ_i distribution.

4. Limits and range of the distribution

A knowledge of the limits and range of a τ distribution (*i.e.* τ_x, τ_y, r in Fig. 2) is not required for the derivation of $\bar{\tau}_c$. It is useful, however, to include these quantities in any summary. For a linear distribution the upper and lower limits are simply established arithmetically during the averaging pass through the data. For a circular distribution these concepts are not applicable: the distribution (*b*) of Fig. 2 (+150 to −110°) will doubtless contain values close to ±180° which would qualify as upper and lower bounds in purely arithmetic terms.

A simple solution is possible for all distributions where $r < 180°$. We assign initial values of (say) −999, +999, −999, +999° to the maximum and minimum absolute values of both the positive and negative torsion angles, denoted as $|\tau^+\text{max}|$, $|\tau^+\text{min}|$, $|\tau^-\text{max}|$, $|\tau^-\text{min}|$ respectively. In general all four will have values ≤ 180° after a single-pass analysis of the data distributions (*a–d*) in Fig. 4. Limits are either $-|\tau^-\text{max}|$ to $|\tau^+\text{max}|$ (*a*,*c*) or $|\tau^+\text{min}|$ to $|\tau^-\text{min}|$ (*b*,*d*). Two special cases exist (distributions *e* and *f* in Fig. 4) in which angles in the distribution are either all positive or all negative. Here, either the second pair (in *e*) or the first pair (in *f*) will retain their initial values at the end of the pass. The values obtained for

$|\tau^+\text{min}|$, $|\tau^-\text{max}|$ in (*e*) and $-|\tau^-\text{min}|$, $-|\tau^-\text{max}|$ in (*f*) are obviously the limiting values for these distributions. The assignment of labels 'upper' and 'lower' to these limits is now only a matter of definition. In this work we define the rotation from lower to upper in a clockwise sense, by analogy with Klyne & Prelog (1960). The limits for (*e*) remain as above, but those for (*f*) must be interchanged to comply with this definition. Derivation of the range r is a simple matter. In the more general case, of which the distributions of Figs. 4(*a–d*) are examples, the range r is either:

$$r_1 = |\tau^-\text{max}| + |\tau^+\text{max}| \qquad (19)$$

or

$$r_2 = 360 - (|\tau^-\text{min}| + |\tau^+\text{min}|). \qquad (20)$$

We simply test that either r_1, *or* r_2 lies in the range $0 \leq r_1, r_2 \leq 180$ in order to assign the correct limits. Distributions with $r > 180°$ generate both r_1 and $r_2 > 180°$ and considerable further analysis is required to establish correct limits. In these rare cases we consider it sufficient to set a range of $-999{\cdot}0°$ with limiting values of $-999{\cdot}0$, $+999{\cdot}0°$ to draw attention to the situation. True limits, based on the largest gap [see, *e.g.*, equation (8)], may be deduced by inspection of the distribution if required.

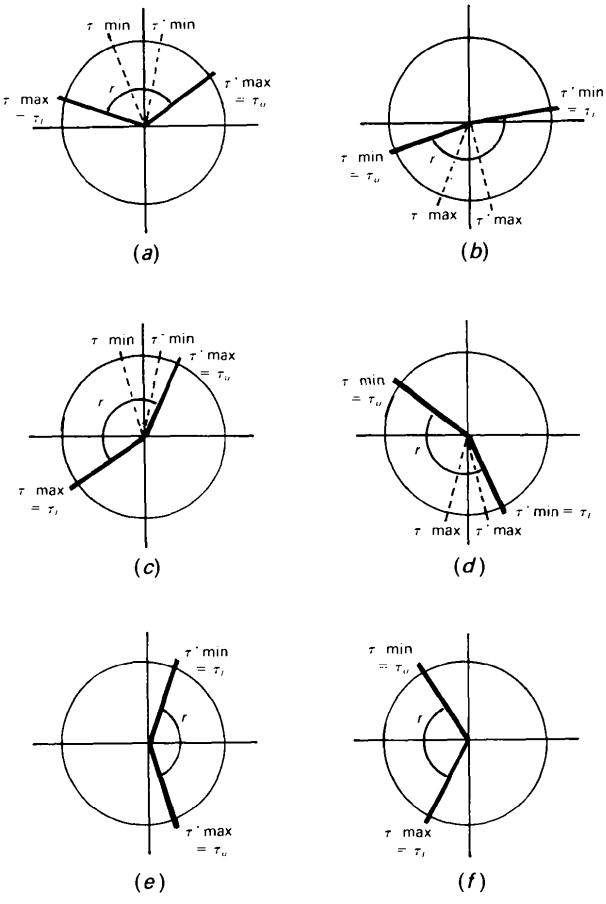

Fig. 4. Determination of upper and lower limits (τ_u, τ_l) and range (r) for six different τ distributions (*a*)–(*f*).

5. Results and discussion

In the current implementation of the cluster analysis routines (ADT3) we calculate both arithmetic and circular summary statistics in a single-pass procedure. The arithmetic results are derived *via* equations (4)–(6) and (1)–(3), *i.e.* we assume a range $r \leq 180°$. The statistics comprise (*a*) the number of observations (N_f), (*b*) the lower and upper limits of the distribution τ_l and τ_u, and its range, r, (*c*) the arithmetic mean $\bar{\tau}_a$, (*d*) the standard error of the mean $\sigma(\bar{\tau}_a)$, (*e*) the sample standard deviation $\sigma(\tau_a)$. Summations for the circular analysis are collected in the same pass. The additional statistics presented are: (*f*) the circular mean $\bar{\tau}_c$, (*g*) the concentration $\bar{R}$, (*h*) the circular standard error, $\hat{\sigma}_c$, (*i*) $\sigma(\bar{\tau}_c)$ from equation (16), (*j*) $\sigma(\tau_c)$ from (17), (*k*) the 99·75% confidence interval of $\bar{\tau}_c$, denoted as $c(\bar{\tau}_c)$, calculated from equation (15) with $u_\alpha = 3{\cdot}0$. If (*i*), (*j*) or (*k*) are undefined, then a value of $-99{\cdot}00$ is reported; if (*i*) lies in the range 15–90° where equation (18) becomes increasingly invalid, then the value is reported with a negative sign.

Representative results for a wide variety of τ distributions (taken from ADT1, ADT3) are presented in Table 1. The arithmetic and circular means are reassuringly similar for all but the broadest distributions. The largest discrepancy occurs for distribution 10 of Table 1, which is taken from the final single-linkage overlay of all 222 six-membered carbocycles of ADT1. This distribution, as well as that of 11 (Table 1) is multimodal in any case. This multimodality is suggested by the very high sample standard deviations for 10 and 11, and, particularly, in their low values of the concentration ($\bar{R}$).

The arithmetic and circular statistical descriptors are also seen to be very similar. The exceptions are the two multimodal examples already noted, and the two smallest distributions (6 and 7; Table 1), where the circular statistics $\sigma(\bar{\tau}_c)$, $\sigma(\tau_c)$ are somewhat low by comparison with their arithmetic equivalents. Examination of Table 1 also shows that $c(\bar{\tau}_c) = 3\sigma(\bar{\tau}_c)$ for most of the examples, but deviations from this equality increase with increasing $\sigma(\bar{\tau}_c)$ as expected from equation (18). The single statistic which appears to convey the maximum information about a distribution is the concentration $\bar{R}$. The physical meaning of $\bar{R}$ is simple to visualize, and it is by far the easiest to calculate, even in manual operations.

Table 1. *Summary statistics for a variety of torsion-angle distributions taken from ADT1 and ADT3*

Distributions 1–4 are from single-linkage clustering of six-membered carbocycles (ADT1). Distributions 5, 8 and 9 are from single-linkage analysis of the steroid C(17) side chain (ADT3). Distributions 6 and 7 are from the Jarvis–Patrick analysis of azacycloheptane (ADT3). Distributions 10 and 11 are taken from final single-linkage overlays of all fragments of six-membered carbocycles (10) and steroid side chains (11).

	N_f	τ_l	τ_u	r	$\bar{\tau}_a$	$\bar{\tau}_c$	$\sigma(\tau_a)$	$\sigma(\tau_c)$	$\sigma(\bar{\tau}_a)$	$\sigma(\bar{\tau}_c)$	$\hat{\sigma}_c$	$c(\bar{\tau}_c)$	$\bar{R}$
1	35	−2·7	0·9	3·6	−0·5	−0·5	0·7	0·7	0·1	0·1	0·002	0·4	1·000
2	51	53·3	65·0	11·7	58·6	58·6	2·7	2·6	0·4	0·4	0·007	1·2	0·999
3	51	−61·6	−36·2	25·4	−50·0	−50·0	5·1	5·0	0·7	0·7	0·012	2·1	0·996
4	11	−56·3	−44·1	12·2	−52·1	−52·1	3·5	3·3	1·1	1·0	0·018	3·0	0·998
5	50	159·8	−159·9	40·3	177·4	177·4	8·2	8·2	1·2	1·2	0·020	3·5	0·990
6	4	−50·2	−7·6	42·6	−25·1	−25·0	18·3	16·1	9·2	8·1	0·140	24·9	0·962
7	6	19·7	66·9	47·7	42·7	42·6	20·1	18·9	8·2	7·7	0·134	23·8	0·950
8	78	3·2	88·4	85·2	56·8	57·3	20·1	19·5	2·3	2·2	0·039	6·6	0·942
9	50	−16·3	88·4	104·7	52·0	53·5	28·9	28·2	4·1	4·0	0·070	12·1	0·883
10	222	−116·2	60·5	176·7	−12·0	−10·2	50·4	60·7	3·4	4·1	0·071	12·3	0·657
11	108	−133·9	119·8	253·7*	10·4	10·1	64·6	83·0	6·2	9·3	0·162	29·1	0·487

* Numerical values obtained by inspection of the distribution. The program will set appropriate default values as decribed in the text.

6. Concluding remarks

This paper has investigated two ways in which a statistical summary may be provided for a distribution of torsion angles. Some, if not all, of the pitfalls associated with the arithmetic approach to circular data have been highlighted and some solutions proposed and tested. The formally correct approach to the problem, *via* the methods of circular statistics, is also described and a number of descriptors of a distribution (assumed to be unimodal) are presented. Comparative results for a number of real distributions, with varying numbers of observations and angular ranges, indicate that the arithmetic and circular means, $\bar{\tau}_a$ and $\bar{\tau}_c$, are not significantly different, at least for the distributions studied. There are, however, some disparities between the arithmetic and circular estimates of (*a*) the standard errors of the means and (*b*) the sample standard deviations. This is especially true for broad distributions and those with small populations.

It is clear that the circular method has a number of computational advantages, especially in reducing the need for irritating phase-shift operations. Hypothesis testing on the value of $\bar{\tau}_c$ is quite straightforward, since confidence limits are readily available (see, *e.g.*, Snedecor & Cochran, 1980). Indeed, the confidence limits based on circular statistics are almost certainly more reliable than their arithmetic counterparts. They are also likely to be more realistic, in view of the limitations inherent in equation (15). One obvious problem is the possibility that the distribution may be multimodal, but this is a problem inherent in all statistical methods. We note also that a wide variety of additional techniques, many of them based on the resultant vector length R [equation (9)] or concentration $\bar{R}$ [equation (14)], are available in circular statistics (Mardia, 1972; Upton & Fingleton, 1989). These include assessments of the shape and modality of a distribution, tests of uniformity, goodness-of-fit, and procedures for the comparison of two or more samples.

We thank Dr Robin Taylor for valuable discussions and Dr Olga Kennard FRS for her interest in this work.

References

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

ALLEN, F. H., DOYLE, M. J. & 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., KENNARD, O., WATSON, D. G., ORPEN, A. G., BRAMMER, L. & TAYLOR, R. (1987). *J. Chem. Soc. Perkin Trans.* 2, pp. 511–519.

BATSCHELET, E. (1981). *Circular Statistics in Biology*. London: Academic Press.

FISHER, N. I. & LEWIS, T. (1983). *Biometrika*, **70**, 333–341.

JEFFREY, G. A. & CRUICKSHANK, D. W. J. (1953). *Q. Rev. Chem. Soc.* **7**, 335–376.

KLYNE, W. & PRELOG, V. (1960). *Endeavour*, **16**, 521–528.

MARDIA, K. V. (1972). *Statistics of Directional Data*. London: Academic Press.

NEWMAN, M. S. (1955). *J. Chem. Educ.* **32**, 344–350.

SNEDECOR, G. W. & COCHRAN, W. G. (1980). *Statistical Methods*, 7th ed. Ames, Iowa: Iowa State Univ. Press.

UPTON, G. J. G. & FINGLETON, B. (1989). *Spatial Data Analysis by Example*, Vol. 2, *Categorical and Directional Data*. Chichester: Wiley.