Acta Cryst. (1996). **B52**, 707–712

Simple Averaging Procedure for Aromatic Ring System: Mean Molecular Fragment of Tryptophan Metabolites

Akio Wakahara[a] and Toshimasa Ishida[b]*

[a]*Fujisawa Pharmaceutical Co., Ltd, 2-1-6 Kashima, Yodogawa-ku, Osaka 532, Japan, and* [b]*Department of Physical Chemistry, Osaka University of Pharmaceutical Sciences, 2-10-65 Kawai, Matsubara, Osaka 580, Japan. E-mail: a61020g@center.osaka-u.ac.jp*

(*Received 29 November 1995; accepted 8 March 1996*)

Abstract

A simple averaging procedure of bond lengths and angles applicable to an aromatic ring system was applied to the crystal structures of the tryptophan metabolites analyzed so far, as an example. Statistical analysis showed no marked effect of the 5-substituent on the molecular dimensions of the 3-substituted indole ring in tryptophan metabolites. The standard bond lengths and angles of the 3-substituted indole ring were calculated according to the statistical averaging procedure outlined here and using data from 35 tryptophan metabolites; the interatomic correlations of the indole ring have been discussed based on these values. The result suggests the validity of the D^2 statistic for averaging the planar ring structure in aromatic systems.

1. Introduction

The exact determination of molecular geometry is of vital importance to our understanding of chemical structure and bonding behavior. It also gives clues to the electronic nature of a molecule and its intrinsic propensity, by which it can play a unique role. The 'standard' bonding parameters, on the other hand, enable evaluation of the effect of a substituent or the accuracy or the significance of differences in experimental data.

Much experimental data have been derived from X-ray and neutron diffraction measurements, and structural parameters of over 140 000 organocarbons have been accumulated in the Cambridge Crystallographic Data Centre (Allen *et al.*, 1991). Methods have been developed to estimate the average molecular dimensions (AMD) for aliphatic organic compounds and the validity of these methods has been tested (Chakrabarti & Dunitz, 1982; Schweizer & Dunitz, 1982; Taylor & Kennard, 1983, 1985, 1986; Allen *et al.*, 1987). There appear to be rather few studies (Taylor & Kennard, 1982) concerning the estimation of AMD for aromatic compounds in which all bonding parameters cannot be treated as independent.

As far as the aromatic ring structures are concerned, there would be considerably strong correlation among bond lengths and angles because of their strict restriction by the ring formation; each of them is closely related with one another in a planar ring. It is supposed that these interrelations are rather inversely proportionate to each other in the closed-ring structure. In statistical analyses for such ring structures, therefore, the parameters (bond lengths/angles) in planar rings should be treated as multivariates. The procedure presented here could detect the ring deformations as multivariate outliers and thus could quantitatively describe the whole deviation of each structure from the mean. This is the main advantage of averaging the ring structures and is superior to other trials so far.

In order to investigate a possible relationship between the metabolic pathway of tryptophan and the conformation of each metabolite, on the other hand, we have been analyzing crystal structures of a series of tryptophan metabolites. A set of standard bond lengths and angles for the tryptophan indole ring is very useful to estimate the accuracy of the analyzed bonding parameters and to consider the accurate electronic/biological propensity of each tryptophan metabolite. However, no systematic and satisfactory AMD estimation method has been reported so far. Under such circumstances we developed an averaging procedure for the indole aromatic ring. In order to estimate the AMD of tryptophan metabolites, the bond lengths and angles of the planar ring structures were summed up *en bloc* and parameters of multivariate outliers were excluded from the estimation.

2. Data set

An atomic numbering of the common moiety of tryptophan metabolites is shown in Fig. 1. A bibliographic list of all structures used in the analysis is given in Table 1, where all trytophan metabolites reported so far, except for the metal complexes and the 4-substituted nonmetabolites, are

Table 1. *Data set of 46 tryptophan metabolites with R^1/R^2 substituents*

No.	R^1	R^2	Form	CSD code	Reference (coden)
1		$-CH_2NH_2$	Free		BBA 543 123
2		$-CH_2NH_3^+$	Salt	TRYPTA10	BCSJA8 46 2481
3		$-CH_2NH_3^+$	Salt	TRYPIC	ACBCAR 30 1841
4		$-CH_2NH_3^+$	Complex	TRYPAC	BCSJA8 51 1123
5		$-CH_2NH_3^+$	Complex	ISXTRA	JCCCAT 358
6		$-CH_2NH_3^+$	Complex	TPATAA	ACBCAR 35 1642
7		$-CH_2NH_3^+$	Complex	TRADAA	BCSJA8 52 2953
8		$-CH_2NH_3^+$	Complex	GANFOQ	ACAPCT 41 539
9		$-CH_2NH_3^+$	Complex	DOTXUF10	JACSAT 110 2286
10		$-CH(COOH)NH_3^+$	Salt	TRYPTC	BCSJA8 39 2369
11		$-CH(COOH)NH_3^+$	Salt	TRYPTD10	CUSCAM 36 139
12		$-CH(COOH)NH_3^+$	Salt	TRYPTF10	ACSAA4 27 471
13		$-CH(COOH)NH_3^+$	Salt	TPTPCM	ACBCAR 30 1841
14		$-CH(COOH)NH_3^+$	Salt	TPYPTB	ACBOCV 34 559
15		$-CH(COOH)NH_3^+$	Free		CPBTAL 41 433
16		$-CH(COO^-)NH_3^+$	Free	QQQBTP01	ACBOCV 34 559
17		$-CH(COO^-)NH_3^+$	Free		CPBTAL 41 433
18		$-CH_2N(CH_3)_2$	Free	DMTRYP	ACBCAR 28 3075
19		$-CH_2N(CH_3)_2$	Free	DMTPYP	ACBCAR 28 3075
20		$-COOH$	Free	INACET	ACCRA9 17 496
21		$-COOH$	Complex	IACNCA	BCSJA8 51 1118
22		$-COOH$	Free	INACET01	ACBCAR 38 2534
23		$-COO^-$	Complex	IAAMTA	ACBCAR 32 3235
24		$-COO^-$	Complex		ABBIA4 200 492
25		$-COO^-$	Complex	BAKMOP	ACBCAR 37 2117
26		$-COO^-$	Complex	MEADIN10	BICHAW 22 3571
27		$-COO^-$	Complex	CEBZUE10	JACSAT 110 2286
28		$-COO^-$	Complex	CEBZUE10	JACSAT 110 2286
29		$-COO^-$	Complex	GATSAV	JACSAT 110 2286
30		$-COO^-$	Complex	GATSAV	JACSAT 110 2286
31	$-OH$	$-CH_2NH_3^+$	Complex	HTRCRS	ACCRA9 19 713
32	$-OH$	$-CH_2NH_3^+$	Salt	SERPIC	ACBCAR 28 82
33	$-OH$	$-CH_2NH_3^+$	Salt	SERHOX	ACAPCT 32 267
34	$-OH$	$-CH(COO^-)NH_3^+$	Free	HTRYPT10	BCSJA8 46 2475
35	$-OH$	$-CH_2N(CH_3)_2$	Free	BUFTEN	ACBCAR 28 3219
36	$-OH$	$-CH_2N(CH_3)_2$	Free	BUFTEN	ACBCAR 28 3219
37	$-OH$	$-COOH$	Free	JEYBUK	ACSCEE 46 2426
38	$-OCH_3$	$-CH_2NH_2$	Free	MXTRYP	ACBCAR 30 95
39	$-OCH_3$	$-CH_2NH_3^+$	Complex	MIAMTA	ACBCAR 32 3235
40	$-OCH_3$	$-CH_2NH_3^+$	Complex	IAAMTA	ACBCAR 32 3235
41	$-OCH_3$	$-CH_2NHCOCH_3$	Free	MELATN	CMLTAG 1139
42	$-OCH_3$	$-CH_2NHCOCH_3$	Free	MELATN01	ACBCAR 30 99
43	$-OCH_3$	$-CH_2NHCOCH_3$	Free	MELATN02	ACBOCV 28 564
44	$-OCH_3$	$-CH_2NH(CH_3)_2^+$	Salt	MOTYPT	ACBCAR 27 411
45	$-OCH_3$	$-COOH$	Free	MXINAC	BCSJA8 48 536
46	$-OCH_3$	$-COO^-$	Complex	MIAMTA	ACBCAR 32 3235

Salt: HCl, HBr, formate, oxalate or picrate.

included.* All tryptophan metabolites used in this paper are 3-substituted (numbers 1–30) and 3,5-disubstituted (numbers 31–46) derivatives. Thus, 46 structures were used for the present statistical treatment.

Table 2. *Distribution of data and statistical values ($n = 46$)*

	q_0	q_1	q_2	q_3	q_4	Mean	σ	QD
N1—C2	1.280	1.369	1.375	1.385	1.463	1.376	0.023	0.008
N1—C8	1.323	1.370	1.373	1.380	1.432	1.376	0.016	0.005
C2—C3	1.309	1.357	1.364	1.371	1.419	1.364	0.016	0.007
C3—C9	1.324	1.428	1.435	1.444	1.513	1.435	0.024	0.008
C4—C9	1.370	1.397	1.403	1.412	1.452	1.405	0.014	0.0075
C4—C5	1.331	1.367	1.376	1.383	1.450	1.377	0.019	0.008
C5—C6	1.332	1.387	1.401	1.408	1.481	1.399	0.023	0.0105
C6—C7	1.298	1.369	1.375	1.383	1.413	1.374	0.019	0.007
C7—C8	1.353	1.386	1.393	1.402	1.474	1.395	0.020	0.008
C8—C9	1.370	1.399	1.406	1.412	1.457	1.405	0.015	0.0065
C2—N1—C8	105.0	108.1	108.6	109.2	110.6	108.6	1.1	0.55
N1—C2—C3	108.7	109.7	110.3	110.7	114.9	110.3	1.1	0.5
C2—C3—C9	102.4	105.6	106.2	106.6	107.4	106.1	0.9	0.5
C3—C9—C8	102.3	106.9	107.3	107.7	110.9	107.4	1.3	0.4
C4—C9—C8	116.4	118.7	119.4	120.0	124.5	119.5	1.5	0.65
C5—C4—C9	113.8	117.3	118.5	119.0	120.2	118.1	1.5	0.85
C4—C5—C6	119.2	120.9	121.4	122.2	125.4	121.7	1.4	0.65
C5—C6—C7	116.4	120.6	121.4	122.3	124.2	121.3	1.4	0.85
C6—C7—C8	112.8	116.9	117.4	118.1	122.7	117.4	1.5	0.6
C7—C8—C9	117.6	121.3	122.0	122.8	127.7	122.0	1.4	0.75
N1—C8—C9	102.9	106.9	107.5	108.0	111.8	107.5	1.5	0.55
N1—C8—C7	125.8	129.9	130.6	131.1	135.5	130.5	1.4	0.6
C3—C9—C4	128.7	132.7	133.3	133.8	135.5	133.1	1.2	0.55

After sorting of the 46 structures for each parameter (bond length and angle), the quartiles, q_0, q_1, q_2, q_3 and q_4, were obtained, which correspond to the minimum, 25, 50 and 75 percentiles, and maximum of each parameter, respectively. A list of these values, together with the mean, standard deviation (σ) and quartile deviation [(QD) = $(q_3 - q_1)/2$ (Read, 1986)], is given in Table 2. The small QD values of the bond lengths (0.005–0.010 Å) and angles (0.4–0.8°) for all the 46 structures indicate that the effect of 5-substitution on the molecular dimensions of the indole ring is not as significant as has been supposed. On the other hand, some bond lengths and angles show threefold larger σ values than the corresponding QD values, indicating that some data unsuitable for the estimation of AMD are included in the present data set. For statistical treatment of reliable data, therefore, it is of utmost importance to develop an appropriate method for the detection and exclusion of such multivariate outliers. This is one of the main objectives of the present study.

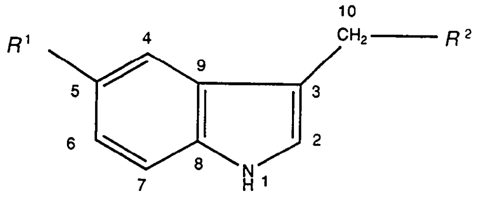

Fig. 1. Chemical structure and atomic numbering of the common moiety of trytophan metabolites.

3. Methods

In order to obtain the mean ring structure of tryptophan metabolites, we used the D^2 statistic as a measure of the distance between the two populations (Mahalanobis, 1936; Pillai, 1985), and this statistic is also applicable as a measure of the distance between each sample structure and its mean in the present study.

When **X** and μ denote the observed and mean vectors of p geometrical variables, respectively, the distance

* A list of bond lengths and bond angles of 46 structures of tryptophan metabolites has been deposited with the IUCr (Reference: OA0001). Copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

Table 3. *Comparison of data from 5-substituted (n = 16) and unsubstituted (n = 30) structures, together with result of t-test*

	5-substituted (n = 16)			Unsubstituted (n = 30)		t-test	
	Range	Mean	σ	Mean	σ	t	P
N1—C2	1.352-1.409	1.376	0.015	1.375	0.027	0.163	0.87
N1—C8	1.362-1.394	1.375	0.008	1.376	0.019	0.247	0.81
C2—C3	1.346-1.389	1.366	0.009	1.363	0.018	0.731	0.47
C3—C9	1.403-1.469	1.438	0.015	1.434	0.028	0.635	0.53
C4—C9	1.370-1.418	1.400	0.013	1.407	0.015	1.593	0.12
C4—C5	1.351-1.416	1.375	0.015	1.377	0.021	0.336	0.74
C5—C6	1.379-1.419	1.405	0.010	1.395	0.027	1.790	0.08
C6—C7	1.352-1.398	1.373	0.013	1.375	0.022	0.394	0.70
C7—C8	1.363-1.431	1.394	0.018	1.396	0.021	0.326	0.75
C8—C9	1.379-1.415	1.403	0.009	1.407	0.018	1.020	0.31
C2—N1—C8	105.6-110.6	108.6	1.1	108.6	1.0	0.0	1.0
N1—C2—C3	108.9-113.4	110.3	1.1	110.4	1.2	0.284	0.78
C2—C3—C9	104.0-107.1	106.1	0.8	106.1	1.0	0.0	1.0
C3—C9—C8	105.3-110.4	107.2	1.1	107.5	1.4	0.746	0.46
C4—C9—C8	116.4-122.9	119.9	1.4	119.2	1.5	1.540	0.13
C5—C4—C9	116.1-119.7	118.0	1.0	118.1	1.7	0.247	0.81
C4—C5—C6	119.6-124.1	121.7	1.2	121.6	1.6	0.221	0.83
C5—C6—C7	118.9-123.1	120.9	1.2	121.5	1.4	1.417	0.16
C6—C7—C8	116.2-118.8	117.8	0.7	117.2	1.7	1.659	0.10
C7—C8—C9	120.6-123.9	121.7	0.9	122.2	1.7	1.353	0.18
N1—C8—C9	104.2-111.0	107.8	1.5	107.4	1.4	0.888	0.38
N1—C8—C7	128.2-131.9	130.5	1.1	130.4	1.6	0.224	0.82
C3—C9—C4	131.8-133.7	132.9	0.5	133.3	1.4	1.379	0.18

(D_i^2) of the ith sample from the mean in p-dimensional space is defined as

$$D_i^2 = (\mathbf{X}_i - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu}), \qquad (1)$$

where $(\mathbf{X}_i - \boldsymbol{\mu})'$ represents the transposed matrix of difference vector $(\mathbf{X}_i - \boldsymbol{\mu})$ and Σ^{-1} is the inverse matrix of the $p \times p$ covariance matrix Σ. Each component of Σ, V_{jk} (the covariance of the jth and kth variables), is defined as

$$V_{jk} = [1/(n-1)]\Sigma(X_j - \mu_j)(X_k - \mu_k), \qquad (2)$$

where n is the number of samples.

The D^2 statistic (D_i^2 in 1) is well known as Mahalanobis' squared distance (Mahalanobis, 1936; Pillai, 1985) and obeys the χ^2 distribution for p degrees of freedom. In this study we attempted to determine multivariate outliers and to quantitatively estimate the deviation from the mean ring structure by applying this D^2 statistic to the indole rings of tryptophan metabolites. A conventional statistical test, the t-test, was also used for the comparison of the bonding parameters between two groups, *i.e.* 5-substituted and unsubstituted structures. For t, χ^2 and r values for the degree of freedom ν at the level of significance α, the expressions of $t(\nu, \alpha)$, $\chi^2(\nu, \alpha)$ and $r(\nu, \alpha)$ were used.

4. Results and discussion

4.1. *Effect of 5-substitution*

In order to assess the effect of 5-substitution on the molecular dimensions of tryptophan metabolites, all bond lengths and angles were classified into two groups (numbers 1-30 and 31-46 in Table 1), and were separately treated for statistical calculations. A comparison of the results from both groups is given in Table 3. From Tables 2 and 3 it is found that all the structural parameters of 5-substituted derivatives are in the range of unsubstituted structures. Moreover, neither $t(44, 0.05)$ value >2.0154 nor p value <0.05 is found in Table 3, indicating that the molecular dimensions of indole rings of both groups are essentially in the same range. This is also clear from histograms of bond angles involving C5 (Fig. 2), where the bond angles of C5-substituted compounds are distributed without specific propensity on the histogram of unsubstituted compounds; bond lengths of the C5-substituted structures also show almost the same distribution as those of the unsubstituted compounds. Thus, it could be concluded that the substitution at C5 does not affect the molecular dimensions of 3-substituted tryptophan metabolites at

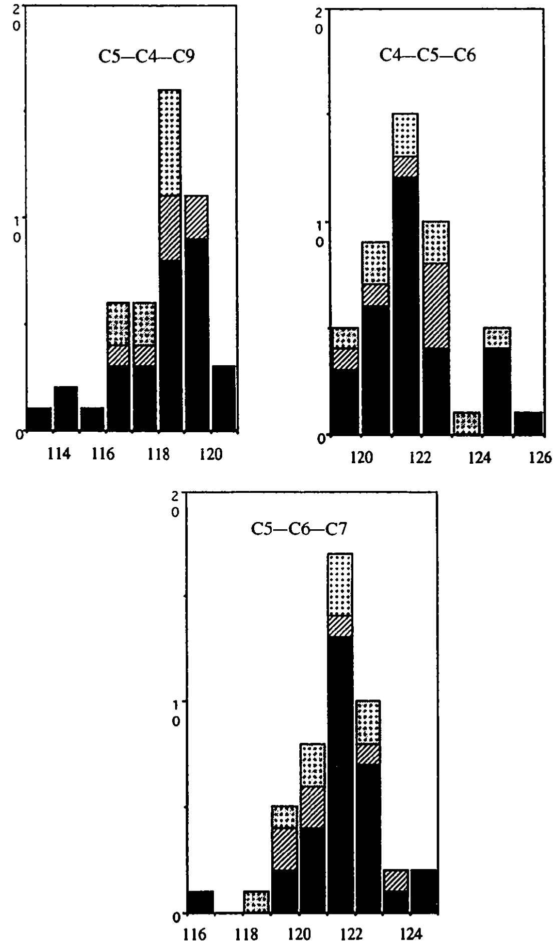

Fig. 2. Distribution of bond angles (C5—C4—C9, C4—C5—C6 and C5—C6—C7). The filled, hatched and dotted bars represent unsubstituted (n = 30), 5-hydroxy- (n = 7) and 5-methoxy- (n = 9) substituted tryptophan metabolites, respectively.

all, so far as the C5-substitution is the hydroxyl or methoxyl group in the tryptophan metabolite. For the subsequent statistical treatment, therefore, data were treated *en bloc*.

4.2. D^2 *statistic and multivariate outliers*

Using equation (1), the D^2 statistic (Mahalanobis' squared distance) was applied to the bond lengths and angles of 46 structures in Table 1. The results are given in Fig. 3(*a*), where ten bond lengths constituting the indole ring were treated as variables. The results for 13 bond angles are given in Fig. 3(*b*). A slight difference in distribution could be observed between the bond lengths and angles, reflecting a certain degree of independence between them. Concerning the bond lengths, there are nine structures with $D^2 \geq 18.307$ [$= \chi^2(10, 0.05)$] and seven structures with $D^2 \geq 23.209$ [$= \chi^2(10, 0.01)$], whereas in the case of bond angles, there are 11 structures with $D^2 \geq 22.362$ [$= \chi^2(13, 0.05)$] and seven structures with $D^2 \geq 27.688$ [$= \chi^2(13, 0.01)$]. In order to estimate the correlation between bond lengths and angles, a scattering diagram is shown in Fig. 3(*c*), where the linear correlation coefficient r for 44 degrees of freedom is 0.874. From these figures, we find that: (i) seven structures, which are judged as outliers at $\alpha = 0.01$, are the same after both calculations of bond lengths and bond angles; (ii) nine structures in D^2(length), which should be excluded at $\alpha = 0.05$, are all included in 11 structures of D^2(angle); (iii) although $r = 0.874$ indicates a relatively high correlation between D^2(length) and D^2(angle), its collinearity is clear in the outlier region (right-upper region of Fig. 3*c*), especially, and the normal region (left-lower region) shows a rather broad distribution (poor correlation) from the collinearity, reflecting also a certain degree of independence between bond lengths and angles. On the basis of the analyses of the D^2 statistic, 11 structures (numbers 2, 5, 11, 15–17, 26, 29–31 and 46 in Table 1), which showed the outlier of D^2(angle) at $\alpha = 0.05$, were excluded from the data set in the subsequent statistical treatment.

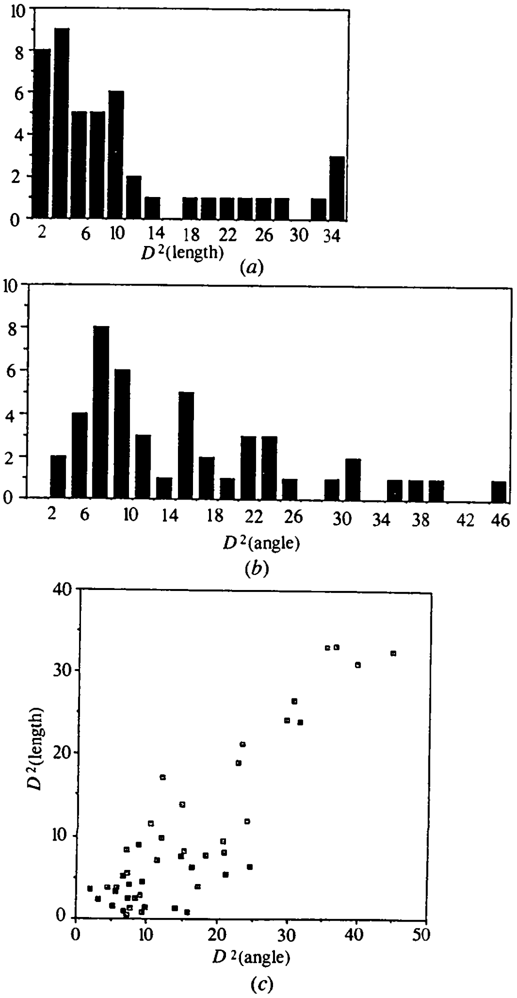

Fig. 3. Histograms of D^2 statistics using (*a*) 10 bond lengths and (*b*) 13 bond angles, where the ordinate and abscissa correspond to the number of observations and D^2(length) or D^2(angle), respectively, and (*c*) a scattering diagram of D^2(length) and D^2(angle). For the calculations, data from all 46 structures were used.

Table 4. *Mean molecular fragment of tryptophan metabolites* ($n = 35$) *and comparison with indole*

	Range	Mean	σ	Indole
N1—C2	1.352-1.409	1.3757	0.0121	1.383
N1—C8	1.362-1.394	1.3738	0.0073	1.364
C2—C3	1.342-1.375	1.3608	0.0082	1.368
C3—C9	1.403-1.471	1.4348	0.0111	1.411
C4—C9	1.370-1.434	1.4027	0.0119	1.383
C4—C5	1.351-1.409	1.3759	0.0128	1.374
C5—C6	1.374-1.419	1.3978	0.0122	1.395
C6—C7	1.352-1.409	1.3749	0.0119	1.372
C7—C8	1.363-1.423	1.3938	0.0124	1.345
C8—C9	1.382-1.420	1.4049	0.0091	1.403
C2—N1—C8	107.3-110.6	108.77	0.74	110.4
N1—C2—C3	108.7-111.5	110.19	0.64	106.4
C2—C3—C9	105.0-107.4	106.27	0.63	109.8
C3—C9—C8	106.1-110.4	107.33	0.78	105.8
C4—C9—C8	116.4-121.1	119.25	0.93	118.1
C5—C4—C9	114.6-120.2	118.44	1.11	120.5
C4—C5—C6	119.6-124.8	121.50	1.16	117.7
C5—C6—C7	118.9-123.1	121.23	1.06	123.9
C6—C7—C8	115.1-118.8	117.51	0.81	116.2
C7—C8—C9	120.6-123.9	122.06	0.82	123.6
N1—C8—C9	104.2-108.5	107.40	0.84	107.7
N1—C8—C7	128.2-132.0	130.53	0.93	128.8
C3—C9—C4	131.1-135.5	133.42	0.80	136.1
C3—C10	1.462-1.530	1.4999	0.0121	
C2—C3—C10	124.9-131.9	127.44	1.61	
C9—C3—C10	121.1-129.5	126.25	1.81	

4.3. *Mean molecular fragment*

Using data from the 35 structures, the bonding parameters for the mean molecular fragment (MMF) were calculated and the results are given in Table 4. The sums of three bond angles around C3, C8 or C9 are all 360.0° and the sums of interior angles of indole five- and six-membered rings are 540.0 and 720.0°, respectively.

In order to estimate the effect of 3-substitution on the bonding parameter of the indole ring, the averaged values were compared with the bond lengths and angles of the indole ring itself. Although there are four reports on the analysis of the indole crystal in the Cambridge Structural Database (CSD), only one analysis [indole: 9-ethyladenine complex (Kaneda & Tanaka, 1976)] is available concerning its bonding parameters (Table 4). Thus, statistical comparison with the AMD of 3-substituted tryptophan metabolites is impossible. However, a general tendency concerning the effect of 3-substitution on the indole ring can be discussed as follows. The bond lengths C3—C9, C4—C9 and C7—C8 appear to increase with the substitution at C3. Furthermore, the increase in the N1—C2—C3 bond angle is concomitant with the decrease in C2—C3—C9 and C3—C9—C4 bond angles. A similar change could also be observed in the C4—C5—C6, C5—C6—C7 and C5—C4—C9 angles.

4.4. *Evaluation of D^2 statistic*

Since values obtained from the D^2 statistic reflect the cumulative deviation of bond lengths or angles ($\mathbf{X}_i - \mu$) of each observation ($\mathbf{X}_i$) from the AMD (μ), these values could be compared with $\delta_{\rm length}$ or $\delta_{\rm angle}$ ($= \Sigma|X_j - \mu_j|$), in order to further evaluate the validity of the D^2 statistic. The scattering diagram of $\delta_{\rm length}$ with respect to D^2(length) for all 46 structures is shown in Fig. 4; a similar tendency is also obtained for $\delta_{\rm angle} - D^2$(angle) plots (data not shown). The correlation coefficients r for $\delta_{\rm length} - D^2$(length) and $\delta_{\rm angle} - D^2$(angle) pairs were 0.925 and 0.840, respectively. These high correlations suggest clearly that the ring deformation and total deviation from the mean ring can be treated by the D^2 statistic, quantitatively. Based on this background, we used effectively the D^2 statistic to detect outliers and exclude them from further averaging procedure.

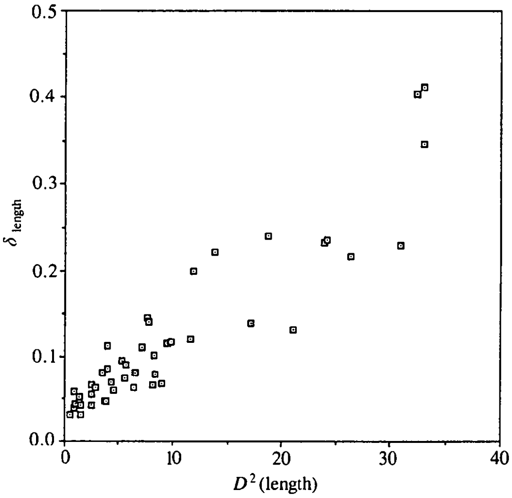

Fig. 4. Scattering diagram of $\delta_{\rm length}$ and D^2(length) using data from 46 structures. Distributions of $\delta_{\rm length}$'s and D^2(length) range from 0.031 to 0.411 and from 0.49 to 33.10, respectively.

On the other hand, Taylor and coworkers (Taylor & Kennard, 1985; Allen *et al.*, 1987) have recommended the prescreening method on the basis of precision (e.s.d. for C—C bonds) of X-ray analysis, in which the data with e.s.d.'s >0.010 Å are eliminated beforehand as unreliable low-precision data. In order to evaluate the relationship between the D^2 outliers and low-precision X-ray analyses, 46 structures were surveyed again and classified into four groups by two features, *i.e.* D^2 statistic (outliers or not) and X-ray precision (less or greater than 0.010 Å). The classification result is given in the following fourfold table as below

31	1	32
4	10	14
35	11	46

$$\chi^2 = 21.360[\gg \chi^2(1, 0.001) = 10.828].$$

The high χ^2 value (= 21.360) obtained from the fourfold table test indicates that a considerably strong interrelation exists between two features. In fact, 10 of 11 D^2 outliers are found later to be from low-precision analyses (e.s.d. > 0.010 Å) by comparison with the original literature or AS-flag (Allen *et al.*, 1987; Allen *et al.*, 1991) in the CSD database. On the contrary, sample number 26 (MEADIN10 in CSD code, AS-flag = 2) is only one D^2 outlier from high-precision X-ray analyses. As shown in the above fourfold table, there are 14 structures from low-precision analyses, in which 10 structures are discriminated as outliers only by the calculation of D^2 values. In other words, the D^2 statistic could detect 10 out of 14 inaccurate structures, based on the total deviation from the mean ring.

4.5. *Interatomic correlation of indole ring*

In order to investigate the possibility of 'concerted' atomic movement in the indole ring, correlations between bond lengths and between bond angles were calculated using data from 35 structures and the results are given in Table 5. There are four bond-length pairs and 20 bond-angle pairs exhibiting correlation coefficients larger than $r(33, 0.001) = 0.532$. Concerning the bond length, the C4—C5 bond showed $r = 0.718$ with respect to the C3—C9 bond. Since the sign of this correlation coefficient is positive, it may indicate easy stretching of the C4—C9 bond in the C3—C9—C4—C5 bond sequence. On the other hand, the highest correlation coefficient ($r = -0.839$)

Table 5. *Correlation coefficient matrix between bond angles (upper triangle) and lengths (lower triangle), calculated using a data set of 35 structures*

	1-2-3	2-3-9	3-9-8	4-9-8	5-4-9	4-5-6	5-6-7	6-7-8	7-8-9	1-8-9	1-8-7	3-9-4	
	−0.658	0.060	0.436	−0.766	0.715	−0.295	−0.151	0.418	0.109	−0.839	0.655	0.450	C2—N1—C8
		−0.607	−0.053	0.365	−0.421	0.195	0.060	−0.210	0.002	0.322	−0.285	−0.379	N1—C2—C3
			−0.680	0.200	0.082	−0.219	0.303	−0.064	−0.356	0.300	0.030	0.449	C2—C3—C9
N1—C8	0.227			−0.583	0.161	0.242	−0.502	0.180	0.575	−0.771	0.202	−0.321	C3—C9—C8
C2—C3	0.113	−0.093			−0.669	0.112	0.189	−0.239	−0.396	0.798	−0.371	−0.581	C4—C9—C8
C3—C9	0.493	0.448	−0.297			−0.706	0.256	0.361	−0.254	−0.513	0.695	0.617	C5—C4—C9
C4—C9	0.181	0.353	−0.241	0.154			−0.763	−0.021	0.398	0.036	−0.398	−0.372	C4—C5—C6
C4—C5	0.323	0.369	−0.322	0.718	0.264			−0.531	−0.240	0.335	−0.080	0.278	C5—C6—C7
C5—C6	0.183	0.213	0.345	0.072	0.019	−0.058			−0.479	−0.332	0.718	0.097	C6—C7—C8
C6—C7	0.514	0.367	−0.125	0.235	0.584	0.216	−0.017			−0.374	−0.543	−0.108	C7—C8—C9
C7—C8	0.259	0.284	−0.283	0.579	0.283	0.572	−0.151	0.184			−0.574	−0.148	N1—C8—C9
C8—C9	0.284	−0.260	0.260	−0.020	−0.333	−0.131	0.112	−0.047	−0.109			0.216	N1—C8—C7
	1-2	1-8	2-3	3-9	4-9	4-5	5-6	6-7	7-8				

was observed between C2—N1—C8 and N1—C8—C9 bond angles. Since the sign of this correlation coefficient is negative, either N1 or C8 in the bond sequence of C2—N1—C8—C9 could be thought to move. Judging from the high correlation coefficient ($r = -0.771$) observed between C3—C9—C8 and C9—C8—N1 bond angles, C8 could be concluded to be the most mobile atom in the indole ring. Similarly, both negative correlation coefficients of $r = -0.706$ and -0.763 between the neighboring angles in the C9—C4—C5—C6—C7 bond sequence indicate the easy movement of C5. A high positive correlation coefficient ($r = 0.798$) between C4—C9—C8 and C9—C8—N1 angles, in contrast, can be interpreted as the easy stretching of the central C8—C9 bond in the N1—C8—C9—C4 bond sequence.

5. Conclusion

The present study demonstrates the applicability of multivariate analysis to averaging of the planar ring structure in aromatic systems, using molecular dimensional data of tryptophan metabolites, and is summarized as follows.

(1) The D^2 statistic procedure, well tested by trial and error, is suitable for averaging the molecular fragments, when the number of observations available is not sufficient for the deviation of meaningful conclusions.

(2) The mean advantage of this approach is its wide range of applications to molecular fragments, particularly to aromatic rings.

(3) The D^2 statistic is valuable for detecting the outliers and expressing the deformation of each molecular structure observed from the mean ring, quantitatively.

References

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187-204.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *J. Chem. Soc. Perkin Trans. 2*, S1-19.

Chakrabarti, P. & Dunitz, J. D. (1982). *Helv. Chim. Acta*, **65**, 1555-1562.

Kaneda, T. & Tanaka, J. (1976). *Bull. Chem. Soc. Jpn*, **49**, 1799-1804.

Mahalanobis, P. C. (1936). *Proc. Natl. Inst. Sci. India*, **2**, 49-55.

Pillai, K. C. S. (1985). *Encyclopedia of Statistical Sciences*, edited by S. Kotz & N. L. Johnson, Vol. 5, pp. 176-181. New York: Wiley.

Read, C. B. (1986). *Encyclopedia of Statistical Sciences*, edited by S. Kotz & N. L. Johnson, Vol. 7, p. 456. New York: Wiley.

Schweizer, W. B. & Dunitz, J. D. (1982). *Helv. Chim. Acta*, **65**, 1547-1554.

Taylor, R. & Kennard, O. (1982). *J. Am. Chem. Soc.* **104**, 3209-3212.

Taylor, R. & Kennard, O. (1983). *Acta Cryst.* B**39**, 517-525.

Taylor, R. & Kennard, O. (1985). *Acta Cryst.* A**41**, 85-89.

Taylor, R. & Kennard, O. (1986). *J. Chem. Inf. Comput. Sci.* **26**, 28-32.