Journal of Computer-Aided Molecular Design, 5 (1991) 119–133
ESCOM

J-CAMD 113

Molecular structure matching by simulated annealing. IV. Classification of atom correspondences in sets of dissimilar molecules

M.C. Papadopoulos and P.M. Dean*

Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.

Received 1 August 1990
Accepted 3 October 1990

Key words: Simulated annealing; Molecular matching; Classification theory

SUMMARY

A set of 6 molecules, active at the benzodiazepine $GABA_A$ site are matched pairwise with one member of the set in turn. Matchings are performed by simulated annealing using null correspondences to reject poorly matched atom positions. Cluster analysis is employed to identify molecular similarities after an optimal molecular superimposition has been discovered. A statistic for the compactness of clustered atom positions is suggested. The introduction of null correspondences causes the clusters of matched atoms to become more compact.

INTRODUCTION

Over recent years the measurement of molecular similarity has become an intensely studied problem and a monograph has been devoted to this rapidly developing field [1]. An expression of molecular similarity is crucial for attempts to derive a 3D QSAR which is able to incorporate some structural dissimilarity. This notion may, at first sight, appear contradictory but there is increasing evidence that the activity of a set of molecules is more related to a three-dimensional pharmacophore and may be less dependent on the structures having common molecular subgraphs. It appears that the spatial disposition of atom properties is the key factor affecting affinity and activity. Drug design may therefore be viewed as a combinatorial problem of structure generation within the three-dimensional constraints of the site atoms. Where the structure of the binding site is known it may be possible to automate parts of the design process [2–5]. However, where the site is unknown less direct methods have to be tried in order to glean structural information about the site. It is these methods which are heavily dependent on precisely how molecular similarity is measured and subsequently searched for.

* To whom correspondence should be addressed.

Similarity may be measured at the molecular surface so that comparisons can be made between van der Waals or accessible surfaces [6,7], or between molecular parameters, such as the electrostatic potential or hydrophobicity, projected on to a surface. Similarity is measured using correlation theory. The intrinsic difficulty here lies in the pivotal problem of how to select the correct superimposition of the molecules; unbiased searching methods are needed [8,9]. For isosteric structures this problem does not occur since the atom assignments are known; the electron density functions can be correlated when the atomic nuclei have been superimposed [10].

Atom correspondences need to be assigned before similarity can be assessed. The assignment process is combinatorial and can be solved by simulated annealing [11,12]. Simulated annealing is an optimization procedure for discrete problems and includes a probabilistic hill-climbing feature to improve the search for the global minimum. Atom correspondences can be found in a pair of molecules A and B; N_A and N_B are the numbers of atoms in A and B, respectively and A and B

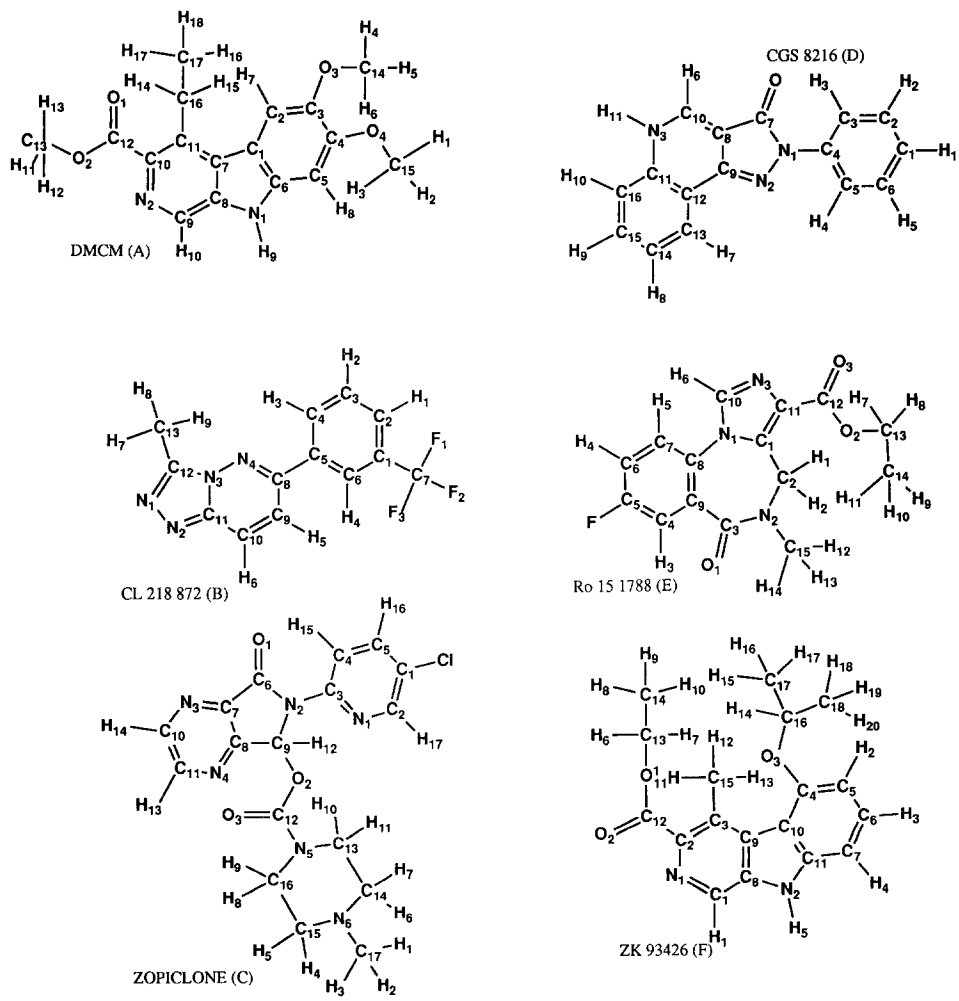


Fig. 1. Molecular structures used in this molecular matching study employing simulated annealing with null correspondences.

are chosen such that $N_A \leq N_B$. Matches can be sought for molecule B if a set of atoms in A is specified; B is searched combinatorially by optimizing the difference-distance matrix for A,B. The recent introduction of null correspondences into the objective function enables good matches to be found between unspecified subsets of atoms in either molecule [13]. For example, suppose that we have two molecules containing 40 and 45 atoms, it is now possible to answer the question: which is the subset of 30 atoms in each molecule that would show the best superimposition? The null correspondences method allows poorly matched atoms to be ignored.

The molecular superposition problem can be divided into two parts. Firstly, if there are n molecules in a set of dissimilar structures it may be of practical importance in drug design to make all pairwise comparisons with a chosen reference molecule. This would provide information about subsets of atoms in the set of molecules which have distinct relationships to a single unmasked subset in the chosen molecule. This information is frequently desired in drug design and would require a separate classification of atom correspondences for each reference molecule. Secondly, a more general question can be asked: is there a subset of atoms in the set of molecules which shows an overall positional similarity? These two parts to the superimposition problem require quite different strategies for their solution. The second problem, not dealt with in this paper, would require the simultaneous annealing of all n molecules and consequently would need an objective function to encompass all possible difference-distance matrices.

In this paper we seek to extend the null correspondence work to investigate the first part of the molecular superimposition problem by using the crystal structures of 6 drugs that bind to the benzodiazepine $GABA_A$ receptor. These molecules have been selected primarily on the grounds of crystal structure availability and for the fact that they exhibit a span of pharmacological activity from agonists, through antagonists to inverse agonists. However, it must be stressed that we do not wish to make any comments here about structure and activity; in this work the molecules selected are used only to test the methodology. Classification theory has been used to identify clusters of commonalities in atom positions. The simulated annealing algorithm used is identical to that described in the previous paper [13].

METHODS

Data

Cartesian coordinate data was taken from the following molecules: DMCM, a β-carboline [14]; CL 218872, a triazolopyridazine [14]; zopiclone, a cyclopyrralone [14]; CGS 8216, a pyrazoloquinoline [14]; Ro 151788, an imidazodiazepine [15]; and ZK 93426, a β-carboline, using an energy-minimized structure based on Ref. 16. Where hydrogen atoms were undefined in the crystal structure, they were added using standard geometries. No rotations were performed round any torsion angles. The molecular structures and numbering schemes are drawn in Fig. 1.

Simulated annealing

Pairwise molecular matching was carried out between the 6 molecules giving 15 pairs in all. A and B were chosen such that $N_A \leq N_B$. Each pair of structures was allowed to anneal using the algorithm described in the previous paper [13]. In the annealing schedule the following parameters were initialized, $T = 2$, $C = 4$ and $\delta = 0.1$. Each annealing run contained 4 annealing attempts; an initial annealing stage followed by a re-annealing step to improve the worst 25% of atom assign-

ments, and a reverse re-anneal stage to reduce possible landscape problems followed by a re-anneal on the worst 25% of assignments. Reversed and re-anneal schedules were performed with T reset to 1.5. The annealing runs were repeated 10 times using a non-repeatable random number generator for each pair of molecules and for each number of nulls. The number of nulls was progressively increased; Table 1 shows the matchings together with the number of nulls used. Nulls were added so that the number of matched atoms were multiples of 5.

Superimposition of molecule pairs

For each annealing run a set of matched atom assignments was produced. Molecules A and B were then superimposed using McLachlan's algorithm [17]. The transformation matrix is composed of 12 elements, 9 for the rotation and 3 for the translation elements. The matrices of all 10 runs could be compared to assess their similarity. Each of the elements was uniformly scaled to achieve equal weighting using the transformation [18]

$$c'_{ip} = (c_{ip} - m_p)/\sigma_p \quad (1)$$

where c_{ip} is the pth element of the ith matrix, m_p and σ_p are the mean and standard deviation of the pth element of all 10 matrices. The 10 scaled transformation matrices were then made into a 10×12 data matrix and examined by cluster analysis in order to discern similarities in the orientational data produced by annealing the same pair of molecules.

Classification

Classification of the molecular superimpositions has been performed, using the CLUSTAN package [19], at two stages in this work. Firstly, cluster analysis has been applied to the 10 runs from each pair of superimpositions. Secondly, atom correspondences from all 6 molecules were superimposed in turn using one molecule as the reference for each comparison. The total number of atom positions for each reference comparison was then forced into 44 clusters using the x, y and

TABLE I
MOLECULAR PAIRINGS FOR SIMULATED ANNEALING

Molecule B	Molecule A																			
	DMCM				CL 218872				CGS 8216			Ro 151788				ZK 93426				
DMCM					*0*	*4*	*9*	*14*	*0*	*6*	*11*	*0*	*6*	*11*	*16*					
					0.97	0.68	0.39	0.42	1.30	0.56	0.56	1.49	0.92	0.78	0.69					
Zopiclone	*0*	*6*	*11*	*16*	*0*	*4*	*9*	*14*	*0*	*6*	*11*	*0*	*6*	*11*	*16*	*0*	*3*	*8*	*13*	*18*
	2.17	1.71	1.36	1.25	1.51	1.16	0.53	0.46	1.53	1.43	0.83	1.62	1.18	1.08	0.71	1.96	2.09	2.09	0.88	1.08
CGS 8216					*0*	*4*	*9*	*14*												
					1.20	0.92	0.68	0.54												
Ro 151788					*0*	*4*	*9*	*14*	*0*	*6*	*11*									
					1.40	1.11	0.76	0.57	1.19	0.98	0.62									
ZK 93426	*0*	*6*	*11*	*16*	*0*	*4*	*9*	*14*	*0*	*6*	*11*	*0*	*6*	*11*	*16*					
	1.68	1.21	0.80	0.41	0.86	0.73	0.70	0.50	0.99	0.70	0.54	1.13	1.52	1.09	0.69					

Number of null correspondences shown in italic together with the best rms (Å) of the 10 runs for each matching.

z coordinates of each atom in their matched orientation. A limit of 44 clusters was set because that is the number of atoms in the largest molecule, zopiclone. A larger number of clusters would encourage some atoms not to be included in any matching cluster and therefore would be undefined outliers.

There are numerous classification methods available. In the case where 10 annealing runs were analysed, three procedures were used: single linkage, Ward's procedure and the centroid method. Each used the squared Euclidean distance metric as a measure of similarity. In the case of atom correspondence in the molecular set Ward's method was used because it maximises cluster isolation and does not violate the ultrametric inequality [20].

Cluster compactness

If in the superimposition step, all the atoms are forced into a fixed number of clusters then we need to determine if particular clusters are compact and contain atoms from separate molecules. The rms distance, P_i, of all atoms belonging to a cluster i from the centroid of i, can be used as a measure of compactness. Its value is determined from

$$P_i = (\sigma_{xi}^2 + \sigma_{yi}^2 + \sigma_{zi}^2)^{1/2} \tag{2}$$

where σ_{xi}, σ_{yi} and σ_{zi} denote the standard deviations in the x, y and z coordinates for all atoms in cluster i. The smaller the value of P_i, the greater is the compactness.

RESULTS

Effect of increasing nulls on molecular matching

The number of nulls, together with the smallest rms value of the 10 repeated runs for the superimposition of the matched atoms, for all pairwise comparisons are shown in Table 1. In general,

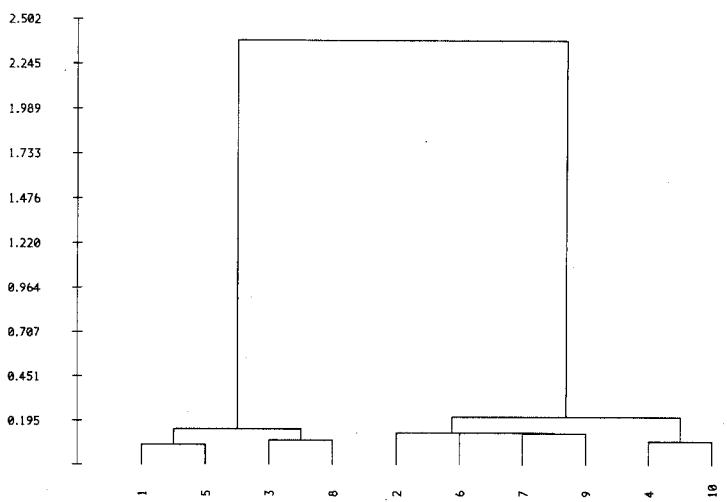


Fig. 2. Dendrogram of differences in the transformation matrices for 10 matchings of CGS 8216 against DMCM with 11 null correspondences. Fusion distances (standardized units) for the clusters are shown on the ordinate and plotted against run number on the abscissa.

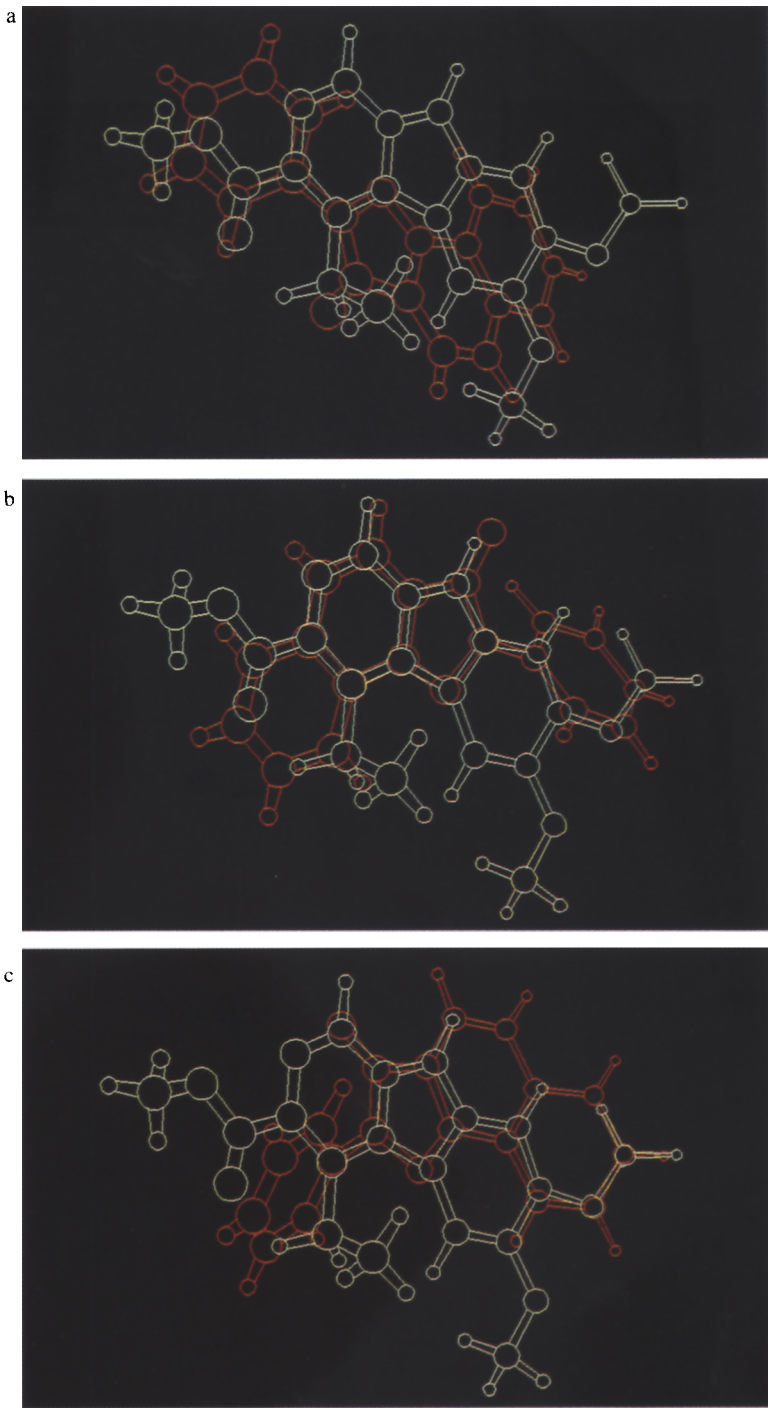


Fig. 3. Molecular superimpositions of DMCM (white) and CGS 8216 (orange). a: The best superimposition with zero nulls; b; The best superimposition with 11 nulls (run 9); c: A superimposition showing a different orientation of CGS 8216 with 11 nulls (run 5).

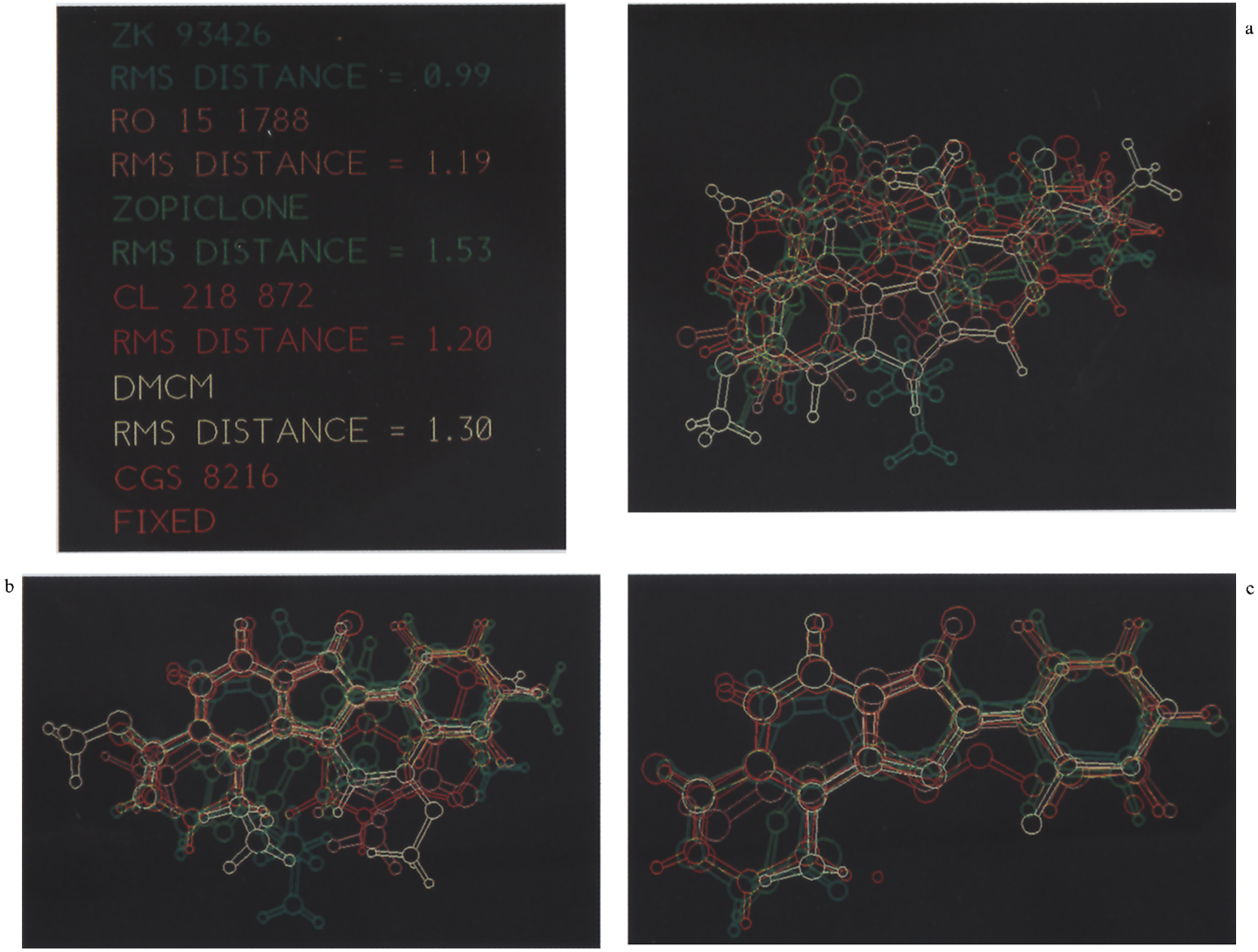


Fig. 4. Molecular superimpositions of 5 molecules (colour code given in panel) onto CGS 8216. a: Zero nulls; b: With nulls to give 20 matched atoms; c: Superimpositions of the matched atoms of the 5 molecules (null correspondences omitted) onto the complete molecule of CGS 8216.

for any pair of molecules the rms deviation between the final matched atoms decreases as the number of nulls is increased. This finding is to be expected since that is the purpose of introducing null correspondences so that poorly matched atoms can be discarded during the annealing process.

Classification of paired molecule superimpositions

The data for a single pair of molecules, CGS 8216 and DMCM, will be described in detail. Essentially similar qualitative results were obtained for the other pairs of molecules. From each annealing run of two molecules, with or without nulls, the best matched atom correspondences were generated. However, as the number of nulls is increased the landscape problem deteriorates and different orientations may occur with similar values for the final objective function. The configuration landscape can be analysed by cluster analysis of the transformation matrix for all superimpositions. The transformation matrix is obtained from the output McLachlan's algorithm [17]. The dendrogram of Fig. 2 shows the classification for 10 runs on two molecules using 11 nulls. All three classification procedures, single-linkage, Ward's method and the centroid method, gave essentially the same results with cophenetic correlation coefficients of >0.9. Two distinct clusters were revealed; runs 1, 5, 3 and 8 showed a large difference in fusion distance from the other runs and represent completely different orientations. However, the matches, although spatially distinct, have similar rms values for the superimposition; run 9 has an rms = 0.47 Å and run 5 has an rms = 0.56 Å. Figure 3a illustrates the best superimposition with zero nulls. The atoms are matched but do not appear to be well superimposed compared with Figs. 3b and c where 11 null correspondences are included. Run 9 has the orientation shown in Fig. 3b; whereas run 5 shows an orientation illustrated in Fig. 3c. The CGS 8216 molecule is rotated through 180° to give a different match but with a similar subset of atoms from DMCM.

Superimposition of 6 molecules

Illustrations are given in Fig. 4 for matches using CGS 8216 as the reference molecule. Figure 4a shows the superimpositions of the best matches for the 5 molecules onto CGS 8216 with zero nulls. With all molecules except zopiclone the ring portions lie approximately in a common plane; with zopiclone the rings lie perpendicular to this plane. The rms deviation for the matched atoms is in the range 0.99 Å and 1.53 Å. CGS 8216 has 31 atoms, if 11 nulls are introduced in all molecules except CL 218872 where 9 nulls are used, then 20 atoms are allowed to be matched in each of the 6 molecules. When nulls are introduced the rms deviation is reduced and is in the range $0.47-0.83$ Å. Figure 4b demonstrates visually that more atoms exhibit a close superimposition. Careful scrutiny of Figs. 4a and b clearly shows striking differences between the superimpositions with and without nulls. With no nulls (Fig. 4a) the atom positions lie predominantly within a triangle of space; at the periphery only a few groups are isolated. Within the matched region the atoms are poorly superimposed and the bonding networks are badly matched. Once nulls are introduced the atom matching is substantially different (Fig. 4b). Firstly, the shape of the periphery is changed, it is no longer triangular but is elongated due to some matches having different orientations. Secondly, there are more isolated groups at the periphery. Thirdly, more atoms appear to be well superimposed. Fourthly, common bonding networks appear. These last two observations are emphasised when the null correspondence atoms are omitted from the drawing in Fig. 4c. The matched atoms show tight clustering and a common composite bonding network can be discerned.

From these observations it would appear that the minimization procedure stresses the best superimposition of the extrema in data sets where nulls are not included. However, where nulls are introduced these appear to be allocated predominantly to unmatched side chains. The matched atoms then become associated with rings and a common bonding pattern is produced.

Where different reference molecules are chosen the superimpositions show differences in detail if the number of nulls is increased. An example is illustrated in Figs. 5a and b for Ro 151788 as the reference molecule. Comparison between Figs. 4b and 5b highlights the differences in the two sets of superimpositions; unmatched side chains are again found at the periphery.

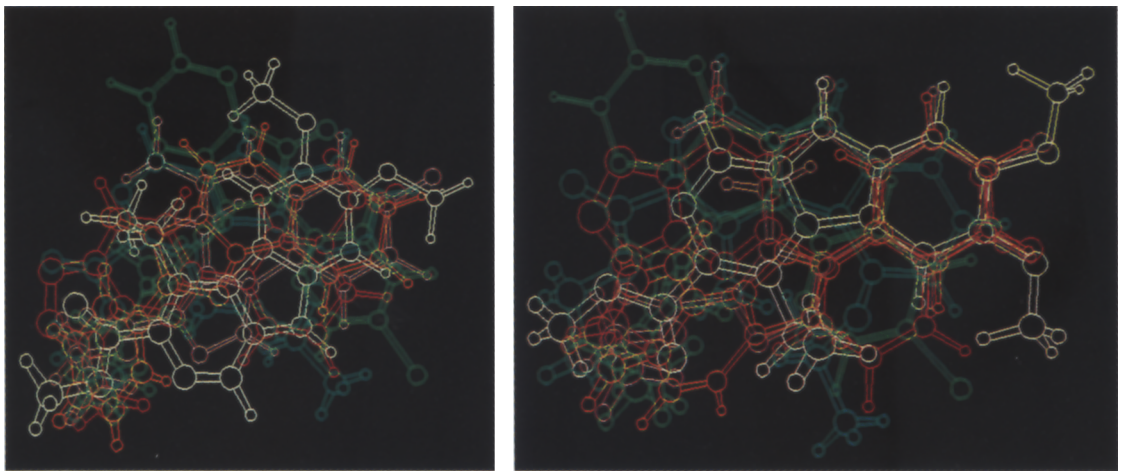
Classification of matched atoms

If we have a set of superimposed molecules, two important questions can be asked: are clusters of atoms readily distinguishable, and which atoms between the molecules lie close to each other? Cluster analysis can be used to answer both questions. Ideally a method needs to be selected which finds isolated, compact and spherical clusters. Single-linkage agglomerative methods have the drawback of chaining; median and centroid methods violate the ultrametric inequality; Ward's method of cluster analysis has been selected because it maximizes cluster compactness by minimizing the within group variance and hence maximizes cluster isolation [20]. This method will also find spherical clusters [19].

After the six molecules have been superimposed, using a particular reference molecule, the 224 atom coordinate positions are subject to cluster analysis using a limit of 44. Figures 6a and b show the atom positions projected onto the xy plane and correspond to the molecular superimpositions of Fig. 4b with CGS 8216 as the reference molecule and with nulls incorporated to give 20 matched atoms. Each cluster has been distinguished by drawing the convex hull round it. This diagram is only a two-dimensional projection so the overlapping convex hulls do not contradict the convex admissibility property of Ward's method. A measure of cluster compactness is given by Eq. 2, where P_i is the rms distance of all atoms of cluster i to the centroid of i; this can be represented as a radius and drawn as a circle round each cluster centroid on the plot. Comparison of cluster radii in Figs. 6a and b shows that compact clusters are present; they are revealed clearly as well isolated clusters when the null corresponding atoms are omitted (Fig. 6b). Table 2 provides information on the compactness of each cluster for CGS 8216 as the reference molecule with zero nulls and Table 3 gives the corresponding information where nulls are included in the annealing. In both tables the clusters are presented in decreasing order of compactness. In general, where the clusters are compact, the introduction of nulls increases the compactness. This clustering procedure also makes it possible to identify the clusters composed only of atoms from different molecules; compact clusters usually contain atoms from different molecules, whereas diffuse clusters may contain a number of atoms from the same molecule.

DISCUSSION

The measurement of molecular similarity is a complex problem and the eventual use to which similarity is put will largely determine which strategy for measurement is adopted. For example, if we wish to determine similarity of a molecular property distributed on two molecular surfaces, then all that needs to be studied is a rotation of the molecules so that the difference between the property at specified positions is minimized. This type of problem can be solved by optimization

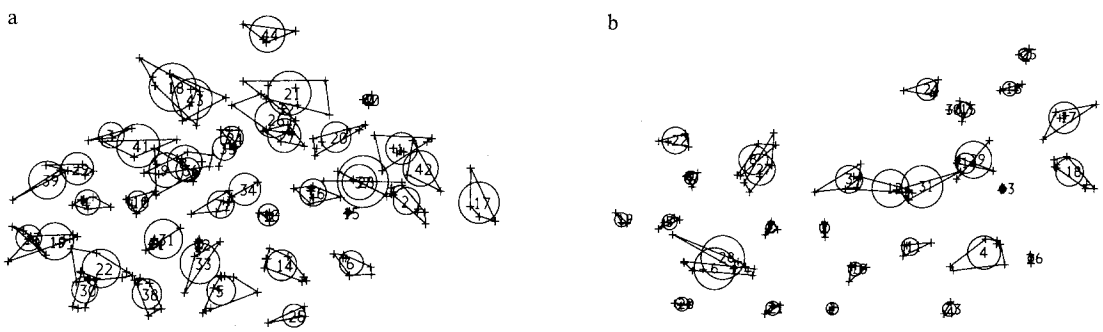


a b

Fig. 5. Molecular superimpositions of 5 molecules onto Ro 151788. Same colour code as shown in Fig. 4. a: Zero nulls; b: Null correspondences included to give 20 matched atoms.

in continuous space [8,9]. On the other hand, if atom positions are to be superimposed then optimization techniques for a discrete space are needed. Simulated annealing is ideal for discrete space optimization because it is capable of approaching close to the global minimum. However, the medicinal chemist may need to know, not whether the whole molecule is similar to another, but whether parts of the molecule show atom positional similarities. Precisely which atoms to include may also be unknown. The introduction of null correspondences into matching by simulated annealing seeks to provide answers to this problem free from any human bias. Null correspondences allow badly matched atoms to be ignored and consequently give preference in the annealing to the better matched atoms.

Two problems arise if we wish to consider more than a pair of molecules. Given a set of molecules; we can ask, firstly, which parts of each molecule have similarities with respect to a chosen



a b

Fig. 6. Cluster analysis of the atom positions of the molecular superimpositions with nulls to give 20 matched atoms. The matches are shown in Fig. 4. Each atom position is shown by +. The convex hulls of each cluster are drawn and the clusters are numbered. The radius of each circle centered on the cluster centroid is a measure of the compactness of each cluster. a: All superimposed atoms forming 44 clusters (compare with Fig. 4b); b: Matched atoms only are superimposed on CGS 8216 to give 31 clusters (compare with Fig. 4c).

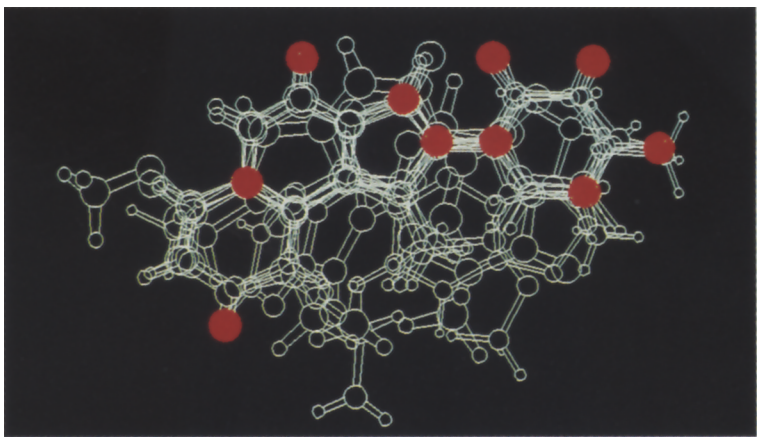


Fig. 7. The superimposition of 5 molecules onto CGS 8216 with null correspondences to give 20 matched atoms (data taken from Fig. 6b). The 10 best matched atoms are indicated in red.

reference molecule and secondly, what are the atom positional similarities within the set of molecules? This paper has outlined a new method for attempting to answer the first question. An answer to the second poses an interesting research problem and would necessitate a much more complex objective function compared with that used here; all the difference-distance matrices would have to be taken into account. Consider a set of p molecules with n being the number of atoms in the smallest molecule. Then any two molecules, l and m from the set p, can each be represented by their distance matrix where the elements, d_{ij}^{l} and d_{ij}^{m}, are the distances between atoms i and j in each molecule. The objective function, E, to be minimized is

$$E = \sum_{l=2}^{p} \sum_{m=1}^{l-1} \left\{ \sum_{i=2}^{n-k} \sum_{j=1}^{i-1} | d_{ij}^{l} - d_{ij}^{m} | \right\} \tag{3}$$

and the number of atoms, n, to be considered in the difference-distance matrices may be reduced by allowing k null correspondences; the term within brackets is the summation of the elements of the difference-distance matrix for molecules l and m. Exchanges would be made either to the order of the atoms, or to the combination of n-k atoms in l and m by the procedure outlined in the previous paper [13].

The introduction of null correspondences into the objective function enables the algorithm to reject poorly matched atoms; the previous paper [13] showed that this procedure worked adequately where the number of null correspondences was less than half the number of atoms. Configurational landscape problems increase with more null correspondences. Here we provided a classification method which helps to analyse the landscape problem and identifies alternative close matches. Alternative matches may vary from minor changes within a set of matched atoms to major differences in orientation or completely different sets of atoms. We have shown that it is necessary to perform multiple runs to identify the alternative matches. If the landscape is complex, that is if the dendograms show many distinct clusters, then many more runs would be necessary to be sure of an adequate sample of the landscape. The possibility of alternative matched structures is

TABLE 2
CLUSTER NUMBER, COMPACTNESS (P_i) AND MEMBERSHIP FOR THE 5 MOLECULES MATCHED AGAINST CGS 8216 WITH ZERO NULL CORRESPONDENCES

Cluster number	P_i	Component atoms[a]
28	0.00	C(CL)
22	0.42	A(H10) B(H2)
7	0.47	A(C1) B(N4) D(C12) E(C8) F(C10)
11	0.49	A(C7) B(C5) D(N2) E(C1) F(C3)
13	0.50	A(C10) B(C1) C(N6) D(C4) E(O2) F(O1)
39	0.53	C(H7) E(O3) E(C12)
8	0.54	A(C2) B(N3) C(N2) C(C9) D(C11) E(C9) F(C11)
27	0.54	B(H1) D(C6) D(H5) F(H9)
32	0.54	C(C2) C(H17) E(H14) F(H1)
33	0.54	C(C4) C(H15)
34	0.54	C(C5) C(H16)
36	0.54	C(C11) C(H13)
37	0.56	C(C14) C(H6) D(C5) D(H4)
5	0.58	A(N1) A(H9) F(C16) F (H14)
21	0.60	A(H7) B(C9) D(C8) E(N2) F(C8)
43	0.60	E(H6) F(C18) F(H18) F(H20)
10	0.61	A(C6) D(C13) D(H7) F(O3) F(C4)
20	0.61	A(H4) B(C11) C(C3) D(N3) D(H11) E(O1) E(C3) F(N2) F(H5)
29	0.62	C(O1) C(C6)
42	0.62	E(N1) E(C10) F(H19)
30	0.66	C(N4) C(C8)
9	0.67	A(C5) A(H8) B(H7) E(C7) E(H5) F(C5)
35	0.68	C(C19) C(H14) D(H8) E(H4) F(H2)
40	0.70	C(H10) E(N3) E(C11) F(H13)
2	0.73	A(O2) A(C12) B(C7) C(C15) D(C3) E(C13) E(H7) F(C13)
3	0.73	A(O3) A(C3) B(N1) B(C12) D(C16) D(H10) E(C4) E(H3) F(C7) F(H4)
12	0.74	A(C8) B(C4) B(H3) C(C13) F(C15) F(H11) F(H12)
16	0.77	A(C14) A(H5) A(H6) B(N2)
38	0.77	C(C16) C(H8) C(H9)
15	0.79	A(C13) A(H11) A(H13)
44	0.79	F(C17) F(H15) F(H16) F(H17)
25	0.80	B(F2) C(C17) C(H1) C(H2) C(H3) E(H8) F(H6)
6	0.81	A(N2) A(C9) B(C2) B(C3) C(H11)
24	0.82	A(H17) B(C10) B(H6) C(N1) D(C10) D(H6) E(C15) F(C1)
14	0.84	A(C11) B(C6) B(H4) D(N1) D(C7) F(C2) F(C12)
31	0.84	C(C1) E(H13)
23	0.85	A(H12) B(F3) C(H4) D(C2) D(H2) E(C14) E(H11) F(H7)
19	0.90	A(C17) A(H16) A(H18) B(H5) D(O) E(H12) F(N1)
41	0.90	D(C1) D(H1) E(H9) E(H10) F(C14) F(H8) F(H10)
17	0.91	A(C15) A(H1) A(H2) A(H3)
4	0.93	A(O4) A(C4) B(C13) B(H8) B(H9) C(N3) C(C7) D(C14) D(C15) D(H9) E(F) E(C5) E(C6) F(C6) F(H3)
26	0.93	B(C8) C(O2) C(H12) D(C9) E(C2) E(H1) F(C9)
18	1.00	A(C16) A(H15) C(O3) C(N5) C(C12) E(H2)
1	1.01	A(O1) A(H14) B(F1) C(H5) D(H3) F(O2)

[a] A = DMCM; B = CL 218872; C = Zopiclone; D = CGS 8216; E = Ro 151788; F = ZK 93426.

TABLE 3
CLUSTER NUMBER, COMPACTNESS (P_i) AND MEMBERSHIP FOR THE 5 MOLECULES MATCHED AGAINST CGS 8216 WITH NULL CORRESPONDENCES TO GIVE 20 MATCHED ATOMS

Cluster number	P_i	Component atoms[a]
32	0.00	C(O1)
15	0.12	A(C10) B(C1) D(C11) F(C11)
12	0.19	A(C6) C(C12) D(N1) E(N1) F(C2)
11	0.20	A(C5) C(N5) D(C4) E(C8) F(C12)
40	0.24	C(H14) D(H8) F(H2)
10	0.47	A(C4) B(N3) C(C13) D(C5) E(C9) F(O1)
13	0.48	A(C7) B(C5) C(C9) D(C9) E(C11) F(C9)
25	0.51	A(H10) B(H2) D(H6) F(H1)
24	0.52	A(H7) B(C9) E(H2) F(H11)
6	0.54	A(N2) B(C2) B(H1) D(N3) D(H11) F(N2) F(H5)
35	0.54	C(C4) C(H15)
36	0.54	C(C5) C(H16)
4	0.55	A(O4) B(C12) C(C14) C(H7) D(C6) E(C4) F(C13) F(H6)
3	0.57	A(O3) B(N2) F(H9)
23	0.58	A(H2) C(C17) C(H3) D(H1) E(F)
30	0.58	B(H8) C(C15) C(H4) D(C2) D(H2) E(C6) E(H4)
2	0.60	A(O2) A(C12) B(F3) B(C7) D(C16) D(H10) E(H8) F(C7) F(H4)
7	0.60	A(C1) B(N4) B(C8) C(O2) D(N2) E(C1) F(C3)
16	0.60	A(C11) B(C6) C(N4) C(C8) D(C12) E(O2) E(C12) F(C10)
5	0.65	A(N1) A(H9) B(H3) C(O3) D(O) D(C7) E(C10) E(H6) F(N1) F(C1)
38	0.67	C(C16) C(H8) C(H9) D(C3) D(H3) E(C7) E(H5)
14	0.68	A(C8) A(C9) B(C3) B(C4) C(H12) D(C8) D(C10) E(N3) F(C8)
20	0.68	A(C16) A(H14) B(H4) C(C19) D(C13) D(C14) F(C4) F(C5)
1	0.71	A(O1) B(F1) D(C15) D(H9) E(C13) E(H7) F(C6) F(H3)
34	0.71	C(N2) C(C3)
29	0.72	B(N1) C(H6) F(C14) F(H8) F(H10)
27	0.75	A(H16) C(C6) C(C7) F(H19)
9	0.77	A(C3) B(C11) C(H11) F(H12)
37	0.78	C(C11) C(H13) E(H9)
8	0.79	A(C2) C(H10) D(H4) E(C2) E(H1) F(C15) F(H13)
42	0.79	E(C14) E(H10) E(H11)
44	0.79	F(C17) F(H15) F(H16) F(H17)
26	0.81	A(H15) B(H5) D(H7) F(O3) F(C16) F(H14)
19	0.83	A(C15) A(H3) C(N6) C(H2) D(C1) E(C5) F(H7)
39	0.84	C(H1) D(H5) E(H3)
22	0.85	A(H1) A(H8) B(C13) B(H7) B(H9) C(H5) F(O2)
31	0.87	C(CL) C(C1)
33	0.90	C(N1) C(C2) C(H17)
17	0.91	A(C13) A(H11) A(H12) A(H13)
43	0.91	E(C15) E(H12) E(H13) E(H14)
21	0.97	A(C17) A(H17) A(H18) C(N3) F(C18) F(H18) F(H20)
41	0.97	E(O1) E(N2) E(C3)
18	1.06	A(C14) A(H4) A(H5) A(H6) B(C10) B(H6)
28	1.15	B(F2) E(O3)

[a] See footnote to Table 2.

important in drug design since they may point towards other subsets of molecular skeletons for drug development.

Classification theory provides a powerful method for analysing multivariate data to enable similar objects within a set to be identified. There is a frequent desire in drug research to try to discover what similarities a group of molecules may have with respect to a particular structure. Where the parameter for similarity measurement is a set of atom positions then the fulfilment of this wish can be obtained by simulated annealing combined with classification analysis. Pairwise comparisons with a reference molecule are made so that superimposed atoms fall into clusters. The compactness of the clusters indicates the goodness of fit and enables the best atom positions to be identified. If a particular match with a specified reference molecule is considered, then the mean value for the compactness, P_i, for the 44 clusters remains approximately constant but the standard deviation, σ_{Pi}, increases as more nulls are added. Therefore, as the number of nulls increases, clusters of matched atoms become more compact but clusters corresponding to null atoms have low compactness (high P_i). Figure 7 is a composite drawing of the best matches against CGS 8216 but with only the best 10 matched atom positions highlighted. The illustration provides visual evidence for the existence of compact clusters. It emphasizes the regions of high overlap and common substructures become apparent. This tight fit is taken from the best 20 matched atoms. If the number of nulls is reduced the compactness of the tightest clusters is reduced and other alignments are possible. These positions might prove useful as a template for searching the Cambridge Structural Database (CSD) for other molecules containing this partial geometry as potential candidates for further research. Different numbers of nulls would be expected to generate different templates for CSD searching.

The structures used in this study have been kept in the rigid conformations found in the crystal structures. If other acceptable low-energy conformations could be determined then each could be matched against a particular conformation of the reference molecule. Simulated annealing methods have been used to determine molecular conformations of small molecules and peptides [21,22]. However, the matching of flexible structures is still a formidable problem for optimization and merits extensive study.

The work described in this paper has drawn attention to common atom positions in sets of superimposed molecules; atom positions have only been considered as points in space. No property has been assigned to the atoms. In an earlier paper [23] a method of combining geometric searching with the hydrogen-bond property for specified atoms was outlined. That procedure used a branch-and-bound method to reduce a brute-force search. If the method outlined here is to be extended similarly to include other properties, as well as atom positions, then extra parameters need to be taken into consideration in the annealing procedure. The atomic properties of interest to a medicinal chemist may include hydrogen bonding capacity, electronic charge and hydrophobicity. Thought needs to be given to how these properties may be included in an objective function. It may be possible simply to add property mismatches as a penalty to the objective function. For example, if μ_{ih} and μ_{jh} are the values for property h on atoms i and j, then for molecules l and m

$$E = \left\{ \sum_{i=2}^{n-k} \sum_{j=1}^{i-1} | d_{ij}^{l} - d_{ij}^{m} | \right\} + \left\{ \sum_{i=1}^{n-k} \sum_{h} f_h | \mu_{ih}^{l} - \mu_{ih}^{m} | \right\} \tag{4}$$

where f_h is a scaling function for the property h. This function would need to be defined carefully so that the penalty would influence the objective function significantly but would not dominate it. Otherwise matching would proceed through parameter minimization and not predominantly by spatial superimposition. The right-hand side of Eq. 4 could then be substituted for the summation terms in brackets in Eq. 3 to attempt to match a set of molecules by atomic positions and by atomic properties.

If the method suggested above is successful, then it might be possible to use the superimpositions, generated by atom positions, together with the property parameters, as the basis for molecular alignments in a 3D QSAR study. Once a satisfactory molecular superimposition has been determined then other dependent parameters, such as the electrostatic potential and the hydrophobic potential, may be constructed in the space surrounding the molecules. Any relationship between the distribution of these parameters in 3D-space with affinity/activity may then be explored knowing that the molecular superimpositions have been derived without human bias.

ACKNOWLEDGEMENTS

M.C.P is grateful to the Cambridge Commonwealth Trust for a scholarship and P.M.D is indebted to the Wellcome Trust for a Principal Research Fellowship. Part of the work was supported by the SERC Cambridge Centre for Molecular Recognition.

REFERENCES

1 Maggiora, G. and Johnson, M., Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.
2 Danziger, D.J. and Dean, P.M., Proc. R. Soc. Lond., B236 (1989) 101.
3 Danziger, D.J. and Dean, P.M., Proc. R. Soc. Lond., B236 (1989) 115.
4 Lewis, R.A. and Dean, P.M., Proc. R. Soc. Lond., B236 (1989) 125.
5 Lewis, R.A. and Dean, P.M., Proc. R. Soc. Lond., B236 (1989) 141.
6 Namasivayam, S. and Dean, P.M., J. Mol. Graphics, 4 (1986) 46.
7 Chau, P.-L and Dean, P.M., J. Mol. Graphics, 5 (1987) 97.
8 Dean, P.M. and Chau, P.-L, J. Mol. Graphics, 5 (1987) 152.
9 Dean, P.M., Callow, P. and Chau, P.-L., J. Mol. Graphics, 6 (1988) 28.
10 Bowen-Jenkins, P.E., Cooper, D.L. and Richards, W.G., J. Phys. Chem., 89 (1985) 2195.
11 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 295.
12 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 317.
13 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 107.
14 Bertolasi, V., Ferreti, V., Gilli, G. and Borea, P.A., J. Chem. Soc. Perkin Trans., 2 (1990) 1.
15 Codding, P.W. and Muir, A.K.S., Mol. Pharmacol., 28 (1985) 178.
16 Bertolasi, V., Ferreti, V., Gilli, G. and Borea, P.A., Acta Crystallogr., B40 (1984) 1981.
17 McLachlan, A.D., Acta Crystallogr., A38 (1982) 871.
18 Ledermann, W., Handbook of Applicable Mathematics, Vol. 6, Wiley, New York, 1980, p. 35.
19 Wishart, D., Clustan User Manual, University of St Andrews, Scotland, 1987.
20 Murtagh, F. and Heck, A., Multivariate Data Analysis, D. Reidel, Dordrecht, 1987.
21 Wilson, S.R., Cui, W., Moscowitz, J.W. and Schmidt, K.E., Tetrahedron Lett., 29 (1988) 4373.
22 Wilson, S.R. and Cui, W., Biopolymers 29 (1990) 225.
23 Danziger, D.J. and Dean, P.M., J. Theor. Biol., 116 (1985) 215.