Journal of Computer-Aided Molecular Design, 11 (1997) 229–242.

J-CAMD 396

An automated method for predicting the positions of hydrogen-bonding atoms in binding sites

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

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

Received 10 December 1996
Accepted 6 February 1997

Keywords: Ligand–receptor interaction; Geometry of intermolecular interactions; Molecular superposition; Receptor model; Drug design

Summary

Hydrogen bonds are the most specific, and therefore predictable of the intermolecular interactions involved in ligand–protein binding. Given the structure of a molecule, it is possible to estimate the positions at which complementary hydrogen-bonding atoms could be found. Crystal-survey data are used in the design of a program, HBMAP, that generates a hydrogen-bond map for any given ligand, which contains all the feasible positions at which a complementary atom could be found. On superposition of ligands, the overlapping regions of their maps represent positions of receptor atoms to which each molecule can bind. The certainty of these positions is increased by the incorporation of a larger number and diversity of molecules. In this work, superposition is achieved using the program HBMATCH, which uses simulated annealing to generate the correspondence between points from the hydrogen-bonding maps of the two molecules. Equivalent matches are distinguished on the basis of their steric similarity. The strategy is tested on a number of ligands for which ligand–protein complexes have been solved crystallographically, which allows validation of the techniques. The receptor atom positions of thermolysin are successfully predicted when the correct superposition is obtained.

Introduction

If the three-dimensional structure of a binding site is unknown, information about the site must be gleaned from the structures of ligands known to bind to it. Salient features of the ligands are extracted by molecular-similarity studies and ligands containing the same features are searched for in databases [1,2] or constructed *de novo* [3] by algorithms. Alternatively, these features are used to construct a model of the site. Site-directed drug-design techniques can then construct *de novo* or identify from databases a number of complementary ligands, which can be put forward as candidates for novel leads.

Molecular superposition

Ligand-based drug design relies on obtaining a superposition for the set of molecules studied. Superposition is most often achieved by manually overlaying selected hydrogen-bonding atoms and/or aromatic rings. While this superposition technique probably suffices for members of congeneric series, it is not practicable when the molecules are ostensibly dissimilar. Less biased methods have been used to superpose molecules on the basis of either atom positions [4–7], molecular volumes [8], molecular skins [9,10], electrostatic potential correlation coefficients [11], electrostatic extrema [12] or hydrogen-bonding site points [13,14]. Any attempt to superpose molecules using a combination of these properties [15,16] is fraught with the problems of weighting, which are usually problem-dependent and therefore not applicable to the general case [17]. In this work, information concerning the hydrogen-bonding properties of the receptor is sought, so molecules are superposed on the basis of their hydrogen-bonding properties.

Hydrogen-bonding and molecular similarity

The hydrogen bond is the predominant intermolecular interaction involved in many ligand–protein complexes. Its properties of strength and directionality provide the anchorage necessary for ligand binding. Indeed, superposition of hydrogen-bonding atoms is often the method of choice even when, for example, deriving information

*To whom correspondence should be addressed.

about electrostatic similarity [18], calculating similarity coefficients [19] or using CoMFA [20]. The problem of unbiased superposition of hydrogen-bonding atoms has been approached by Danziger and Dean [4], using a branch-and-bound search method. However, in order for molecules to bind to the same site, the hydrogen-bonding atoms need not be superposed, but only project to the same positions in the site. In the program AUTOFIT [13], each hydrogen-bonding group is represented by a site position, the optimum position for a receptor atom, together with a vector representing the direction of the hydrogen bond. Each possible permutation and combination of points is superposed by least-squares methods in order to determine the optimum superposition. DISCO [14] also calculates one site point for each hydrogen-bonding group (two or three for rotatable groups) and uses these in addition to aromatic ring centroids as a basis for superposition. However, recent crystal surveys of hydrogen bonds in small molecules [21,22] and ligand–protein hydrogen bonds [21] show that hydrogen bonds vary a great deal in their geometry, although most commonly occurring in the orthodox directions. This suggests that representing the hydrogen-bonding characteristics of a group with one site point is inadequate. A more realistic approach is provided by HSITE [23,24], which represents the hydrogen-bonding properties of proteins with maps consisting of shells of points about each hydrogen-bonding group. These points are assigned probability values based on crystal-survey data and represent positions at which ligand hydrogen-bonding atoms would be expected to be found. This concept has been adapted in this work to provide the basis for hydrogen-bond map superposition.

Strategy

Given the structure of a ligand, the possible positions of complementary hydrogen-bonding atoms on the receptor can be predicted. In this work, crystal-survey data [22] have been used in the design of a program, HBMAP, that generates hydrogen-bond maps for any given ligand. Each hydrogen-bonding group gives rise to a large number of potential site points, representing the positions of atoms to which the ligand could form a hydrogen bond. Following superposition onto another molecule, the overlapping regions of their maps represent the possible positions of atoms on the receptor to which both ligands can form hydrogen bonds. The introduction of more ligands into the superposition refines these possible positions to give a more accurate prediction of the relative locations of the receptor atoms, which can then be used in *de novo* structure generation algorithms or database searches in the design of novel lead compounds.

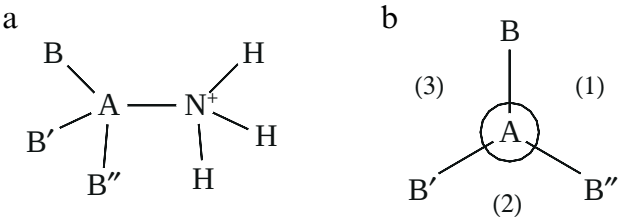

Fig. 1. Torsion faces used to describe the geometry of hydrogen bonds to rotatable regions, illustrated by (a) the charged primary amine group. In the Newman projection down the A-N bond (b), the torsion faces (1), (2) and (3) are delineated by the lines joining A to B, B′ and B″. Ideally, the primary amine adopts a staggered conformation and a hydrogen atom is positioned at the centre of each torsion face.

Although loosely based on the program HSITE [23,24], HBMAP differs fundamentally in a number of ways. First, the generated maps consist of clouds of points rather than shells, to allow a more complete representation of the 3D distribution of hydrogen bonds. Secondly, HBMAP recognizes and generates maps for all of the hydrogen-bonding groups of interest to the drug designer, rather than being restricted to the groups present in proteins. Thirdly, the method for probability calculation differs. In HSITE, the calculations were based on 1D frequency distributions, relying on the assumption that the distributions in the three dimensions were independent, an assumption which has been shown to be invalid [22]. In contrast, HBMAP makes use of 3D crystal-survey information to calculate hydrogen-bond probability values.

Molecular superposition is carried out using a novel program, HBMATCH, to superpose the hydrogen-bond maps. This incorporates a simulated annealing algorithm to optimize the selection of the appropriate combinations and permutations of points and thereby superpose the maps. Alternative superpositions that give rise to equal hydrogen-bonding similarity are subsequently distinguished by their steric similarity.

Molecules used in the study

The superposition program, HBMATCH, was tested with pairs of ligands that bind to three different sites: *retro*-thiorphan and thiorphan, which bind to thermolysin; folate and methotrexate, which bind to dihydrofolate reductase; and N^{α}-(4-toluene-sulphonyl)-DL-*m*-amidinophenylalanyl-piperidine (TAPAP) and (2*R*,4*R*)-4-methyl-1-[N^{α}-(3-methyl-1,2,3,4-tetrahydro-8-quinolinesulphonyl)-L-arginyl]-2-piperidine carboxylic acid (MQPA), which bind to thrombin. Site-point derivation was attempted using six ligands that bind to thermolysin: *N*-{1-(2(*R*,*S*)-carboxy-4-phenylbutyl)-cyclopentylcarbonyl}-(*S*)-tryptophan (CCT); *N*-(*S*)-(1-carboxy-3-phenylpropyl)-(*S*)-Leu-(*S*)-Trp (CLT); *N*-phosphoryl-L-leucinamide (NPL); HONH-benzyl-malonyl-L-Ala-Gly-*p*-nitroanilide (HBAGN); Val-Trp (VT); and carbobenzoxy-Gly^{P}-(*O*)-Leu-Leu (ZGPLL). In each case, the coordinates of the ligands were taken from the crystal structures of protein–ligand complexes in order to allow validation of the results, both in terms of superposition and the site points obtained.

All CPU times are given for a single run on a Silicon

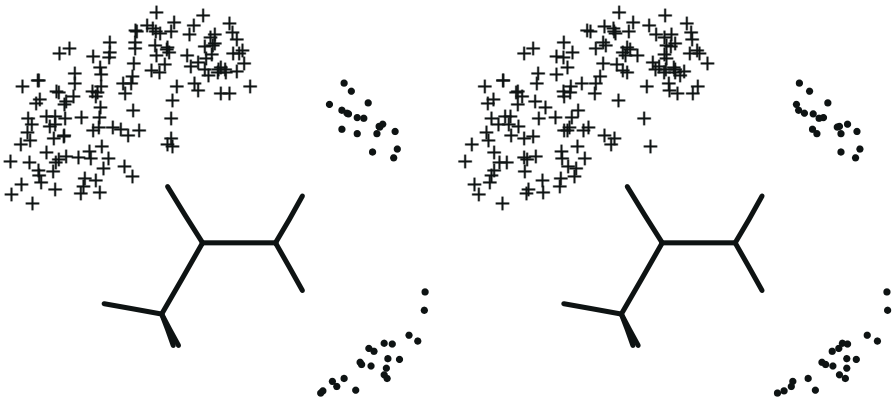

Fig. 2. Stereoview of the hydrogen-bond map for ethanamide. Donor regions are shown as crosses and acceptor regions as dots.

Graphics Indy R5000, compiled with optimization level 2 and mips2 flags.

Methods

Hydrogen-bond map generation

For a given ligand, all hydrogen-bonding atoms are identified and classified according to previously derived criteria [22]. Each hydrogen-bonding atom is placed at the centre of an 8 Å cube and a number of points (usually 3000) are randomly plotted within this cube to generate a hydrogen-bond region, which represents the possible positions of a complementary receptor atom bonding to the ligand atom. Two simple bump checks are used to preclude the majority of these points for steric reasons. First, if a point is closer to the hydrogen-bonding atom than the closest observed contact in the crystal survey, it is discarded. Secondly, since the points represent the positions of complementary hydrogen-bonding heavy atoms, they cannot approach closer than the sum of the van der Waals radius of a nitrogen or oxygen atom and that of any ligand atom. However, points are permitted to approach complementary hydrogen-bonding atoms. For example, a point in an N-H region is permitted to be close to the H atom of a C-H group. Such a point represents the position of a receptor atom forming multicentre hydrogen bonds with the two donor groups of the ligand and certainly should not be discarded. C-H donor groups are included amongst these secondary contacts but are not considered as primary hydrogen-bonding groups in map generation because they would otherwise dominate the superposition procedure and receptor model, through sheer weight of numbers, despite being very weak hydrogen-bonding groups.

A sterically accessible point is only considered as a point in the hydrogen-bond map if the hydrogen-bond probability value for the point is nonzero. In the crystal survey [22], the space about each hydrogen-bonding group was divided into bins of equal volume such that the number of observed hydrogen-bonding atoms in each bin represented the probability of finding a hydrogen-bonding atom in that bin. The probability value is then simply calculated by locating the bin in which the point is positioned and looking up the normalized number of crystal-survey observations within it.

The points about sp^2-hybridized oxygen atoms are divided into two regions, one for each lone pair. However, this is not possible when the oxygen atom is bonded to a sulphur or phosphorus atom because the lone pair positions are unknown, so in these cases all the points remain in one region. The space about the rotatable charged primary amine group can be divided into three torsion faces, as illustrated in Fig. 1. Each torsion face is derived on the premise that it contains either a hydrogen atom or a lone pair attached to the hydrogen-bonding heteroatom. In this case, each torsion face contains a hydrogen atom, so there are three hydrogen-bonding regions, each corresponding to a torsion face. Three torsion faces are also set up for the points of rotatable hydroxyl and amine groups and two torsion faces are set up for the phenol and carboxylic acid hydroxyl groups. However, separate regions are not defined by individual torsion faces because the position of each hydrogen atom with respect to the torsion faces is unknown. In each of these cases, there is one donor and one acceptor region

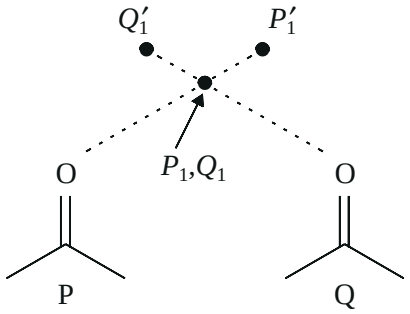

Fig. 3. Schematic of possible superposition for hydrogen-bond maps of two molecules of propanone, P and Q, in which points P_1 and Q_1 are superposed. P_1' and Q_1' are created by projecting P_1 and Q_1 further away from their hydrogen-bonding atoms by a user-defined distance.

with points distributed throughout the faces, so in effect terminal hydrogen atoms are allowed to rotate.

Some ligands do not contain many hydrogen-bonding groups, so there is an option to generate aromatic regions in addition to donor and acceptor regions. Two points are placed above and below the plane of the ring such that their midpoint lies at the centroid of the ring. This is used more as an aid for superposition than as a representation of the hydrogen-bonding properties of aromatic rings. Although aromatic rings have been proposed to act as hydrogen-bond acceptors [25] and have been shown to partake in ligand–receptor interactions [26], the interaction has been shown to be weak [25] and, as has been observed in preliminary observations in a crystal survey (data not shown), nondirectional with respect to the ring. Aromatic rings play a more important role in terms of stacking interactions [27] and cation–π interactions [28], justifying their incorporation into the superposition.

Hydrogen-bond map superposition

When superposing hydrogen-bond maps, donor regions must be superposed onto donor regions and acceptors onto acceptors (and aromatics onto aromatics if this option is selected). Intuitively, this should perhaps entail searching for the largest possible volume of overlap of the regions. However, the degree of region overlap turns out to be irrelevant because the hydrogen-bonding atom on the receptor can only occur at one point (assuming 1:1 hydrogen-bond formation). Any attempt to overlap whole regions would be equivalent to overlaying the hydrogen-bonding atoms, which is not necessary for the formation of hydrogen bonds to the same receptor atoms. Instead, single points representing each region are superposed. This is a combinatoric problem, involving selection of the correct point to represent each region and the superposition of the points using the optimum correspondence. In addition, there is the problem of null correspondences; not all of the regions are involved in the optimum overlay so the correct subset of regions (donor and acceptor) must be selected. Hence, in moving from one configuration to another, there are three possible types of transition: a change of correspondence between the current selection of points; a change in the selection of regions from either molecule; and a change in the selection of the representative point from any of the currently selected regions in either molecule.

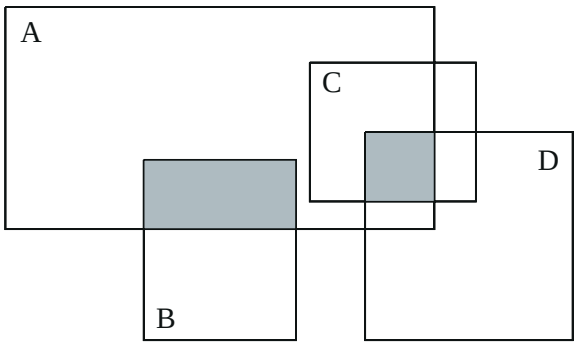

Fig. 4. Schematic of superposition of hydrogen-bond regions from four molecules, A–D. Three maps overlay in the shaded region on the right, so the points within have a tag value of 3 and the region contains contributions from A, C and D. Only two maps overlap in the shaded region on the left, so the points have a tag value of 2, which is the highest value for the points from B only.

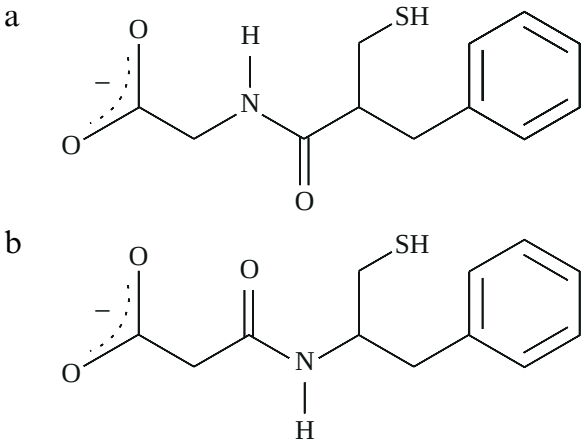

Fig. 5. The structures of (a) thiorphan and (b) *retro*-thiorphan.

Superposition of the maps by the program HBMATCH is based on the atom-superposition algorithm of Barakat and Dean [5], in which simulated annealing generates the optimum correspondence between points by minimizing the sum of the elements of the difference-distance matrix (DDM). The map for molecule A is represented by a distance matrix, M^A, given by

$$M_{ij}^A = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2} \qquad (1)$$

where M_{ij}^A is the distance between points i and j, and x, y and z are the Cartesian coordinates of the points. For a given selection of points from maps of molecules A and B, the objective function, E, becomes

$$E = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \left| M_{kl}^A - M_{mn}^B \right| \qquad (2)$$

where k and l are the points from the map for molecule A being used in the ith and jth pair of superposed points, respectively. The points m and n from the map for molecule B are the points corresponding with k and l, respectively. N is the number of matching pairs of points. A least-squares superposition procedure [29] uses the resultant optimum correspondence to create the rotation matrix and translation vector. Previous work has shown that dynamically varying the degree of configuration change improves the efficiency of annealing [30]. Fewer large con-

TABLE 1
RESULTS OBTAINED ON MATCHING THE HYDROGEN-BOND MAPS OF *RETRO*-THIORPHAN AND THIORPHAN, WITH NO NULL REGIONS; AROMATIC REGIONS ARE NOT INCLUDED IN THE MATCH

Trial	1	2	3	4	5	6	7	8	9	10	Crystal
D, A[a]	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2
ξ[b]	0.700	0.777	0.776	0.795	0.755	0.768	0.755	0.837	0.734	0.755	0.857
Rmsd[c] (Å)	1.092	0.621	0.642	0.516	0.694	0.596	0.720	0.337	0.744	0.618	

[a] D is the number of matching donor regions and A is the number of matching acceptor regions.
[b] ξ is the steric overlap score, as calculated by PLM.
[c] Rmsd represents the root-mean-square deviation between the superposition and the crystal superposition.

formational changes should occur as the annealing proceeds, to allow the algorithm to settle towards the global minimum. Interregion swaps involve much greater changes in configuration than intraregion swaps, so the ratio of interregion to intraregion swaps is reduced in proportion with the acceptance ratio during the course of the annealing.

Systematic problems arise when hydrogen-bond maps are to be superposed. Firstly, superposition of the points does not take into account hydrogen-bond directionality with respect to the receptor. As Fig. 2 (a typical hydrogen-bond map) shows, the distribution of complementary hydrogen-bonding atoms about donor groups is much tighter; there is much more directionality about donor groups. As illustrated in Fig. 3, although two donor regions (from an acceptor atom) may overlap at the same point (P_1 and Q_1), a donor heavy atom at that position may not be able to form hydrogen bonds to both molecules because the donor hydrogen atom cannot be positioned such that its hydrogen-bond map encompasses both acceptor atoms (unless the donor group is rotatable, which cannot be assumed). For this reason, each point P is projected out from its hydrogen-bonding atom to generate another point, P'. If point P_1 is superposed onto point Q_1, P'_1 must also be superposed onto Q'_1. Therefore, there are two distance matrices for each molecule, one for the map (M) and one for the projected points (M'), and the objective function becomes

$$E = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \left|M_{kl}^{A} - M_{mn}^{B}\right| + \left|M_{kl}'^{A} - M_{mn}'^{B}\right| \qquad (3)$$

The degree by which the directionality is taken into account is determined by the distance by which the points are projected. Large distances of projection encourage the acceptor atoms to superpose. Preliminary experiments showed a distance of 1 Å to be large enough to deter the algorithm from generating spurious superpositions but small enough to prevent biasing it towards the superposition of the hydrogen-bonding atoms. The elements in M' for acceptor regions are repeats of the equivalent elements in M. This prevents HBMATCH from biasing the selection of points towards those from acceptor regions, which would otherwise only have one DDM contributing to the objective function. The second problem is encountered when superposing regions from rotatable amphiprotic groups. For example, a rotatable hydroxyl group is considered to accept and donate one hydrogen bond, and if two hydrogen bonds are formed, the complementary atoms must be positioned in two different torsion faces because the lone pair and hydrogen atom are in different torsion faces.

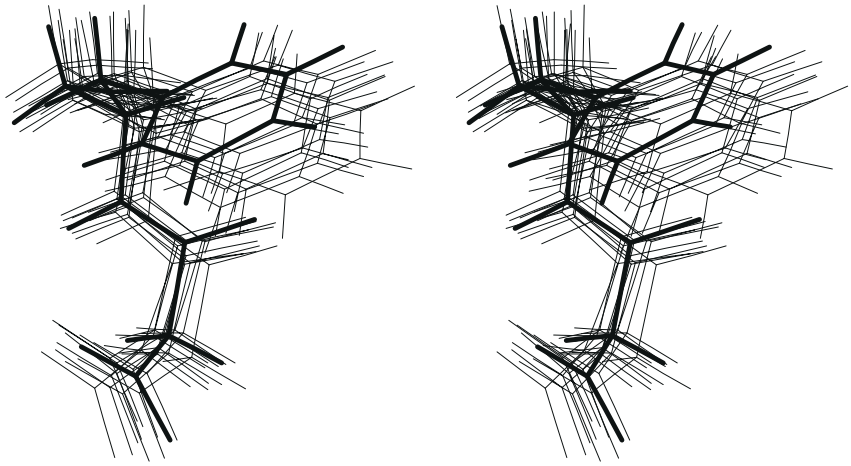

Fig. 6. Stereoview of the 10 orientations of *retro*-thiorphan generated by superposing its hydrogen-bond map onto that of thiorphan without the inclusion of aromatic regions. The crystal orientation of *retro*-thiorphan is shown in bold.

TABLE 2
RESULTS OBTAINED ON MATCHING THE HYDROGEN-BOND MAPS OF *RETRO*-THIORPHAN AND THIORPHAN, WITH NO NULL REGIONS; AROMATIC REGIONS ARE INCLUDED IN THE MATCH

Trial	1	2	3	4	5	6	7	8	9	10	Crystal
D, *A*[a]	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2	7, 2
ξ[b]	0.828	0.828	0.826	0.864	0.816	0.817	0.837	0.832	0.851	0.851	0.857
Rmsd[c] (Å)	0.359	0.439	0.389	0.340	0.358	0.354	0.282	0.299	0.242	0.261	

[a–c] As in Table 1.

The optimum match should possess good steric and hydrogen-bonding similarity. The steric similarity is calculated by the program PLM [10] as the fraction of the surface volume of the smaller molecule that overlaps with the surface volume of the larger molecule. The hydrogen-bond score is the number of regions in the resultant superposition that overlap. The match with the lowest sum of the ranks for the steric and hydrogen-bond scores is deemed the best match.

Two alternatives present themselves when the superposition of more than two molecules is considered. In the first, all the molecules are superposed onto a base molecule. The best base molecule is determined by carrying out the superposition for each possible base molecule and assessing each resulting superposition. In the alternative method, each molecule is superposed onto its most similar counterpart. Invariably, this will not lead to a superposition of all the molecules in the set, because, for example, if molecules A and B are superposed and molecules C and D are superposed, there is no information on how to incorporate all four molecules into the same superposition. In such cases, the two most similar molecules that are not currently superposed are overlaid. If B and C were the next two most similar molecules, they would be overlaid and the chain of superposition would run A-B-C-D. Both approaches have been tested in this study.

Hydrogen-bond site point generation

Having superposed two hydrogen-bonding maps, the overlapping regions of the maps represent the possible positions of receptor atoms that could form hydrogen bonds to both molecules. Although the process of superposition has identified a number of points that contribute to similar DDMs, there could be a different number of regions that actually overlap when the maps are superposed. If the points selected by the minimization did not superpose well, it is possible that their respective regions would not overlap. Conversely, if there were many null regions, regions that actually overlap would be excluded from the DDM as null correspondences. Hence, overlapping points are determined by searching through the superposed maps for points from different maps that are less than a threshold distance apart. The threshold distance for overlapping points is taken as the mean distance between nearest-neighbour points in the map. For points to be considered as overlapping, their projected counterparts must also be within the threshold distance for projected points.

When more than two maps are overlaid, using the regions where all the maps overlap would not take into account the fact that different molecules may form hydrogen bonds to a different subset of receptor groups. Hence, each map is considered individually and each overlapping point is tagged with the number of other maps with

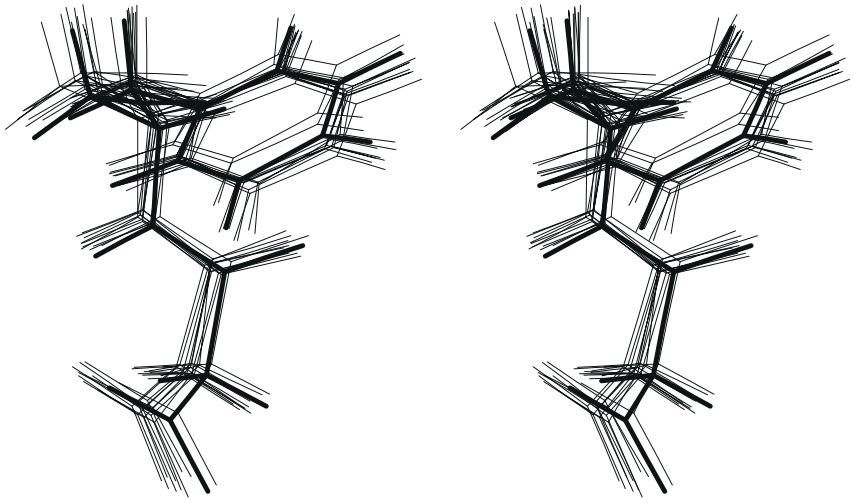

Fig. 7. Stereoview of the 10 orientations of *retro*-thiorphan generated by superposing its hydrogen-bond map onto that of thiorphan with the inclusion of aromatic regions. The crystal orientation of *retro*-thiorphan is shown in bold.

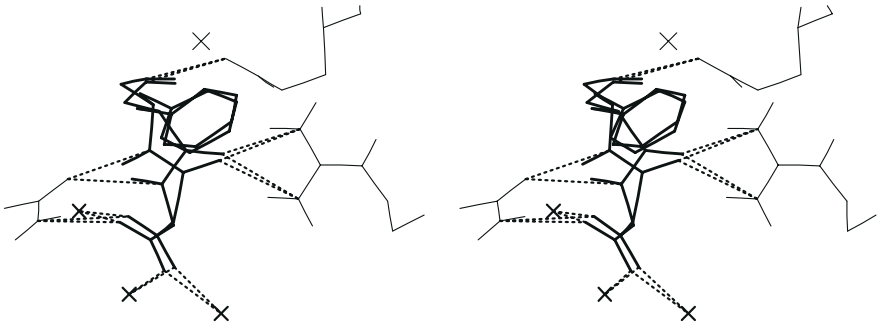

Fig. 8. Stereoview of the crystal orientations of *retro*-thiorphan and thiorphan, shown with the thermolysin binding site (feint). Hydrogen bonds between ligand and site are shown as dotted lines between the heavy atoms.

which it overlaps. From each region, only the points with the highest value of this tag are chosen as overlapping points. As a result of this method, a number of subregions are defined, as shown in schematic form in Fig. 4. This number of subregions is reduced by only including those that contain points from more than one map.

Results

Thiorphan and retro-*thiorphan*

Hydrogen-bond maps were constructed from the crystal coordinates [31] of thiorphan and *retro*-thiorphan (Fig. 5), each map containing seven donor regions, two acceptor regions and one aromatic region. These maps were superposed by HBMATCH, using no null regions, with and without use of the aromatic regions. The results of 10 trials without the incorporation of aromatic regions are shown in Table 1 and in Fig. 6. Each trial took around 25 s CPU time. In terms of hydrogen-bonding similarity, the 10 trials were indistinguishable because all nine regions overlaid in each case. The best match was identified as that giving rise to the highest steric similarity, in this case trial 8, which was also the closest match to the crystal orientation. The results of the 10 trials in which aromatic rings were incorporated are shown in Table 2 and Fig. 7. Again, each trial took around 25 s. The inclusion of aromatic rings fine-tuned the superposition, so an accurate solution was obtained consistently. However, the quality of the best solution was no better than when aromatic regions were ignored. Although the best match according to the steric score was not the same as the best match in terms of root-mean-square deviation (rmsd) from the crystal orientation, all 10 orientations could be considered to be correct, given that the crystal structures of other thermolysin binding sites overlap with an rmsd value of about 0.3 Å.

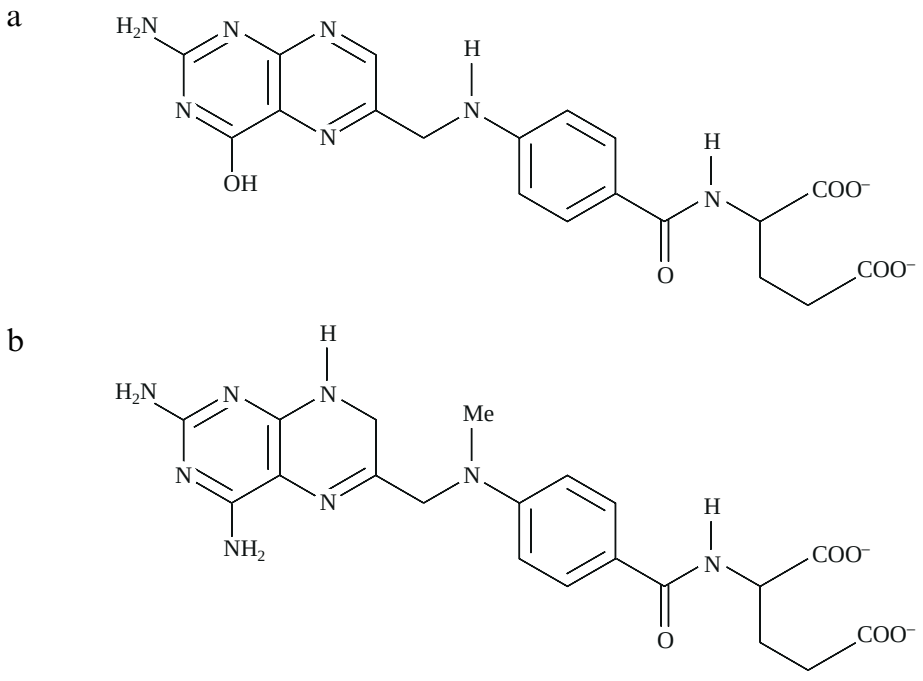

Fig. 9. The structures of (a) folate and (b) methotrexate.

TABLE 3
TEN BEST RESULTS OBTAINED ON MATCHING THE HYDROGEN-BOND MAPS OF FOLATE AND METHOTREXATE

Rank	D, A[a]	ξ[a]	n[b]	Rmsd[c] (Å)
1	11, 3 (8)	0.723 (5)	5	0.540
2	11, 3 (8)	0.697 (12)	6	0.690
3	11, 4 (2)	0.668 (19)	5	0.644
4	10, 4 (8)	0.689 (14)	6	0.892
5	11, 4 (2)	0.662 (21)	6	0.925
6	10, 4 (8)	0.680 (16)	6	0.644
7	10, 4 (8)	0.679 (17)	5	0.634
8	11, 3 (8)	0.665 (20)	3	0.744
9	11, 4 (2)	0.647 (27)	5	0.850
10	11, 3 (8)	0.660 (22)	1	0.873
Crystal	10, 3	0.740		

[a] Numbers in parentheses indicate the rank for the score.
[b] n represents the number of null regions used to obtain the match.
[c] Rmsd represents the root-mean-square deviation between the superposition and the crystal superposition.

These ligands are a good example of two molecules whose hydrogen-bonding maps superpose despite the fact that the hydrogen-bonding atoms are not coincident. Superposition of the molecules on the basis of merely their hydrogen-bonding-atom positions yields an rmsd value of 1.3 Å from the crystal orientation if the heavy atoms are used in the superposition, 1.2 Å if the amide hydrogen atom is used instead of the amide nitrogen atom, and 1.2 Å if both the amide nitrogen and hydrogen atoms are used. These results are significantly worse than the rmsd value of 0.3 Å obtained by superposing their hydrogen-bond maps. In the crystal orientation (Fig. 8), the nitrogen atoms of the amide groups of the two molecules are about 1.2 Å apart, yet they form hydrogen bonds to the same site atom. The carbonyl atoms are much closer (0.2 Å in the crystal orientation and 0.4 Å in the best-match orientation) to each other. This is interesting in the light of the earlier observation that acceptor atoms must be positioned more consistently than donor atoms because hydrogen-bond directionality is more strict about receptor donor groups (Fig. 2).

Folate and methotrexate

The atom coordinates of folate and methotrexate (Fig. 9) in the conformations with which they bind to *Escherichia coli* dihydrofolate reductase (DHFR) in the absence of cofactor were extracted from the Brookhaven Protein Databank (PDB) [32] files with reference codes 1dyi and 4dfr, respectively. Each of their hydrogen-bond maps contained 14 donor and five acceptor regions. Ten trials were carried out for each of 0–6 null regions, each trial taking approximately 60 s. The best 10 matches from all 70 trials are shown in Table 3, and all are within 1 Å of the crystal orientation, which was generated by superposing all the heavy atoms from residues 5–8, 27, 28, 31, 32, 44–46, 50, 52, 54, 57, 94, 96, 100 and 113 (rmsd value 0.3 Å). The best match (Fig. 10) had an rmsd value of 0.54 Å from the crystal orientation, which was not much larger than the 0.3 Å obtained on superposing the site atoms. The top 10 matches all contained more overlapping hydrogen-bonding regions than the crystal orientation, which is to be expected given that the procedure aims to maximize this quantity. The incorporation of aromatic regions did not improve the superposition, either in terms of hydrogen-bond and steric score or rmsd from the crystal orientation (data not shown).

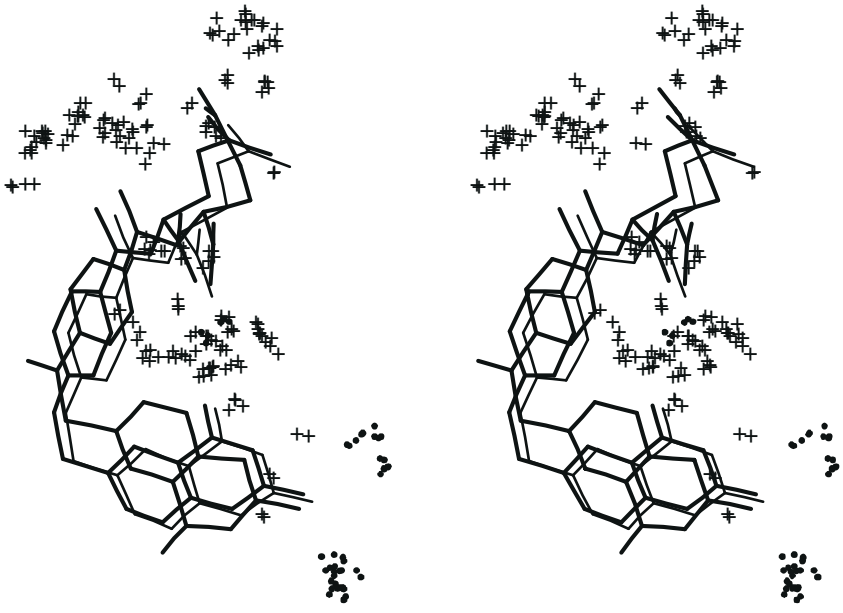

Fig. 10. Stereoview of the best match between folate and methotrexate and their overlapping hydrogen-bonding points. The crystal orientation of folate is shown in feint.

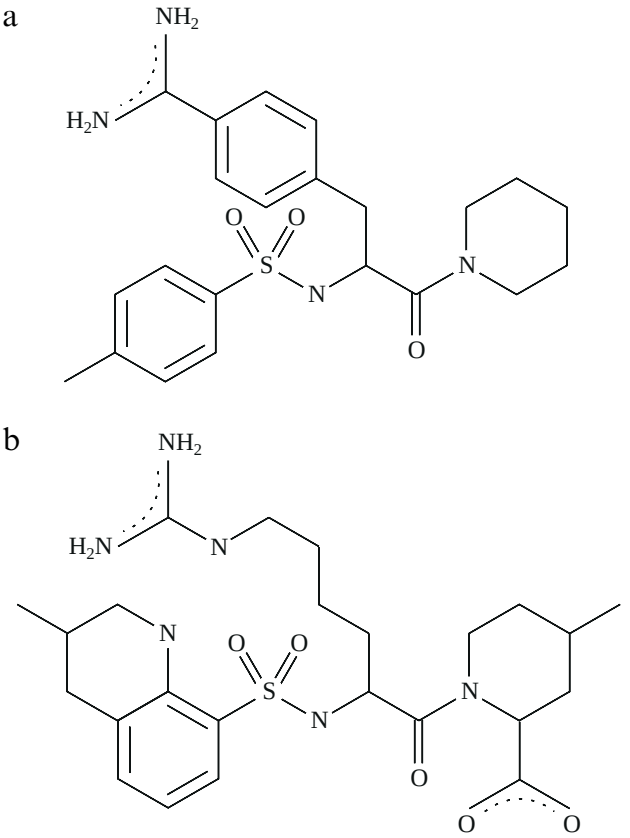

Fig. 11. The structures of (a) TAPAP and (b) MQPA.

TAPAP and MQPA

The atom coordinates of TAPAP and MQPA (Fig. 11) in their binding conformations were extracted from PDB files 1ett and 1etr, respectively. The hydrogen-bond map of TAPAP contained six donor and five acceptor regions and that of MQPA contained 11 donor and six acceptor regions. The best 10 matches, having carried out runs with 0–6 null regions, are shown in Table 4. Each trial took around 30 s. The best of these matches is shown in Fig. 12. There was a much greater discrepancy between these results and the crystal alignment (generated by superposing the heavy atoms from residues 57, 60A–60I, 195, 214–217 and 219) than for the DHFR ligands, despite the fact that the hydrogen-bonding and steric scores were higher than for the crystal alignment. There are two reasons for this. First, on further examination of the crystal structures (Fig. 13), MQPA forms bridging hydrogen bonds to the site via water molecules, which are not consistent in their positions between the two crystals. Secondly, there are only three common hydrogen-bonding atoms on the receptor to which both molecules bind whereas the algorithm attempts to maximize the number of overlapping regions. Despite this, the algorithm has still managed to identify respectable superpositions, implying that other information (steric and electrostatic) is implicitly contained in the hydrogen-bonding maps.

Six thermolysin ligands

The thermolysin inhibitors used in this study are shown in Fig. 14. The site-point generation procedure was tested by carrying out the crystal superposition of the binding sites (using all the heavy atoms from residues 112–116, 130, 133, 139, 142, 143, 146, 157, 166, 202, 203, 226, 231 and 232) and comparing the overlapping regions with the actual site atoms, as shown in Fig. 15. There was significant overlap of the predicted and actual site points and the positions of five receptor atoms were predicted to within 0.8 Å. These atoms were Asn^{112} $O^{\delta 1}$ (0.23 Å, 0.30 Å), Asn^{112} $N^{\delta 2}$ (0.76 Å), Arg^{203} $N^{\eta 1}$ (0.30 Å, 0.79 Å), Arg^{203} $N^{\eta 2}$ (0.53 Å, 0.78 Å) and His^{231} $N^{\varepsilon 2}$ (0.44 Å). Additionally, the positions of two conserved water molecules were predicted to within 0.8 Å. These uncertainties in predicting the precise positions compare favourably with the rmsd of about 0.3 Å obtained on superposing the crystal structures of the binding sites. However, there are also a number of false positive predicted positions. If these data were to be used as input to either database search or *de novo* design algorithms, the correct selection of points would have to be made in order to allow the identification of complementary ligands.

The results of the pairwise superpositions are listed in Table 5. Each trial took around 30 s. As a result of selecting the best partner for each molecule, VT was superposed onto CCT, which was superposed onto CLT, which was superposed onto ZGPLL. HBAGN was superposed onto NPL. This did not produce a consensus superposition of all six molecules. The next most similar pair of molecules was VT and CLT, but they were already superposed (indirectly). The next most similar pair was NPL and CLT, which were not already superposed, so NPL was overlaid onto CLT to complete the superposition of all six ligands. Each of the superpositions are compared in terms of the rmsd value between the generated superposition and the crystal superposition, using ZGPLL as the reference molecule, in Table 6. The best base molecule is CLT, but the consensus match appears to achieve bet-

TABLE 4
TEN BEST RESULTS OBTAINED ON MATCHING THE HYDROGEN-BOND MAPS OF TAPAP AND MQPA

Rank	D, A^a	ξ^a	n^b	Rmsd^c (Å)
1	6, 2 (1)	0.649 (7)	2	1.39
2	6, 1 (9)	0.672 (2)	6	1.06
3	6, 1 (9)	0.666 (3)	2	1.13
4	3, 3 (19)	0.662 (4)	5	1.04
5	5, 1 (19)	0.659 (5)	1	1.13
6	5, 1 (19)	0.659 (6)	2	1.17
7	5, 3 (1)	0.535 (27)	3	2.73
8	5, 3 (1)	0.535 (28)	5	2.40
9	3, 4 (9)	0.552 (20)	6	2.22
10	3, 4 (9)	0.551 (21)	6	2.27
Crystal	5, 1	0.663		

[a–c] As in Table 3.

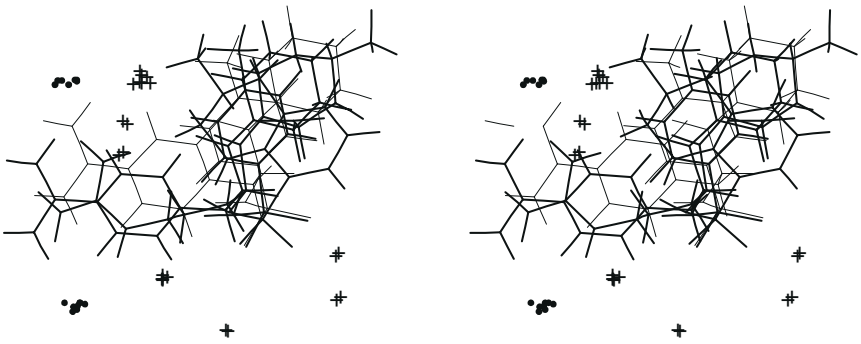

Fig. 12. Stereoview of the best match between TAPAP and MQPA and their overlapping hydrogen-bonding points. The crystal orientation of TAPAP is shown in feint.

ter results in terms of rmsd values. Hence, this consensus superposition was used to generate the overlapping hydrogen-bond map points, and is shown in Fig. 16.

The predicted site points are shown, along with the structure of the site, in Fig. 17. As can be seen, the method significantly narrowed down the number of overlapping points to allow a reasonable prediction of the relative dispositions of the receptor atoms. The predicted atoms were Asn^{112} $O^{\delta 1}$ (0.31 Å, 0.92 Å), His^{146} $N^{\varepsilon 2}$ (0.63 Å, 0.73 Å, 1.02 Å, 1.32 Å), Tyr^{157} OH (0.72 Å, 0.79 Å, 0.79 Å, 0.93 Å, 1.00 Å, 1.01 Å), Glu^{166} $O^{\varepsilon 1}$ (0.75 Å, 0.99 Å, 1.09 Å), Glu^{166} $O^{\varepsilon 2}$ (0.86 Å, 0.99 Å), Arg^{203} $N^{\eta 1}$ (0.65 Å), Arg^{203} $N^{\eta 2}$ (0.87 Å, 1.02 Å, 1.26 Å, 1.29 Å) and His^{231} $N^{\varepsilon 1}$ (0.68 Å, 0.71 Å). Additionally, the positions of two water molecules were predicted to within 0.8 Å, one of which was the same as the water molecule predicted using the crystal superposition. Although a larger number of atom positions have been predicted, the proximity of the points to the atoms is lower and the uncertainty in prediction is higher in that there are more points distributed about each atom. The increased number of predicted atom positions is to be expected, given that the approach aims to maximize this quantity. However, not only is this at the expense of accuracy in prediction, but also there are a larger number of false positive results. In terms of applying the data to drug design, this makes it more complicated to select the correct subset of points for use by design algorithms.

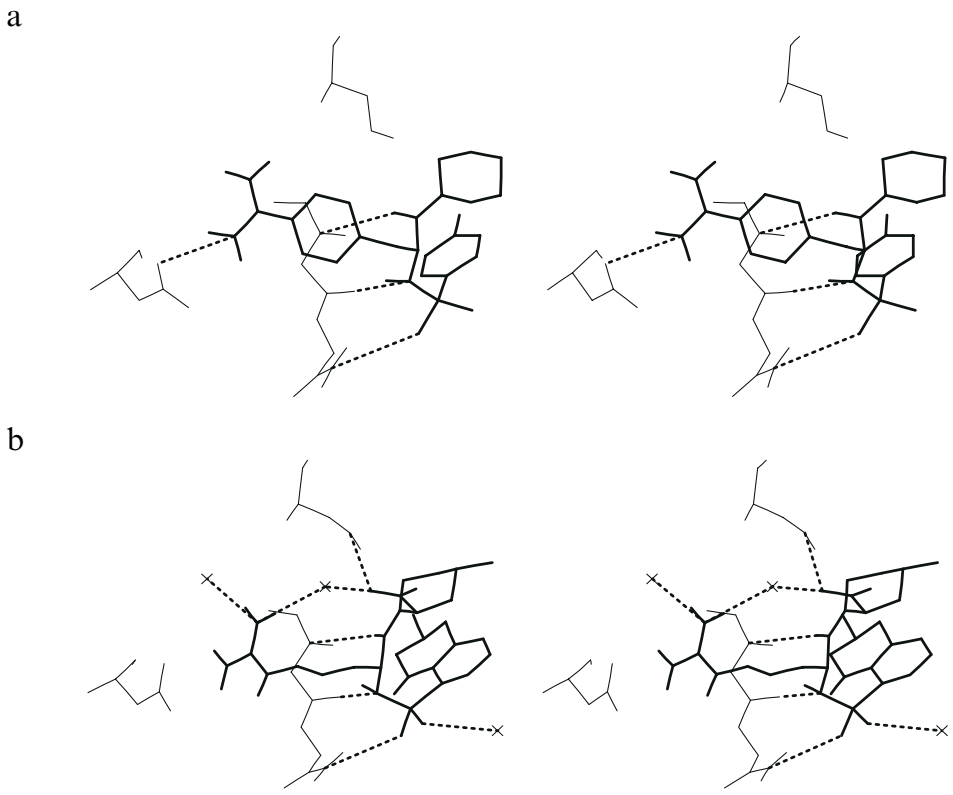

Fig. 13. Stereoview of the hydrogen bonds between (a) TAPAP and (b) MQPA and the binding site in thrombin.

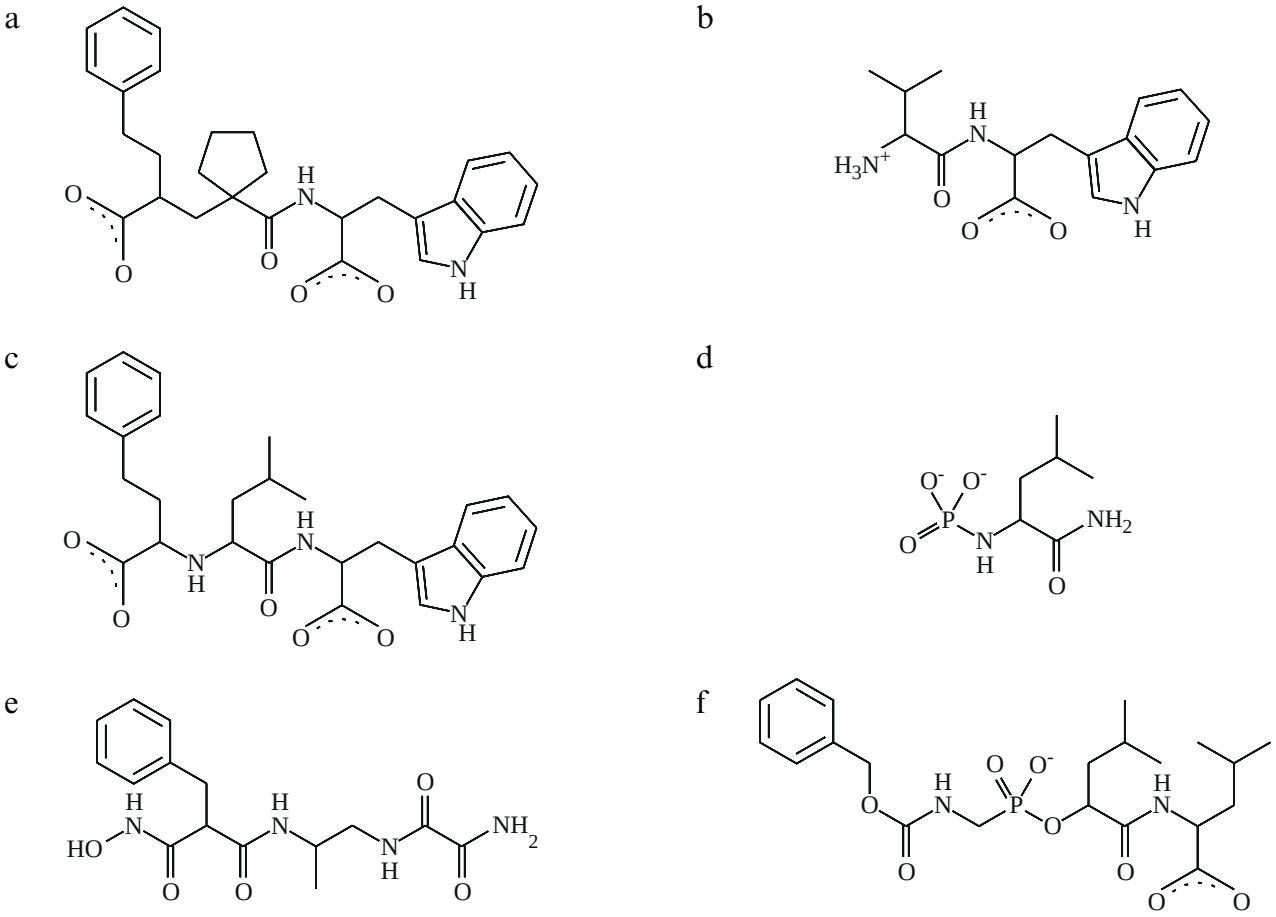

Fig. 14. The structures of the six ligands for thermolysin used in this study: (a) CCT; (b) VT; (c) CLT; (d) PLN; (e) HBAGN; (f) ZGPLL.

Discussion and Conclusions

A method has been presented to aid in the prediction of the relative positions of receptor hydrogen-bonding atoms from the structures of ligands. An algorithm has been written to generate hydrogen-bond maps for any given ligand; the overlapping regions of these maps obtained on molecular superposition represent positions to which molecules involved in the superposition can bind. As more molecules are incorporated into the superposition, the overlapping regions reduce in size, increasing the certainty in the prediction of the relative positions of the receptor hydrogen-bonding groups. Clearly, this method is superposition-dependent.

In this work, the superposition is carried out by searching for orientations whereby the maximum number of hydrogen-bonding regions were overlaid. For any pair of molecules, there is a good deal of degeneracy about the optimum solution, owing to the nature of the maps. Small changes around the global minimum orientation do not affect the number of overlapping hydrogen-bonding regions. This complicates the superposition, so the incorporation of another criterion, in this case steric overlap, is required to determine the orientation to be used in site-point generation. In such cases, it would seem that hydrogen-bond maps are more useful to assess a given superposition rather than as a basis for superposition. However, this flexibility in superposition could prove useful in, for example, docking studies or *de novo* design, where small changes in chemical configuration will not affect the ability of the ligand to bind to the same site. There are also cases where there are gross differences between orientations that give rise to the same number of overlapping hydrogen-bonding regions. For example, HBAGN is a

TABLE 5
RESULTS OF THE PAIRWISE MATCHES BETWEEN THE SIX THERMOLYSIN LIGANDS[a]

	ZGPLL	CLT	HBAGN	CCT	NPL	VT
ZGPLL		**10**, **2**	7, 2	10, 1	4, 1	6, 1
CLT	**0.549**		6, 1	**10**, **2**	5, 2	6, 4
HBAGN	0.532	0.454		10, 0	**5**, **2**	4, 3
CCT	0.527	**0.770**	0.441		4, 1	**6**, **2**
NPL	0.524	0.640	**0.709**	0.578		3, 3
VT	0.530	0.692	0.381	**0.788**	0.626	

[a] The upper half contains the number of common hydrogen-bonding regions and the lower half contains the ξ value calculated by PLM. The bold values represent the best match for each molecule.

TABLE 6
COMPARISON BETWEEN GENERATED SUPERPOSITIONS AND CRYSTAL SUPERPOSITION FOR SIX THERMOLYSIN LIGANDS

Base molecule	$Rms_{crystal}$ (Å)				
	CLT	HBAGN	CCT	NPL	VT
ZGPLL	0.751	1.008	1.166	1.931	1.163
CLT	0.751	2.148	0.775	0.966	0.626
HBAGN	2.164	1.008	8.413	0.866	3.031
CCT	1.270	7.374	1.166	0.972	1.184
NPL	2.378	1.943	2.897	1.931	4.821
VT	1.243	3.802	0.877	3.619	1.163
Consensus	0.751	1.588	0.775	0.966	0.856

linear molecule that is almost symmetrical from a hydrogen-bonding perspective and, in this study, was observed to adopt either orientation on superposition by HBMATCH. Hence, this method could prove useful in the identification of multiple binding modes for ligands.

In this work, simulated annealing was run repeatedly to generate a number of feasible solutions. It may be the case that a more efficient method to generate multiple solutions would be to use a genetic algorithm, which provides a population of solutions at any given time. Simulated annealing was used on account of its proven ability to deal with the problem of null correspondences, but the performance of genetic and evolutionary algorithms in such a problem would be interesting to study.

One issue not yet addressed by this work is conformational flexibility. The hydrogen-bond maps only take account of the rotation of terminal bonds in hydroxyl and amine groups. A number of representative conformers for each molecule could be used, an approach that has proved successful before [10,15]. This is possible, although time consuming, given the number of superpositions to be carried out. Any attempts to incorporate flexibility into the optimization algorithm of HBMATCH would require recalculation of the map points for each change in conformation, which would be time consuming owing to the rigorous treatment of hydrogen bonding. Although the hydrogen-bond map superposition presented here is a slow process when compared with previous methods, no assumptions are made about how hydrogen-bond regions superpose. The alternative approaches use only the high-probability points from each region in the superposition, whereas crystal-survey studies [21,22] have shown that there is great variability in hydrogen-bond geometry. This implies that a large amount of information is lost when a hydrogen-bond region is collapsed onto one point. For example, if a point with the ideal hydrogen-bond directionality is precluded by steric interaction, no alternative hydrogen-bond geometry is available for that group. Furthermore, the current method would be expected to provide a more realistic simulation of hydrogen-bonding properties when the position of the lone pair is unknown (e.g. sulphone and phosphate oxygen atoms). However, if rapid pharmacophore generation from flexible molecules is required, other methods [14–16] could prove more suitable for generating the superposition, although at the expense of a loss of accuracy. They would be expected to work in congenerically similar molecules, in which the hydrogen-bonding atoms (and therefore the maximum hydrogen-bond probability points) overlay, but could be expected to perform less well in cases of ostensibly dissimilar molecules.

This work has also shown that, although it is possible to superpose molecules on the basis of their hydrogen-bonding properties, there are a number of other problems that could be encountered. The superposition of the thrombin inhibitors provides an example for which superposition was not successful. The main reason for this was that there are few atoms on thrombin to which both TAPAP and MQPA bind. Also, the remaining hydrogen bonds anchoring these ligands to the site involve water molecules, which are not consistently positioned when comparing complexes of proteins with different ligands. Similar failings would be expected if the ligands did not

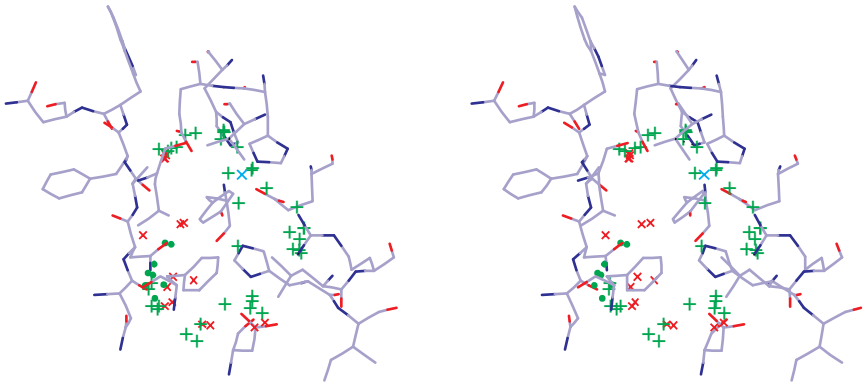

Fig. 15. Stereoview of the binding site within thermolysin and the overlapping points obtained on superposing the hydrogen-bond maps of the six thermolysin ligands in their crystal orientations. The site atoms shown are those from the complex with ZGPLL (PDB reference code 6tmn). Carbon atoms are shown in grey, oxygen atoms in red, nitrogen atoms in blue and hydrogen atoms in black. Solvent water molecules are shown as red crosses and the zinc atom as a cyan cross.

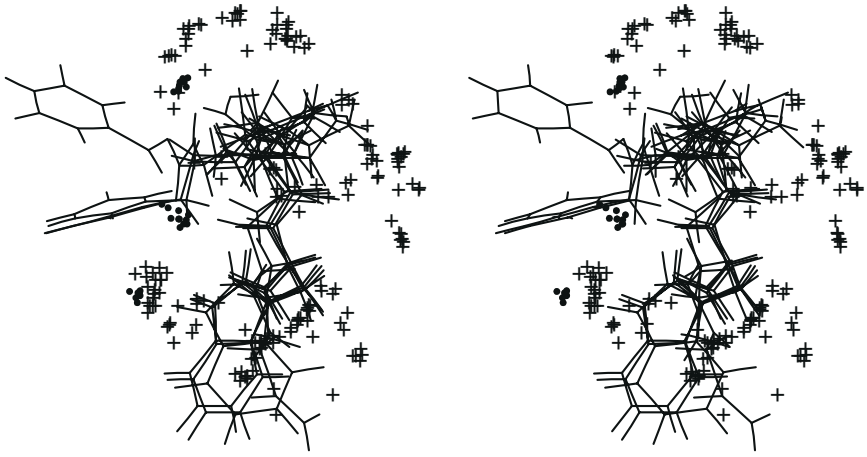

Fig. 16. Stereoview of the superposition of all six thermolysin ligands obtained by HBMATCH along with all the overlapping hydrogen-bonding points.

contain many hydrogen-bonding groups, or if the receptor hydrogen-bonding atoms were inconsistently placed when bound to different ligands. For example, a serine residue in the binding site of thrombin is oriented differently in the two complexes (top of Fig. 13), forming a hydrogen bond to MQPA, but not to TAPAP. The fact that the algorithm found the relative orientation of the ligands to within about 1 Å shows that the hydrogen-bonding information also implicitly includes steric and electrostatic information.

Two approaches for the incorporation of more than two molecules into the superposition were tested. It was found that generating a consensus superposition using data from all pairwise matches gave better results than superposition onto the same base molecule. This is expected, given that the consensus superposition makes use of more information in generating the optimum superposition of all the ligands. Each molecule is superposed onto its most similar counterpart so there is no dependence on the choice of base molecule. Additionally, it is unlikely that a base molecule will superpose well with all the remaining molecules in the set because different molecules in the set usually bind to different subregions within the receptor binding site.

Once the optimum superposition has been obtained, the method for predicting receptor-atom positions from the overlapping regions of the maps is successful, where the crystal orientations of six ligands binding to the same site are used. The search for novel ligands could proceed by identifying ligands whose hydrogen-bond maps overlap with the derived overlapping hydrogen-bond map points. This could be carried out either by database searching or using *de novo* design algorithms and would rely on the selection of the appropriate subset of derived site points.

From a given ligand superposition, a number of approaches to receptor modelling have previously appeared in the literature. For example, GERM [33] sets up around 50 points at random about the ligands and optimizes the choice of atom types to place there such that the energies of interaction with the model correlate with known activities of the ligands. YAK [34] clusters the peaks in hydrogen-bond and hydrophobic potential from each ligand and selects suitably disposed amino acids to occupy the clusters. In both cases, the resulting pseudoreceptor com-

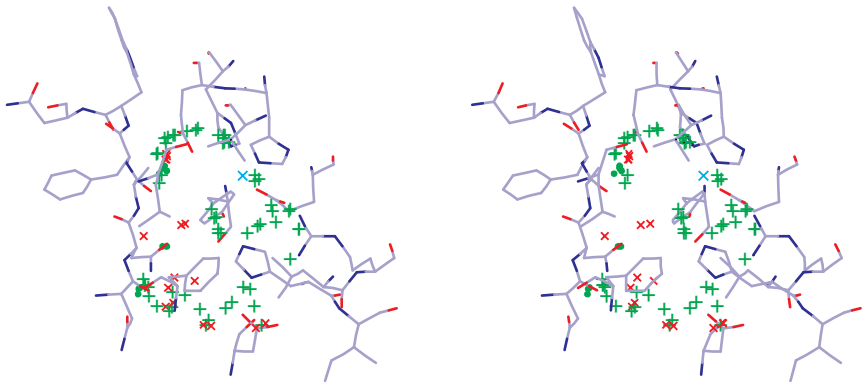

Fig. 17. Stereoview of the binding site within thermolysin and the selected overlapping points obtained on superposing the hydrogen-bond maps of the six thermolysin ligands with HBMATCH.

plements the ligand field without necessarily bearing any resemblance to the actual receptor. In the approach of Hahn [35], the mean values of various properties of the ligands are projected onto the supersurface of the ligand (electrostatic potential and partial atomic charge values are reversed in sign) to create a model of the field pattern exerted by the receptor. The approach presented here differs from these three methods in a number of ways. Firstly, it predicts the positions of atoms in the receptor, rather than describing the nature of the receptor at a selection of points in space about the ligands. Coupled with mutation and primary sequence data, this information could prove useful in building accurate and realistic 3D models of a binding site, allowing site-directed drug-design strategies to be used in lead generation. Secondly, only hydrogen-bonding interactions are used in the model generation. Although other interactions are not considered explicitly, hydrogen bonding also contributes greatly to partial charge and electrostatic and hydrophobic potential values, i.e. there is interdependence between the different parameters. Furthermore, it has previously been shown that electrostatic complementarity is not always observed and partial-charge complementarity is rarely observed between ligand and receptor [36]. Thirdly, hydrogen bonds are represented more realistically because positions of maximum likelihood are not assumed to be the positions of hydrogen-bond formation. Fourthly, like YAK, points common to a subset of the binding ligands are also derived, whereas the other two approaches tend to search only for properties common to all the binding ligands.

In conclusion, two programs have been written to aid ligand-based drug design. HBMAP predicts all the possible positions of complementary hydrogen-bonding atoms about any organic molecule. HBMATCH superposes two molecules on the basis of their hydrogen-bond maps, generating a large number of feasible superpositions for a given pair of molecules. Once the correct superposition has been obtained, the overlapping regions of the hydrogen-bond maps provide a good prediction of the positions of receptor atoms to which the ligands bind.

Acknowledgements

The authors wish to thank Rhône-Poulenc Rorer (J.E.J.M., T.D.J.P.) and the Wellcome Trust through the PRF scheme (P.M.D.) for personal financial support. Part of this work was carried out in the Cambridge Centre for Molecular Recognition, funded by the BBSRC.

References

1 Manallack, D.T., Drug Discov. Today, 1 (1996) 231.
2 Finn, P.W., Drug Discov. Today, 1 (1996) 363.
3 Böhm, H.-J., Curr. Opin. Biotechnol., 7 (1996) 433.
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 Masek, B.B., Marchant, A. and Matthew, J.B., J. Med. Chem., 36 (1993) 1230.
9 Masek, B.B., Marchant, A. and Matthew, J.B., Proteins, 17 (1993) 193.
10 Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 479.
11 Chau, P.-L. and Dean, P.M., J. Mol. Graph., 5 (1987) 97.
12 Apaya, R.P., Lucchese, B., Price, S.L. and Vinter, J.G., J. Comput.-Aided Mol. Design, 9 (1995) 33.
13 Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 6 (1992) 475.
14 Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1993) 83.
15 Barnum, D., Greene, J., Smellie, A. and Sprague, P., J. Chem. Inf. Comput. Sci., 36 (1996) 563.
16 Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9 (1995) 532.
17 Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 8 (1994) 751.
18 Prendergast, K., Adams, K., Greenlee, W.J., Nachbar, R.B., Patchett, A.A. and Underwood, D.J., J. Comput.-Aided Mol. Design, 8 (1994) 491.
19 Good, A.C., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic and Professional, London, U.K., 1995, pp. 24–56.
20 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
21 Klebe, G., J. Mol. Biol., 237 (1994) 212.
22 Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 10 (1996) 607.
23 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 101.
24 Danziger, D.J. and Dean, P.M., Proc. R. Soc. London, B236 (1989) 115.
25 Levitt, M. and Perutz, M.F., J. Mol. Biol., 201 (1988) 751.
26 Fong, T.M., Cascieri, M.A., Yu, H., Bansal, A., Swain, C. and Strader, C.D., Nature, 365 (1993) 350.
27 Mitchell, J.B.O., Nandi, C.L., McDonald, I.K., Thornton, J.M. and Price, S.L., J. Mol. Biol., 239 (1994) 315.
28 Dougherty, D.A., Science, 271 (1996) 163.
29 McLachlan, A.D., J. Mol. Biol., 128 (1979) 49.
30 Szu, H. and Hartley, P., Phys. Lett., A122 (1987) 157.
31 Roderick, S.L., Fournie-Zaluski, M.C., Roques, B.P. and Matthews, B.W., Biochemistry, 28 (1989) 1493.
32 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
33 Walters, D.E. and Hinds, R.M., J. Med. Chem., 37 (1994) 2527.
34 Vedani, A., Zbinden, P., Snyder, J.P. and Greenidge, P.A., J. Am. Chem. Soc., 117 (1995) 4987.
35 Hahn, M., J. Med. Chem., 38 (1995) 2080.
36 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 8 (1994) 513.