Journal of Computer-Aided Molecular Design, 9 (1995) 479–490
ESCOM



J-CAMD 306

# Molecular surface-volume and property matching to superpose flexible dissimilar molecules

T.D.J. Perkins*, J.E.J. Mills and P.M. Dean

*Drug Design Group, Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.*

Received 13 May 1995
Accepted 2 July 1995

*Keywords:* Molecular similarity; Simulated annealing; Hydrogen bonding; Electrostatic potential; Molecular skin; Drug design

**Summary**

Steric complementarity is a prerequisite for ligand–receptor recognition; this implies that drugs with a common receptor binding site should possess sterically similar binding surfaces. This principle is used as the basis for an automatic and unbiased method that superposes molecules. One molecule is rotated and translated to maximize the overlap between the two molecular surface volumes. A fast grid-based method is used to determine the extent of this overlap, and this is optimized using simulated annealing. Matches with high steric similarity scores are then sorted on the basis of both hydrogen-bond and electrostatic similarity between the matched molecules. Flexible molecules are treated as a set of rigid representative conformers. The algorithm has correctly predicted superpositions between a number of pairs of molecules, according to crystallographic data from ligands that have been co-crystallized at common enzyme binding sites.

## Introduction

Molecular matching is a potentially useful tool for drug design, particularly for cases where receptor sites have not been solved crystallographically. There are two main ways in which the determination of optimal molecular matches and assessment of molecular similarity can be used. First, in lead-compound generation, similarities can be extracted from structurally divergent molecules that bind to the same site (typically competitors' compounds) and the information can be used to design novel compounds. Second, in lead-compound optimization, three-dimensional quantitative structure–activity relationship (3D QSAR) studies using the technique of comparative molecular field analysis (CoMFA) [1] can be used to investigate molecular similarities. This latter procedure is very sensitive to the alignment of the molecules compared; its success is therefore highly dependent on obtaining optimal superpositions [2]. For both approaches, there are often a large number of ways in which the molecules can be superposed; even for apparently very similar molecules, the most obvious superposition may not be optimal. The problem becomes considerably greater if the inherent flexibility of the matched molecules is considered as well. This paper describes a novel method for molecular matching that superposes molecular surface volumes of two molecules. A preliminary report of this work has been presented elsewhere [3].

Previously described molecular matching methods have considered a range of molecular features and properties as a basis for matching. These include hydrogen-bonding atoms and pharmacophoric points [4], best unspecified set of atoms [5–8], molecular electrostatic potential [9,10], electron density [11], molecular volumes [12] and molecular skins [13]. A number of methods have also been proposed that attempt to deal with flexibility, in a variety of ways [14–18]. Each of these methods may prove valuable for particular cases, although many have limitations. Pharmacophore-based methods often require equivalent pharmacophoric groups to be nominated before matching is attempted. This may not be feasible if, for example, the molecules possess a number of equivalent groups, such as a number of hydrogen-bond acceptors.

The concept of matching molecular surface volumes or skins (i.e., a molecular surface with thickness) is an attractive one, since it includes only the area of interaction

*To whom correspondence should be addressed.

between the ligand and its receptor. By an appropriate choice of parameters, superpositions can attempt to mimic the overlay of predicted receptor-atom positions, rather than ligand atoms, during the superposition procedure. This should be particularly useful when matching two structures that have no obvious atom–atom correspondences but that bind to the same receptor with high affinity. Masek et al. [13] have described a method for superposing molecular 'skins' on the basis of volume overlap of the surface skins, in which volumes were calculated analytically. The method was able to match successfully molecules that differed significantly in size. However, the procedure was rather slow, taking 12 days of CPU time to match two serine protease inhibitors, DFKi and TOMI. It is unlikely that such molecules could ever be superposed using total volume methods, since any superposition that placed the smaller molecule entirely within the larger molecule would exhibit a perfect similarity score. In this paper, we describe an alternative approach to molecular surface volume matching that uses the combination of a grid representation of the molecular surface and optimization by simulated annealing to produce a large improvement in speed of the surface volume matching procedures.

*Strategy*

The superposition of two molecules requires optimal alignment of steric, electrostatic and hydrogen-bonding properties of the molecules. Probably the most important feature is steric similarity, since without steric complementarity between the ligand and its receptor, binding is impossible. The method described below considers each of the molecular similarity properties separately, but generates its alignments on the basis of steric similarity. Those matches that possess high steric similarity are then assessed in terms of additional surface features to filter out those that do not have high overall similarity.

The procedures described in this paper use a grid system within the surface volume, which is constructed between the van der Waals surface and a solvent-accessible surface. The methods have been implemented in a FORTRAN 77 program PLM (peptide–ligand match), which uses simulated annealing to rotate and translate one molecule onto the other to maximize the surface volume overlap. The program generates a set of points that define the common matched surface between the two molecules. Molecular surface properties related to both molecules are calculated at these points and are compared to provide the additional surface measures of molecular similarity, given the alignment between the molecules determined by the steric superposition.

Flexibility is dealt with using the FMATCH suite of programs [16–18], in which a small number of representative low-energy conformers are selected using simulated annealing and cluster analysis. These representative structures are then matched combinatorially as rigid molecules.

## Methods

*Steric match*

The procedures described here superpose the smaller of two molecules (defined as that with fewer atoms and designated molecule A) onto the larger molecule (molecule B). Molecule B is first placed in a standard orientation within a box, and a grid system is constructed around it. Each point in the grid is marked as one of INTERIOR_POINT (within the van der Waals surface), EXTERIOR_POINT (outside the outer, solvent-accessible surface), or SURFACE_POINT (within the surface volume). The positioning of these layers is determined by the radius of the 'solvent' probe, which is user-specified with a default value of 1.4 Å. The grid boundaries are chosen such that when the centroid of molecule A is positioned on any part of the van der Waals surface of molecule B, the solvent-accessible surface of molecule A remains within the grid. The surface volume points for molecule A are then selected from a rectangular grid. These surface points are positioned so that their centroid is coincident with that of molecule B, before a random rigid-body transformation is applied to them. The entire system is then translated and scaled to ensure that the grid points (from molecule B) coincide with a unit-spaced grid; this simplifies grid-point identification during assessment of the objective function, since 3D coordinates map directly to grid-point indices. The calculation of the grid system is a relatively CPU-intensive process. In order to speed up multiple matches between the same pair of molecules, a compressed representation of the grid is written to a file, which can be read rapidly for future runs.

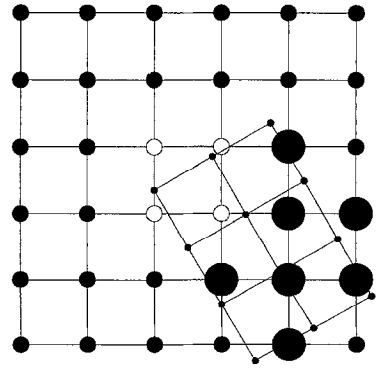

Fig. 1. A simplified representation of the procedure used to determine the extent of surface volume overlap. The horizontal/vertical grid represents the grid for molecule B; filled and open circles represent surface points and interior points, respectively. The tilted grid represents only the surface points on molecule A. For each of these points, the closest grid point of molecule B is determined. If this is a surface point, it is added to the common matched surface, and is represented by a large filled circle.

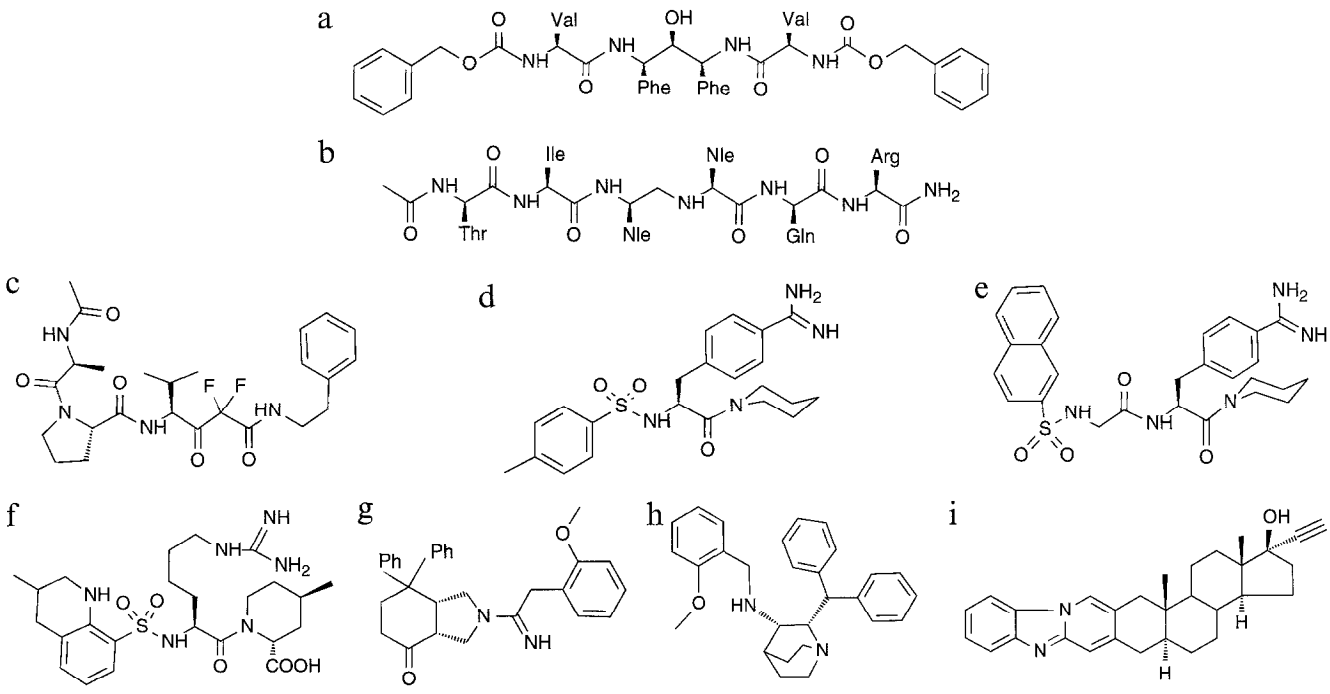


Fig. 2. Chemical structures of the molecules used in matching. (a) A-74704; (b) MVT-101; (c) DFKi; (d) TAPAP; (e) NAPAP; (f) MQPA; (g) RP-67580; (h) CP-96345; and (i) WIN-51708.

The objective function that is optimized by the program is the extent of overlap of the surface volumes between the two molecules; this is determined using a simple counting procedure (see Fig. 1). For each point within the surface volume of molecule A, the closest grid point on molecule B is found as the integer-rounded coordinates of that point. If this grid point is marked SURFACE_POINT, the overlap score is incremented by −1 and the grid point of molecule B becomes part of the common matched surface. The score is negative, by analogy with energy measurements, for the benefit of the simulated annealing algorithm. Interior and exterior points do not contribute towards the score.

Optimization is performed by making random changes in the position of molecule A. These can be either a translation along one of the Cartesian axes, or a rotation around one of the Euler axes. The maximum translational shift allowed along any axis is half the length of molecule B in that dimension. The maximum rotation allowed is $\pi$ radians for axes $a$ and $c$, and $\pi/2$ radians for axis $b$; this covers all rotational space without degeneracy. Any translational move that places the centroid of molecule A outside the van der Waals surface of molecule B (i.e., the closest grid point to the centroid is not marked INTERIOR_POINT) is rejected immediately (before the objective function is calculated), and a new configuration is generated. The choice between a translation or rotation step is made at random but with a user-defined probability, the default being a probability of 0.1 that a translation is chosen (selected as the optimum value from preliminary experiments). The choice of axis for both rotation and translation is made with equal probability for each axis.

The objective function is optimized using a standard simulated annealing algorithm [19], utilizing dynamic cooling [20] and fixed-length Markov chains. Briefly, the difference in the objective function ($\Delta E$) between subsequent configurations is calculated. All changes that have a negative $\Delta E$ are accepted; positive changes are accepted with a probability of $\exp(-\Delta E/T)$, where $T$ is an annealing control temperature. The initial value of $T$ is set such that the initial acceptance ratio is around 0.8 [5]. The value of $T$ remains fixed during each Markov chain; at the end of Markov chain $i$ it is decreased by an exponential factor $f$:

$$T_{i+1} = T_i f \tag{1}$$

The value of $f$ depends on the progress of the algorithm, and in particular the mean value of the objective function $\bar{E}_i$ in Markov chain $i$, and its standard deviation, $\sigma_E$:

$$f = (1 + (T_i \bar{E}_i \ln(1 + \delta))/3\sigma_E)^{-1} \tag{2}$$

where $\delta$ is the temperature decrement factor; this is a small positive number, typically 0.002. Increasing the value of $\delta$ will increase the rate of cooling within the algorithm. By default, the simulated annealing uses 50 Markov chains, each of length 500. It will terminate early if the acceptance ratio falls below 0.002, indicating that a deep minimum has been found.

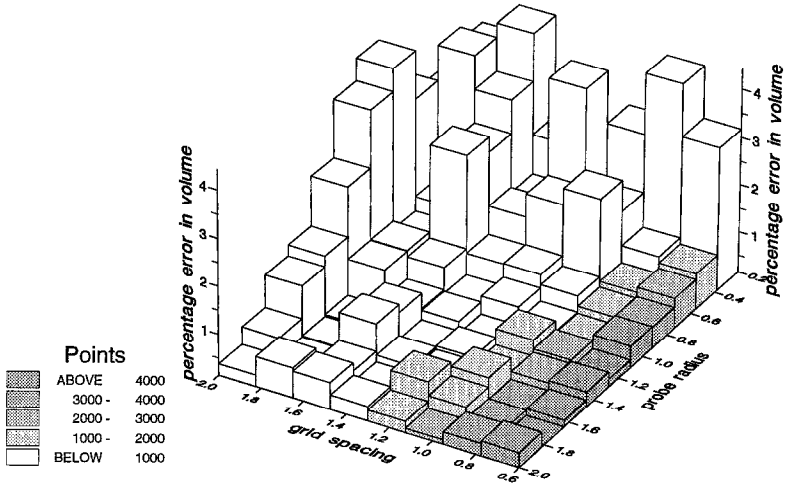


Fig. 3. Graph of percentage error in a surface volume calculation on 20 random start calculations for a match between A-74704 and MVT-101, plotted against both grid spacing (Å) and probe radius (Å), used in the accessible surface calculation.

After the optimization has been completed, the surface around molecule A is regenerated using a grid coincident with that of molecule B, and the similarity score is recalculated. This overcomes any errors that occur owing to the granular nature of grid-based matching. This new grid system is used to generate the final common matched surface. The score is represented as the steric score coefficient ($\xi$), defined as the number of common matched points as a fraction of the number of points in the surface volume of molecule A. This implies that if molecule A is a substructure of molecule B, and has been perfectly superposed onto it, a $\xi$ value of 1.0 should be achieved.

These steric-matching procedures have been implemented in the PLM program. All annealing and grid parameters are under user control via the command-line interface to the program. Separate programs have been written to determine the hydrogen-bonding similarity and electrostatic potential similarity between the matched molecules at the common matched surface, and are described below.

*Hydrogen-bonding similarity*

At each point in the common matched surface, the probability of finding a hydrogen-bond donor or acceptor atom at the receptor is calculated for both molecules. The methods used are a reimplementation for non-protein molecules of the program HSITE [21], and will be described in more detail elsewhere. The program uses data from crystal surveys of intermolecular hydrogen-bond geometry to predict the probability of a receptor hydrogen-bonding atom at any point in space around a ligand.

If the same type of receptor hydrogen-bonding atom (i.e., donor or acceptor) is predicted for both molecules at a particular point in the common matched surface, that point becomes a common matched hydrogen-bonding point and the hydrogen-bonding score ($N_{hbp}$) is incremented by one.

*Electrostatic potential similarity*

For all molecules, partial atomic charges were calculated as described previously [16]. At each point in the common matched surface, the electrostatic potential was calculated for both molecules. Two scoring methods were used to compare electrostatic potentials in two molecules. In the first, banding, method each point in the distribution is defined as either positive, negative or neutral, by ranking the distribution and selecting cutoffs such that both positive and negative tails of the distribution contain 35% of the total number of points. If a point falls into either the positive or the negative band for both molecules, it becomes a common matched electrostatic potential point, and the electrostatic similarity score ($N_{ev}$) is incremented by 1. The second approach calculates the root-mean-square (rms) difference between the electrostatic potential at each equivalent point on the surface, and is termed $R_{ev}$.

*Incorporation of flexibility*

The inherent flexibility of matched molecules is incorporated by considering them as a number of rigid representative conformers. These conformers are generated using the FMATCH procedure, which has been fully described elsewhere [16–18] and is summarized below. The aim of this approach to flexibility is to ensure that the minimum possible number of conformers are selected that cover the feasible conformational space available to the ligand.

In this procedure, the conformational space of the molecules is searched using simulated annealing by rotation around acyclic bonds; for each conformation, the potential energy is calculated using the COSMIC [22] force field. During optimization, a history file is stored that records all conformers encountered. All conformers that have an energy greater than a threshold level (typically 50 kJ mol$^{-1}$) above the global minimum-energy conformation

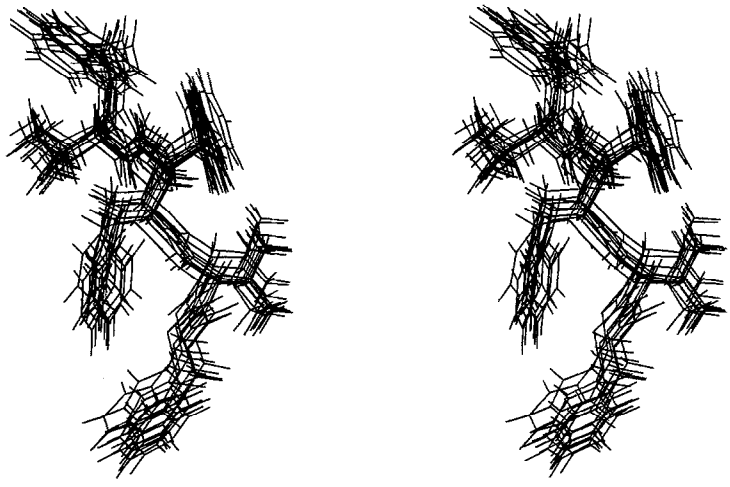

Fig. 4. Stereoview of the 10 resulting orientations of A-74704 obtained by superposition onto MVT-101.

(which is determined by the annealing) are discarded. The remaining conformers are then subjected to cluster analysis using the difference-distance matrix of atom positions as a measure of conformer separation. The clustering procedures incorporate stopping rules, so that the optimum number of clusters is determined automatically. The conformer closest to the centre of each cluster is selected as the representative conformer of that cluster.

*Overall similarity*

A combined ranking scheme was used to select matches that possess high electrostatic and hydrogen-bond similarities in addition to high steric similarity. This was particularly valuable when matching flexible molecules, since many feasible steric matches were generated. The matches were ranked individually on the basis of each of the steric, hydrogen-bonding and electrostatic scores to generate three lists. A sliding threshold scheme was used to select matches for the new, overall ranking. The threshold rank was gradually decreased until a match was found that had each of its steric, hydrogen-bond and electrostatic scores above the threshold. This approach ensures that only superpositions in which all three properties match well are considered to be good matches.

*Molecules matched*

This paper describes molecular matching of molecules from four receptor sites. Chemical structures for all small molecules are shown in Fig. 2. In the case of ligands taken from the Brookhaven Protein Data Bank (PDB) [23], idealized matches were generated by least-squares fitting [24] of the enzyme active sites, and the use of the same transformation matrix to align the ligands.

The HIV protease inhibitors A-74704 and MVT-101 were taken from the PDB files 9HVP and 4HVP, respectively. Main-chain atoms from residues 25–27 in chains A and B (marked as SITE in the PDB file) were used to fit the active site of 9HVP onto that of 4HVP with an rms error in atom positions of 0.18 Å, generating an X-ray crystal alignment between A-74704 and MVT-101.

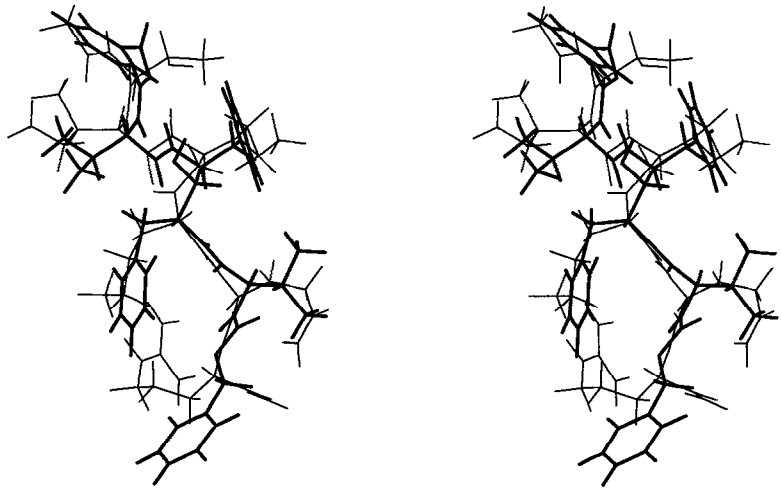

Fig. 5. Stereoview of the best matched orientation of A-74704 (thick lines) superposed onto MVT-101 (thin lines).

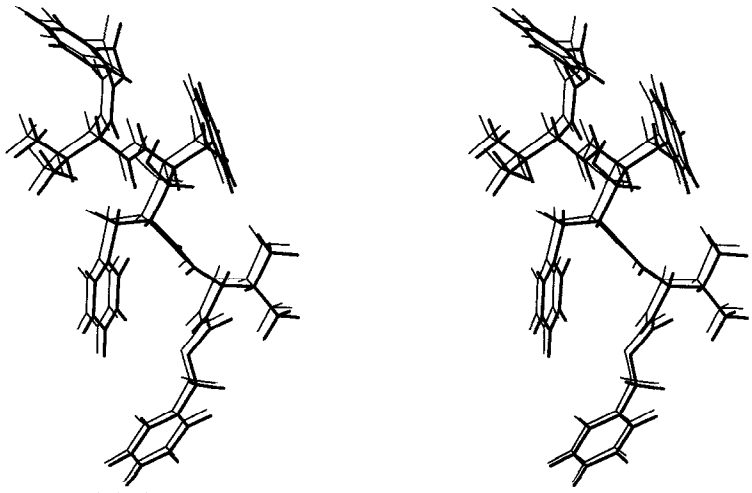

Fig. 6. Stereoview comparison of the best matched orientation of A-74704 (thick lines) with its crystal orientation (thin lines).

The serine protease inhibitors Ac-Ala-Pro-Val-difluoro-*N*-phenylethylacetamide (DFKi) and the third domain of turkey ovomucoid inhibitor (TOMI) were taken from the PDB files 4EST and 1PPF, respectively. The main-chain atoms from residues 41, 57, 102, 189–195, 213–216 and 226–228 (rms 0.34 Å) [13] were used to align the enzymes and hence their inhibitors.

Three thrombin inhibitors, $N^{\alpha}$-(4-toluene-sulfonyl)-DL-*m*-amidinophenylalanyl-piperidine (TAPAP), $N^{\alpha}$-(2-naphthyl-sulfonyl-glycyl)-DL-amidinophenylalanyl-piperidine (NAPAP) and (*2R, 4R*)-4-methyl-1-[$N^{\alpha}$-3-methyl-1,2,3,4-tetrahydro-8-quinolinesulfonyl)-L-arginyl]-2-piperidine carboxylic acid (MQPA), were taken from the PDB files 1ETT, 1ETS and 1ETR, respectively. Using the main-chain atoms from residues 189–195, 214–220, 57 and 102 (by analogy with the serine protease inhibitors), both 1ETS and 1ETT were fitted to 1ETR, with rms deviations of 0.26 and 0.03 Å, respectively. This generated the experimental alignments of TAPAP and NAPAP onto MQPA.

Three tachykinin $NK_1$ receptor antagonist molecules were also matched using these procedures. There are no crystal structures available that define their relative orientations, nor any that describe their 3D structures. These molecules were therefore modelled using available fragments from the Cambridge Structural Database (CSD) [25]. RP-67580 was based on the CSD structure SINXUI [26], with additional parts added using SYBYL v. 6.1 [27], followed by 200 cycles of Powell/simplex optimization [28] within SYBYL. Similarly, CP-96345 was built from KIYRAL [29] with the same treatment in SYBYL. WIN-51708 was built by combining the two CSD structures WEGREV [30] and VEWZAY [31]. Four equivalent atoms were overlaid and the structures were merged without minimization.

Any CPU timings given relate to a single run on a Silicon Graphics Indy R4000SC, compiled with optimization level 2 and mips2 flags.

## Results

*Optimum grid-system parameters*

One of the major problems when using any grid-based system is choosing an appropriate size for the grid spac-

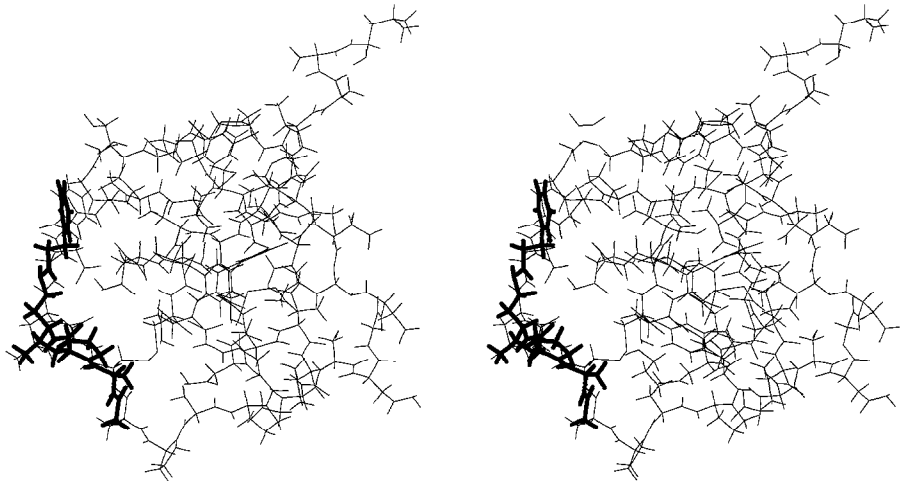

Fig. 7. Stereoview of the best match of DFKi (thick lines) superposed onto TOMI (thin lines).

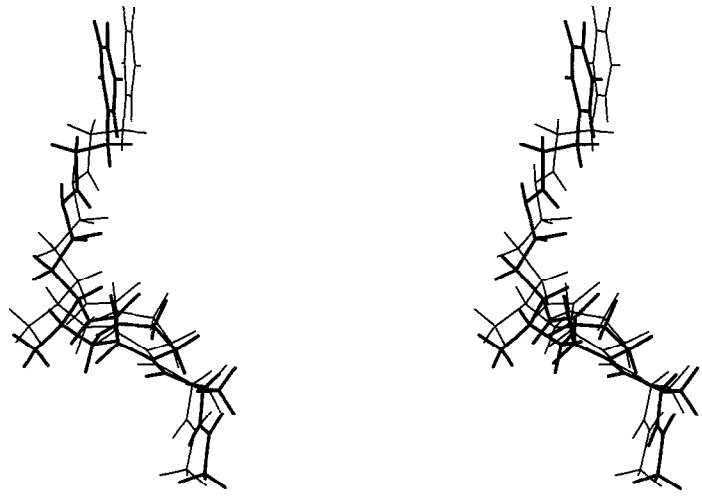

Fig. 8. Stereoview comparison of the best matched orientation of DFKi (thick lines) with its crystal orientation (thin lines).

ing. Too coarse a grid will ignore subtle features, but too large a grid will increase the number of points (for an $n$-fold decrease in the grid spacing, the number of points increases by a factor of $n^3$). The other important variable in a molecular surface volume calculation is the probe radius. In an attempt to determine optimum values for these variables, the surface points were calculated for the HIV protease inhibitor A-74704 over a range of values for both parameters. Each calculation was repeated 20 times, with a random displacement in the $x$-, $y$- and $z$-axes so that 20 different grid systems were built around the molecule. The mean ($\bar{n}$) and standard deviation ($\sigma_n$) of the number of points were calculated; the percentage error in the volume calculation was then represented as $\sigma_n / \bar{n} \times 100\%$. This number is a measure of the repeatability of the surface volume calculation and is plotted against both grid size and probe radius in Fig. 3. As expected, the combination of a small probe radius and a coarse grid results in less accuracy in volume calculation. There appears to be a plateau having fairly good accuracy when the probe radius is greater than 1 Å and the grid spacing is less than 1.4 Å. The values selected for the remainder of the calculations in this paper were a probe radius of 1.4 Å (as a compromise value that mimics a water molecule probe) and a grid size of 1.0 Å, giving an error in the surface volume calculation of less than 0.5%.

*HIV protease inhibitors*

The X-ray crystal conformations of A-74704 and MVT-101 were matched 10 times, fitting the smaller molecule, A-74704, onto MVT-101. The computation required around 60 s of CPU time for each run (an additional 10 s was required to calculate the grid system for MVT-101 for the first run). The 10 resulting alignments are essentially the same (see Fig. 4); the orientation of A-74704 for the match with the highest score ($\xi = 0.58$) (see Fig. 5) was compared with the crystal orientation, and had an rms difference in atom positions of 0.64 Å (see Fig. 6). For the X-ray crystal alignment, the steric score $\xi$ also had a value of 0.58. Hydrogen-bond and electrostatic potential comparisons were not made, since the resulting orientations of A-74704 were so similar.

*Serine protease inhibitors*

These molecules provide an example where the matched molecules differ considerably in size: TOMI has roughly 10 times as many atoms as DFKi. Again, the program was run 10 times using the rigid structures in their crystal conformations; a Markov chain length of 1000 was used because of the increased size of the translational search space. The CPU time was around 95 s (plus 120 s to create the grid system for TOMI on the

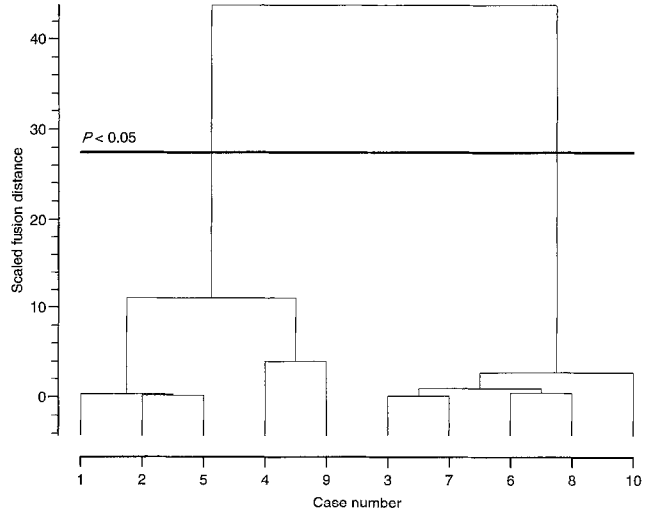


Fig. 9. Dendrogram illustrating the clustering of the 10 orientations of DFKi, using the rms difference in equivalent atom positions as the measure of similarity between matches. The solid horizontal line represents the cutoff that generates significantly different clusters with $p < 0.05$.

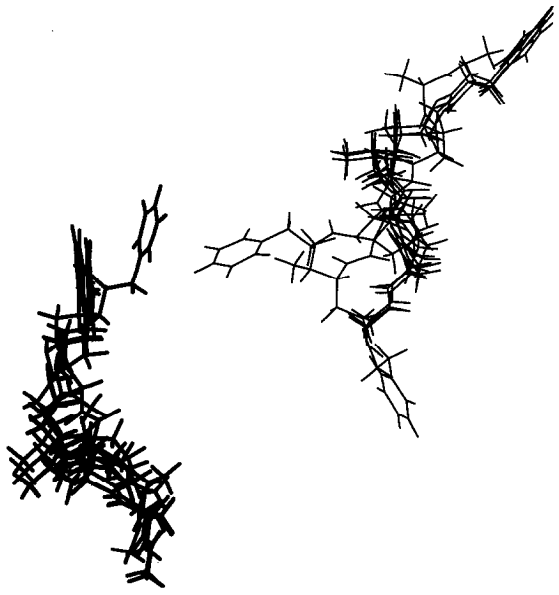

Fig. 10. The 10 matched orientations of DFKi illustrated in the relative positions found on the surface of TOMI; those in cluster 1 are shown with thin lines, those in cluster 2 with thick lines. The two clusters are well separated; within each cluster there are only small differences in orientation. See Table 1 and the text for an explanation of the clusters.

TABLE 1
STERIC AND ELECTROSTATIC SCORES FOR 10 MATCHES OF DFKi ONTO TOMI

| Match | Cluster | $\xi$ | $N_{ev}$ | $R_{ev}$ | $N_{hbp}$ |
|---|---|---|---|---|---|
| 1 | 1 | 0.57 | 80 | 326.5 | 36 |
| 2 | 1 | 0.59 | 120 | 324.5 | 27 |
| 5 | 1 | 0.58 | 112 | 320.8 | 31 |
| 4 | 1 | 0.56 | 45 | 321.7 | 37 |
| 9 | 1 | 0.55 | 63 | 312.8 | 32 |
| 3 | 2 | 0.66 | 88 | 308.5 | 45 |
| 7 | 2 | 0.64 | 93 | 308.0 | 43 |
| 6 | 2 | 0.61 | 73 | 295.8 | 36 |
| 8 | 2 | 0.55 | 78 | 303.6 | 32 |
| 10 | 2 | 0.58 | 86 | 281.5 | 48 |
| X-ray | | 0.62 | 77 | 311.6 | 35 |

'Cluster' is the cluster number into which the match is placed, $\xi$ is the steric score, $N_{ev}$ is the number of common electrostatic potential points, $R_{ev}$ is the rms difference in the electrostatic potential of the matched points, and $N_{hbp}$ is the number of common hydrogen-bonding points. 'X-ray' shows the data for the X-ray alignment. The order of the matches is that shown in the cluster diagram in Fig. 9.

first run). The best match of DFKi onto TOMI had a steric score $\xi$ of 0.66 (see Fig. 7); when compared with the crystal alignment of DFKi (see Fig. 8), it had an rms difference in atom positions of 0.75 Å. The 10 alignments are not identical: cluster analysis on the basis of the rms differences between the atom positions of all matched orientations of DFKi, combined with the use of stopping rules, indicates that there are two significantly different clusters (see Fig. 9). The 10 matches are shown, by cluster, in Fig. 10.

Electrostatic potential similarities were calculated for each match, using the two methods of calculation defined above, and the results are shown in Table 1. By comparing this table with the cluster diagram, it appears that the $R_{ev}$ measure of similarity in electrostatic potential between equivalent points is a better indication of the cluster membership than $N_{ev}$, the number of similar electrostatic points. For the former, the values are consistently lower for cluster 2; this is the cluster that contains the orientations close to the crystal orientation. The number of common hydrogen-bonding points, $N_{hbp}$, is also shown in Table 1; this value is generally higher for matches that belong to cluster 2.

*Thrombin inhibitors*

These molecules were first subjected to a conformational analysis using FMATCH to generate a number of representative conformations for each molecule. The difference-distance matrix of all atoms was used as the measure of distance between conformations, and conformers within 50 kJ mol$^{-1}$ of each global minimum-energy conformation were retained. The numbers of representative conformers generated were: 12 for MQPA; 11 for TAPAP; and 9 for NAPAP. Each conformation of one molecule was then matched 10 times with each conformation of the matched molecule. Both TAPAP and NAPAP were fitted onto the largest molecule, MQPA.

For each combination of conformers, the match with the highest steric score from the 10 replicates was selected

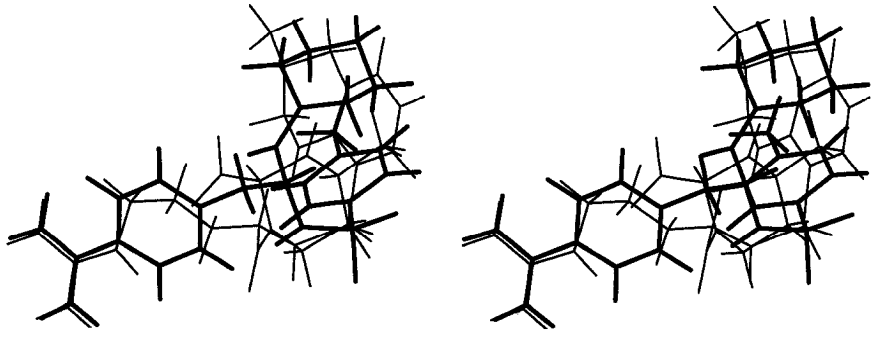

Fig. 11. Stereoview of the best overall match of TAPAP (thick lines) onto MQPA (thin lines).

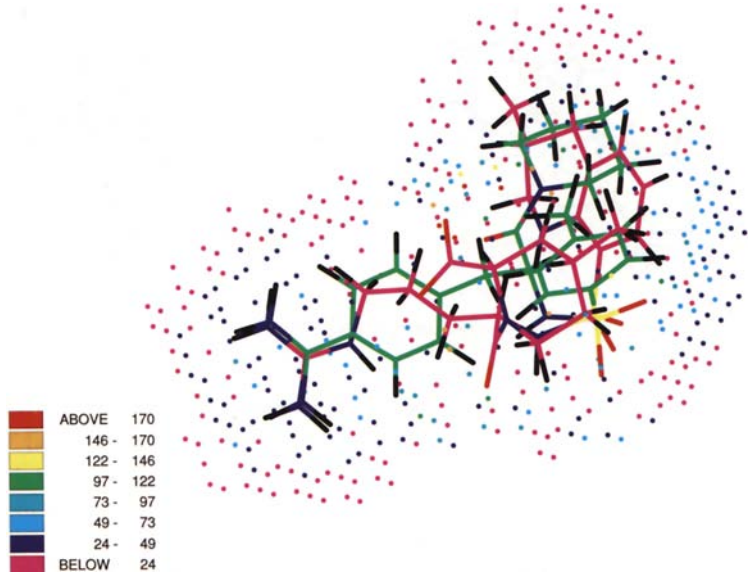


Fig. 12. Best overall match between TAPAP (green carbon atoms) and MQPA (purple carbon atoms), together with the electrostatic rms difference map (in kJ mol$^{-1}$). Heteroatom colours: blue = N; red = O; yellow = S; and black = H.

and the hydrogen-bond and electrostatic potential similarity scores were calculated. The matches were then ranked by each of the similarity criteria. A combined ranking procedure was then used to identify matches that scored well on all criteria. In essence, a match must be above a particular rank for each parameter under consideration to be included in the new ordered list.

For TAPAP matched with MQPA, the match with the best overall score according to the combined ranking procedure is shown in Fig. 11; it is ranked 11th in the steric score, 7th in the $N_{hbp}$ hydrogen-bond score, 4th in the $N_{ev}$ electrostatic potential number score, and 1st in the $R_{ev}$ electrostatic potential rms score. The majority of the common surface has a small difference in the electrostatic potential of the two molecules (see Fig. 12), and a number of separate regions of hydrogen-bonding atoms at the receptor are predicted (see Fig. 13). Conformational differences between the matched and crystal structures for both TAPAP and MQPA make simple atom–atom quantitative comparisons of the computational and experimental alignments of TAPAP inappropriate. However, the match has a $\xi$ value of 0.64, and appears to be qualitat-

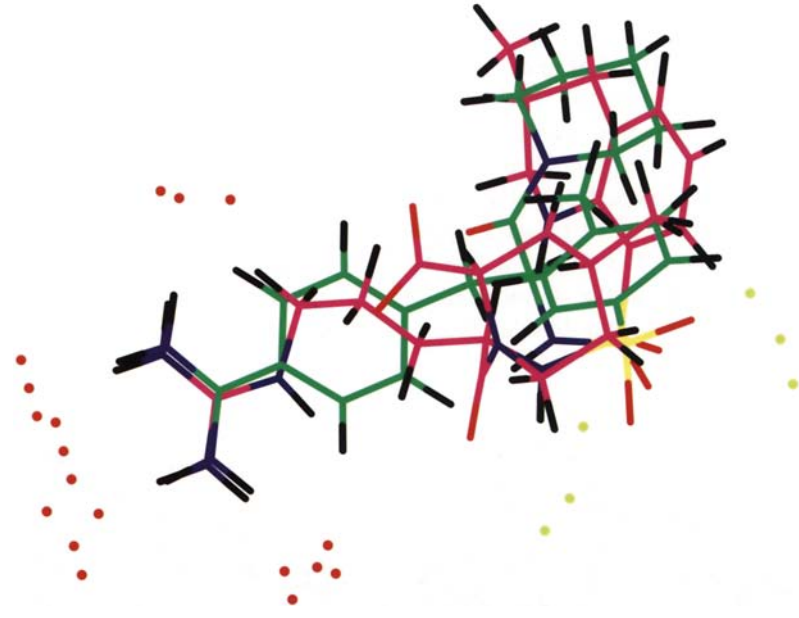

Fig. 13. Best overall match between TAPAP (green carbon atoms) and MQPA (purple carbon atoms), together with the common hydrogen-bond predicted points. Predicted receptor hydrogen-bond donor atom positions are shown in red and acceptor atom positions in green. Heteroatom colours are as listed in the legend to Fig. 12.

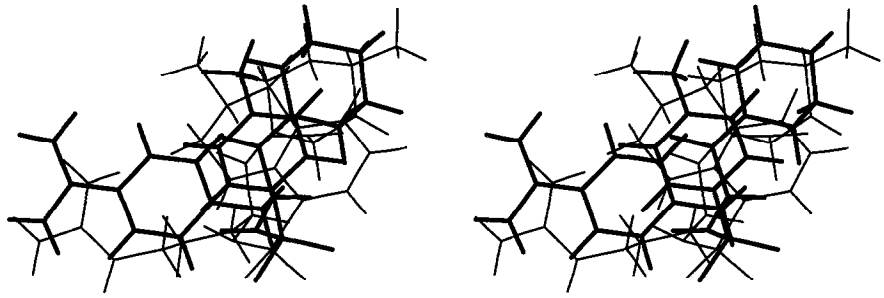

Fig. 14. Stereoview of the crystal alignment of TAPAP (thick lines) and MQPA (thin lines).

ively similar to the crystal alignment ($\xi = 0.67$), which is shown in Fig. 14 for comparison.

The match with the best steric score had a $\xi$ of 0.67. However, visual comparison revealed little chemical similarity; this match also scored poorly on the hydrogen-bond and electrostatic potential similarity scores.

For NAPAP and MQPA, the match with the best overall score is shown in Fig. 15. This match has a $\xi$ score of 0.61 (compared with 0.59 for the crystal alignment), and appears 2nd in the steric rank, 10th in the hydrogen-bonding rank and 3rd in the electrostatic rms rank. Again, the functional groups that are superposed are the same as in the crystal overlap.

*Tachykinin $NK_1$ receptor antagonists*

The three $NK_1$ antagonists CP-96345, RP-67580 and WIN-51708 differ from the structures discussed previously in that there are no crystal data that illustrate their relative orientations at the $NK_1$ receptor. There is some information from site-directed mutagenesis as to the residues involved in binding [32,33], although there appear to be major differences between rat and human forms of the receptor.

Conformational analysis of RP-67580 and CP-96345 was carried out in the same way as for the thrombin inhibitors. This produced 11 representative conformers for both molecules; WIN-51708 is a rigid structure. Both CP-96345 and RP-67580 were matched onto the largest molecule of the set, WIN-51708. The best overall match between RP-67580 and WIN-51708 is shown in Fig. 16. In this match, RP-67580 is placed at the steroid 'end' of WIN-51708, and two common hydrogen-bonding points were found, indicating that the OH group of RP-67580 could form a hydrogen bond to the same receptor atom as the OH group of WIN-51708. The $\xi$ value is 0.53, i.e., significantly less than that obtained with the thrombin inhibitors. This difference is presumably related to the unmatched phenyl ring in RP-67580.

Figure 17 illustrates the best overall match between CP-96345 and WIN-51708, which also has a $\xi$ score of 0.53. No common hydrogen-bonding points were found, although this match scores very well on the $R_{ev}$ score; this may be due to the generally nonpolar nature of these compounds. As with the RP-67580 match, it is the steroid end of WIN-51708 that is superposed in the match.

## Discussion and Conclusions

The ability to determine unbiased superpositions between dissimilar molecules remains an important, and only partially solved, problem for drug design and 3D QSAR. The problem can be divided into two components: first, defining a measure of similarity for a particular overlap, and second, searching through all the possible superpositions to find the global maximum similarity. This paper presents a method of determining similarity based on both steric and chemical features of a surface volume common to the two matched molecules.

It is a common maxim that steric complementarity is a prerequisite for ligand binding, but that hydrogen-bond and electrostatic complementarity, together with hydrophobic similarity, are also required for specific high-affinity ligands. The methods described here attempt to follow these principles by first searching for superpositions that

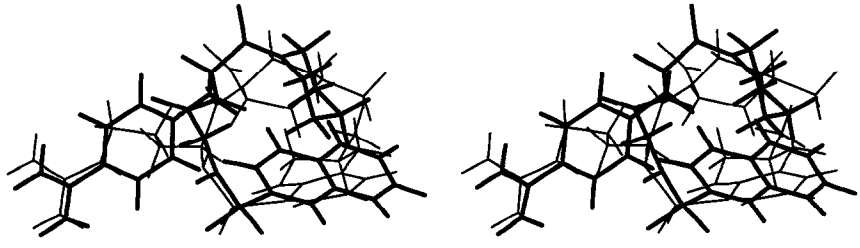

Fig. 15. Stereoview of the best overall match between NAPAP (thick lines) and MQPA (thin lines).

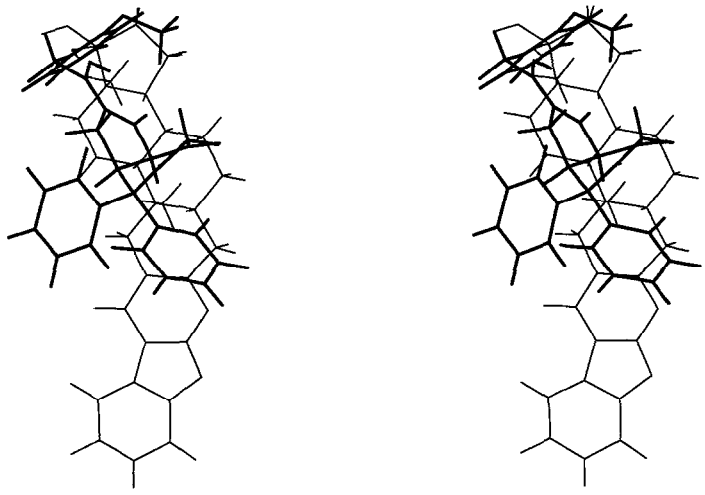

Fig. 16. Stereoview of the best overall match of RP-67580 (thick lines) onto WIN-51708 (thin lines).

exhibit good steric similarity. These are then filtered to leave only those that also show high hydrogen-bonding and electrostatic similarity between the matched molecules.

An alternative strategy would be to use the steric, hydrogen-bond and electrostatic similarity scores to obtain a composite score, and to optimize this value. This, however, has the problem that scaling factors would have to be devised to relate the three values. These factors would have to normalize the values to a common scale (such as interaction energy), but also describe the relative importance of each; the latter factor may well differ between different receptor sites. To sidestep these issues, we have adopted a simpler approach that applies a filter to superpositions with high steric similarity, removing those with poor hydrogen-bond or electrostatic similarities.

The calculation of steric similarity described here uses a relatively simple grid system to allow the rapid lookup of surface point coordinates in terms of grid positions. This provides a significant increase in speed compared with the more analytical procedures employed in the surface volume matching described by Masek et al. [13]. It might be expected that this speed may have to be paid for with a loss of accuracy. However, the results in this paper have demonstrated that the algorithm can reproduce alignments found from X-ray crystal structure information. It is significant that, in nearly all cases, the alignment with the highest $\xi$ score from a set of repeated annealing runs is that closest to the X-ray experimental alignment. This suggests that the grid-based method is sufficiently accurate to be used for molecular matching; the precise alignment can be modified, either manually or by alternative methods if necessary.

The search through all possible matches in Cartesian space is a combinatorial process in six dimensions, i.e., three rotational and three translational dimensions. Other procedures have used distance-matrix methods [5–8]. However, the use of Cartesian coordinates allowed the use of a rapid grid lookup procedure, offsetting the time required for calculation of the transformed coordinates.

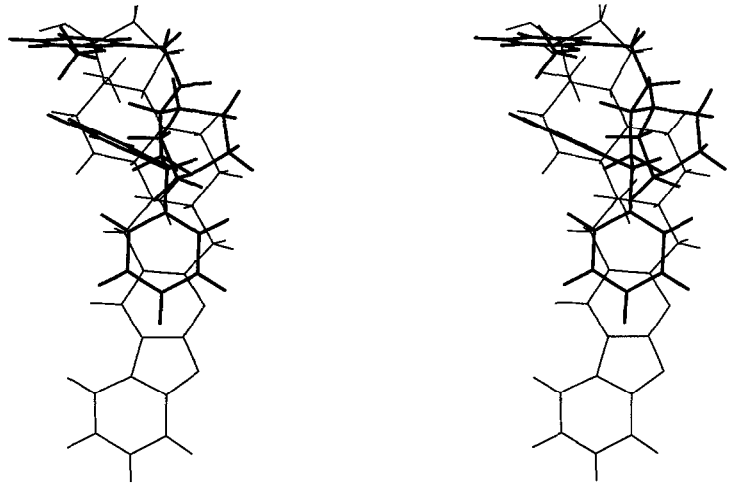

Fig. 17. Stereoview of the best overall match of CP-96345 (thick lines) onto WIN-51708 (thin lines).

One of the major problems with any molecular matching procedure is its handling of conformational flexibility. This work has used the concept of representative conformers, using our previously developed methods [16–18]. In the use of representative conformers, a balance always has to be found between a large number of conformers, giving good coverage of conformational space but forcing a very large number of matches to be carried out, and a smaller number of conformers that may miss some important areas of conformational space. Future methods should perhaps allow a combined conformational/matching search, or use a larger number of representative conformers, perhaps with a filter to rapidly remove unpromising conformer pairs. Naturally, the problem is eased if additional information, such as pharmacophore distances of rigid analogues, is available.

One of the advantages of the procedures described in this paper is that a range of feasible superpositions are presented to the user. These may include alignments that would be overlooked both by manual superposition procedures and by methods that attempt to find a single best superposition through modification of conformations during matching. This ability to uncover multiple potential superpositions should also be valuable when dealing with sets of compounds that bind to more than one receptor, but with different relative orientations; one of the forms of multiple binding modes. In conclusion, we have presented a novel and rapid method for the superposition of dissimilar and flexible molecular structures of differing sizes. The method uses a grid-based procedure to maximize the overlap between the surface volume of the matched molecules.

## Acknowledgements

The authors wish to thank Rhône-Poulenc Rorer (T.D.J.P., J.E.J.M.) and the Wellcome Trust through the PRF scheme (P.M.D.) for personal financial support. We wish to acknowledge the use of EPSRC's Chemical Database Service at Daresbury. Part of this work was carried out in the Cambridge Centre for Molecular Recognition, funded by the BBSRC.

## References

1 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
2 Kim, K.H., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie, Glasgow, 1995, pp. 291–331.
3 Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., In Sanz, F. (Ed.) Trends in QSAR and Molecular Modelling 94 (Proceedings of the 10th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling), Prous, Barcelona, 1995, in press.
4 Danziger, D.J. and Dean, P.M., J. Theor. Biol., 116 (1985) 215.
5 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 295.
6 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 317.
7 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 107.
8 Papadopoulos, M.C. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 119.
9 Chau, P.-L. and Dean, P.M., J. Mol. Graphics, 5 (1987) 97.
10 Dean, P.M. and Chau, P.-L., J. Mol. Graphics, 5 (1987) 152.
11 Meyer, A.Y. and Richards, W.G., J. Comput.-Aided Mol. Design, 5 (1991) 427.
12 Masek, B.B., Marchant, A. and Matthew, J.B., J. Med. Chem., 36 (1993) 1230.
13 Masek, B.B., Marchant, A. and Matthew, J.B., Proteins, 17 (1993) 193.
14 Hodgkin, E.E., Miller, A. and Whittaker, M., J. Comput.-Aided Mol. Design, 7 (1993) 515.
15 Burt, C. and Richards, W.G., J. Comput.-Aided Mol. Design, 4 (1990) 231.
16 Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 7 (1993) 155.
17 Dean, P.M. and Perkins, T.D.J., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings of the 9th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, pp. 207–215.
18 Perkins, T.D.J., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie, Glasgow, 1995, pp. 89–109.
19 Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., Science, 220 (1983) 671.
20 Van Laarhoven, P.J.M. and Aarts, E.H.L., Simulated Annealing: Theory and Applications, Reidel, Dordrecht, 1987.
21 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, Ser. B, 236 (1989) 101.
22 Vinter, J.G., Davis, A. and Saunders, M.R., J. Comput.-Aided Mol. Design, 1 (1987) 31.
23 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasmui, M., J. Mol. Biol., 112 (1977) 535.
24 McLachlan, A.D., J. Mol. Biol., 128 (1979) 49.
25 Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M. and Watson, D.G., J. Chem. Inf. Comput. Sci., 31 (1991) 187.
26 Smits, J.M.M., Bosman, W.P., Beurskens, P.T., Thijs, L., Magnus, E. and Zwanenburg, B., J. Cryst. Spectrosc., 16 (1986) 333.
27 SYBYL, v. 6.1, Tripos Associates Inc., St. Louis, MO.
28 Lowe III, J.A., Drozda, S.E., Snider, R.M., Longo, K.P. and Bordner, J., Bioorg. Med. Chem. Lett., 1 (1991) 129.
29 Clark, M., Cramer III, R.D. and Van Opdenbosch, N., J. Comput. Chem., 10 (1989) 982.
30 Dorokhov, V.A., Gordeev, M.F., Dem'yanets, Z.K., Vorontsova, L.G., Kurella, M.G., Andrianov, V.I. and Rekhlova, O.A., Bull. Acad. Sci. USSR, Div. Chem. Sci., 40 (1991) 2262.
31 Christiansen, R.G., Bell, M.R., D'Ambra, T.E., Mallamo, J.P., Herrmann, J.L., Ackerman, J.H., Opalka, C.J., Kullnig, R.K., Winneker, R.C., Snyder, B.W., Batzold, F.H. and Schane, H.P., J. Med. Chem., 33 (1990) 2094.
32 Sachais, B.S. and Krause, J.E., Mol. Pharmacol., 46 (1994) 122.
33 Fong, T.M., Yu, H., Cascieri, M.A., Underwood, D., Swain, C.J. and Strader, C.D., J. Biol. Chem., 269 (1994) 14957.