Molecular recognition: identification of local minima for matching in rotational 3-space by cluster analysis

P.M. Dean and P. Callow*

Department of Pharmacology, University of Cambridge, Hills Road, Cambridge CB2 2QD, UK
*Computer Laboratory, Corn Exchange Street, Cambridge CB2 3QG, UK

This paper outlines a method for identifying discrete structural matches between molecular parameter surfaces by cluster analysis. The method is an integral part of a minimization search procedure for surface comparisons between dissimilar molecules described in the preceding paper. Priority is given in this study to the discovery of orientations having parameter matches near to the global minimum in dissimilarity.

Keywords: pattern-matching, cluster analysis, accessible surface, tetrodotoxin, saxitoxin

Received 1 July 1987
Accepted 14 July 1987

INTRODUCTION

The search for structural matches between molecules belonging to a congeneric series is usually carried out by superimposing common structural moieties and optimizing the fit between specified atom positions. Molecules compared in this manner have readily identifiable structural fragments in common. A completely different problem arises if the two molecules being compared bear *no* structural resemblance. Furthermore, if some other parameter needs to be matched, such as the shape of the accessible surface or the molecular electrostatic potential computed on a three-dimensional surface surrounding the molecules, then exhaustive search procedures are needed to scan the rotational *n*-space for orientations leading to a good match. An optimization method to perform the search in 3-space has been described in the accompanying paper.[1]

Steepest descent minimization procedures find the nearest local minimum; in this paper the minimum corresponds to a pattern match. One usually wants to find the global minimum, although in drug research, minima near to the global minimum value could indicate important comparative orientations for quantitative-structure activity studies. This paper tackles the problem of how to identify numerous local minima near to the global minimum. In other words, we strive to answer the question: are there closely equivalent molecular matches that have distinctly different orientations?

The procedure outlined here is an application of cluster analysis to minimization data that have been modified to reduce the fuzzy nature of the distribution; the objective is to extract discrete clusters that lead to the identification of matched molecular orientations. This technique is used as a step in the minimization search for matching. Densely clustered points from a brief minimization step (level 1) are approximated by the position of the local minimum; this minimum is then used as input in a further extensive minimizaton step (level 2) to define accurately the positions of local minima. An agglomerative, single-linkage clustering method has been chosen, since examination of scattergrams revealed nonspherical clusters elongated along the minimization trajectories.

METHODS AND RESULTS

The molecules selected for comparison are the neurotoxins tetrodotoxin and saxitoxin. This account is confined to a description of the method of searching for matches between the accessible surfaces. It is equally applicable to searching any other parameter surface and has been used elsewhere to search the molecular electrostatic potential surfaces. Minimization data were provided from the study described in the previous paper.[1]

Molecule A (tetrodotoxin) is kept in a fixed orientation, and the accessible surface is computed by gnomonic projection of a hemisphere[2]; 90 points in a semiregular distribution are used to compute the surface. Molecule B (saxitoxin) is rotated around its center of mass, and its accessible surface is calculated by the same procedure as that for molecule A. Both gnomonically projected faces are compared by computing the residuals between them. Molecule B is then rotated around the Euler angles x_1, x_2, x_3 to minimize the residuals. A uniformly distributed random set of starting angles is used to orientate molecule B; these values form the input for the minimization algorithm NAG E04JBF.[3] After a brief period of minimization (level 1), limited to 40

 0263-7855/87/030159-06 $03.00 © 1987 Butterworth Publishers

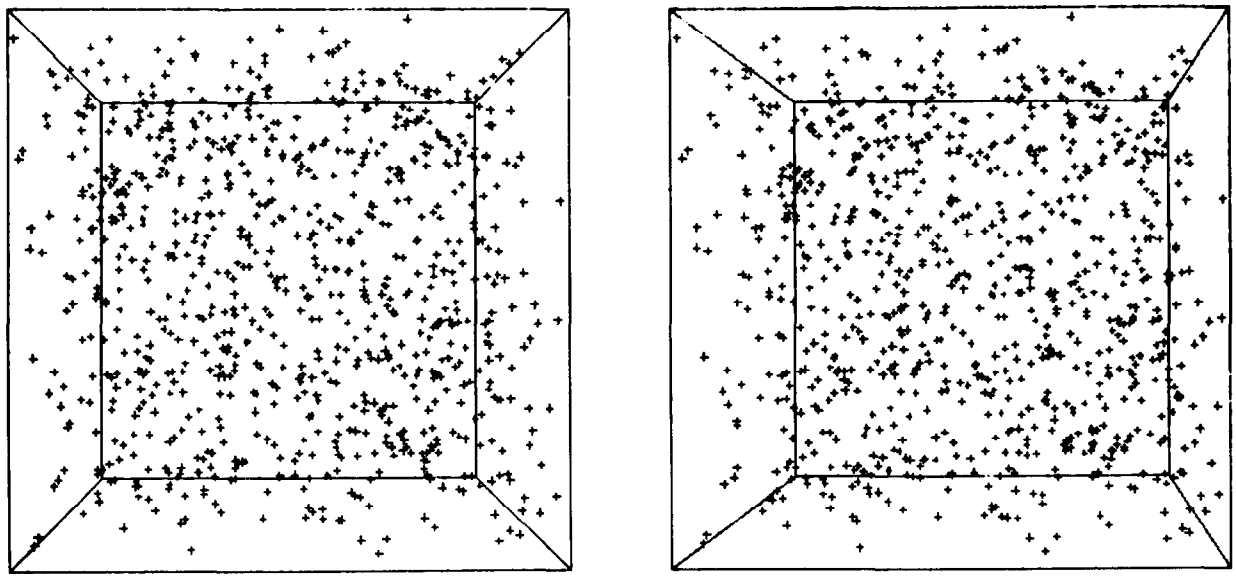

Figure 1. A stereo scattergram of 831 randomly distributed initial positions for minimization. The axes of the block represent the Euler angles x_1,x_2,x_3. The dimensions of the block are 2π for x_1 and x_2 in the plane of the paper, and π for x_3 normal to this plane

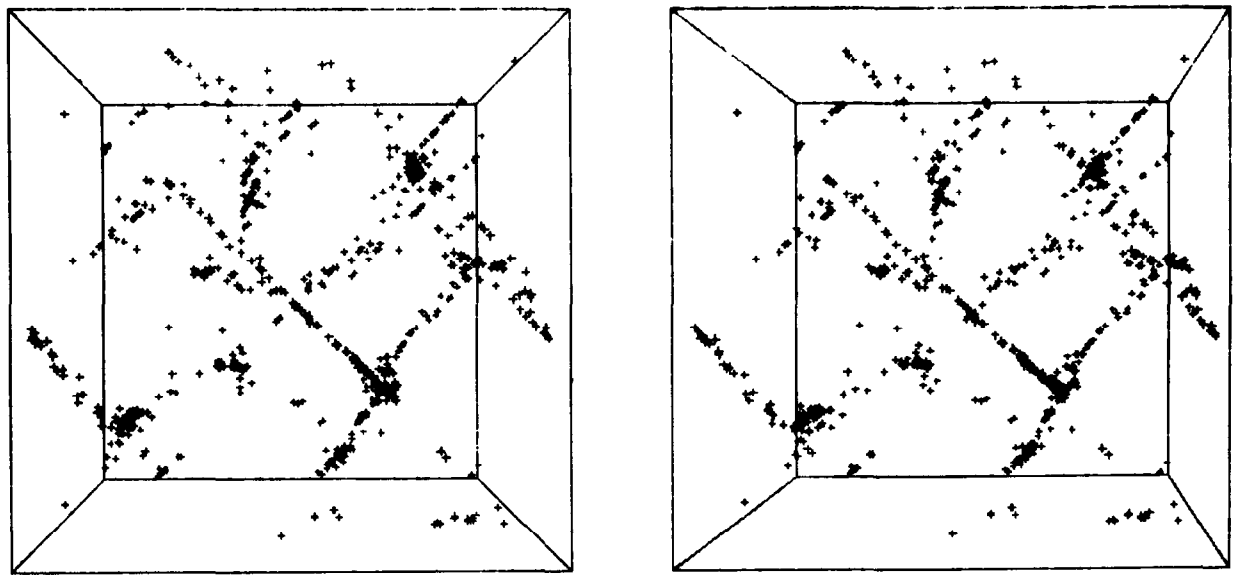

Figure 2. Stereo scattergram of the distribution of end-points after level 1 optimization (831 points)

function calls, the algorithm is stopped and the generated Euler angles are written as output.

Figure 1 is a stereo scattergram of randomly distributed points between 0 and 2π for coordinate positions x_1, x_2 and 0 to π for x_3. These positions are used as input to the first minimization step. After minimization, at level 1, the distribution of 831 points is shown in Figure 2. The distribution is no longer random; the points form clusters. Many of the clusters appear elongated and are not clearly defined. Careful scrutiny of Figure 2 in stereo reveals that the great majority of points lie approximately on a number of distinct curvilinear pathways. This observation indicates that in minimization, the algorithm rapidly moves the points into a valley in 3-space before moving them along a complicated curved trajectory toward the nearest local minimum. Cluster methods that assume spherical agglomerations of points are unlikely to be applicable in this case; single-linkage nearest-neighbor methods are more appropriate, given that the structures are both chainlike and well isolated.

Filtering noise from the data

Data points clustered in the neighborhood of a local minimum will have lower residual values than those found far from a feasible minimum. The histogram in Figure 3 shows the distribution of residuals associated with the data points in Figure 2. If a cut-off value for the residuals, R_p, is chosen close to the peak in the distribution, say 47 Å^2, so that only those points with residuals $R(x_1,x_2,x_3) \leqslant R_p$ are included, the scattergram is changed substantially (Figure 4). Most of the noise, including much of that along the trajectory lines, is removed. The fuzzy appearance of the data in Figure 2 is lost, and the remaining 236 points appear to be

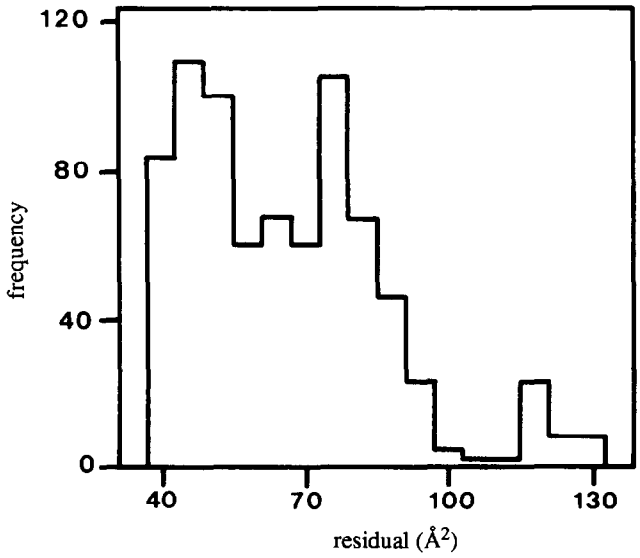

Figure 3. Histogram of the frequency distribution of residuals in the match between tetrodotoxin and saxitoxin at level 1 optimization (831 points)

clustered tightly in discrete regions. Filtration has reduced the number of data points by 70%. All points with these low residuals are used as input to the cluster routine.

Agglomerative single-linkage clustering

This method operates on a distance matrix, **D**, containing $n \times n$ elements, d_{ij}, between the points i and j with coordinates x_{i1}, x_{i2}, x_{i3} and x_{j1}, x_{j2}, x_{j3}. The distance d_{ij} is obtained from the Euclidean metric.

$$d_{ij} = \{ \sum_{r=1}^{3} (x_{ir} - x_{jr})^2 \}^{1/2}$$

The following algorithm for clustering is given by Mardia, Kent and Bibby.[4] Order the distance matrix elements into ascending order. Let there be $C_1, \ldots C_n$ starting clusters each containing a single point. Two steps, employing recursion, allocate points to the clusters.

1. Since after ordering, $d_{12} = \min(d_{ij})$, join the pair 1,2 to form C_2^* so that there are $C_2^*, \ldots C_n^*$ groups.
2. Obtain a new distance matrix $\mathbf{D}^* = (d_{ij}^*)$ with $(n-1) \times (n-1)$ elements, $d_{2j}^* = \min(d_{1j}, d_{2j})$ for $j = 3, \ldots n$; $d_{ij}^* = d_{ij}$ for $i,j = 3, \ldots n$. Find $\min(d_{ij}^*)$ for $i,j = 2, \ldots n$; then go to step 1.

It is computationally quicker to use the squared Euclidean distance d_{ij}^2 rather than the distance d_{ij}; the results are unaffected by the change. On completion of the algorithm we are left with a single cluster as all the clusters are successively agglomerated. Clusters can be identified by inspection of the dendrogram, or a specified number of clusters, g, can be obtained and their cluster membership defined.

Cluster analysis

Cluster analysis on the reduced number of data points is carried out by the SPSSx[5] package with the number of desired clusters set to 20; the results are shown in Table 1. This number of clusters appears to be too many; nine clusters contain only a single point. The first eight clusters contain 92% of the points. However, since our problem is not to define the number of clusters exactly, these excess clusters may be retained for further minimization. The standard deviation of the residuals in each cluster is small, less than 5% of the mean value.

The structure of each cluster, containing 12 or more points listed in Table 1, was subjected to a principal components analysis. The cumulative percentage of the variance of each principal component is given in Table 2. More than 66% of the variance is accounted for by the first component. Inclusion of the second component accounts for more than 98% of the variance in the data for four of the five clusters. This analysis indicates that the clusters are composed predominantly of points distributed in a local metric plane. Careful examination of the stereo scattergrams of Figures 2 and 4 reveals that the clusters are found at the confluence of two nearly linear trajectories. A scattergram of points for the two

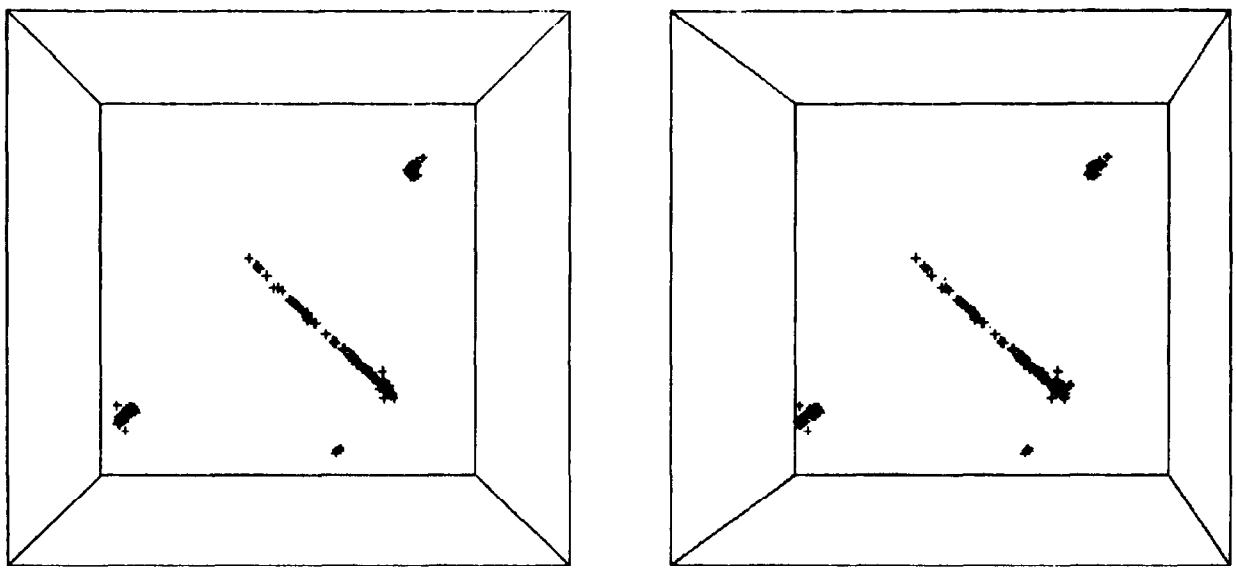

Figure 4. Stereo scattergram of the data from Figure 2 after removal of points whose residual is > 47Å² (236 points)

Table 1. Cluster data after level 1 optimization (236 data points)

Cluster number	Residual (Å^2)	Points in cluster	Cumulative percentage	Rank order of residual
1	34.1	78	33.1	1
2	46.2	12	38.1	18
3	38.9	43	56.4	4
4	37.3	13	61.9	2
5	45.2	2	62.7	16
6	39.1	61	88.6	5
7	46.5	6	91.1	20
8	45.0	2	91.9	15
9	42.6	1	92.4	9
10	44.5	1	92.8	13
11	44.1	1	93.2	12
12	40.0	3	94.5	6
13	43.3	3	95.8	10
14	40.4	1	96.2	7
15	41.7	1	96.6	8
16	43.7	1	97.0	11
17	46.3	1	97.5	19
18	44.8	4	99.2	14
19	45.6	1	99.6	17
20	37.6	1	100.0	3

The accessible surface of saxitoxin is rotated against a fixed accessible surface face of tetrodotoxin. The cluster number is given for identification purposes in subsequent tables and figures; it has no special significance on its own

principal components of cluster 1 plotted against their residuals is shown in Figure 5. The points lie on the surface of a conventional well with a concentration of points having low residuals distributed at the bottom of the well marked **a**; a minor well, **b**, can also be discerned with a larger residual than the local minimum **a**. This scatter conforms to what is often expected from a truncated minimization close to a feasible region.

The minimum position in each cluster can be calculated readily; these 20 positions are displayed in Figure 6. Each position is then used as input for the second level of optimization with 2000 function calls. Positions of the final end-points are drawn in Figure 7 and the Euler angles with the associated residuals are given in Table 3. With 2000 function calls the algorithm failed to find satisfactory stable minimum positions for nine of the 20 clusters. Eight discrete local minima were generated for matched rotations of saxitoxin against tetrodotoxin. Comparison between Figures 6 and 7, taken with data from Table 3, shows that the final positions of the minima fall into three groups. A large group with similar Euler angles is formed from cluster numbers 1, 4, 10, 12, 13, 14, 15, and 20; this group contains the global minimum and positions with residuals less than 35 Å^2. Clusters 8 and 18 form a second group; cluster 16 suggests the presence of a third rotational position. The predominant effect of optimization at level 2 has been to condense the string of clusters 19, 13, 10, 12, 4, 14, 20, 15 found at level 1 optimization.

DISCUSSION

Cluster analysis has been employed successfully to identify minima in rotational space for the pattern match between two dissimilar molecules. When sandwiched in a two-level optimization process, cluster analysis can lead to a massive reduction in data needing to be minimized. In the example shown here, only 20 points near to the global minimum value were finally minimized, although the initial search started from a larger number (831) of positions. Eight separate discrete minima were located and suggest a number of good matches between the fixed face of tetrodotoxin and a freely rotating molecule of saxitoxin. Of the 20 feasible regions studied, 11 converged to an acceptable minimum at level 2 optimization. The remainder would require further minimization to assess whether they are true local minima or saddlepoint regions.

The appropriate choice of a particular cluster procedure is dependent on the distribution of data to be examined and the objective of the clustering exercise. Hierarchical techniques, of which agglomerative single-linkage is a popular method, eventually reduce the data into a single cluster. The user has to decide how many clusters are needed. Furthermore, there is no procedure for reallocating a point to another group. In single-linkage methods points are frequently chained; in the case of tetrodotoxin and saxitoxin this property was advantageous, but in others it could obscure the presence of poorly separated spherical clusters for which other cluster techniques would be more effective.

The problem of deciding how many clusters, g, are present in a distribution is a complex one, for which no general and reliable answer is available. A possible practical solution has been employed here by using a value for g large enough to produce a number of clusters each containing only a single point. The rationale has been to attempt to overdetermine the number of clusters; the approach may not be successful if the volume occupied by the domain around a minimum is smaller than the mean volume associated with each starting point. In fact, there is evidence from our data that this approach performs weakly; cluster numbers 10, 14, 15 and 20 contain only one point, but their residual values lie close to the global minimum.

Clusters shown in Figure 2 have a pronounced fuzzy nature; there is no clear demarcation between them. Rigorous methods for the allocation of points in fuzzy regions to a particular cluster can be carried out by computing a membership function, which denotes a degree of belonging to a particular cluster. Various fuzzy clustering techniques are discussed by Gordon.[6] Our method of handling the fuzzy clusters has been more drastic and is linked directly to the nature of the problem and the desired solution. In the neighborhood of a cluster minimum, and along the line of the local trajectory, the residual values for each point are determined by the topography of the rotational surface. We wished only to consider strong matches between the molecular surfaces that correspond to deep minima in the minimization. If a search starts from a random distribution, the probability of a point being near to an acceptable minimum is an inverse function of its associated residual value at the end of the minimization step. Points with large residuals can thus be discarded. This simple inequality $R(x_1,x_2,x_3) \leqslant R_p$ can therefore be used as a hard partition of the data.

The cluster method described here is valuable for finding numerous possible matches between molecular surface parameters. At the same time it can be employed to improve substantially the search for the best match

Table 2. Principal components analysis of the distribution of points within clusters after filtration at level 1 optimization

Cluster number	Points in cluster	Percentage of variance of principal components			Cumulative percentage of the two principal components
		1	2	3	
1	78	86.6	13.3	0.1	99.9
2	12	78.4	20.0	1.6	98.4
3	43	91.2	6.9	1.9	98.1
4	13	71.5	27.8	0.7	99.3
6	61	66.5	23.3	10.2	89.8

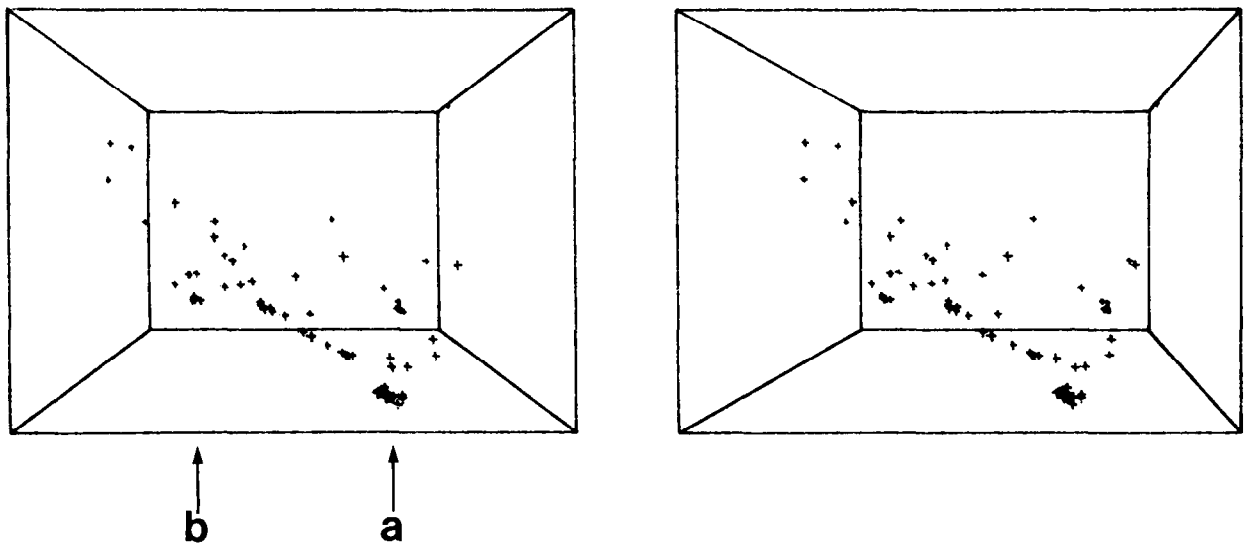

Figure 5. Cluster number 1; stereo scattergram of the residuals (scaled between the minimum and the maximum on the vertical y-axis) plotted against the two principal components (scaled on the x- and z-axes). In this case the size of the box is determined by the limits of the cluster itself. The local minimum is marked **a**, *a minor well is found at* **b**

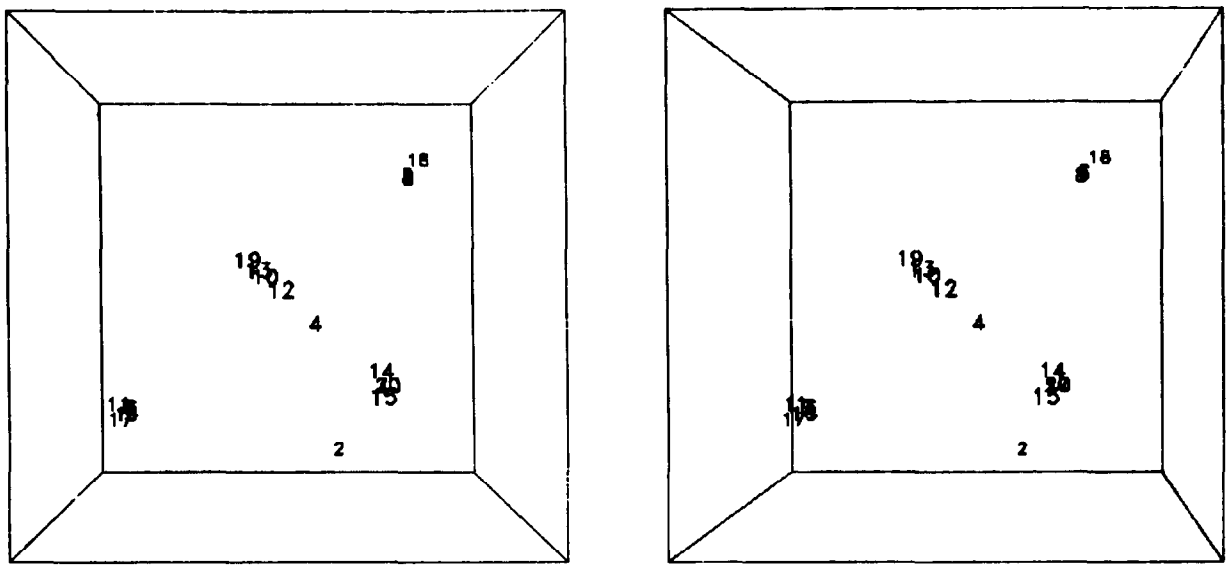

Figure 6. A stereo drawing of the positions of the minima of 20 clusters taken from Figure 4. The numbers plotted represent the cluster number

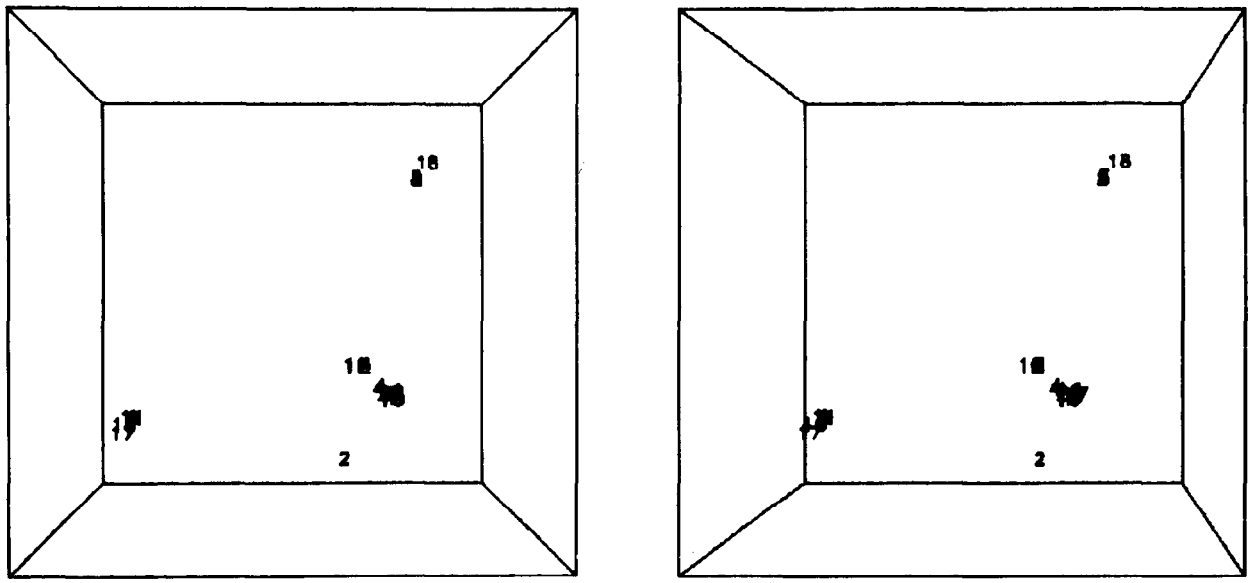

Figure 7. Final positions of the local minima after level 2 optimization

Table 3. Positions of discrete local minima after level 2 minimization

Cluster number	Euler angles x_1	x_2	x_3	Residual (Å^2)	Order of local minimum
1	4.32	1.92	0.56	34.09	1
4	4.17	2.06	0.49	34.64	4
8	4.82	4.73	1.45	40.07	6
10	3.89	2.34	0.42	34.85	5
12	3.89	2.34	0.42	34.85	5
13	3.89	2.34	0.42	34.85	5
14	4.33	1.97	0.78	34.45	2
15	4.32	1.92	0.56	34.09	1
16	0.78	1.35	1.63	41.48	7
18	5.11	5.11	2.05	44.82	8
20	4.34	1.95	0.75	34.48	3

in an optimized search procedure. It should prove useful in the search for similarities in a non-congeneric series of molecules believed to have certain ligand surface characteristics in common. The clustering procedure is not limited to three-dimensional problems; higher dimensional blind-searching in rotational 6-space, to reveal matched surface patterns, could become possible with this method.

ACKNOWLEDGEMENT

P.M.D. wishes to thank the Wellcome Trust for continued financial support through the Senior Lectureship Scheme.

REFERENCES

1 Dean, P. M. and Chau, P.–L. Molecular recognition: optimized searching through rotational 3-space for pattern matches on molecular surfaces. *J. Mol. Graph.* 1987, **5**, 152–158
2 Chau, P.–L. and Dean, P. M. Molecular recognition: 3D surface structure comparison by gnomonic projection. *J. Mol. Graph.* 1987, **5**, 97–100
3 EO4JBF: a comprehensive quasi-Newton algorithm for finding an unconstrained minimum of several variables. Numerical Algorithms Library Routine
4 Mardia, K. V., Kent, J. T. and Bibby, J. M. *Multivariate analysis.* Academic Press, London (1979)
5 SPSSx SPSS inc. McGraw-Hill Company, New York (1985)
6 Gordon, A. D. *Classification.* Chapman and Hall, London (1981)