AD-A113 786   PRINCETON UNIV NJ DEPT OF STATISTICS   F/G 6/1
ROBUST COMPARISON OF THREE-DIMENSIONAL SHAPES WITH AN APPLICATI--ETC(U)
MAR 82 A F SIEGEL, J R PINKERTON   DAAG29-79-C-0205
UNCLASSIFIED   TR-217-SER-2   ARO-16669.17-M   NL

1 of 1
AD A
113786

END
DATE
FILMED
5-82
DTIC

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
16669.17-M

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Robust Comparison of Three-Dimensional Shapes with an Application to Protein Molecule Configurations

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Andrew F. Siegel
John R. Pinkerton

8. CONTRACT OR GRANT NUMBER(s)
DAAG29 79 C 0205

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Princeton University
Princeton, NJ 08540

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
Mar 82

13. NUMBER OF PAGES
18

MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

shape comparison
resistance
pattern matching

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Robust statistical methods have recently been shown to have desirable properties when used for the identification of similarities and differences in shape. We present a generalization of the two-dimensional repeated median algorithm to three and higher dimensions. The extension is achieved using a duality between orthogonal and skew-symmetric matrices, which permits the definition of a median of a collection of orthogonal matrices. The methods are illustrated by comparing the predicted three-dimensional configuration of a protein molecule to a refined structure that had been found using nuclear magnetic resonance techniques.

ROBUST COMPARISON OF THREE-DIMENSIONAL SHAPES WITH AN APPLICATION TO PROTEIN MOLECULE CONFIGURATIONS

by

Andrew F. Siegel and John R. Pinkerton
Department of Statistics, Princeton University

Technical Report No. 217, Series 2
Department of Statistics
Princeton University
March, 1982

This work was supported in part by U.S. Army Research Office Contract No. DAAG-29-79-C-0205. We are grateful to S.K. Dower for sharing his data on protein molecule configurations, and to G. Easton for some helpful suggestions on the exposition.

ABSTRACT

Robust statistical methods have recently been shown to have desirable properties when used for the identification of similarities and differences in shape. We present a generalization of the two-dimensional repeated median algorithm to three and higher dimensions. The extension is achieved using a duality between orthogonal and skew-symmetric matrices, which permits the definition of a median of a collection of orthogonal matrices. The methods are illustrated by comparing the predicted three-dimensional configuration of a protein molecule to a refined structure that had been found using nuclear magnetic resonance techniques.

SOME KEY WORDS: Shape comparison, Resistance, Pattern matching.

1. INTRODUCTION

Many quantitative methods for the comparison of shape and form in two dimensions have been proposed since the fundamental descriptive work of Thompson (1917). Sneath (1967) used least squares as a basis for establishing a common frame of reference for the comparison of two objects and for drawing inferences about their similarities and differences. Robust estimation for this problem using the technique of repeated medians was proposed by Siegel and Benson (1982), and some real advantages of robust methods over least squares were demonstrated. Related theory and examples may be found in Siegel (1982a and 1982b) and in Olshan, Siegel, and Swindler (1982). Some additional contributions to the study of shape and form include Gould (1966), Mosimann (1970), Gower (1975), and Bookstein (1977).

Robust methods are often superior to least squares in the comparison of shape because a localized difference in shape between two objects can be thought of as an outlier in the fitting process. Due to its high sensitivity to outliers, a least squares fit will tend to underplay the size of such a shape difference, and thereby render it difficult to detect. At the same time, differences may tend to be exaggerated at points that would otherwise have been fitted closely.

For example, Figure 1.1 shows the comparison of two hypothetical three-dimensional geometric shapes by rotation and translation. The fitting process acts on the nine homologous pairs of points, one point in each shape, and tries to bring point i of shape 1 close to point i of shape 2. The shapes are identical in 8 of their 9 points, which are placed at the vertices of a cube, while the last point is different. As a result of trying to bring the outlying points closer together, the least squares fit suggests the existence of shape differences at all 9 points, while the robust fit (computed using the methods to be developed here) correctly indicates the closeness of the correspondence at the vertices of the cube, and also indicates the full size of the difference at the last point.

The main difficulty involved in extending the repeated median technique for shape comparison from two to three dimensions is that the componentwise median of a set of orthogonal matrices need not itself be an orthogonal matrix. By working with angles instead of matrices this problem can be avoided in two dimensions. In Section 2 we show how the three-dimensional rotational component of the fit can be obtained by medians using a duality between orthogonal and skew symmetric matrices. These methods are illustrated in Section 3 using data from the three-dimensional configurations of related protein molecules that have been studied by Dover (1979) using least squares techniques.

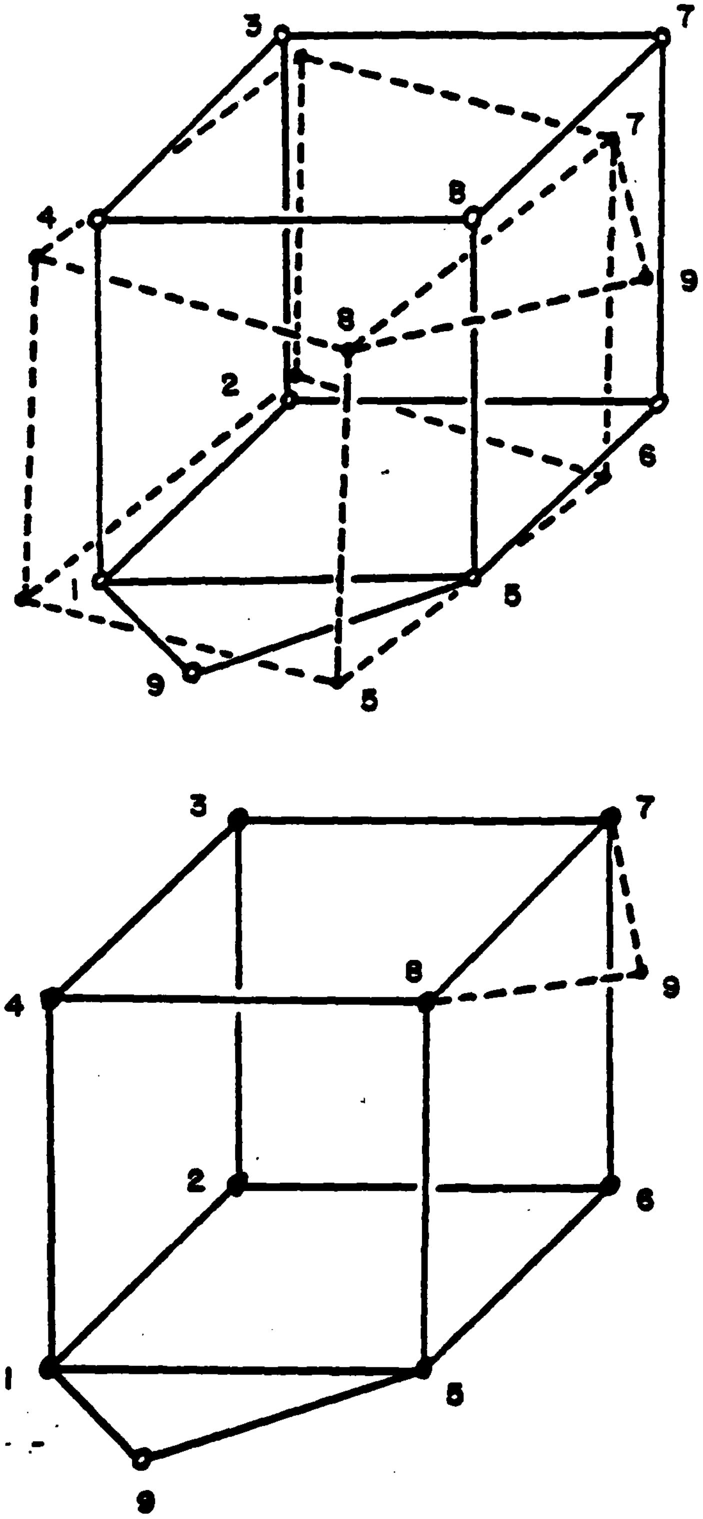

FIGURE 1.1. Two hypothetical three-dimensional geometric shapes, superimposed by least squares fitting (top) and by repeated median fitting (bottom).

We will assume that our data set consists of n homologous points in k dimensions, denoted $X_1, \ldots, X_n$ for shape 1 and $U_1, \ldots, U_n$ for shape 2. In order to transform the points of shape 2 to a close fit with the corresponding points of shape 1, we will allow rotations and translations, estimating an orthogonal rotation matrix O and a translation vector T so that the residual vectors

$$(1.1) \qquad U_i - (T + O\,X_i)$$

are small in magnitude. A magnification factor m can be included, in which case we would make the residual vectors

$$(1.2) \qquad U_i - m\,(T + O\,X_i)$$

small in magnitude by estimating O, T, and m.

The least squares solution, which minimizes the sum of the squared lengths of the terms in (1.1), can be computed using the singular value decomposition (Huber, 1980).

2. THE THREE-DIMENSIONAL REPEATED MEDIAN ALGORITHM

The repeated median algorithm, like a U-statistic (Hoeffding, 1948), proceeds one parameter at a time. We first present details for estimation of the orthogonal matrix, then summarize the steps for obtaining the translation and magnification. A preliminary least squares fit is used as a point of departure.

A subset of two pairs of homologous points, say points with indices i and j, from each of the two shapes (i.e. X_i and X_j of shape 1 with points U_i and U_j of shape 2) is not sufficient to uniquely determine a three-dimensional rotation. Three pairs of homologous points, for example i, j, and k, generally are sufficient to determine such a rotation, although different methods will result in slightly different rotation matrices. One method that generalizes easily to higher dimensions is based on the least squares fit of the three points. However, this will not usually match anything exactly. In order to match some aspects of the data exactly, we will choose a three by three orthogonal matrix O_{ijk} (one matrix for each ordered triple i, j, and k) so that

(2.1) the directions of $O_{ijk}(X_j - X_i)$ and $U_j - U_i$ are the same

and

(2.2) the transformed point $O_{ijk} X_k$ is in the same plane as the points U_i, U_j, and U_k, and is on the same side of the line through U_i and U_j as is U_k.

To find O_{ijk} we will define vectors

$$(2.3) \qquad X_{ij} = \frac{X_j - X_i}{\|X_j - X_i\|} \qquad\qquad U_{ij} = \frac{U_j - U_i}{\|U_j - U_i\|}$$

$$(2.4) \qquad X_{ijk} = \frac{(X_k - X_i) - [(X_k - X_i)\cdot(X_j - X_i)]\ X_{ij}}{\|(X_k - X_i) - [(X_k - X_i)\cdot(X_j - X_i)]\ X_{ij}\|}$$

$$(2.5) \qquad U_{ijk} = \frac{(U_k - U_i) - [(U_k - U_i)\cdot(U_j - U_i)]\ U_{ij}}{\|(U_k - U_i) - [(U_k - U_i)\cdot(U_j - U_i)]\ U_{ij}\|}$$

It can be verified that the rotation matrix O_{ijk} satisfying conditions (2.1) and (2.2) for points i, j, and k is

$$(2.6) \qquad O_{ijk} = (U_{ij}, U_{ijk}, U_{ij} \times U_{ijk}) \cdot \begin{bmatrix} X'_{ij} \\ X'_{ijk} \\ (X_{ij} \times X_{ijk})' \end{bmatrix}$$

where "x" denotes the cross product of two vectors.

The repeated median process computes a single matrix from these n(n-1)(n-2) orthogonal matrices using a duality between orthogonal and skew symmetric matrices, details of which may be found in Eves (1966). The skew symmetric matrix corresponding to O_{ijk} is

$$(2.7) \qquad S_{ijk} = (O_{ijk}+I)^{-1}(O_{ijk}-I)$$

where I denotes the identity matrix. Taking triply repeated medians of each entry, we obtain the skew symmetric matrix S:

$$(2.8) \qquad S = \underset{i}{\text{median}} \, [\underset{j \neq i}{\text{median}} \, [\underset{k \neq i,j}{\text{median}} \; S_{ijk}]]$$

where the median of a set of matrices is defined as the matrix of medians computed at each entry. The skew symmetric matrix S is then transformed back to the orthogonal matrix O using the inverse relation

$$(2.9) \qquad O = (I+S)(I-S)^{-1}$$

which completes the definition of the repeated median orthogonal rotation matrix O.

The translation vector, T, should be computed by finding a robust estimate of the three dimensional location of the data $U_i - O\ X_i$ (i=1,...,n), using the value for O from (2.9). This location might be found using the mediancentre (Bedall and Zimmermann, 1979), which is the point that minimizes the sum of the Euclidean distances from it. A simpler method is to use the vector of univariate medians computed separately for each coordinate.

The magnification factor m, which is needed in some applications but omitted in others, can be found using the same technique used by Siegel and Benson (1982). Because m can be estimated as a U-statistic based on pairs of points regardless of the dimensionality of the data, this procedure is no more complicated in higher dimensions than it is in two dimensions. The doubly repeated median of the ratios of the lengths of homologous line segments is

$$(2.10) \qquad m = \underset{i}{\text{median}} \left(\underset{j \neq i}{\text{median}} \frac{\|U_j - U_i\|}{\|X_j - X_i\|} \right)$$

The breakdown and resistance properties of repeated median procedures as outlined in Siegel and Benson (1982) and in Siegel (1982a) still hold with these procedures: the breakdown value is approximately 50%. In particular, if more than $(n+2)/2$ of the points can be fitted closely, then this repeated median procedure will do so. If a single overall median is used instead, then the breakdown value is approximately 21% (this is $1-.5^{1/3}$), indicating that the overall median technique may not indicate clearly a localized distortion involving more than one fifth of the points.

3. COMPARING PROTEIN MOLECULAR STRUCTURES

The determination and comparison of the three-dimensional configuration of protein molecules provides a setting for illustration of the repeated median fitting method and how it relates to least squares. Dower (1979) studied the structure of "the Fv fragment of protein 315, a Dnp-binding BALB/c mouse IgA(λ_2) myeloma protein." Dower started with a predicted structure based on previous studies of related proteins. This initial configuration was modified and refined until nuclear magnetic resonance properties computed for the modified structure matched laboratory data from the protein fragment itself. The comparison of the initial predicted configuration to the final refined structure is of interest, and Dower used least squares techniques as a basis for interpreting the differences.

We fitted all 50 points of the two homologous protein molecules, each point being the center of the alpha carbon atom of an amino acid in the protein chain. Rotation and translation were allowed, but no magnification was fitted due to the nature of this problem. After fitting, residual vectors were computed, representing the direction and amount of shape change or distortion which would be necessary at each point to deform one shape into the other.

Histograms of the lengths of these residuals are shown in Figure 3.1, both for the least squares and for the robust fitting methods. The expected relationship between least squares and robust methods is evident; the robust method can tolerate a few larger residuals in order to achieve a closer fit elsewhere thereby resulting in more small residuals than least squares could achieve.

Figure 3.2 shows a plot of the least squares residuals against the repeated median residuals, allowing us to see how the residual sizes have changed on an individual basis with the 45 degree line indicated for reference. This overall picture shows that the fitting methods agree on the identification of the largest two residuals, although they do not identify the same point as the third largest.

The 50 amino acids are classified in Dower as belonging in six distinct subgroups. Because the two largest residuals both belong to the sixth group, this was examined separately. Table 3.1 lists the coordinate and residual data for this group analyzed by itself. Figure 3.3 displays the residual lengths for this subgroup under the two fitting methods. By reference to the 45 degree line, we can see that all but one residual has been reduced by the robust fit as compared-to least squares.

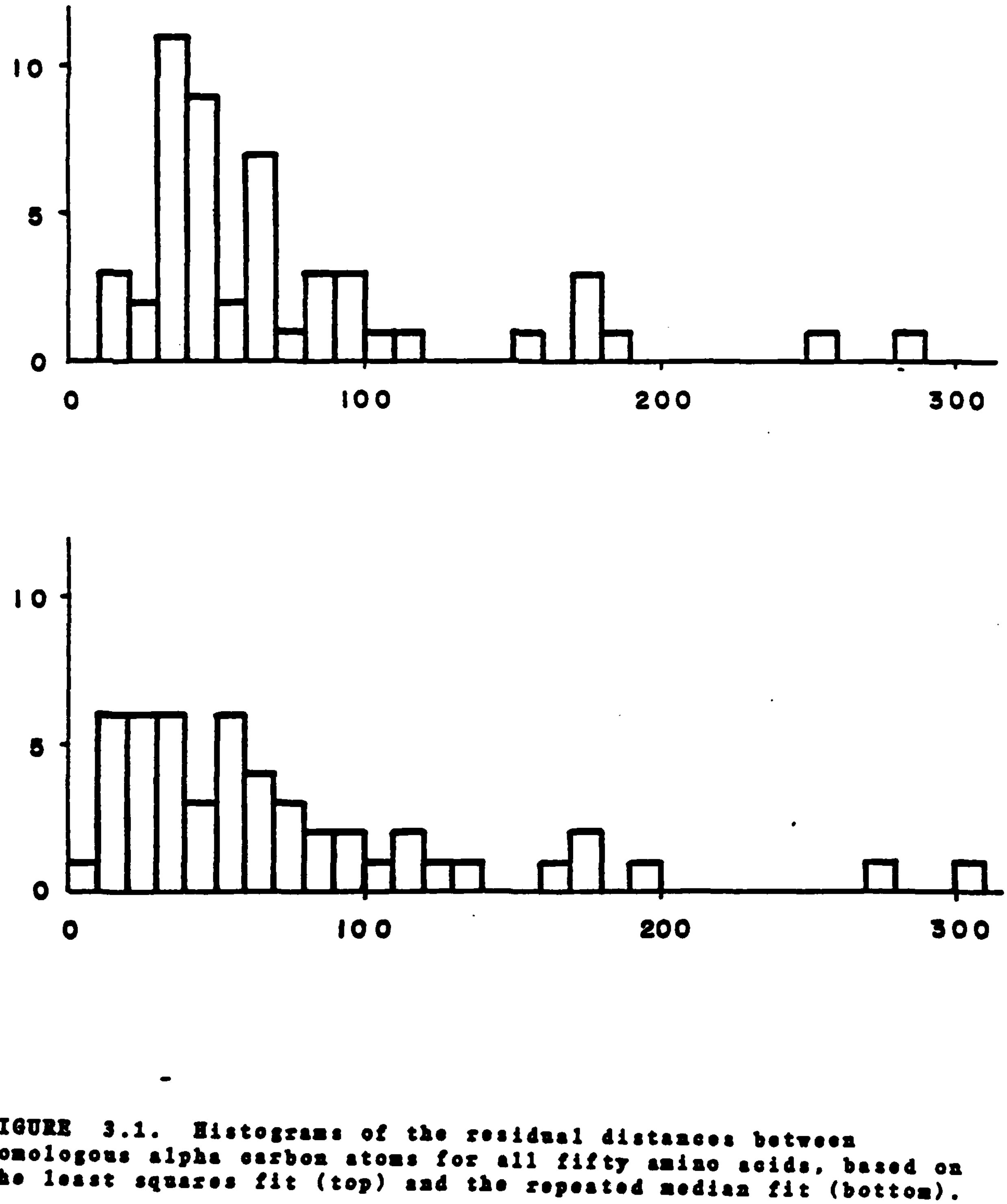

FIGURE 3.1. Histograms of the residual distances between homologous alpha carbon atoms for all fifty amino acids, based on the least squares fit (top) and the repeated median fit (bottom).

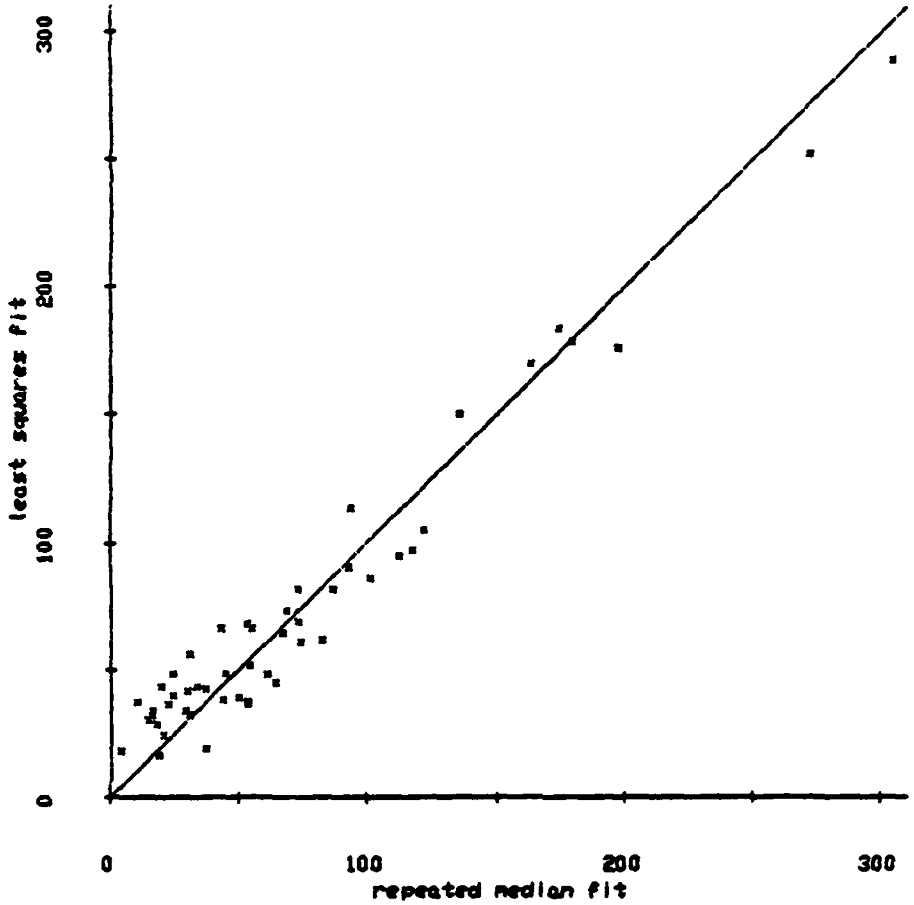

FIGURE 3.2. Least-squares residuals plotted against repeated median residuals for all fifty amino acids. Note that small residuals are primarily above the 45 degree line, indicating that repeated medians have achieved a closer fit in these areas.

TABLE 3.1. Cartesian coordinates of the alpha carbon atoms of the six amino acids of group 6, and the orthogonal matrix 0 estimated by repeated medians.

Initial configuration, centered at the origin	Modified by Dower to match nuclear magnetic resonance data, after least squares fit
$U_1' = (297, -225, 364)$	$X_1' = (276, -122, 347)$
$U_2' = (220, -195, -4)$	$X_2' = (158, -221, -3)$
$U_3' = (224, 32, -309)$	$X_3' = (204, -10, -316)$
$U_4' = (-145, 107, -377)$	$X_4' = (-163, 92, -337)$
$U_5' = (-294, 101, -23)$	$X_5' = (-149, 229, 17)$
$U_6' = (-304, 180, 351)$	$X_6' = (-326, 32, 290)$

$$0 = \begin{bmatrix} .98 & -.17 & -.08 \\ .16 & .98 & -.13 \\ .10 & .12 & .99 \end{bmatrix}$$

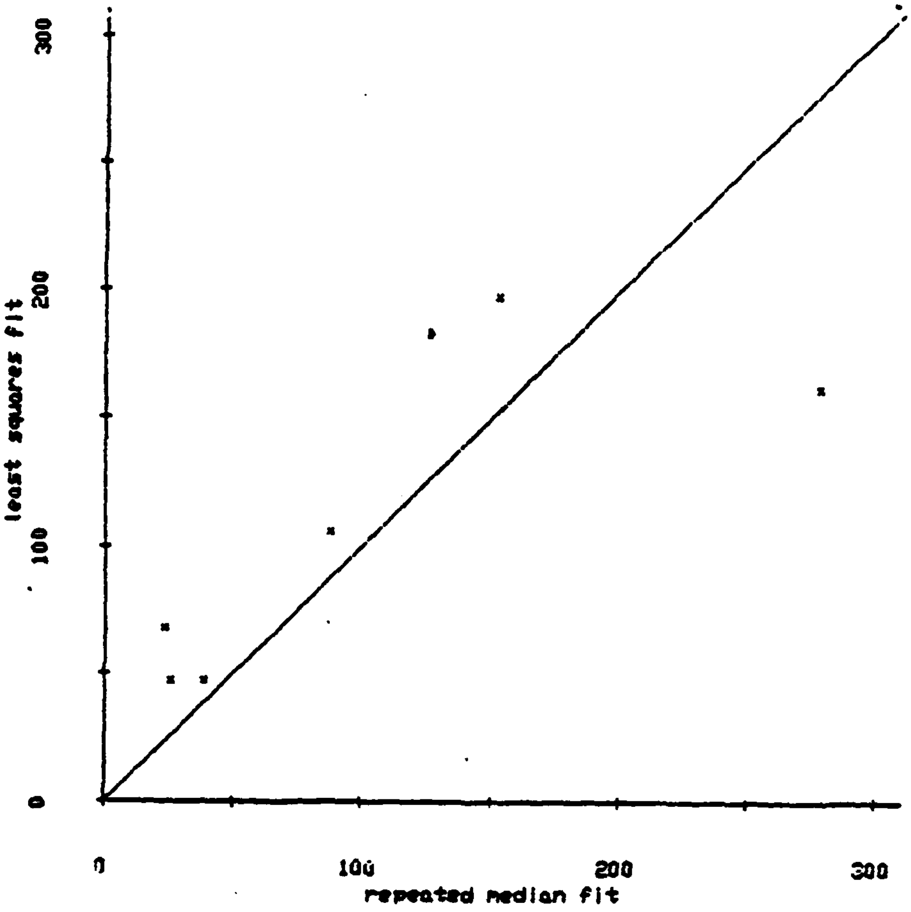

FIGURE 3.3. Least-squares residuals plotted against repeated median residuals for subgroup six fitted by itself.

The two fitting methods suggest different interpretations of the relationships of the amino acids in group six, as indicated by the histograms in Figure 3.4. Least squares suggests a continuous but skewed distribution of residuals with no clear outliers, whereas the robust fit might suggest the presence of at least one outlier. Curiously, the amino acid corresponding to the largest robust residual does not correspond to the largest least squares residual.

One interpretation of this configuration can be given if the two shapes in group six differ primarily at one point. In this case the robust fit will probably correctly identify this point by its large residual. Because the sum of squares could not be minimized in the presence of such a large residual, the least squares method would probably select a rotation that distorts the relationship among the other points while bringing the outlying points closer together.

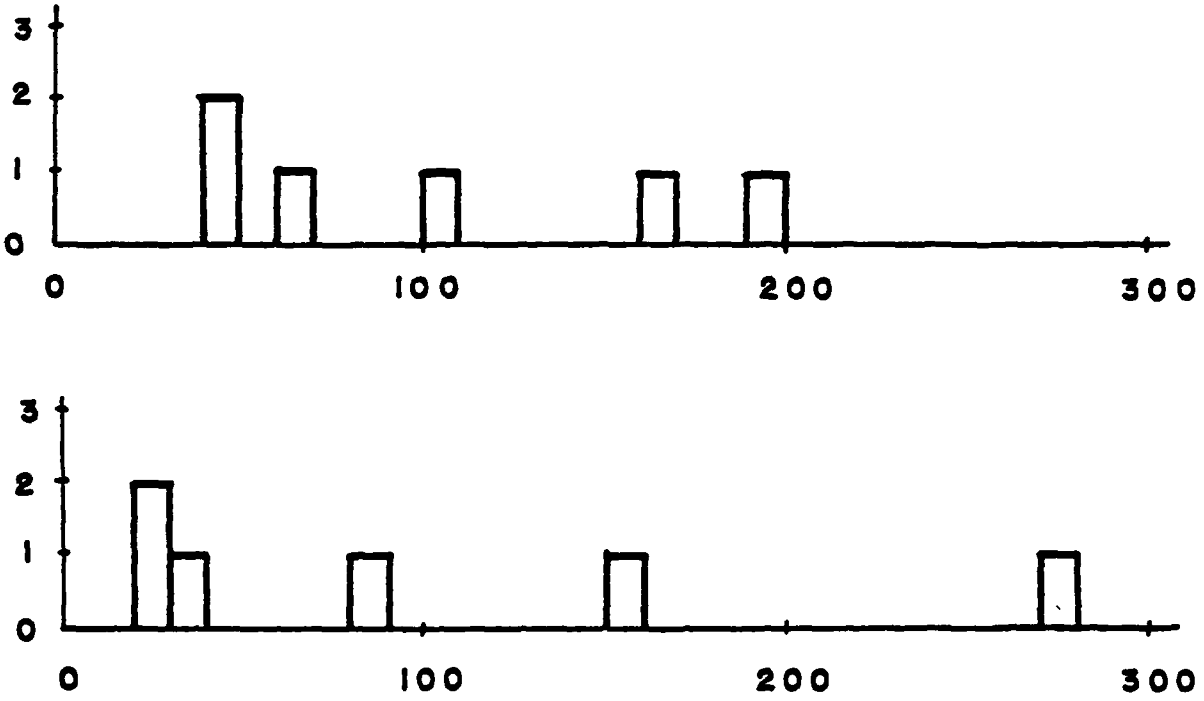

FIGURE 3.4. Histograms of the residual distances between homologous alpha carbon atoms for subgroup six fitted by itself, based on the least squares fit (top) and the repeated median fit (bottom).

REFERENCES

Bedall, F.K., and Zimmerman, H. (1979). The Mediancentre. Applied Statistics 28, 325-328.

Bookstein, F.L. (1977). The Study of Shape Transformations After D'Arcy Thompson. Mathematical Biosciences, 34, 177-219.

Dover, S.K. (1979). Structural Studies on Antibodies. Ph.D. Thesis, Worcester College, Oxford.

Eves, H. (1966). Elementary Matrix Theory. Boston: Allyn and Bacon

Gould, S.J. (1966). Allometry and Size in Ontogeny and Phylogeny. Biological Reviews, 41, 587-640.

Gower, J.C. (1975). Generalized Procrustes Analysis. Psychometrika, 40, 33-52.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematical Statistics, 19, 293-325.

Huber, P.J. (1980). Comparison of Point Configurations. Technical Report PJH-1, Department of Statistics, Harvard University.

Mosimann, J.E. (1970). Allometry: Size and Shape Variables with Characterizations of the Lognormal and Generalized Gamma Distributions. Journal of the American Statistical Association, 65, 928-945.

Olshan, A.F., Siegel, A.F., and Swindler, D.R. (1982). Robust and Least Squares Orthogonal Mapping: Methods for the Study of Cephalofacial Form and Growth. Technical Report No. 222, Series 2, Department of Statistics, Princeton University.

Siegel, A.F. (1982a). Robust Regression by Repeated Medians. Biometrika, in press.

Siegel, A.F. (1982b). Geometric Data Analysis: an Interactive Graphics Program for Shape Comparison. In Modern Data Analysis (Launer, R.L. and Siegel, A.F., editors), New York, Academic Press, in press.

Siegel, A.F., and Benson, R.H. (1982). A Robust Comparison of Biological Shapes. Biometrics, in press.

Sneath, P.H.A. (1967). Trend Surface Analysis of Transformation Grids. Journal of Zoology, Proceedings of the Zoological Society of London, 151, 65-122.

Thompson, D.W. (1917). On Growth and Form. Cambridge University Press.

DATE
ILME