Acta Cryst. (1980). A**36**, 975–978

Pattern Recognition in Molecules from a General Linear Transformation

By D. J. Watkin*

Laboratoire de Cristallographie et de Physique Cristalline associé au CNRS, 351 *cours de la Libération,*
33405 *Talence, France*

(*Received* 16 *April* 1980; *accepted* 27 *May* 1980)

Abstract

The comparison of molecular structures is facilitated by a combination of graphical and numerical techniques. Conversion to a molecular reference frame makes the linear relationships between the molecules or fragments easily visualized, and can be used to produce diagrams clearly displaying the molecular similarities and differences.

Introduction

There are several ways of comparing two molecules for which atomic coordinates are known. Graphical representations are very efficient since a lot of information can be assimilated at a glance. However, information must be suppressed in producing a drawing, and the results are thus open to misinterpretation. Purely numerical representations, such as lists of torsion angles, are less likely to be misleading, but at the same time may fail to make some of the relationships between the data clear. Statistical analyses will reveal the presence of differences between structures, but without pin-pointing them (De Camp, 1973; Albertsson & Svensson, 1978). For these reasons we have broken down the extraction of the linear function relating the two coordinate sets into easily visualized steps.

The set of coordinates B can be fitted to the set A by letting its frame of reference translate, rotate,† dilate and curve. A full analysis of the problem when all these changes are allowed has been given (Diamond, 1976), together with suggestions on how to visualize the properties of the 30 independent elements needed to quantify these effects. Discussions in terms of translation, rotation and dilation (12 parameters) have been given (Fletterick & Wyckoff, 1975; Mackay, 1977), and in terms of translation and pure rotation (6 parameters) only (McLachlan, 1972; Nyburg, 1974; Ferro & Hermans, 1977; Kabsch, 1978; Yuen & Nyburg, 1979). The purpose of this paper is to show that combined with other procedures these analyses provide powerful methods for examining conformational changes.

Method

The matrix of coordinates $\mathsf{B}(3,n)$ is to be fitted to the coordinates $\mathsf{A}(3,n)$, in which each column of B is identified with the corresponding column of A. The fitting is achieved by an origin shift **c** and a linear transformation. It can be shown (Hamilton, 1964) that the best fit will occur when the centroids coincide, though there may be good reasons for constraining the centre of rotation to be at some other point (Nyburg, 1974). The problem of fitting B to A then becomes one of solving

$$\mathsf{A} = \mathsf{D}_1 . \mathsf{B} \qquad (1)$$

where D_1 is a (3×3) rotation–dilation matrix. If A and B are in the same coordinate system, as for subunits of a molecule or independent molecules in the same cell, then **c** and D_1 may be of interest since they will describe the internal symmetry (Hendrickson, 1979). In general, skew and dilation effects due to non-orthogonal coordinate systems are first removed, giving orthonormal coordinates. A and B are usually arbitrarily orientated with respect to the orthogonal axes and become much easier to visualize when put into a coordinate system defined by their principal axes of inertia,

$$\mathsf{S}_a \mathsf{L}_a \mathsf{A} = \mathsf{D}_2 \mathsf{S}_b \mathsf{L}_b \mathsf{B}. \qquad (2)$$

The subscripts *a* and *b* indicate that the matrices have been computed from data proper to A and B respectively. L is the usual orthogonalization matrix, and S a pure rotation matrix. Projections of SLA and SLB are commonly used to display the molecules, since one axis is perpendicular to the best plane, and another is parallel to the best line. The rotation–dilation matrix D_2 will now be defined with respect to the shape of the molecules.

* On leave from the Chemical Crystallography Laboratory, 9 Parks Road, Oxford, England.

† Rotate will be used to mean either pure rotation or rotary inversion unless explicitly stated. Dilations will be positive unless stated otherwise.

0567-7394/80/060975-04$01.00

The solution of (2), or $A_2 = D_2 B_2$, where A_2 and B_2 are now in the corresponding inertial systems, is:

$$D_2 = A_2 B_2^T [B_2 B_2^T]^{-1}.$$

The matrix D_2 can be resolved into a rotation matrix R, and a symmetric dilation tensor $\mathbf{T}$:

$$S_a L_a A = R\mathbf{T}S_b L_b B. \qquad (3)$$

The nine elements of R can be expressed as three independent variables in either Eulerian or polar spherical coordinates (Rossman & Blow, 1962); and the strain tensor, **T**, analysed into its eigenvalues and -vectors, which give the magnitudes and directions of the dilations.

Since (2) is solved without the constraint that D be orthogonal, the value found for R will not be the same as that found by processes that fit a rotation matrix only. Although RT gives a best fit between A_2 and B_2, it will not preserve the bond lengths and bond angles. Undistorted figures are obtained by plotting RB_2.

Results

A brief example of the use of these calculations is their application to a steroid which crystallizes in $P2_1$ with $Z = 4$ (Busetta, Hospital & Precigoux, 1979). The bond lengths and angles for the two independent molecules are in substantial agreement, but the torsion angles indicate that there are widespread conformational changes. In order to determine the overall effect of these differences, we first fit the whole of molecule (II) to molecule (I). Column (i) in Table 1 summarizes the results. The eigenvalues of the inertial tensor show that the length and breadth of the molecules are substantially the same, but that molecule (II) is substantially thicker. The rotation matrix relating (II) to (I) shows that the two molecules are related by an approximate twofold axis. The eigenvalues and -vectors of the strain matrix show that (II) has to be compressed approximately parallel to z to make it match (I). The r.m.s. linear displacements between atoms in the two molecules show that the effect of the dilation is quite small, as would be expected since it is perpendicular to the plane of the molecule. Fig. 1 is a projection of the rotationally adjusted molecules. Two things are clear: that it is not possible to fit the *A* rings simultaneously with the rest of the molecule, and that the centre of gravity is probably not the best point to make the centre of rotation.

Column (ii) of Table 1 shows equivalent figures when the atoms of the *A* ring and the ketonic O atom have been omitted from the fitting. The residues being fitted together are now very similar, as can be seen from the inertial eigenvalues and the dilation eigenvalues. Fig. 2 suggests that there is a progressive bending and twisting of (II), and that a full Diamond-type analysis must be made if this twist is to be analysed. However, for descriptive purposes it is sufficient to note that the

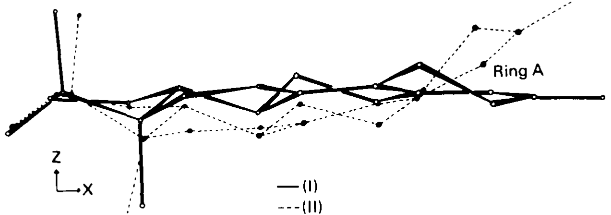

Fig. 1. Projection of molecule (I) and molecule (II) after rotation to obtain the best fit between all equivalent atoms, on to the plane in (I) perpendicular to the eigenvector corresponding to the medium inertial eigenvalue.

Table 1. *Molecular and intermolecular parameters characterizing the relationship between two steroid molecules*

(i) Whole molecule fitted; (ii) rings *B*, *C*, *D* and substituents fitted; (iii) rings *A*, *B* and substituent fitted; (iv) rings *A*, *B* and substituent fitted and proper rotation extracted.

	(i)			(ii)			(iii)			(iv)		
Inertial eigenvalues (I)	206·3	29·5	7·7	57·5	18·0	7·1	42·6	12·4	0·5	42·6	12·4	0·5
Inertial eigenvalues (II)	195·5	30·1	11·7	57·0	18·7	7·3	41·7	12·4	0·9	41·7	12·4	0·9
Rotation matrix (II) into (I)	−0·992 −0·016 0·124	0·000 −0·992 0·129	0·125 −0·128 0·984	−0·999 −0·005 0·035	0·005 −1·000 0·001	0·035 0·001 0·999	−0·999 0·003 −0·052	−0·007 −0·997 0·083	0·052 −0·083 −0·995	−0·997 0·000 0·076	−0·009 −0·993 −0·115	0·075 −0·116 0·990
Eigenvalues and eigenvectors of strain matrix	1·052 0·922 −0·277 −0·271	0·990 0·312 0·945 0·095	0·532 0·230 −0·172 0·958	1·021 0·834 −0·070 −0·548	0·978 0·039 0·997 −0·068	0·961 0·551 0·035 0·834	1·013 0·997 −0·026 −0·067	1·003 0·033 0·995 0·097	0·149 0·064 −0·099 0·993	1·013 0·997 −0·026 −0·067	1·003 0·033 0·995 0·097	−0·149 0·064 −0·099 0·993
Number of atoms fitted		21			14			11			11	
R.m.s. deviation (Å) with dilation		0·465			0·117			0·223			0·223	
Rotation only		0·684			0·134			0·355			0·435	

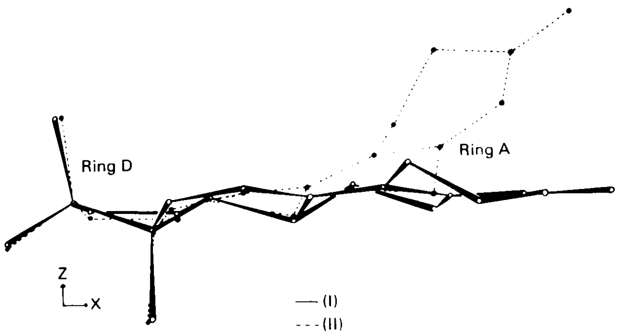

Fig. 2. Projection as in Fig. 1, but omitting ring *A* and the ketonic O atom from the fitting.

configuration at the *D* ring is very stable. Table 2 lists the inertial coordinates for molecule (I), and the coordinates for molecule (II) after each of these fitting procedures. The angular deviations are measured from the centre of rotation.

The final two columns of Table 1 show the result of trying to fit two regions that are quite incompatible. For column (iii) a rotary inversion was permitted, and for column (iv) a negative dilation was selected in order to preserve a pure rotation. The r.m.s. deviation after the application of the rotation matrix has risen, but this has to be accepted if the chirality of C(17), which atom was not included in the fitting, is to be preserved. The

Table 2. *Atomic parameters and deviations for molecules* (I) *and* (II), *in the molecular coordinate systems*

All atoms included in fitting

	$x_{\rm I}$	$y_{\rm I}$	$z_{\rm I}$	$x_{\rm II}$	$y_{\rm II}$	$z_{\rm II}$	Δ (Å)	Δ (°)
C(1)	2·65	−1·56	−0·53	2·69	−1·61	−0·06	0·47	8·69
C(2)	3·97	−1·59	0·21	3·66	−1·46	−1·21	1·47	20·08
C(3)	4·73	−0·30	0·09	4·48	−0·21	−1·11	1·23	15·09
O(83)	5·96	−0·26	0·13	5·62	−0·14	−1·57	1·73	16·90
C(4)	3·93	0·90	0·01	3·82	0·94	−0·50	0·53	7·57
C(5)	2·59	0·89	−0·03	2·59	0·89	0·03	0·07	1·39
C(10)	1·82	−0·34	−0·19	1·85	−0·37	0·15	0·33	10·24
C(6)	1·84	2·20	0·10	1·92	2·14	0·53	0·44	8·78
C(7)	0·41	2·11	−0·41	0·45	2·12	0·10	0·52	13·57
C(8)	−0·30	0·92	0·24	−0·26	0·91	0·68	0·43	23·03
C(9)	0·47	−0·36	−0·07	0·53	−0·38	0·45	0·52	43·19
C(11)	−0·30	−1·59	−0·22	−0·24	−1·63	0·52	0·74	25·64
C(12)	−1·63	−1·67	−0·04	−1·56	−1·70	0·57	0·62	15·14
C(13)	−2·43	−0·47	0·37	−2·40	−0·47	0·66	0·29	6·66
C(14)	−1·74	0·78	−0·22	−1·63	0·71	0·06	0·31	9·49
C(15)	−2·71	1·69	−0·00	−2·67	1·84	0·07	0·17	3·10
C(16)	−4·11	1·22	−0·09	−4·02	1·11	−0·17	0·16	2·11
C(17)	−3·86	−0·32	−0·21	−3·72	−0·41	−0·18	0·18	2·64
C(18)	−2·48	−0·45	1·91	−2·77	−0·27	2·15	0·42	7·23
C(27)	−3·97	−0·81	−1·64	−3·54	−0·92	−1·60	0·44	6·01
O(97)	−4·86	−0·98	0·57	−4·80	−1·12	0·42	0·21	2·46
						R.m.s. deviation 0·68		
Not all atoms included in fitting								
*C(1)	3·66	−3·08	−0·72	3·50	−2·98	−1·38	0·69	8·16
*C(2)	4·90	−3·59	−0·00	4·31	−2·94	−2·65	2·79	27·09
*C(3)	6·04	−2·61	−0·08	5·51	−2·06	−2·54	2·58	22·89
*O(83)	7·21	−2·99	−0·06	6·54	−2·28	−3·18	3·27	24·47
*C(4)	5·68	−1·21	−0·08	5·38	−0·89	−1·68	1·66	16·58
*C(5)	4·41	−0·78	−0·10	4·28	−0·65	−0·95	0·86	11·12
*C(10)	3·28	−1·67	−0·30	3·17	−1·60	−0·87	0·58	9·07
C(6)	4·14	0·70	0·11	4·15	0·64	−0·17	0·29	3·96
C(7)	2·76	1·12	−0·38	2·72	1·17	−0·33	0·08	1·62
C(8)	1·70	0·20	0·23	1·71	0·17	0·19	0·05	1·66
C(9)	2·00	−1·25	−0·16	1·98	−1·25	−0·34	0·18	4·46
C(11)	0·86	−2·14	−0·35	0·84	−2·17	−0·29	0·07	1·75
C(12)	−0·42	−1·79	−0·16	−0·41	−1·82	−0·02	0·15	4·52
C(13)	−0·77	−0·42	0·33	−0·76	−0·42	0·38	0·06	3·41
C(14)	0·30	0·56	−0·21	0·29	0·53	−0·18	0·04	3·35
C(15)	−0·32	1·73	0·07	−0·29	1·92	0·16	0·21	6·46
C(16)	−1·80	1·75	−0·02	−1·82	1·71	0·07	0·10	2·29
C(17)	−2·07	0·23	−0·22	−2·07	0·20	−0·20	0·04	1·10
C(18)	−0·81	−0·46	1·87	−0·85	−0·40	1·93	0·10	2·62
C(27)	−2·34	−0·12	−1·66	−2·26	−0·06	−1·68	0·10	2·07
O(97)	−3·24	−0·10	0·55	−3·24	−0·22	0·49	0·13	2·29
						R.m.s. deviation 1·13		

* Atoms not included in the fitting.

production of a negative dilation may indicate that the sense of one of the molecules has been incorrectly assigned, that there is curvature in one of the molecules, or an attempt has been made to fit non-equivalent atoms. The rejection from subsequent fittings of atoms deviating by more than 3 times the r.m.s. deviation provides an elementary form of pattern recognition (Rao & Rossmann, 1973).

Conclusions

Usually, physical conditions are quantified only after they have been qualitatively identified, and this identification itself requires the recognition of a pattern in the condition. Diagrams and graphs are traditional aids to this process, and the above combination of several well known calculations is intended to provide such aids.

In the examination and comparison of discrete molecules, or parts of discrete molecules, the actual values of the rotations and dilations producing the best fit are of little direct value, but serve to give a view of the molecules that enhances their similarities and their differences. More complete analyses exist for those conditions where the differences must be quantified. Since general linear transformations produce changes in molecular parameters, pure rotations must be used for model building. However, the additional information obtained from the linear fitting makes this procedure more suitable for conformational analysis.

I thank Ian Mackay for valuable correspondence and the Royal Society for a European Science Exchange Fellowship.

APPENDIX

By treating the coordinates for each molecule as a matrix, the whole calculation can be conveniently coded on to a computer providing the usual matrix operations. The eigenvalue–eigenvector routine must handle equal or zero roots and return vectors defining a right-handed coordinate system.

If W is a diagonal matrix with each element the uncorrelated isotropic weight to be given to the corresponding point, then the inertial matrix S is the transpose of the matrix of eigenvectors of $[\mathsf{A}_2\mathsf{W}\mathsf{A}_2^{\mathbf{T}}]$. The eigenvalues are the sums of the squares of the deviations parallel to the corresponding axis (Rollett, 1965). Equation (2) can readily be solved, with or without weights, or Diamond's (1976) equation (68):

$$\mathsf{A} = (d|\mathsf{D}|\mathsf{E})\left(\begin{array}{c} 1 \\ \hline \mathsf{B} \\ \hline \mathsf{b}_i\,\mathsf{b}_j \end{array}\right)$$

if details of twist and curvature are sought. E defines the curvature, D the rotation—dilation and d is an origin which must now be determined simultaneously with E. B is the matrix of coordinates and $\mathsf{b}_i\mathsf{b}_j$ is the matrix of second-order terms. If V is the matrix of eigenvectors of $\mathsf{D}^{\mathbf{T}}\mathsf{D}$ and U is a matrix with the square roots of the corresponding eigenvalues as the diagonal elements, then $\mathbf{T} = \mathsf{VUV}^{\mathbf{T}}$, and $\mathsf{R} = \mathsf{DT}^{-1}$ (Diamond, 1976). A negative determinant for R corresponds to a rotary inversion if both S_a and S_b are proper. The best proper rotation for molecules that are really related by an improper rotation presents interpretive problems. One solution is to make the dilation most nearly parallel to the z axis negative. For an approximately planar molecule this corresponds to inversion across the plane, and hence the smallest change in the coordinates. Note that if one or other of the molecules is exactly planar or linear one or more of the rotations and dilations will not be defined.

References

Albertsson, J. & Svensson, S. (1978). *Acta Cryst.* A**34**, S17.

Busetta, B., Hospital, M. & Precigoux, G. (1979). Unpublished.

De Camp, W. H. (1973). *Acta Cryst.* A**29**, 148–150.

Diamond, R. (1976). *Acta Cryst.* A**32**, 1–10.

Ferro, D. R. & Hermans, J. (1977). *Acta Cryst.* A**33**, 345–347.

Fletterick, R. J. & Wyckoff, H. W. (1975). *Acta Cryst.* A**31**, 698–700.

Hamilton, W. C. (1964). *Statistics in Physical Science.* New York: Ronald Press.

Hendrickson, W. A. (1979). *Acta Cryst.* A**35**, 158–163.

Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.

Mackay, A. L. (1977). *Acta Cryst.* A**33**, 212–215.

McLachlan, A. D. (1972). *Acta Cryst.* A**28**, 656–657.

Nyburg, S. C. (1974). *Acta Cryst.* B**30**, 251–253.

Rao, S. T. & Rossman, M. G. (1973). *J. Mol. Biol.* **76**, 241–256.

Rollett, J. S. (1965). *Computing Methods in Crystallography.* Oxford: Pergamon Press.

Rossman, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Yuen, P. S. & Nyburg, S. C. (1979). *J. Appl. Cryst.* **12**, 258.