Journal of Computer-Aided Molecular Design, 7 (1993) 155–172
ESCOM

J-CAMD 192

An exploration of a novel strategy for superposing several flexible molecules

T.D.J. Perkins and P.M. Dean

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

Received 14 October 1992
Accepted 11 November 1992

Key words: Molecular matching; Molecular similarity; Simulated annealing; Conformational analysis; Cluster analysis; Angiotensin II antagonists; Superposition

SUMMARY

This paper describes a computational strategy for the superposition of a set of flexible molecules. The combinatorial problems of searching conformational space and molecular matching are reduced drastically by the combined use of simulated annealing methods and cluster analysis. For each molecule, the global minimum of the conformational energy is determined by annealing and the search trajectory is retained in a history file. All the significantly different low-energy conformations are extracted by cluster analysis of data in the history file. Each pair of molecules, in each of their significantly different conformations, is then matched by simulated annealing, using the difference-distance matrix as the objective function. A set of match statistics is then obtained, from which the best match taken from all different conformations can be found. The molecules are then superposed either by reference to a base molecule or by a consensus method. This strategy ensures that as wide a range of conformations as possible is considered, but at the same time that the smallest number of significantly different conformations is used. The method has been tested on a set of six angiotensin II antagonists with between 7–11 rotatable bonds.

INTRODUCTION

In the pharmaceutical industry, a drug designer may wish to know what structural similarities exist within a set of molecules in order to perform a full three-dimensional (3D) quantitative study of structure–activity relationships (QSAR). One of the tools that is emerging to help with this design problem is the much-discussed Comparative Molecular Field Analysis (CoMFA) program from TRIPOS [1–7]. If the molecules can be aligned, then the molecular fields can be compared in 3D space. A partial least-squares analysis can be used to relate biological activity to the disposition of fields and thus reveal the pharmacophore. Novel molecules can then be built that contain this pharmacophore, and these can be expected to have biological activity. The rationale for this scheme is highly attractive and has stimulated much research into deriving new methods for expressing molecular similarity and determining molecular superposition. The CoMFA pro-

cedure is dependent on the molecular superpositions used as input to the program. If the molecules have a fixed conformation and are more-or-less bio-isosteres, their structural superposition is obvious and needs only least-squares fitting between equivalent atom positions. Once superposed, the structural similarity between the molecules can be expressed readily by a variety of methods [8–16]. However, if the molecules are flexible, and ostensibly dissimilar, their superposition presents a more complex problem. In this paper we outline and test a novel method for the superposition of a set of flexible dissimilar molecules in which the atom correspondences are initially unknown. A preliminary account of this work has been presented elsewhere [17].

A procedure frequently used to match a set of flexible molecules is to use a rigid analogue and to fold the remainder of the molecules to fit this shape (the shape reference method) [18–20]. However, where there is no rigid analogue, the choice of any single conformer as a base effectively forces an arbitrary superposition, since the receptor-bound conformation is generally unknown. This conformation need not be the global minimum energy conformation, as determined by either theoretical or experimental techniques, but it is expected to be accessible and therefore within approximately 30 kJ mol^{-1} of the global minimum energy conformation. Thus, any general method of matching flexible molecules should, ideally, consider all possible conformations of the molecules to be matched. However, this procedure becomes unmanageable for any molecule with more than a few rotatable bonds. If each molecule has t rotatable bonds each with j rotamers, then j^t conformers have to be considered. For example, four rotatable bonds with 10° intervals give $36^4 \approx 10^6$ possible conformers, while ten bonds produce 10^{15} conformers. Many of these conformations would be improbable due to steric and/or electronic clashes. There are, thus, two problems to be solved: firstly, conformations with too high energies must be excluded; secondly, the total number of conformations to be examined must be reduced to a manageable number. The combinatorial explosion that results from attempting to match a very large number of conformers is described below, before a strategy for solving the problem is presented.

In this paper we are not concerned with matching molecular fields, thus the easiest type of unbiased molecular superposition that can be performed on dissimilar molecules is based on atom positional matching. For two n-atom molecules, there are $n!$ possible atom positional assignments. If we wish to match p n-atom molecules, each containing t rotatable bonds each with j rotamers, then the total number of assignments that have to be considered is $n!j^{tp}$. For six 40-atom molecules, each containing six rotatable bonds with 10° rotation steps, about 10^{104} atom assignments have to be made. Many molecular conformations will vary by only small, but energetically feasible, angular displacements about the rotatable bonds; these will give essentially similar atom correspondences, and therefore similar superpositions, when compared with the other molecules. Our philosophy has been to reduce the problem to tractable proportions, by removing all but one of the similar feasible conformations. This reduction has been achieved by identifying conformational clusters in torsion space that are significantly different from each other. Thus only a small number of conformers, one from each cluster instead of j^{tp}, are considered for matching.

Strategy

The problem of matching a set of flexible molecules is tackled by using a reduced representation of the conformational space of each of the molecules. A small number of diverse conformers are determined, and these are used to represent the entire allowed conformational space. The

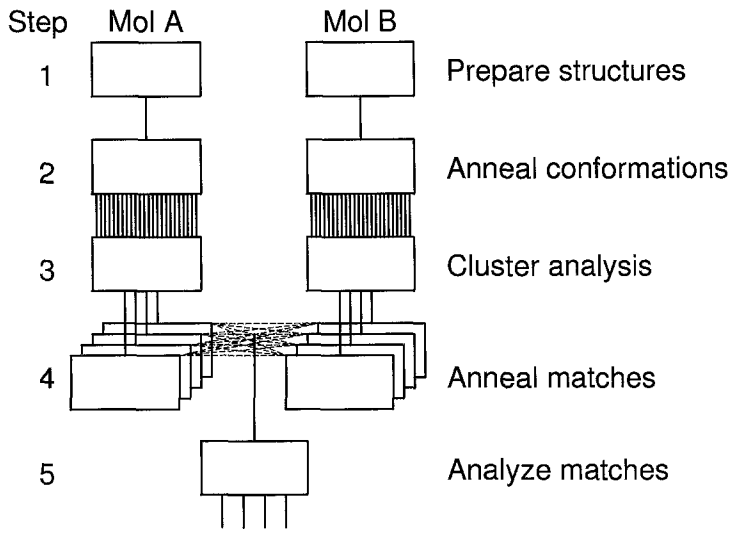

Fig. 1. Flow diagram showing the overall strategy used for matching one pair of molecules.

complete process of matching one pair of molecules is illustrated by the flow diagram in Fig. 1, and the step numbers below refer to that figure.

The starting structures are prepared (Step 1) for simulated annealing [21] to search for the global minimum conformational energy of each molecule (Step 2). Conformations along the trajectory are saved, and a random subset of low-energy conformations is classified, using cluster analysis, into a small number of significantly distinct clusters (Step 3). A single conformation is chosen as being representative for each cluster. Thus, each molecule is described by a small number of significantly different, but representative, conformers. These conformers are then matched, pairwise, by minimizing their difference-distance matrix, using simulated annealing (Step 4), as described previously [22–25]. Different procedures can be used to analyse the matches generated (Step 5). When there are only two molecules, it is possible to use cluster analysis to determine significantly different classes of matches. However, for a larger set of molecules, it is more informative to determine the conformations of each molecule that give an optimum superposition. There are two possible approaches to this problem. First, each molecule can be taken in turn and used as a base onto which the other molecules are superposed, and all conformations of the base molecule and the fitted molecules are searched for the maximum similarity. Second, the best consensus match between the molecules in all their representative conformations can be found independently of any base molecule. The usefulness of either method will depend on the eventual practice to which the matching procedure is put.

We propose a new computational method for matching a small set of flexible dissimilar molecules, using the atom positional method. The algorithm is tested on six angiotensin II antagonists but no comment is made on their QSAR. The methods described attempt to reduce the combinatorial process of matching a set of flexible molecules, and inevitably, a large number of good matches are excluded. Nevertheless, this type of approximation is necessary to begin to tackle the problem. It should then be possible to refine the solution presented in the light of experience with

these algorithms. Our hope is that this strategy may provide a way forward for CoMFA users who feel thwarted by not being able to include flexible molecules in their 3D QSAR studies.

METHODS

Computational experiments were carried out on a Sun Microsystems SPARCstation IPX. All programs were written in FORTRAN 77 and were compiled and linked with Sun f77 1.4. Any timings relate to codes compiled and linked at optimization level 4. The numbers of the sections below correspond to those in the flow diagram shown in Fig. 1.

1 Structure preparation

Six angiotensin II receptor antagonists were used to illustrate the molecular matching methods described in this paper: 2-*n*-butyl-4-chloro-5-(hydroxymethyl)-1-[[2'-(1H-tetrazole-5-yl)biphen-4-yl]methyl] imidazol (DuP 753 from DuPont); (E)-3-[2-butyl-1-[4-(carboxyphenyl)methyl]-1H-imidazol-5-yl]-2-[(2-thienyl)methyl]propanoic acid (SKB 108,566 from SmithKline Beecham); 1-[[3-chloro-2-[2-(1H-tetrazol-5-yl)-phenyl]-5-benzo-[b]-thiophenyl]methyl]-2-butyl-4-chloro-1H-imidazole-5-carboxylic acid (GLAXO from Glaxo Group Ltd., EP 0430709 A); 2-butyl-1-[[2'-(1H-tetrazol-5-yl)biphenyl-4-yl]methyl]-benzimidazole-7-carboxylic acid (TAK from Takeda, EP 425921 A); 2-butyl-5,7-dimethyl-3-[[2'-(tetrazol-5-yl)biphen-4-yl]methyl]-3H-imidazol-[4,5-b]pyridine (L-158,809 from Merck, Sharp and Dohme); 4'-[5,7-dimethyl-2-butyl-6(1H)-cycloheptimidazolon]-1-yl]-[1,1'-biphenyl]-2-carboxylic acid (SEARLE from Searle and Co., EP 0432737 A) (see Fig. 2). 3D coordinates were calculated (using the program CONCORD [26]) and supplied by Dr. I.L. McLay, Rhône-Poulenc Rorer, Dagenham, Essex. The alkyl chain groups marked as X or Y in Fig. 2 were truncated to a methyl group in each case. The structures were not optimized, although the structures were checked manually to ensure reasonable stereochemistry. If the precise geometry is thought to be critical for a particular molecule, it would be prudent to refine the coordinates further by using molecular mechanics or semiempirical methods. Atomic partial charges (later used in the assessment of potential energy) were calculated using the program for Complete Neglect of Differential Overlap (CNDO/2) [27].

2 Conformational search by simulated annealing

In a conformational search, the only way to ensure that the entire conformational space is covered is to use a grid or 'brute force' search. However, the number of calculations and the computer time required increase exponentially with the number of rotatable bonds. This makes a grid search impracticable for all but the smallest molecules. However, simulated annealing appears to be the optimization method of choice for this problem. It should be able to determine the global (or a near-global) minimum energy conformation rapidly, without becoming trapped in a local minimum of the conformational energy hyperspace [28].

In this study, the conformational space of each molecule was searched by using a simulated annealing program based on that of Wilson et al. [21]. Minimization was attempted only in terms of a rigid rotor model. All acyclic bonds in the molecules were designated as being rotatable (see Fig. 2). Potential energies were calculated using the Coulombic, van der Waals, and periodic torsion angle components of the COSMIC force field [29]. Bond length and bond angle components were neglected since these internal coordinates remained constant in this study.

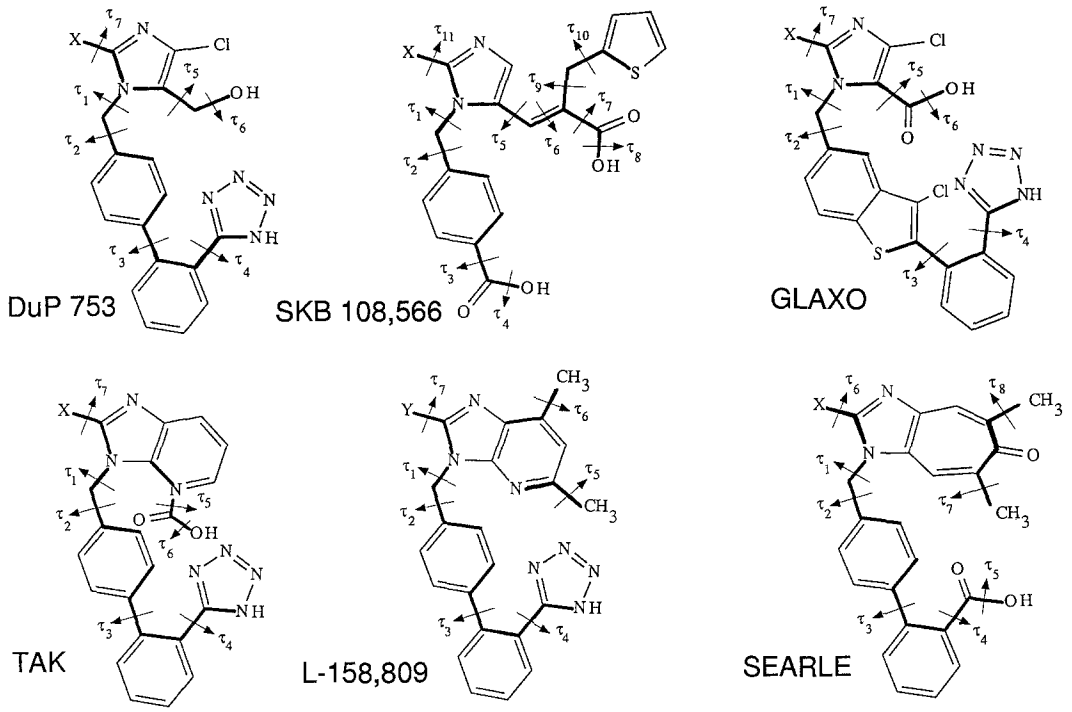

Fig. 2. Chemical structures of the molecules used in matching: DuP 753, SKB 108,566, GLAXO, TAK, L-158,809, SEARLE. Torsion angles designated as rotatable are also indicated as τ_1–τ_{11} for each molecule. Bonds that describe torsion angles are drawn as thick lines. X represents a butyl group, and Y a methyl group.

The annealing algorithm proceeds as follows. The energy, E_i, of the initial conformation, i, is calculated, and a new conformation, $i+1$, is generated by choosing one torsion angle at random and rotating it by a random amount in the range ± 180° in steps of 1°. For rotationally symmetrical terminal groups (e.g., methyl, phenyl), degeneracy is taken into account by reducing the maximum rotational step. Thus, for a terminal methyl group where the rotational degeneracy, r, is 3, the maximum allowed rotational step is ± 180°/r = ± 60°. The energy, E_{i+1}, of the new conformation, $i+1$, is then calculated. If the change in energy, $\Delta E_{i+1}=E_{i+1}-E_i$, is negative, the new conformer, $i+1$, is accepted immediately. Otherwise, it is accepted according to the Metropolis condition [30], with a probability of $\exp(-\Delta E_{i+1}/T)$, where T is an annealing control temperature. If accepted, conformer $i+1$ becomes the current conformer from which any further changes are propagated.

For all six molecules, the conformational search consisted of 40 Markov chains, each comprising 500 trial conformations, with exponential cooling at a rate of 0.9 between Markov chains. The initial temperature was set at $100R$ (831 K), chosen so that the initial acceptance rate was roughly 80%. During the simulated annealing run, the conformations and potential energies of all conformers – both accepted and rejected – were stored in a history file.

3 Cluster analysis of generated conformers

The conformers recorded in the history file were taken as a reasonable random sample of the

conformational space of the molecule (although with a possible bias towards low-energy structures). The global minimum energy conformation (in terms of the rigid rotor model used throughout this paper) was then used to determine the probability of existence of all other conformers in a Boltzmann distribution at 290 K. All conformers with a probability less than 10^{-6} (corresponding to those with potential energies greater than 33.3 kJ mol^{-1} above the global minimum energy conformation) were discarded. A random subset of 250 of the remaining conformers was subjected to cluster analysis. This generous energetic cut-off should have ensured that all energetically accessible conformations were included.

Two different methods were used in the calculation of the 'distance' between conformers: a distance in torsion angle space and a measure based on the difference between interatomic distance matrices. Preliminary studies indicated that, at least for the molecules presented here, the choice of distance measure did not significantly affect the results produced by the clustering procedure. The torsion angle space method was adopted and is described in the rest of this paper.

For any two conformers, j and k, of a molecule, each with t torsion angles, the squared Euclidean torsion angle distance, d_τ^{jk}, is calculated by taking into account the periodicity in torsion angle space:

$$\Delta\tau_i^{jk} = \tau_i^j - \tau_i^k,$$

$$\text{if } \Delta\tau_i^{jk} > 180°/r_i \text{, then } \Delta\tau_i^{jk} = \Delta\tau_i^{jk} - 360°/r_i$$

$$\text{if } \Delta\tau_i^{jk} \leq -180°r_i \text{, then } \Delta\tau_i^{jk} = \Delta\tau_i^{jk} + 360°/r_i$$

$$d_i^{jk} = \sum_{i=1}^{t} (\Delta\tau_i^{jk})^2$$

where τ_i is the value of the torsion angle, i, r_i is the rotational degeneracy of the torsion angle, i, and superscripts j and k represent the conformer.

Cluster analysis was carried out using Ward's minimum variance method [31], and the optimum number of clusters was chosen according to the method of Mojena [32]. This method considers the sequence of $N-1$ criterion values of α_j, corresponding to each fusion, j, in the cluster hierarchy of N conformations. A classification is significant if the next fusion criterion, α_{j+1}, is in the upper part of the distribution of α values. The mean ($\bar{\alpha}$) and standard deviation (σ_α) of the criterion values for each fusion step are calculated and compared to the individual values for each fusion, α_{j+1}, to give the number of k_j deviates between them:

$$k_j = (\alpha_{j+1} - \bar{\alpha})/\sigma_\alpha.$$

At each fusion step, the probability, p, of α_{j+1} being in the upper part was calculated using the incomplete β function [33], and the first classification where the subsequent fusion had $p < 0.05$ was selected as the most significant classification. This provided the number of significantly different clusters and was then used to determine the most representative conformer in the cluster. The structure chosen as being representative was the structure whose sum of the elements in the squared distance torsion space was the smallest. This conformer is closest to the cluster centre.

4 Matching conformers by simulated annealing

For all pairs of molecules, all combinations of representative conformers were matched as rigid molecules. The Barakat–Dean [22–25] algorithm was used to determine atom assignments. This program calculates interatomic distance matrices for the two molecules to be matched, and the positive difference between equivalent distances is calculated. The sum of this difference-distance matrix (Σd_{DDM}) represents the objective function to be minimized by simulated annealing. Changes in the configuration of the system were made by swapping pairs of atoms in one of the molecules. Dynamic cooling was used, and the parameters [22] were set at $\delta = 0.05$, $C = 8$, $T = 2.0$, with no null matches allowed. All combinations of pairs of the six molecules were matched five times, each starting with different random number seeds.

5 Analysis of matches

The matches produced in the previous step were analysed in three different ways.

a. Cluster analysis of all pairwise matches. The difference in assignments between matches of the same pair of molecules, independent of the conformation of either, was established using a binary scoring scheme: identical assignments score zero and non-identical assignments score one. The sum of the binary assignments is a numerical measure of the similarity between replicate matches of the same pair of molecules. The matrix of total scores for all the matches between a pair of molecules constructed from these data was then subjected to cluster analysis with Mojena's stopping rules as described in Section 3. For each cluster, the member with the lowest value of Σd_{DDM}, the match statistic from the simulated annealing, was considered to be representative. This new method of identifying different assignments is much faster than that previously described [25], which forms clusters on the basis of the superposition transformation matrix. In addition, it allows the comparison of different conformations of molecules since it does not require any 3D frame of reference.

b. All molecules: matching onto a base molecule. Matches between all combinations of pairs of molecules were examined to determine the set of conformations that gives the best match onto a base molecule. The sum of the difference-distance matrix, Σd_{DDM}, was normalized to give a novel metric, $\bar{d}_{DDM}$. This is the mean value of each element in the difference-distance matrix calculated in simulated annealing matching. Normalization allows an unbiased comparison between matches involving differing numbers of atoms. Each molecule was used in turn as the base, and all combinations of the representative conformers of each of the molecules were examined to determine the set with the minimum $\bar{d}_{DDM}$ statistic between the base molecule and all other molecules. 3D coordinates were generated by least-squares fitting [34] onto the appropriate base molecule, using the atom assignments generated by the simulated annealing match. The top five conformer sets were retained to allow a variety of good matches to be examined.

c. All molecules: consensus matching independent of a base molecule. The best combination of conformations independent of any base molecule was selected, using a method similar to that described above. However, the match statistic to be minimized, $\bar{d}_{DDM}$, was calculated over all of the matches in the set, a total of $n(n-1)/2$ matches for n molecules. This should produce a match that is not biased towards any particular molecule. Thus, the conformations determined by this procedure may possibly provide a clue to the conformation adopted by these molecules, if they all bind to the same receptor site. The superposition of the set of molecules was determined using

each of the molecules in turn for reference, and calculating the root-mean-square (r.m.s.) over all molecular pairs. The superposition with the lowest overall r.m.s. was then chosen for display.

RESULTS

1 Structures

Six angiotensin II antagonist molecules were used to test the method, but since we were not attempting to describe their QSAR, only representative parts of the data are presented here. The complete results are presented for DuP 753 and SKB 108,566.

2 Conformational search by simulated annealing

The global minimum energy conformation determined by one run of the simulated annealing conformational search is shown in Table 1 for each of the six structures, together with its final potential energy. The torsion angle key is that shown in Fig. 2. The large positive potential energy for SEARLE appears to be due primarily to the strain energy within the cycloheptimidazolone ring. The CPU time required for this step for SKB 108,566 was about 8 min, which included the generation of 20 000 conformers.

3 Cluster analysis of generated conformers

From the 20 000 conformers stored in the history file, 7719 DuP 753 and 2160 SKB 108 566 conformers were found with a probability $>10^{-6}$. Dendrograms representing the cluster analysis of 250 conformers randomly selected from these are shown for DuP 753 and SKB 108,566 in Fig. 3. The solid line represents the hierarchical arrangement that is considered significant at the level $p < 0.05$. The number of significant clusters can thus be read off as the number of vertical lines that cut the thick horizontal line. The energies and conformations of the representative conformers of the significant clusters of these two molecules are shown in Table 2. Note the wide range of conformational energies that are represented. For most of the torsion angles a spread of values is observed, although some (e.g., τ_5, τ_6, and τ_8) in SKB 108,566 appear locked in a *trans* conformation. The numbers of representative clusters for the remaining structures were GLAXO, 9, TAK, 8, L-158,809, 8 and SEARLE, 7.

Figure 4 shows a stereoview of the eight representative conformers of DuP 753, aligned such

TABLE 1
GLOBAL MINIMUM ENERGY CONFORMATIONS

	E^a	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}
DuP 753	−5.33	60	−108	49	−162	84	61	155				
SKB 108,566	−95.15	97	−93	175	−179	168	162	180	−171	−145	63	73
GLAXO	−62.67	−113	49	125	163	−177	179	49				
TAK	−1.06	−118	−159	47	−161	−144	177	−63				
L-158,809	−10.46	153	−60	50	−167	−112	141	−109				
SEARLE	426.05	133	−46	−128	−166	−176	35	118	17			

[a] Global minimum energy (kJ mol^{-1}) found during conformational analysis.

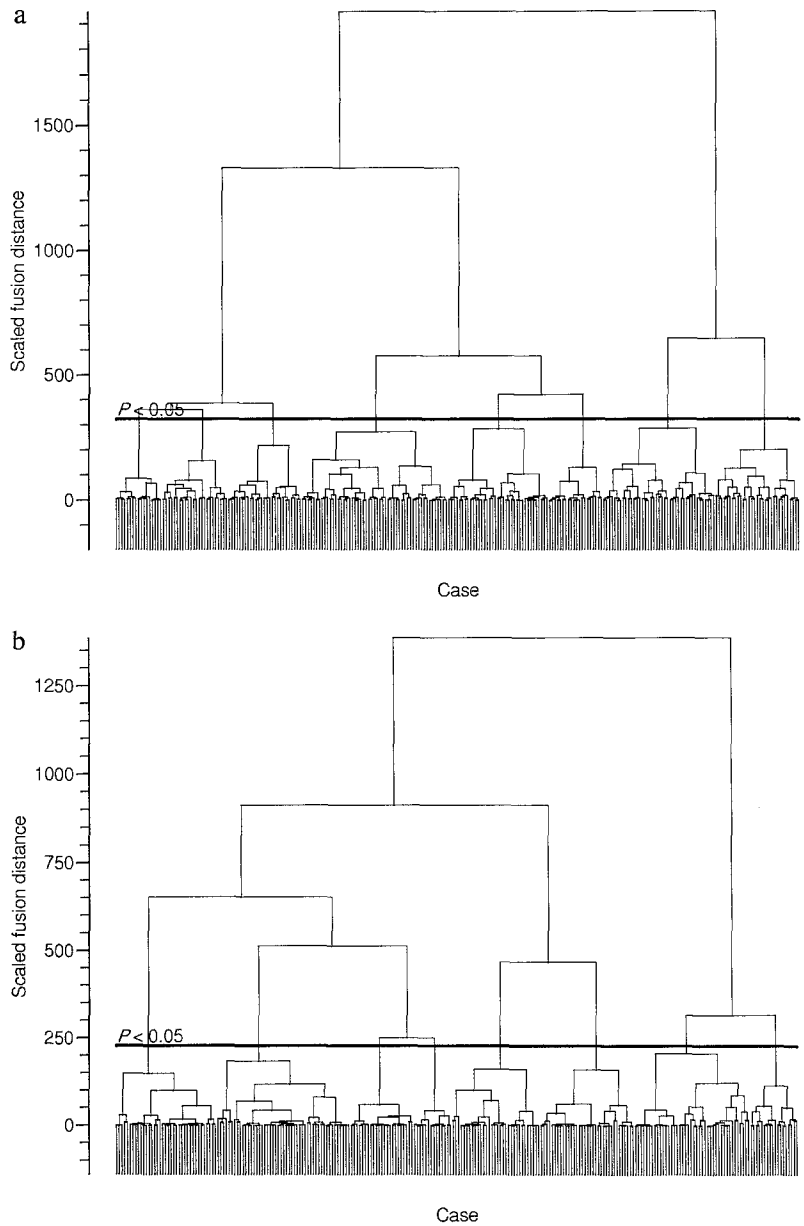

Fig. 3. Dendrogram illustrating the clustering of the conformational searching: (a) DuP 753; (b) SKB 108,566.

that a common imidazole ring is superposed exactly. This illustrates the wide distribution of the tetrazole ring in 3D space, indicating the extent of conformational space available to this molecule. The CPU time required for this step, including sorting, clustering and determination of representative conformers for SKB 108,566, was about 25 s.

4 Matching conformers by simulated annealing

The matches between DuP 753 and SKB 108,566 are described by the data in Table 3. For each pair of representative conformers, the best match between the atom correspondences of the five replicates is shown as the normalized difference-distance metric, $\bar{d}_{DDM}$, and the r.m.s. difference following least-squares fitting. Note that the two measurements yield differing matches as maxima and minima. For all 320 individual pair-wise matches, the total CPU time required for DuP 753 and SKB 108,566 was a little under 6 h.

5 Analysis of the matches

a. Cluster analysis of all pairwise matches. Cluster analysis of the 64 matches (each carried out five times) between DuP 753 and SKB 108,566 is shown as a dendrogram in Fig. 5. The line showing the level of significance indicates that there are six significantly different matches. The superposition with the lowest r.m.s. distance (1.70 Å) had a mean atomic difference-distance element, $\bar{d}_{DDM}$, of 0.80 Å.

b. All molecules: matching onto a base molecule. The best conformations for superposition of each molecule onto each of the others are shown in Table 4; together with the r.m.s. distance

TABLE 2
RELATIVE POTENTIAL ENERGIES OF THE EIGHT REPRESENTATIVE CONFORMERS OF TWO COMPOUNDS WITH THEIR CORRESPONDING TORSION ANGLES

A. Representative conformers of DuP 753

Rep	E^a	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
1	18.07	−107	46	−127	−167	117	133	−57
2	28.57	−100	73	49	14	166	−100	3
3	15.21	−69	−31	−119	−165	−112	−46	1
4	30.14	72	−83	57	−166	−79	−6	23
5	7.74	−119	−125	49	−166	69	158	15
6	13.22	57	58	−121	−162	94	85	17
7	4.05	75	−108	−134	−164	−64	−66	56
8	25.08	85	178	−127	2	−154	180	−52

B. Representative conformers of SKB 108,566

Rep	E^a	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}
1	7.27	94	−45	174	−173	171	162	177	171	−148	75	−25
2	11.65	−142	59	−173	173	171	162	178	−174	−145	−111	10
3	11.60	−136	46	−169	−172	168	162	7	−172	−144	83	−24
4	2.93	115	−72	4	−180	171	158	−178	172	−142	−126	48
5	10.02	104	130	4	174	171	158	179	165	−139	32	44
6	9.79	−91	110	13	177	171	162	178	−175	−142	48	−10
7	9.71	−93	29	171	−163	168	162	−176	−177	−144	46	−8
8	9.16	−48	145	−12	−178	171	158	0	−174	−142	58	−30

[a] Energy (kJ mol^{-1}) above global minimum found in conformational analysis.

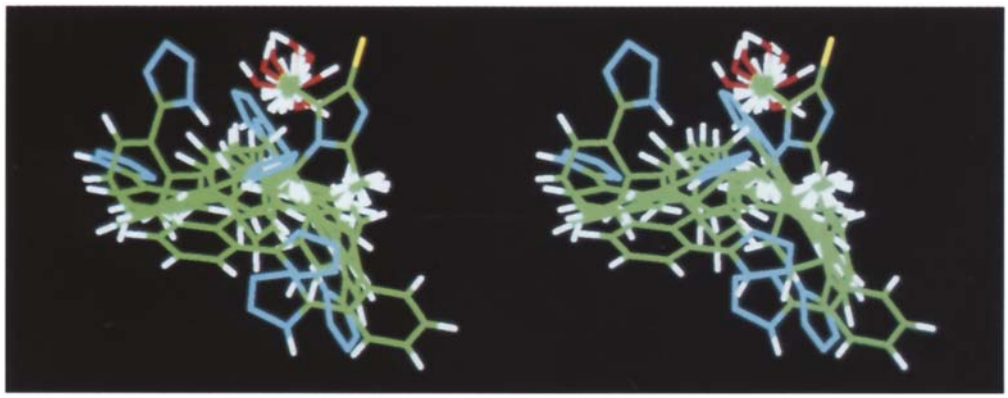

Fig. 4. Colour stereoview of the eight representative conformers of DuP 753. Atom colours: green (C); blue (N); red (O); yellow (Cl); white (H).

between the molecule and the base onto which it is being fitted. The value of $\bar{d}_{DDM}$ varies only slightly (0.5–0.7 Å) between different base molecules. However, each base molecule is matched by

TABLE 3
THE BEST MATCH OF FIVE REPLICATES FOR EACH PAIR OF EIGHT REPRESENTATIVE CONFORMERS BETWEEN DuP 753 AND SKB 108,566

A. DDM statistic, $\bar{d}_{DDM}$[a], for best match of the five replicates

DuP 753 SKB 108,566	1	2	3	4	5	6	7	8
1	0.85	0.81	0.85	0.79	0.86	0.86	0.78	0.81
2	0.95	0.92	1.06	0.98	0.95	0.99	1.00	0.91
3	0.96	0.92	1.03	0.95	0.94	0.98	0.97	0.86
4	0.94	0.92	1.03	0.94	0.95	0.95	0.98	0.93
5	0.89	0.86	0.94	0.86	0.89	0.92	0.87	0.87
6	0.92	0.86	0.86	0.86	0.90	0.85	0.82	0.89
7	0.91	0.88	0.86	0.83	0.89	0.88	0.80	0.88
8	1.06	0.97	0.94	0.97	1.03	0.96	0.94	1.04

B. RMS differences[a] for above matches

DuP 753 SKB 108,566	1	2	3	4	5	6	7	8
1	2.09	1.84	2.14	2.98	2.56	2.84	2.82	2.57
2	2.24	2.46	3.06	2.13	2.23	2.23	2.33	2.56
3	2.31	2.47	2.54	2.35	2.41	2.01	2.54	2.52
4	2.15	1.95	2.33	2.98	2.46	1.92	3.02	2.79
5	2.03	1.90	2.10	2.80	2.56	1.98	3.00	2.53
6	2.12	2.08	2.18	1.80	2.02	2.06	1.83	2.12
7	1.96	2.32	2.46	1.58	1.77	2.35	1.70	2.49
8	2.66	2.59	2.65	1.92	2.12	2.08	2.07	2.72

[a] Measurements in Å.

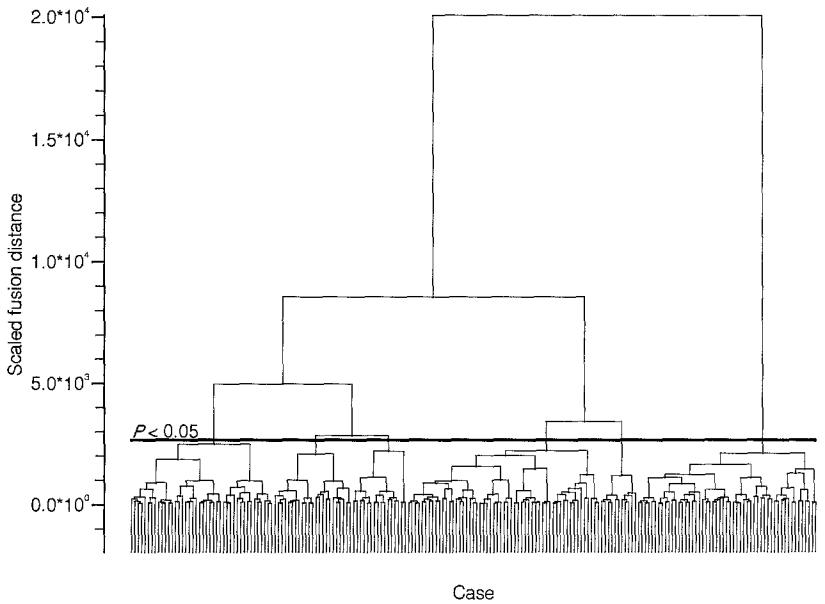

Fig. 5. Dendrogram illustrating the clustering of different atom matches between replicate matches of DuP 753 and SKB 108,566, over all conformations.

a different set of conformations. The base molecule that gives the best $\bar{d}_{DDM}$ is DuP 753, and the superposition of the remaining five molecules onto this molecule is shown in Fig. 6.

c. All molecules: consensus matching independent of a base molecule. Table 5 shows the case where no molecule is considered to be a base, and so the difference-distance metric, $\bar{d}_{DDM}$, is calculated over all combinations of pairs of molecules. The lowest value for this sum was 0.609 Å. The conformations found in the consensus match show a strong similarity with those for matching with SKB 108,566 as a base molecule (see Table 4). The superposition reference molecule that gives the best overall r.m.s. distance is SEARLE, and the superpositions of the remaining five molecules onto this molecule are shown in Fig. 7. The entire set of molecules has been transformed so that DuP 753 is in the same orientation as found in Fig. 6.

In order to assess the effect of changing the reference molecule for consensus superposition, superpositions were determined taking each molecule in turn, and then transforming the entire set of superposed molecules so that DuP 753 maintained its original orientation. The relative orientations of the remaining molecules were then drawn. The molecule that had the smallest range of relative orientations was SKB 108,566 (Fig. 8); L-158, 809, which appeared to change the most, is shown in Fig. 9.

DISCUSSION

The problem of molecular superposition is of crucial importance in 3D QSAR studies, as the utility of any results obtained depends implicitly on the quality of the superpositions. For molecules of very similar shapes and stereochemistry, the superposition is generally straightforward.

TABLE 4
CONFORMATIONS AND MATCH STATISTICS FOR MATCHING ALL SIX MOLECULES USING EACH IN TURN AS A BASE[a]

Structure Base	DuP 753	SKB 108,566	GLAXO	TAK	L-158,809	SEARLE
DuP 753	(0.49 Å)					
rep	(7	1	3	5	1	4
r.m.s./Å	–	2.82	2.45	0.61	2.38	1.96
SKB 108,566	(0.72 Å)					
rep	(7	7	5	5	7	6
r.m.s./Å	1.70	–	2.57	1.83	2.36	2.34
GLAXO	(0.66 Å)					
rep	(1	7	3	5	7	7
r.m.s./Å	1.70	1.90	–	2.93	3.30	2.76
TAK	(0.51 Å)					
rep	(1	1	9	8	8	7
r.m.s./Å	0.84	1.72	2.55	–	2.77	2.74
L-158,809	(0.53 Å)					
rep	(7	7	6	5	1	1
r.m.s./Å	2.38	2.52	1.60	2.75	–	1.52
SEARLE	(0.52 Å)					
rep	(5	4	3	7	8	7
r.m.s./Å	2.27	1.60	2.76	2.72	1.03	–

[a] The number in brackets after each base molecule is the match statistic $\bar{d}_{DDM}$ for stacking the remainder of the molecules onto that particular base.

However, for apparently dissimilar molecules, the problem is far more formidable. An experienced medicinal chemist may be able to surmise potential pharmacophoric atoms and then attempt to overlap those considered equivalent. As larger and more flexible molecules are encountered, and consequently the number of possible superpositions increases, it is not practicable to consider all possible superpositions manually. Molecular flexibility increases the difficulty of the problem by many orders of magnitude since each new pair of conformers considered may lead to different molecular superpositions.

Previous automated methods for molecular matching have, in general, suffered from the disadvantage that equivalent atom pairs must be specified by the user before any attempt can be made at the alignment – this exemplifies the superposition problem. In addition, most methods described hitherto deal with only a single rigid conformation of a particular molecule. Methods

TABLE 5
THE BEST CONFORMATIONS TO SUPERPOSE, WITHOUT USING ANY MOLECULE AS A BASE, AND CONSIDERING MATCHES BETWEEN ALL COMBINATIONS OF PAIRS OF MOLECULES[a]

Structure	DuP 753	SKB 108,566	GLAXO	TAK	L-158,809	SEARLE
rep	7	7	3	5	7	6
r.m.s./Å	2.10	2.33	2.92	2.43	1.53	–

[a] The overall r.m.s. value was 2.35 Å, and the superpositions were calculated with respect to SEARLE.

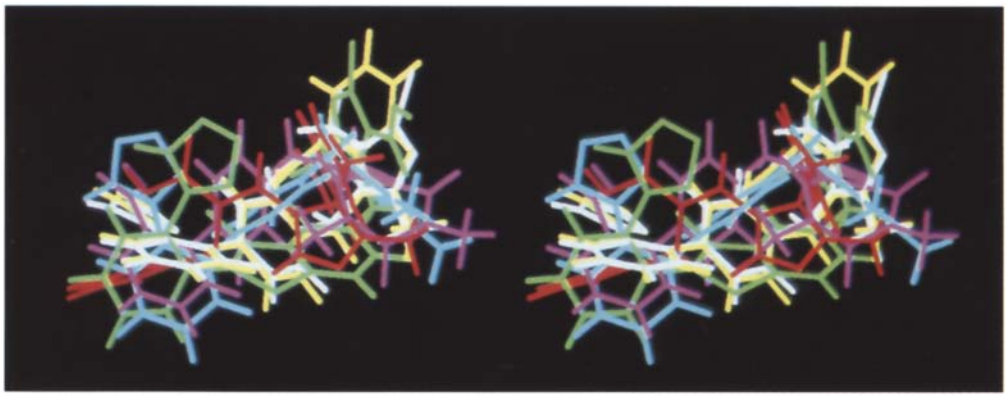

Fig. 6. Colour stereoview with DuP 753 as a base and the remainder of molecules stacked onto it. Molecule colours: white (DuP 753); red (SKB 108,566); green (GLAXO); yellow (TAK); cyan (L-158,809); magenta (SEARLE).

for superposing flexible molecules inevitably attempt to reduce the search of available conformational space.

The fundamental problem in matching flexible structures is how to decide on the number and spread of conformations to try to superpose. The number of possibilities increases with the number of rotatable bonds. In certain cases, flexibility may be restricted severely and bounded distance matrices may have to be used to derive a superposition [35]. This strategy cannot be applied to highly flexible structures. It is possible to achieve a certain reduction of the conformational space by using an expert system, such as WIZARD [36–38], to generate the acceptable conformers and to match all possible conformer pairs between the molecules. We think that this procedure would still produce too many conformers each with slightly different torsion angles, for matching. The method described here provides a drastically different approach. The conformational energies are computed by using a rigid rotor model with the COSMIC force field for expediency. More accurate methods that use a full molecular mechanics calculation are desirable, but not essential to demonstrate the strategy. We sample the entire conformational space by a simulated annealing procedure for the feasible conformations and then seek only those representative conformations in a flexible molecule that are significantly different from each other. However, the conformational search by annealing may not be the best procedure [39]. In general, for the structures studied here with 7–11 rotatable bonds, only a small number of conformations are selected for matching and eventual superposition. Simplification of the problem space on this scale inevitably has a cost – conformations midway between the representatives will be ignored. An extension to the method might be to start from the significantly different representative conformers and explore the match between different molecules by optimizing a search for better superpositions within the cluster, using conformers saved in the history files.

The procedure that we describe does not take into account the variation in the acceptable conformational energy of the conformer used as the representative conformer. Indeed, the energy values shown in Table 2 indicate that a very large spread in conformational energy is present. It is possible that the cut-off in acceptable conformational energy is too large ($\approx$33 kJ mol^{-1}) and a smaller value would be more appropriate. An alternative strategy might be to use the conformational energy as another dimension in the cluster analysis to derive only clusters of low-energy

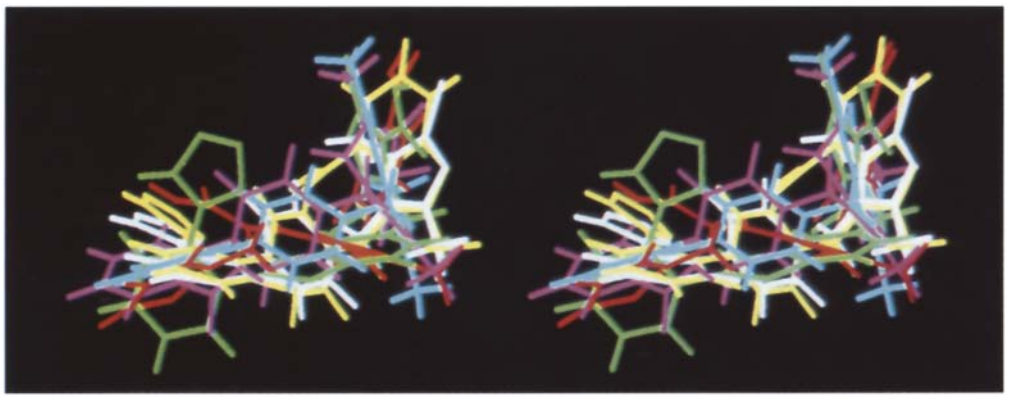

Fig. 7. Colour stereoview of the consensus match, with the molecules fitted onto SEARLE, oriented such that DuP 753 has the same orientation as in Fig. 6. Molecule colours: white (DuP 753); red (SKB 108,566); green (GLAXO); yellow (TAK); cyan (L-158,809); magenta (SEARLE).

conformers, but some sort of weighting scheme would be needed for an appropriate metric for clustering. Reduction of the comparison to low-energy conformers may be too restrictive. There is evidence from conformationally flexible 3D databases that the derived molecular matches are not necessarily those with the nominal low-energy conformations [40]. For this reason, we would hesitate to restrict the conformational space to much below the cut-off value chosen. Nevertheless, further experiments are needed to gain experience with the strategy and explore the cut-off level for acceptance.

One of the problems encountered with cluster analysis is that a decision has to be taken about how many clusters are important. This is an essential element of our strategy, and we used Mojena's stopping rules so that the number of clusters is determined by the data, at a given level of statistical significance. The method thus permits completely automated use of the entire conformation-matching algorithm, and does not force the choice of an arbitrary number of clusters. A note of caution should be introduced. The use of stopping rules to derive a cut-off in significance between clusters is dependent on the distribution of points within the dataset being examined. Inspection of the dendrograms reveals that there are some clearly separate clusters. None-

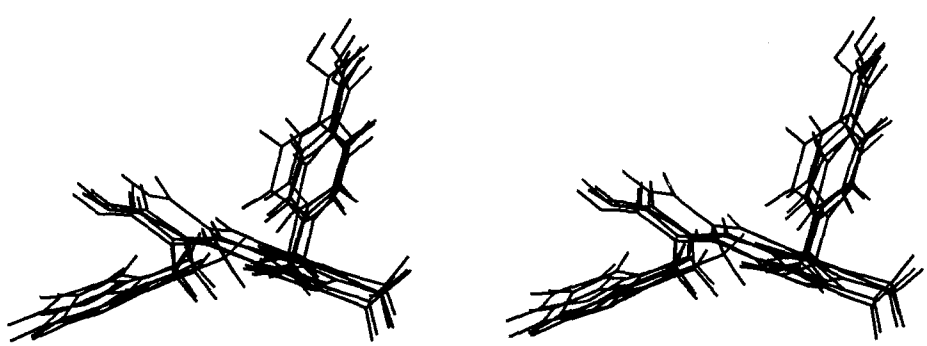

Fig. 8. Stereoview of the range of orientations of SKB 108,566 in the consensus match.

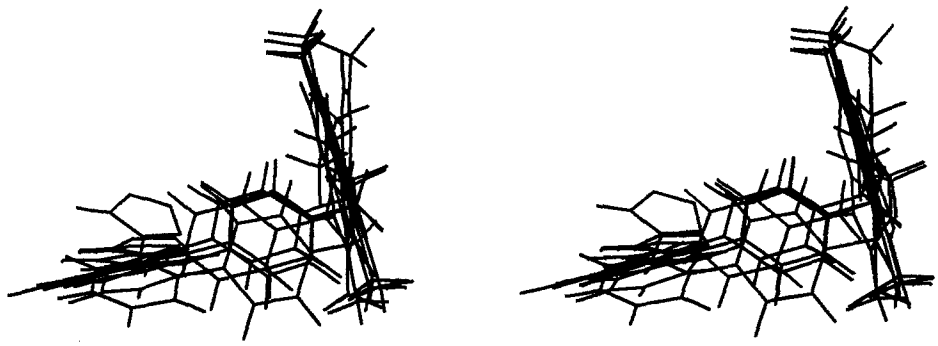

Fig. 9. Stereoview of the range of orientations of L-158,809 in the consensus match.

theless, in some cases smoother transition between agglomerations in the dendrogram was observed and the partition into definite clusters would seem to be more arbitrary. However, some decision has to be made and, in the absence of a better method, the stopping rule does at least have the advantage of enabling a cut-off to be made with a chosen level of significance. A strong argument for retaining this feature is that the procedure definitely identifies the major conformational differences. This ensures that a wide range of conformational space is selected for matching. The minimum variance method for clustering, recommended for use with the stopping rule, has the added advantages that it maximizes cluster compactness and does not violate the ultrametric inequality [41]. We also describe a new and rapid method for detecting different classes of molecular superposition, with different atom sets paired together being revealed directly by cluster analysis. The method is independent of the determination of the transformation matrix for the atom correspondences described in an earlier approach [25] and consequently is not affected by differences in conformation. This procedure may be useful for studying the problem of multiple binding modes in 3D QSAR.

The two types of superposition used here for a set of molecules are the base reference method and the consensus base-independent method. The base reference method is simple to achieve because the corresponding pairs of atoms are defined with respect to the base molecule. This method may be of value if the key pharmacophoric groups are known in one particular molecule and data relative to that molecule is required. However, the consensus method may be of greater potential value for drug design, since it does not bias the search towards any one molecule. The choice of the set of conformers with the minimum value of $\bar{d}_{DDM}$ is relatively straightforward. The 3D superposition is, however, more complex. The r.m.s. value needs to take into account all the pairwise atomic matches in each of the $n(n-1)/2$ alignments between a set of n molecules. The superposition cannot be achieved with either the McLachlan [34] or Gerber and Müller [42] methods because they require a consistent correspondence between the atoms in the set of molecules. For example, consider three molecules, *A*, *B* and *C*. If atom 6 in *A* is paired both with atom 3 in *B* and atom 5 in *C*, then for consistent correspondence atom 3 in *B* should be paired with 5 in *C*. A detailed examination of all correspondences between atom pairs in the six molecules studied here did not show a single example of consistent correspondence. If consistent correspondence could be attained, then superposition of the set of molecules could be made relative to the observed consistent correspondence. In the absence of any consistent correspondence, at the

moment we can only relate the consensus superposition to one molecule of the set. One might expect that if null correspondences had been used in the matching algorithm [24], a more consistent correspondence would have been attained, since poorly matched atom positions would be omitted by the algorithm. The effect of null correspondences on matching flexible molecules has yet to be investigated. Their inclusion in matching should be elementary.

We have presented a novel strategy for the unbiased superposition of a set of flexible molecules on the basis of their atomic positions. This strategy reduces the enormous combinatorial difficulties by applying simulated annealing and cluster analysis to a number of components of the problem. The method has been tested on six angiotensin II antagonists that appear to be structurally diverse, and the results show encouragingly good superpositions involving all six molecules. This algorithm has many potential uses in molecular modelling and drug design, particularly as a tool for molecular alignment for flexible molecules in CoMFA.

ACKNOWLEDGEMENTS

The authors wish to thank Rhône-Poulenc Rorer Central Research, Dagenham, Essex for coordinates. TDJP is a Rhône-Poulenc Rorer Research Fellow, and PMD is a Wellcome Principal Research Fellow. Part of this work was carried out in the Cambridge Centre for Molecular Recognition, supported by the SERC.

REFERENCES

1 Tripos Associates Inc, St Louis, MO, USA.
2 Harpalani, A.D., Egorin, M.J. and Callery, P.S., Abstr. Am. Chem. Soc., 204 (1992) 124.
3 Kim, K.H., Quant. Struct.-Act. Relat., 11 (1992) 127.
4 Kellogg, G.E., Semus, S.F. and Abraham, D.J., J. Comput.-Aided Mol. Design., 5 (1992) 545.
5 Diana, G.D., Kowalczyk, P., Treasurywala, A.M., Oglesby, R.C., Pevear, D.C. and Dutko, F.J., J. Med. Chem., 35 (1992) 1002.
6 Vaz, R.J., Hecht, P. and Kong, S.B., Abstr. Am. Chem. Soc., 203 (1992) 32.
7 Greco, G., Novellino, E., Silipo, C. and Vittoria, A., Quant. Struct. Act. Rel., 10 (1991) 289.
8 Johnson, M.A. and Maggiora, G., Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.
9 Bowen-Jenkins, P.E., Cooper, D.L. and Richards, W.G., J. Phys. Chem., 89 (1985) 2195.
10 Richards, W.G. and Hodgkin, E.E., Chem. Britain, 24 (1988) 1141.
11 Meyer, A.Y. and Richards, W.G., J. Comput.-Aided Mol. Design, 5 (1991) 427.
12 Namasivayam, S. and Dean, P.M., J. Mol. Graph., 4 (1986) 46.
13 Chau, P.-L. and Dean, P.M., J. Mol. Graph., 5 (1987) 97.
14 Dean, P.M. and Chau, P.-L., J. Mol. Graph., 5 (1987) 152.
15 Good, A.C., J. Mol. Graph., 10 (1992) 144.
16 Borea, P.A., Dean, P.M., Martin, I.L. and Perkins, T.D.J., Mol. Neuropharmacol., 2 (1992) 261.
17 Dean, P.M. and Perkins, T.D.J., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings 9th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, in press.
18 Hopfinger, A.J. and Burke, B.J., In Johnson, M.A. and Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, Wiley, New York, 1990, p. 173.
19 Burt, C. and Richards, W.G., J. Comput.-Aided Mol. Design, 4 (1990) 213.
20 Kolossvary, I. and Guida, W.C., J. Chem. Inf. Comput. Sci., 32 (1992) 191.
21 Wilson, S.R., Cui, W., Moskowitz, J.W. and Schmidt, K.E., Tetrahedron Lett., 29 (1988) 4373.
22 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 295.

23 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 317.
24 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 107.
25 Papadopoulos, M.C. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 119.
26 Rusinko, A., Skell, J.M., Balducci, R. and Pearlman, R.S., Abstr. Am. Chem. Soc., 192 (1986) 12.
27 Pople, J.A. and Beveridge, D.L., Approximate Molecular Orbital Theory, McGraw-Hill, New York, 1970.
28 Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., Science, 220 (1983) 671.
29 Vinter, J.G., Davis, A. and Saunders, M.R., J. Comput.-Aided Mol. Design, 1 (1987) 31.
30 Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E., J. Chem. Physics, 21 (1953) 1087.
31 Ward, J.H., J. Am. Stat. Assoc., 58 (1963) 236.
32 Mojena, R., Comp. J., 20 (1977) 359.
33 Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., Numerical Recipes: The Art of Scientific Programming, Cambridge University Press, Cambridge, 1987.
34 McLachlan, A.D., Acta Crystallogr. A, 38 (1982) 871.
35 Streich, W.J., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings 9th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, in press.
36 Leach, A.R. and Smellie, A.S., J. Chem. Inf. Comput., Sci., 32 (1992) 379.
37 Leach, A.R. and Smellie, A.S., Abstr. Am. Chem. Soc., 202 (1991) 35.
38 Leach, A.R., Pesticide Sci., 33 (1991) 87.
39 Goodman, J.M. and Still, W.C., J. Comp. Chem., 12 (1991) 1110.
40 Mason, J., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings 9th European Symposium on Structure-Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, in press.
41 Murtagh, F. and Hecht, A., Multivariate Data Analysis, Reidel, Dordrecht, 1987.
42 Gerber, P. and Müller, K., Acta Crystallogr. A, 43 (1987) 426.