RESEARCH PAPERS

Acta Cryst. (1995). D**51**, 127–135

Coordinate-Based Cluster Analysis

BY R. DIAMOND

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, England

(*Received* 13 *June* 1994; *accepted* 19 *September* 1994)

Abstract

A new approach to cluster analysis of structures based on collective superpositions rather than pairwise superpositions is presented. The method is fast and rigorous and is illustrated by application to 21 structures derived from NMR experiments. Source code, suitable for most laboratory machines, is available from the author, and a *CCP*4 version is in preparation.

Introduction

Superposition of coordinates of proteins is a commonly used aid to the comparison of related structures. Pairwise superpositions have been commonplace for many years, but simultaneous superposition of many structures is a problem which has only relatively recently been solved: Kearsley (1990); Shapiro, Botha, Pastore & Lesk (1992); Diamond (1992). In this paper these techniques are extended to show how structures may be grouped into clusters on the basis of the similarity of their coordinates, so as to identify families of similar structures which may exist within an ensemble of structures, in a manner which is both fast and rigorous.

Previous solutions to this problem have been of two main types: those that generate a tree of structures with similar clusters united at each node, and those that represent the structures as a distribution of points in two or three dimensions, such that the distances between such points are a direct measure of the r.m.s. coordinate differences between the corresponding structures. Both types of method have been limited in the past by being based on pairwise superpositions of structures to measure the r.m.s. distances between structures, rather than collective superpositions.

In the first type (*e.g.* Russell & Barton, 1992), clusters are represented by the average coordinates of structures within a cluster, and successive unions are performed by superimposing such averages. This is not ideal because the averaging step may degrade stereochemical features, and because the average coordinates have to be calculated at each step.

In the second type of method (*e.g.* Sutcliffe, 1993) the use of pairwise superpositions is also not ideal because it implies that each structure is allowed a multiplicity of coordinate sets, one for each other structure with which it is paired, which can lead to negative eigenvalues and imaginary coordinates in the resulting representation. Although this is not likely to arise when the structures being compared are closely similar, as many as eight negative eigenvalues out of 29 have been encountered for this reason using this technique.

It is shown here how clustering may be performed rigorously without any reference to coordinates, (except during the initialization stage) and without introducing any distortions to the structures, by processes which involve only transformations and additions of 4×4 matrices. The process leads to a dendrogram, or tree, and also, optionally, to one or more constellations of structures which are certainly free of negative eigenvalues.

Theory

The theory underlying this method is an extension of the theory presented by Diamond (1992) for multiple simultaneous superpositions and it is necessary, therefore, to quote extensively from that paper. In this paper, where two numbers are attached to an equation, the first is the number of the corresponding equation in that paper, to which reference should be made for the relevant derivation.

Diamond (1988) showed that the weighted sum of squares of coordinate differences, E, between a vector set $\mathbf{X}$ and a rotated vector set $\mathbf{Rx}$, is given by

$$E = E_0 - 2\boldsymbol{\rho}^T \mathbf{P} \boldsymbol{\rho}, \qquad (18,1)$$

in which E_0 is the value associated with $\mathbf{X}$ and the unrotated $\mathbf{x}$, *i.e.* for $\mathbf{R} = \mathbf{I}$, $\mathbf{P}$ is a real symmetric 4×4 matrix bilinear on $\mathbf{X}$ and on $\mathbf{x}$, and $\boldsymbol{\rho}$ is the rotation vector specifying the rotation effected by the orthogonal 3×3 matrix $\mathbf{R}$. $\boldsymbol{\rho}$ is defined by

$$\boldsymbol{\rho} = \begin{pmatrix} \lambda \\ \mu \\ \nu \\ \sigma \end{pmatrix} = \begin{pmatrix} l \sin(\theta/2) \\ m \sin(\theta/2) \\ n \sin(\theta/2) \\ \cos(\theta/2) \end{pmatrix}, \qquad (1,2)$$

for axial direction cosines l, m, n and rotation angle θ.

Evidently E is minimized if $\boldsymbol{\rho}$ is the top eigenvector of $\mathbf{P}$. Subsequently, Kearsley (1989) developed an equivalent solution using a matrix $\mathbf{K}$ given by

$$\mathbf{K} = E_0\mathbf{I} - 2\mathbf{P}, \tag{3}$$

(Diamond, 1989), in terms of which

$$E = \boldsymbol{\rho}^T\mathbf{K}\boldsymbol{\rho}, \tag{4}$$

so that the optimal $\boldsymbol{\rho}$ is the bottom eigenvector of $\mathbf{K}$. $\mathbf{P}$ and $\mathbf{K}$ are equally effective agents for the determination of optimal superpositions, but in the current work there are advantages to be gained from the use of Kearsley's form, which provides for the monitoring of E values through many transformations and unions, as well as for the determination of $\boldsymbol{\rho}$ vectors.

It was shown previously that if structure B is to be rotated by $\boldsymbol{\rho}_B$ from its original orientation to superimpose on a structure A which has already been rotated by $\boldsymbol{\rho}_A$, then $\boldsymbol{\rho}_B$ should be the top eigenvector of

$$[\boldsymbol{\rho}_A]\mathbf{P}_{BA}[\boldsymbol{\rho}_A]^T, \tag{23,5}$$

or the bottom eigenvector of

$$[\boldsymbol{\rho}_A]\mathbf{K}_{BA}[\boldsymbol{\rho}_A]^T, \tag{6}$$

in which

$$[\boldsymbol{\rho}] = \begin{pmatrix} \sigma & -\nu & \mu & \lambda \\ \nu & \sigma & -\lambda & \mu \\ -\mu & \lambda & \sigma & \nu \\ -\lambda & -\mu & -\nu & \sigma \end{pmatrix}. \tag{8,7}$$

Similarly, if structure A has already been rotated by $\boldsymbol{\rho}_A$ and structure B has already been rotated by $\boldsymbol{\rho}_B$, and if a further rotation, $\boldsymbol{\rho}$, is to be applied to structure B, so that B's rotation vector relative to its original orientation is then

$$[\boldsymbol{\rho}]\boldsymbol{\rho}_B \tag{12,8}$$

then this further rotation, $\boldsymbol{\rho}$, optimizes the fit of B on A if

$$\boldsymbol{\rho}_B^T[\boldsymbol{\rho}]^T[\boldsymbol{\rho}_A]\mathbf{K}_{BA}[\boldsymbol{\rho}_A]^T[\boldsymbol{\rho}]\boldsymbol{\rho}_B \tag{9}$$

is minimized.

Now, the compound rotation $[\boldsymbol{\rho}]\boldsymbol{\rho}_B$ may be reversed by applying $\bar{\boldsymbol{\rho}}$ followed by $\bar{\boldsymbol{\rho}}_B$, *i.e.*,

$$\bar{\mathbf{I}}[\boldsymbol{\rho}]\boldsymbol{\rho}_B = [\bar{\boldsymbol{\rho}}_B]\bar{\boldsymbol{\rho}} = [\boldsymbol{\rho}_B]^T\bar{\mathbf{I}}\boldsymbol{\rho}, \tag{6,10}$$

so that

$$[\boldsymbol{\rho}]\boldsymbol{\rho}_B = \bar{\mathbf{I}}[\boldsymbol{\rho}_B]^T\bar{\mathbf{I}}\boldsymbol{\rho}, \tag{11}$$

in which

$$\bar{\mathbf{I}} = \begin{pmatrix} -\mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix}, \tag{5,12}$$

where $\mathbf{I}$ is the 3×3 identity, and it is convenient to define a further type of matrix given by

$$\langle\boldsymbol{\rho}\rangle = \bar{\mathbf{I}}[\boldsymbol{\rho}]\bar{\mathbf{I}} = \begin{pmatrix} \sigma & -\nu & \mu & -\lambda \\ \nu & \sigma & -\lambda & -\mu \\ -\mu & \lambda & \sigma & -\nu \\ \lambda & \mu & \nu & \sigma \end{pmatrix} \tag{13}$$

in terms of which the minimization of (9) becomes the minimization of

$$\boldsymbol{\rho}^T\langle\boldsymbol{\rho}_B\rangle[\boldsymbol{\rho}_A]\mathbf{K}_{BA}[\boldsymbol{\rho}_A]^T\langle\boldsymbol{\rho}_B\rangle^T\boldsymbol{\rho}, \tag{14}$$

so that the further rotation required for structure B is the bottom eigenvector of

$$\langle\boldsymbol{\rho}_B\rangle[\boldsymbol{\rho}_A]\mathbf{K}_{BA}[\boldsymbol{\rho}_A]^T\langle\boldsymbol{\rho}_B\rangle^T. \tag{15}$$

Like $[\boldsymbol{\rho}]$, $\langle\boldsymbol{\rho}\rangle$ is orthogonal with positive determinant and it is shown in the *Appendix* that $\langle.\rangle$ matrices concatenate like $[.]$ matrices and that any $\langle.\rangle$ matrix commutes with any $[.]$ matrix. Note that in the development of (15) it is immaterial whether $\boldsymbol{\rho}_A$ or $\boldsymbol{\rho}_B$ is applied first provided only that both are applied before the further rotation $\boldsymbol{\rho}$ is sought. The counterpart of this in (15) is to note that $\langle\boldsymbol{\rho}_B\rangle$ and $[\boldsymbol{\rho}_A]$ may appear in either order because these matrices commute. Some further properties of the eigenvectors of (15) are also discussed in the *Appendix*.

Suppose that clusters C_I and C_J have already been formed containing n_I and n_J structures, respectively, and let

$$E_J = \sum_{p\in C_J}\sum_{q\in C_J}^{q<p}\sum_{a=1}^{m}|\mathbf{r}_{ap} - \mathbf{r}_{aq}|^2, \tag{16}$$

in which $\mathbf{r}_{ap}$ and $\mathbf{r}_{aq}$ are the position vectors of atom a in structures p and q in their current orientations within the cluster C_J, and m is the number of atoms in each and every structure. Note that (after summation over a) there are $\frac{1}{2}n_J(n_J - 1)$ terms in (16). Suppose that it is proposed to form a new cluster, C_K, by rotating the entire cluster C_I onto cluster C_J, then

$$C_K = C_I \cup C_J \tag{17}$$

$$n_K = n_I + n_J, \tag{18}$$

and

$$E_K = E_I + E_J + E_{IJ}, \tag{19}$$

in which

$$E_{IJ} = \boldsymbol{\rho}_I^T\mathbf{K}_{IJ}\boldsymbol{\rho}_I, \tag{20}$$

where

$$\mathbf{K}_{IJ} = \sum_{i\in C_I}\sum_{j\in C_J}\langle\boldsymbol{\rho}_i\rangle[\boldsymbol{\rho}_j]\mathbf{K}_{ij}[\boldsymbol{\rho}_j]^T\langle\boldsymbol{\rho}_i\rangle^T, \tag{21}$$

in which $\boldsymbol{\rho}_I$ is the rotation to be applied to the entire cluster, C_I, $\boldsymbol{\rho}_i$ and $\boldsymbol{\rho}_j$ being the rotations already

applied to the individual structures i and j within the clusters C_I and C_J in constructing them from their initial orientations. Note that (because $\bar{\mathbf{I}}^2 = \mathbf{I}$),

$$\begin{aligned}\bar{\mathbf{I}}\mathbf{K}_{IJ}\bar{\mathbf{I}} &= \sum_{i\in C_I}\sum_{j\in C_J}\bar{\mathbf{I}}\langle\boldsymbol{\rho}_i\rangle\bar{\mathbf{I}}.\bar{\mathbf{I}}[\boldsymbol{\rho}_j]\bar{\mathbf{I}}.\bar{\mathbf{I}}\mathbf{K}_{ij}\bar{\mathbf{I}}.\bar{\mathbf{I}}[\boldsymbol{\rho}_j]^T\bar{\mathbf{I}}.\bar{\mathbf{I}}\langle\boldsymbol{\rho}_i\rangle^T\bar{\mathbf{I}}\\ &= \sum_{i\in C_I}\sum_{j\in C_J}[\boldsymbol{\rho}_i]\langle\boldsymbol{\rho}_j\rangle\mathbf{K}_{ji}\langle\boldsymbol{\rho}_j\rangle^T[\boldsymbol{\rho}_i]^T\\ &= \sum_{i\in C_I}\sum_{j\in C_J}\langle\boldsymbol{\rho}_j\rangle[\boldsymbol{\rho}_i]\mathbf{K}_{ji}[\boldsymbol{\rho}_i]^T\langle\boldsymbol{\rho}_j\rangle^T\\ &= \mathbf{K}_{JI}. \qquad (20,22)\end{aligned}$$

Evidently E_K is minimized if $\boldsymbol{\rho}_I$ is the bottom eigenvector of $\mathbf{K}_{IJ}$, E_I and E_J may be supposed to be already known and E_{IJ} is the least eigenvalue of $\mathbf{K}_{IJ}$. When the cluster C_K is formed in this way the rotations $\boldsymbol{\rho}_j$, $j \in C_J$, are unchanged, whereas the vectors $\boldsymbol{\rho}_i$, $i \in C_I$, are replaced by $[\boldsymbol{\rho}_I]\boldsymbol{\rho}_i$.

The enantiomorphism test of Diamond (1990) still applies to $\mathbf{K}_{IJ}$ and takes the form

$$\begin{aligned}E'_{\min} - E_{\min} &= p_1 - p_2 - p_3 + p_4\\ &= -(k_1 - k_2 - k_3 + k_4)/2, \qquad (23)\end{aligned}$$

in which $p_1 \ldots p_4$ and $k_1 \ldots k_4$ are the eigenvalues of $\mathbf{P}$ and $\mathbf{K}$, respectively, arranged in decreasing order. Thus, the enantiomorphism of an entire cluster relative to another may be detected.

Clustering consists of determining at each stage which two clusters currently existing are the most similar to each other, and combining these by superposition. If, for a given criterion R, combining clusters I and J to form cluster K produces an R value R_K, and if combining clusters L and M to produce cluster N similarly leads to R_N, then if $R_K < R_N$, cluster K should be formed in preference to N because I is more similar to J than L is to M. There are, however, several measures which may serve as the clustering criterion R, and three possibilities are discussed here. Two of these, R_1 and R_2, are measures of the r.m.s. fit of the entire resulting cluster K, whilst the third, R_4, measures only the r.m.s. inter-cluster distance between C_I and C_J to the exclusion of intra-cluster terms within C_I and C_J. (This notation avoids contention with an R_3 defined elsewhere by D. Neuhaus.) R_1 and R_2 were defined by Diamond (1992) as

$$\begin{aligned}R_{1K} &= \left[\frac{1}{mn_K(n_K-1)}\sum_{p\in C_K}\sum_{q\in C_K}^{q\neq p}\sum_{a=1}^{m}|\mathbf{r}_{ap}-\mathbf{r}_{aq}|^2\right]^{\frac{1}{2}}\\ &= \left[\frac{1}{\frac{1}{2}mn_K(n_K-1)}\sum_{p\in C_K}\sum_{q\in C_K}^{q<p}\sum_{a=1}^{m}|\mathbf{r}_{ap}-\mathbf{r}_{aq}|^2\right]^{\frac{1}{2}}\\ &= \left[\frac{E_K}{\frac{1}{2}mn_K(n_K-1)}\right]^{\frac{1}{2}}, \qquad (42,24)\end{aligned}$$

which is the r.m.s. inter-structure distance, and with

$$\begin{aligned}\overline{\mathbf{r}_a} &= \frac{1}{n_K}\sum_{p\in C_K}\mathbf{r}_{ap}\\ R_{2K} &= \left(\frac{1}{mn_K}\sum_{p\in C_K}\sum_{a=1}^{m}|\mathbf{r}_{ap}-\overline{\mathbf{r}_a}|^2\right)^{\frac{1}{2}}\\ &= \left(\frac{E_K}{mn_K^2}\right)^{\frac{1}{2}}, \qquad (25)\end{aligned}$$

which is the standard deviation of the structures about their mean. Evidently

$$\frac{R_1^2(n-1)}{2n} = R_2^2 = \overline{|\mathbf{r}|^2} - |\bar{\mathbf{r}}|^2 \qquad (43,26)$$

so that, for $n \geq 2$, $R_2 2^{\frac{1}{2}} < R_1 \leq 2R_2$. R_4 is given by

$$R_{4K} = \left(\frac{E_{IJ}}{mn_I n_J}\right)^{\frac{1}{2}}. \qquad (27)$$

The selection of I and J is done by scanning I and J to find the least available value of R_{1K}, R_{2K} or R_{4K}, according to preference, the structure of the resulting tree being somewhat dependent on the choice in ways which are outlined below. This scan is limited to the lower triangle of arrays such as (33), so that the array address of cluster I always exceeds that of cluster J.

The clustering algorithm which has been implemented exists in two variants. The first and simplest form forms C_K by rotating C_I onto C_J without disturbing the internal structures of C_I and C_J, treating these as rigid. Thus, in (19), E_I and E_J are regarded as constants, only E_{IJ} being negotiable to minimize E_K. However, E_I was determined at some previous stage when C_I was formed, and its value was optimized in the absence of the structures in C_J. In the second variant of the algorithm it is recognised that a deeper minimum for E_K may be found if all three terms in (19) are optimized together. In this form (19) is still used to determine which cluster pair to unite at each stage, but, following union, E_K is further reduced using the algorithm of Diamond (1992) in which the rotations are initialized to $\boldsymbol{\rho}_j$, $j \in C_J$ and to $[\boldsymbol{\rho}_I]\boldsymbol{\rho}_i$, $i \in C_I$. This step will be referred to as annealing. In software terms, it is controlled by setting an upper limit to the number of cycles of annealing which, if zero, gives the first variant.

In the simple form the matrices $\mathbf{K}_{IJ}$ and the vectors $\boldsymbol{\rho}$ can be developed cumulatively, as illustrated below for the case of five structures. We begin by assembling an array

		1	2	3	4	5
		1	1	1	1	1
1	1	.	$\mathbf{K}_{12}$	$\mathbf{K}_{13}$	$\mathbf{K}_{14}$	$\mathbf{K}_{15}$
2	1	$\mathbf{K}_{21}$	.	$\mathbf{K}_{23}$	$\mathbf{K}_{24}$	$\mathbf{K}_{25}$
3	1	$\mathbf{K}_{31}$	$\mathbf{K}_{32}$	.	$\mathbf{K}_{34}$	$\mathbf{K}_{35}$
4	1	$\mathbf{K}_{41}$	$\mathbf{K}_{42}$	$\mathbf{K}_{43}$	.	$\mathbf{K}_{45}$
5	1	$\mathbf{K}_{51}$	$\mathbf{K}_{52}$	$\mathbf{K}_{53}$	$\mathbf{K}_{54}$	.

(28)

in which the outer border is the cluster number and the inner border is the number of structures in each cluster. We also evaluate the smallest eigenvalue of each matrix and similarly tabulate these

$$\begin{array}{ccc|ccccc} & & & 1 & 2 & 3 & 4 & 5 \\ & & & 1 & 1 & 1 & 1 & 1 \\ & & & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & . & E_{12} & E_{13} & E_{14} & E_{15} \\ 2 & 1 & 0 & E_{21} & . & E_{23} & E_{24} & E_{25} \\ 3 & 1 & 0 & E_{31} & E_{32} & . & E_{34} & E_{35} \\ 4 & 1 & 0 & E_{41} & E_{42} & E_{43} & . & E_{45} \\ 5 & 1 & 0 & E_{51} & E_{52} & E_{53} & E_{54} & . \end{array} \tag{29}$$

If we suppose, for example, that cluster 6 is formed from the union of clusters 2 and 5, then the arrays then stand as

$$\begin{array}{cc|ccccc} & & 1 & 6 & 3 & 4 & . \\ & & 1 & 2 & 1 & 1 & 0 \\ \hline 1 & 1 & . & \mathbf{K}_{16} & \mathbf{K}_{13} & \mathbf{K}_{14} & . \\ 6 & 2 & \mathbf{K}_{61} & . & \mathbf{K}_{63} & \mathbf{K}_{64} & . \\ 3 & 1 & \mathbf{K}_{31} & \mathbf{K}_{36} & . & \mathbf{K}_{34} & . \\ 4 & 1 & \mathbf{K}_{41} & \mathbf{K}_{46} & \mathbf{K}_{43} & . & . \\ . & 0 & . & . & . & . & . \end{array} \tag{31}$$

which (omitting the last row and column) is

$$\begin{array}{cc|ccc} & & 1 & 6 & 3 & 4 \\ & & 1 & 2 & 1 & 1 \\ \hline 1 & 1 & . & (\mathbf{K}_{12}+[\boldsymbol{\rho}_5]\mathbf{K}_{15}[\boldsymbol{\rho}_5]^T) & \mathbf{K}_{13} & \mathbf{K}_{14} \\ 6 & 2 & (\mathbf{K}_{21}+\langle\boldsymbol{\rho}_5\rangle\mathbf{K}_{51}\langle\boldsymbol{\rho}_5\rangle^T) & . & (\mathbf{K}_{23}+\langle\boldsymbol{\rho}_5\rangle\mathbf{K}_{53}\langle\boldsymbol{\rho}_5\rangle^T) & (\mathbf{K}_{24}+\langle\boldsymbol{\rho}_5\rangle\mathbf{K}_{54}\langle\boldsymbol{\rho}_5\rangle^T) \\ 3 & 1 & \mathbf{K}_{31} & (\mathbf{K}_{32}+[\boldsymbol{\rho}_5]\mathbf{K}_{35}[\boldsymbol{\rho}_5]^T) & . & \mathbf{K}_{34} \\ 4 & 1 & \mathbf{K}_{41} & (\mathbf{K}_{42}+[\boldsymbol{\rho}_5]\mathbf{K}_{45}[\boldsymbol{\rho}_5]^T) & \mathbf{K}_{43} & . \end{array} \tag{32}$$

in which the two outer borders are as before, and the third border contains the E_I values. Initially, each cluster contains only one structure and all E_I values are zero. Scanning the criterion R_1, R_2 or R_4 is based on this array and its descendents.

The E_{IJ} values tabulated in (29) each represent the best possible fit of one structure on one other, without regard to the remaining $(n-2)$ structures, and are not necessarily all attainable simultaneously. Consequently they should be regarded as potential E values. E values actually achieved will be denoted by e.

Having determined $\boldsymbol{\rho}_I$, the optimal value of which is the bottom eigenvector of the corresponding $\mathbf{K}_{IJ}$, to superimpose cluster I on cluster J the Jth column of $\mathbf{K}$ matrices is then replaced by

$$\mathbf{K}_{LK} = \mathbf{K}_{LJ} + [\boldsymbol{\rho}_I]\mathbf{K}_{LI}[\boldsymbol{\rho}_I]^T \tag{30}$$

for all L except $L = I$, $L = J$ and on abandoned rows, where K is the cluster number for the newly-formed cluster. The Jth row is then replaced using (20,22) and the Ith row and column are abandoned. $\mathbf{K}_{LK}$ then has the property that its lowest eigenvalue is E_{LK} and the corresponding eigenvector, $\boldsymbol{\rho}_L$, optimally rotates cluster L (still a single structure) onto cluster K (currently two structures). Similarly $\mathbf{K}_{KL}$ provides for the rotation of cluster K by $\boldsymbol{\rho}_K$ onto cluster L, involving now the further rotation of those structures within C_K which have already been rotated.

The expression (19) then replaces E_J in both borders of (29), n_J is replaced by $n_I + n_J$ and n_I is set to zero.

and the eigenvalue array becomes

$$\begin{array}{ccc|ccccc} & & & 1 & 6 & 3 & 4 & . \\ & & & 1 & 2 & 1 & 1 & 0 \\ & & & 0 & E_6 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & . & E_{16} & E_{13} & E_{14} & . \\ 6 & 2 & E_6 & E_{61} & . & E_{63} & E_{64} & e_{25} \\ 3 & 1 & 0 & E_{31} & E_{36} & . & E_{34} & . \\ 4 & 1 & 0 & E_{41} & E_{46} & E_{43} & . & . \\ . & 0 & 0 & . & e_{52} & . & . & . \end{array} \tag{33}$$

in which $E_6 = E_{52}$ and other entries involving cluster 6 are the least eigenvalues of the new entries in (32). If constellations are also to be calculated (see below) then the achieved value $e_{25} = E_{25}$ is also recorded as shown.

Suppose that by repeated application of this process cluster 7 is formed by the union of clusters 1 and 3 and cluster 8 is formed by the union of clusters 4 and 6, then the $\mathbf{K}$ array will become

$$\begin{array}{cc|ccccc} & & 7 & 6 & . & 4 & . \\ & & 2 & 2 & 0 & 1 & 0 \\ \hline 7 & 2 & . & \mathbf{K}_{76} & . & \mathbf{K}_{74} & . \\ 6 & 2 & \mathbf{K}_{67} & . & . & \mathbf{K}_{64} & . \\ . & 0 & . & . & . & . & . \\ 4 & 1 & \mathbf{K}_{47} & \mathbf{K}_{46} & . & . & . \\ . & 0 & . & . & . & . & . \end{array} \tag{34}$$

then

$$\begin{array}{cc|ccccc} & & 7 & 8 & . & . & . \\ & & 2 & 3 & 0 & 0 & 0 \\ \hline 7 & 2 & . & \mathbf{K}_{78} & . & . & . \\ 8 & 3 & \mathbf{K}_{87} & . & . & . & . \\ . & 0 & . & . & . & . & . \\ . & 0 & . & . & . & . & . \\ . & 0 & . & . & . & . & . \end{array} \qquad (35)$$

in which, for example,

$$\begin{aligned} \mathbf{K}_{67} =& \mathbf{K}_{61} + [\boldsymbol{\rho}_3]\mathbf{K}_{63}[\boldsymbol{\rho}_3]^T \\ =& \mathbf{K}_{21} + \langle\boldsymbol{\rho}_5\rangle\mathbf{K}_{51}\langle\boldsymbol{\rho}_5\rangle^T \\ &+ [\boldsymbol{\rho}_3](\mathbf{K}_{23} + \langle\boldsymbol{\rho}_5\rangle\mathbf{K}_{53}\langle\boldsymbol{\rho}_5\rangle^T)[\boldsymbol{\rho}_3]^T \\ \mathbf{K}_{76} =& \bar{\mathbf{I}}\mathbf{K}_{67}\bar{\mathbf{I}} = \mathbf{K}_{12} + [\boldsymbol{\rho}_5]\mathbf{K}_{15}[\boldsymbol{\rho}_5]^T \\ &+ \langle\boldsymbol{\rho}_3\rangle(\mathbf{K}_{32} + [\boldsymbol{\rho}_5]\mathbf{K}_{35}[\boldsymbol{\rho}_5]^T)\langle\boldsymbol{\rho}_3\rangle^T \\ \mathbf{K}_{47} =& \mathbf{K}_{41} + [\boldsymbol{\rho}_3]\mathbf{K}_{43}[\boldsymbol{\rho}_3]^T \\ \mathbf{K}_{78} =& \mathbf{K}_{76} + [\boldsymbol{\rho}_4]\mathbf{K}_{74}[\boldsymbol{\rho}_4]^T \\ =& \mathbf{K}_{12} + [\boldsymbol{\rho}_5]\mathbf{K}_{15}[\boldsymbol{\rho}_5]^T \\ &+ \langle\boldsymbol{\rho}_3\rangle(\mathbf{K}_{32} + [\boldsymbol{\rho}_5]\mathbf{K}_{35}[\boldsymbol{\rho}_5]^T)\langle\boldsymbol{\rho}_3\rangle^T \\ &+ [\boldsymbol{\rho}_4](\mathbf{K}_{14} + \langle\boldsymbol{\rho}_3\rangle\mathbf{K}_{34}\langle\boldsymbol{\rho}_3\rangle^T)[\boldsymbol{\rho}_4]^T, \end{aligned} \qquad (36)$$

and the eigenvalue array becomes

$$\begin{array}{ccc|ccccc} & & & 7 & 6 & . & 4 & . \\ & & & 2 & 2 & 0 & 1 & 0 \\ & & & E_7 & E_6 & 0 & 0 & 0 \\ \hline 7 & 2 & E_7 & . & E_{76} & e_{13} & E_{74} & . \\ 6 & 2 & E_6 & E_{67} & . & . & E_{64} & e_{25} \\ . & 0 & 0 & e_{31} & . & . & . & . \\ 4 & 1 & 0 & E_{47} & E_{46} & . & . & . \\ . & 0 & 0 & . & e_{52} & . & . & . \end{array} \qquad (37)$$

then

$$\begin{array}{ccc|ccccc} & & & 7 & 8 & . & . & . \\ & & & 2 & 3 & 0 & 0 & 0 \\ & & & E_7 & E_8 & 0 & 0 & 0 \\ \hline 7 & 2 & E_7 & . & E_{78} & e_{13} & . & . \\ 8 & 3 & E_8 & E_{87} & . & . & e_{24} & e_{25} \\ . & 0 & 0 & e_{31} & . & . & . & . \\ . & 0 & 0 & . & e_{42} & . & . & e_{45} \\ . & 0 & 0 & . & e_{52} & . & e_{54} & . \end{array} \qquad (38)$$

in which $E_7 = E_{31}$ and other E values involving cluster 7 are the corresponding least eigenvalues in (34), $E_8 = E_{46} + E_6$, (E_4 being zero) and E_{87} is the least eigenvalue of $\mathbf{K}_{87}$. Note that in forming cluster 8, one structure, 4, is rotated onto two others, 2 and 5, (which comprise cluster 6) and the value of E_{64} contained in (33) and (37) becomes $e_{24}+e_{54}$. Such sums are sufficient to control clustering, but if constellations are also to be calculated then the individual e values are required, and, on the formation of C_K, e_{ij} values are determined from

$$e_{ij} = \boldsymbol{\rho}_i^T[\boldsymbol{\rho}_j]\mathbf{K}_{ij}[\boldsymbol{\rho}_j]^T\boldsymbol{\rho}_i \qquad i \in C_I, j \in C_J, \qquad (39)$$

and such values are included in (38).

Cluster 8 is then rotated onto cluster 7 using $\boldsymbol{\rho}_8$ which is the bottom eigenvector of $\mathbf{K}_{87}$, giving a final mean-square residual, mR_1^2, over the entire ensemble of $(E_{87}+E_8+E_7)/10$. The expressions thus developed for $\mathbf{K}_{87}$ and $\mathbf{K}_{78}$ correspond to (21) with $[\boldsymbol{\rho}_1] = [\boldsymbol{\rho}_2] = \mathbf{I}$.

In the second variant of the algorithm, revisions are made to the internal orientations of each structure within a cluster during the annealing step, which means that cumulative procedures cannot be used, the rotations $\boldsymbol{\rho}_j$, $j \in C_J$ and $[\boldsymbol{\rho}_I]\boldsymbol{\rho}_i$, $i \in C_I$ must be replaced by the revised values and matrices $\mathbf{K}_{IJ}$ must be calculated from (21) after each cluster is formed and annealed, for all the remaining cluster pairs for which the newly-formed cluster, C_K, is one of the pair, because the least eigenvalues of these matrices must be scanned to select clusters for subsequent unions. The commutation property of [.] and ⟨.⟩ matrices enables (21) to be evaluated with n_I sums of n_J terms followed by one sum of n_I terms, or *vice versa*, regardless of the order of events leading to the current situation.

If annealing is being done and constellations are to be calculated then, on formation of C_K, e_{ij} must be evaluated for all $i \in C_K$ and $j \in C_K$. With or without annealing, the E array is ultimately fully populated with e values if constellations are being calculated, and is vacated if not.

Constellations

A structure of m atoms, normally represented by m vectors in three dimensions, may alternatively be represented by a single vector in $3m$ dimensions. Furthermore, n structures may be represented by a constellation of n points in $3m$ dimensions, and the entire constellation may be represented without distortion in a space of $(n-1)$ dimensions. For n values up to about 10 it is also frequently the case that a projection of such a constellation into two or three dimensions involves little loss of information, and may be graphically displayed. This principle has been exploited in a different context, for example, by Frank (1990), and in the present context by Sutcliffe (1993).

The value of such constellations arises because distances between points in the constellation, when suitably scaled, are the r.m.s. differences between the corresponding structures, and because the cosine of the angle between two vectors in the constellation is the correlation coefficient of the inter-structure differences represented by those vectors. The calculation of such constellations

is described by Diamond (1974) (§4) for the case in which one of the structures is placed at the origin of the constellation. However, for cluster C_K, the centroid of the structures may be placed at the origin by evaluating

$$f_i = \frac{1}{n_K} \sum_{j \in C_K} e_{ij} \qquad i \in C_K$$
$$\bar{f} = \frac{1}{n_K} \sum_{i \in C_K} f_i, \tag{40}$$

and a matrix $\mathbf{S}^T\mathbf{S}$ with elements

$$\frac{1}{2}(f_i + f_j - \bar{f} - e_{ij}). \tag{41}$$

Then if $\mathbf{A}$ is orthogonal and $\mathbf{\Lambda}$ diagonal such that

$$\mathbf{\Lambda} = \mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{A}, \tag{42}$$

and

$$\mathbf{C} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{A}^T, \tag{43}$$

then $\mathbf{C}$ contains, as columns, the position vectors of

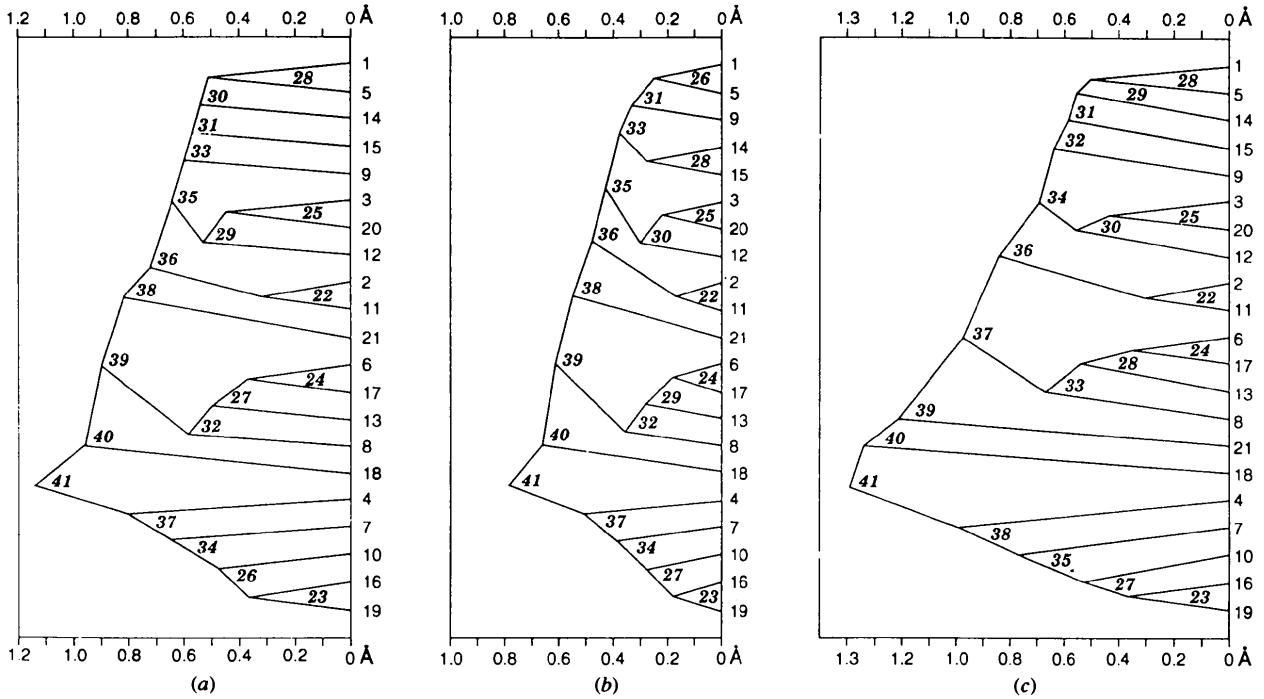

Fig. 1. (*a*), (*b*) and (*c*). Dendrograms for the clustering of 21 structures by each of the criteria R_1, R_2 and R_4, respectively, the values of which are plotted horizontally in Å. Structures are identified by numbers on the right and resulting clusters are identified by cluster numbers to the right of each node, beginning with cluster 22. See the text for a discussion of the differences. [The criteria are defined by (42,24), (25) and (27).]

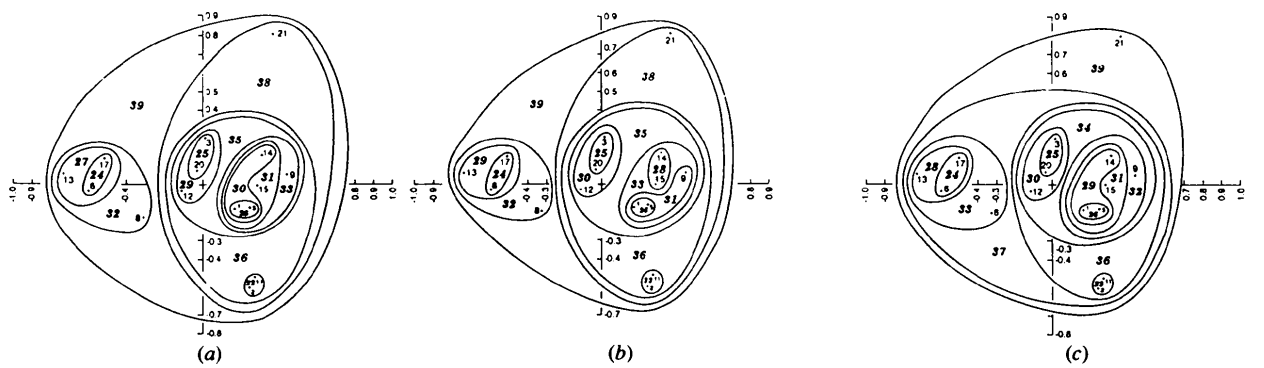

Fig. 2. (*a*), (*b*) and (*c*). Principal projections of the constellation corresponding to cluster 39 in each of Figs. 1(*a*), 1(*b*) and 1(*c*). Numbered points correspond to the corresponding structures, which are enclosed in numbered regions corresponding to the clusters formed. Although the numbered points are in the same positions in each of these diagrams the resulting groupings differ. Horizontal and vertical scales are in Å with the origin at the centroid.

each of the points in the constellation representing C_K. It is convenient to arrange the columns of $\mathbf{A}$ so that the eigenvalues in $\mathbf{\Lambda}$ are presented in decreasing order, so that the first p components of each vector in $\mathbf{C}$ provide the dominant projection of the constellation in p dimensions, and the ratio $100\sum_{i=1}^{p}\mathbf{\Lambda}_{ii}/\sum_{i=1}^{n_K}\mathbf{\Lambda}_{ii}$ is a percentage expressing the extent to which the projection represents the entire constellation.

From expression (41) it is clear that $\mathbf{S}^T\mathbf{S}$ is rank deficient, [any row (column) is minus the sum of all other rows (columns)] and, therefore, has at least one vanishing eigenvalue, and the n_Kth component of each column of $\mathbf{C}$ is, therefore, zero.

The matrix $\mathbf{S}$, though never computed, has $3m$ rows and n_K columns, the ith column containing the coordinates of structure i expressed relative to the mean structure. Thus, $\mathbf{S}^T\mathbf{S}$ is positive semi-definite for any set of coordinates. Replacing e values by E_0 values, for example, would enable a constellation to be computed which would express similarities among the initial orientations. Using the E values in (29), however, may lead to negative eigenvalues in (42), because these E values are not simultaneously attainable and do not correspond to a situation in which each structure is represented by a single set of coordinates.

Examples

The examples are taken from studies of proteins known as the high-mobility group, HMG-D [Jones *et al.* (1994) to whom I am indebted for the use of the coordinates] and the related *B* domain of rat HMG-1 (Weir *et al.*, 1993). Sixteen HMG-D and five HMG-1 structures together form a single ensemble in this example.

Figs. 1(*a*), 1(*b*) and 1(*c*) show trees developed without annealing for 21 structures derived from n.m.r experiments using 117 main-chain atoms (Cα, C and N only) from each structure, basing the clustering on R_1, R_2 and R_4, respectively. The trees are similar but not identical. The n-dependence given in (43,26) means that clustering based on R_2 tends to favour the initiation of new small clusters in preference to augmenting existing large ones, as illustrated by structures 14 and 15 forming cluster 28 by themselves in Fig. 1(*b*), whereas in Fig. 1(*a*) these two structures form successive addends to a pre-existing cluster. This tendency becomes more marked as the size of the tree increases.

With R_4 the similarity or otherwise of clusters I and J, rather than the compactness of the resulting cluster K, determines clustering. This engenders a tendency to postpone the incorporation of 'outliers'. For example, in Fig. 1(*a*), structure 21 unites with cluster 36 (already containing ten structures) to produce cluster 38, with a resulting increase in R_1 from 0.71 to 0.81 Å which is a substantial change to attribute to a single structure. However, in cluster 38 the ten interactions between structure 21 and the structures within cluster 36 are diluted by the 45 interactions within cluster 36 when evaluating R_1 for cluster 38. When R_4 is the clustering criterion no such dilution takes place, only the ten interactions between structure 21 and the structures in cluster 36 are considered, all of which are large, as a result of which structure 21 is not united with cluster 36, and its incorporation is postponed.

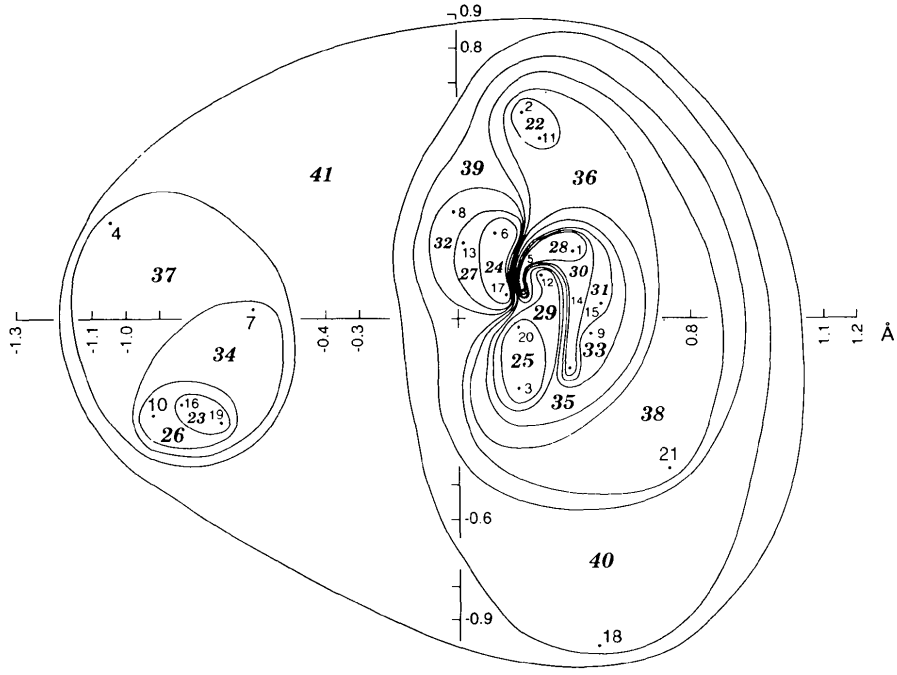

Fig. 3. Principal projection of the constellation of cluster 41 in Fig. 1(*a*). The orientation of the constellation is dominated by the distance between cluster 37, containing the rat structures, and cluster 40, containing all others.

All three trees indicate that the five structures 4, 7, 10, 16 and 19, in the lower part of the tree, are substantially different from all other structures. These five structures have a different origin, being those of the *B* domain of rat HMG-1 (Weir *et al.*, 1993) and the clustering has detected that they form a separate family within the ensemble.

In all three figures cluster 39 contains the same 15 structures (all from HMG-D) and, therefore, if annealed, can be represented by the same 14-dimensional constellation. In this instance the structures are sufficiently similar that the annealing step makes a negligible difference, and the three constellations may be regarded as the same. The dominant two-dimensional projections of these constellations are shown in Figs. 2(*a*), 2(*b*) and 2(*c*), for which 62.3% of the squared coordinates lie in the plane of the paper. In these diagrams each structure is represented by a numbered point, and the envelopes enclosing each cluster are also shown and numbered. These diagrams serve to show how trees with differing connectivities may yet be consistent with a single constellation. In this instance the structures involved are sufficiently similar to present almost the same principal projection when calculated from the pairwise residuals of (29).

In Fig. 3 the two-dimensional projection of the 20-dimensional constellation corresponding to cluster 41 of Fig. 1(*a*) (all 21 structures) is shown. In this figure the separateness of cluster 37, containing the rat structures, is evident and its distance from all other structures is such that its position in 20-dimensional space is the principal determinant of the long axis of the constellation, and hence of its dominant projection. Consequently cluster 39 is seen somewhat 'edge-on' in this projection with consequential overloading of detail in the diagram. In this instance 59.0% of squared coordinates are in the plane of the paper, and the 41% not represented in the diagram is sufficient to obscure much detail and, thus, to limit its usefulness.

Cross-term errors for each of the 117 atoms in the structures are shown in Fig. 4. The figures on the left are cluster numbers for the tree of Fig. 1(*a*), whilst the two rows of figures on the right are lists of structures in each of the two clusters being united to form the new cluster. Only differences as between structures on the upper row and on the lower row contribute to this measure. The graph for cluster 34, for example, shows that structure 7 differs from structures 10, 16 and 19 around atoms 18, 45 and 116 particularly. The differences between structure 21 and ten other structures at cluster 38, discussed above, exceed 3 Å at atom 65. Such a plot may draw attention to regions where stereochemical differences would be worth examining visually. I am indebted to Dr A.D. MacLachlan for suggesting this form of output.

APPENDIX

Using $\boldsymbol{\lambda}$ to represent the three-dimensional vector consisting of the first three components of $\boldsymbol{\rho}$, [*c.f.* (1,2)], and defining

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0 & \nu & -\mu \\ -\nu & 0 & \lambda \\ \mu & -\lambda & 0 \end{pmatrix} \tag{A1}$$

we find that the product

$$\begin{pmatrix} (\sigma_p \mathbf{I} - \boldsymbol{\Lambda}_p) & \boldsymbol{\lambda}_p s_p \\ -s_p \boldsymbol{\lambda}_p^T & \sigma_p \end{pmatrix} \begin{pmatrix} (\sigma_q \mathbf{I} - \boldsymbol{\Lambda}_q) & \boldsymbol{\lambda}_q s_q \\ -s_q \boldsymbol{\lambda}_q^T & \sigma_q \end{pmatrix} =$$

$$\begin{pmatrix} \left((\sigma_p\sigma_q - \boldsymbol{\lambda}_p^T\boldsymbol{\lambda}_q)\mathbf{I} - \sigma_q\boldsymbol{\Lambda}_p - \sigma_p\boldsymbol{\Lambda}_q + \boldsymbol{\lambda}_q\boldsymbol{\lambda}_p^T - \boldsymbol{\lambda}_p s_p s_q \boldsymbol{\lambda}_q^T\right) & \left((\boldsymbol{\lambda}_p \times \boldsymbol{\lambda}_q)s_q + \boldsymbol{\lambda}_q\sigma_p s_q + \boldsymbol{\lambda}_p\sigma_q s_p\right) \\ -\left((\boldsymbol{\lambda}_p \times \boldsymbol{\lambda}_q)s_p + \boldsymbol{\lambda}_q\sigma_p s_q + \boldsymbol{\lambda}_p\sigma_q s_p\right)^T & (\sigma_p\sigma_q - s_p\boldsymbol{\lambda}_p^T\boldsymbol{\lambda}_q s_q) \end{pmatrix}. \tag{A2}$$

Setting $s_p = s_q = +1$ shows that the product of two [.] matrices is a [.] matrix with $\boldsymbol{\rho}$ vector given by (11) of Diamond (1992).

Setting $s_p = s_q = -1$ shows that the product of two $\langle . \rangle$ matrices is also a $\langle . \rangle$ matrix with the same $\boldsymbol{\rho}$.

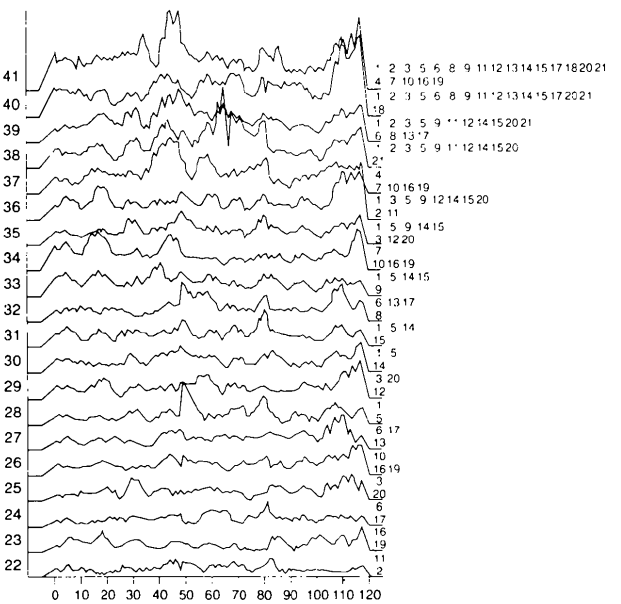

Fig. 4. The quantity $[(n_I n_J)^{-1} \sum_{p \in C_I} \sum_{q \in C_J} |\mathbf{r}_{ap} - \mathbf{r}_{aq}|^2]^{\frac{1}{2}}$ plotted as a function of the atom number, a, as each cluster C_K is formed by the criterion R_1. K values are shown on the left and p and q values (relating to individual structures) are listed in two rows on the right of each plot. The base line spacing is 1 Å. Such plots serve to identify regions of difference.

Setting $p = 2, q = 1, s_p = +1, s_q = -1$ gives an expression for $[\boldsymbol{\rho}_2]\langle\boldsymbol{\rho}_1\rangle$, and setting $p = 1, q = 2, s_p = -1, s_q = +1$ gives an expression for $\langle\boldsymbol{\rho}_1\rangle[\boldsymbol{\rho}_2]$ which is found to be the same expression. Therefore any [.] matrix commutes with any $\langle.\rangle$ matrix.

Setting $\boldsymbol{\rho}_p = \boldsymbol{\rho}_q$, $|s_p| = 1$ and $s_p = -s_q$ gives (32) of Diamond (1992).

It has been shown that the bottom eigenvector of the expression (15) provides the further rotation of the rotated B structure to fit the rotated A structure. The remaining three eigenvectors, which are not required for clustering purposes, provide further rotations of the rotated B structure onto the rotated A structure which result in stationary values of E, namely a maximum, a pass and a pale. However, it may be of interest in other applications to use all four eigenvectors to control rotations of *both* structures simultaneously in the manner outlined below.

Let $\mathbf{A}$ be orthogonal with positive determinant such that

$$\boldsymbol{\mathcal{K}}_{BA} = \mathbf{A}\langle\boldsymbol{\rho}_B\rangle[\boldsymbol{\rho}_A]\mathbf{K}_{BA}[\boldsymbol{\rho}_A]^T\langle\boldsymbol{\rho}_B\rangle^T\mathbf{A}^T \qquad (A3)$$

is diagonal, rows of $\mathbf{A}$ being the eigenvectors. $\mathbf{A}$, being orthogonal, has six degrees of freedom because, for example, specifying the six off-diagonal elements in the lower triangle of $\mathbf{A}$ is sufficient to determine the remaining ten elements. $\langle.\rangle$ and [.] matrices, by contrast, each have only three degrees of freedom, being functions only of λ, μ, ν and σ which are not independent. The product $\langle\boldsymbol{\rho}_B\rangle[\boldsymbol{\rho}_A]$ has six degrees of freedom and no special structure other than being orthogonal, and may therefore be regarded as representing the general form of a 4×4 orthogonal matrix with positive determinant. Consequently $\mathbf{A}$ may be factorized according to

$$\mathbf{A} = \langle\boldsymbol{\rho}_b\rangle[\boldsymbol{\rho}_a] = [\boldsymbol{\rho}_a]\langle\boldsymbol{\rho}_b\rangle = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \qquad (A4)$$

and $\boldsymbol{\rho}_a$ and $\boldsymbol{\rho}_b$ may be derived from $\mathbf{A}$ by forming the product as in (A2) and inverting. The result is,

$$\begin{pmatrix} \lambda_a\lambda_b & \mu_a\nu_b & \nu_a\sigma_b & \sigma_a\mu_b \\ \mu_a\mu_b & \lambda_a\sigma_b & \sigma_a\nu_b & \nu_a\lambda_b \\ \nu_a\nu_b & \sigma_a\lambda_b & \lambda_a\mu_b & \mu_a\sigma_b \\ \sigma_a\sigma_b & \nu_a\mu_b & \mu_a\lambda_b & \lambda_a\nu_b \end{pmatrix} =$$

$$\tfrac{1}{4}\begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} a_{11} & a_{23} & a_{34} & a_{42} \\ a_{22} & a_{14} & a_{43} & a_{31} \\ a_{33} & a_{41} & a_{12} & a_{24} \\ a_{44} & a_{32} & a_{21} & a_{13} \end{pmatrix}, \qquad (A5)$$

from which both $\boldsymbol{\rho}_a$ and $\boldsymbol{\rho}_b$ may be extracted, with a single ambiguity of sign, using $\boldsymbol{\rho}^T\boldsymbol{\rho} = 1$. Hence, (A3) becomes

$$\begin{aligned} \boldsymbol{\mathcal{K}}_{BA} &= \langle\boldsymbol{\rho}_b\rangle[\boldsymbol{\rho}_a]\langle\boldsymbol{\rho}_B\rangle[\boldsymbol{\rho}_A]\mathbf{K}_{BA}[\boldsymbol{\rho}_A]^T\langle\boldsymbol{\rho}_B\rangle^T[\boldsymbol{\rho}_a]^T\langle\boldsymbol{\rho}_b\rangle^T \\ &= \langle\boldsymbol{\rho}_\beta\rangle[\boldsymbol{\rho}_\alpha]\mathbf{K}_{BA}[\boldsymbol{\rho}_\alpha]^T\langle\boldsymbol{\rho}_\beta\rangle^T, \end{aligned} \qquad (A6)$$

in which

$$\begin{aligned} \langle\boldsymbol{\rho}_\beta\rangle &= \langle\boldsymbol{\rho}_b\rangle\langle\boldsymbol{\rho}_B\rangle \\ [\boldsymbol{\rho}_\alpha] &= [\boldsymbol{\rho}_a][\boldsymbol{\rho}_A]. \end{aligned} \qquad (A7)$$

Thus, the eigenvectors of $\boldsymbol{\mathcal{K}}_{BA}$, which are the columns of the identity, specify the further rotations of the (twice) rotated B structure onto the (twice) rotated A structure which lead to the four stationary values of E, and the rotations so specified are 180° rotations about each of the axes of the coordinate system, and the identity. Thus, $\mathbf{A}$ may be used to orient both structures simultaneously with these special rotation axes aligned on the coordinate axes. Note that because $\boldsymbol{\mathcal{K}}_{BA}$ is diagonal $\boldsymbol{\mathcal{K}}_{BA} = \boldsymbol{\mathcal{K}}_{BA}$ by (20,22), so that the foregoing statements concerning the further rotation of structure B are equally true of structure A.

References

DIAMOND, R. (1974). *J. Mol. Biol.* **82**, 371–391

DIAMOND, R. (1988). *Acta Cryst.* A**44**, 211–216.

DIAMOND, R. (1989). *Acta Cryst.* A**45**, 657–657

DIAMOND, R. (1990). *Acta Cryst.* A**46**, 423–423.

DIAMOND, R. (1992). *Protein Sci.* **1**, 1279–1287.

FRANK, J. (1990). *Quart. Rev. Biophys.* **23**, 281–329.

JONES, D. N. M., SEARLES, M. A., SHAW, G. L., CHURCHILL, M. E. A., NER, S. S., KEELER, J., TRAVERS, A. A. & NEUHAUS, D. (1994). *Structure*, **2**, 609–627.

KEARSLEY, S. K. (1989). *Acta Cryst.* A**45**, 208–210.

KEARSLEY, S. K. (1990). *J. Comput. Chem.* **11**(10), 1187–1192.

RUSSELL, R. B. & BARTON, G. J. (1992). *Proteins Struct. Funct. Genet.* **14**, 309–323.

SHAPIRO, A., BOTHA, J. D., PASTORE, A. & LESK, A. M. (1992). *Acta Cryst.* A**48**, 11–14.

SUTCLIFFE, M. J., (1993). *Protein Sci.* **2**, 936–944.

WEIR, H. M., KRAULIS, P. J., HILL, C. S., RAINE, A. R. C., LAUE, E. D. & THOMAS, J. O. (1993). *EMBO J.* **12**, 1311–1319.