

A new method for estimating the importance of hydrogen-bonding groups in the binding site of a protein

Matthew D. Kelly^{a,*} & Ricardo L. Mancera^b

^a*Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, United Kingdom;* ^b*De Novo Pharmaceuticals, Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9ZR, United Kingdom*

Received 19 February 2003; accepted 27 May 2003

Key words: hydrogen bond, binding site, structure-based drug design

Summary

We introduce a new method to estimate the importance of hydrogen-bonding sitepoints in the binding site of a protein as part of a structure-based design strategy. Our method identifies hydrogen-bonding sitepoints within a binding pocket and ranks them according to both the accessibility of their hydrogen bonding regions to incoming ligands and their hydrogen-bonding strength. The combination of these components produces a prioritised list of sitepoints that are more likely to be involved in hydrogen bonding with an incoming ligand. A dataset of known protein-ligand interactions was used to compare the prioritisation of sitepoints identified by our method with those observed to be engaged in hydrogen bonding in their crystal structures. Our method was able to remove those sitepoints unable to bind the ligand due to a low accessibility or an unfavourable orientation and to award significantly higher hydrogen-bonding ranking values to those sitepoints observed to form hydrogen bonds. Our method can thus be used to identify hydrogen-bonding sitepoints that should be targeted preferentially in a drug design strategy.

Introduction

Structure-based drug design [1–7] takes, as a starting point, a protein's ligand-binding site for designing compounds that will bind. *De novo* design methods following an outside-in approach [8] begin by analysing the binding site of any target protein to identify regions that may play an important role in ligand binding. Following the identification of these key regions, functional groups are placed at complementary positions within the binding site and connected by 'linker groups' creating a molecular scaffold to be used in lead generation [9]. On the other hand, an inside-out algorithm [10] 'grows' a molecular structure within the binding site of a protein by proposing a number of possible modifications to a current molecular structure. Such modifications may include the addition of

a specific functional group or the re-positioning of an existing group to satisfy particular interactions in the binding site of the protein. Each proposed modification is then evaluated and the most favourable accepted. Hybrid methods [11] combine the two approaches by piecing together and modifying molecular fragments within the confines of a binding site guided by a set of key protein interacting groups, chemical rules and a scoring function.

An important component of structure-based drug design is the identification of the key interacting groups within the binding site of the protein. In particular, hydrogen-bonding interactions [12] are known to play a vital role in the molecular recognition between protein and ligand, a feature that is central to the high degree of selectivity observed in protein-ligand interactions. In knowledge-based methods, the binding site may be analysed to identify hydrogen-bonding atoms (sitepoints) that are likely to play an important role in ligand binding based solely upon atom type

*To whom correspondence should be addressed. E-mail: mdk27@cam.ac.uk

and position relative to neighbouring atoms within the binding site. Such hydrogen-bonding sitepoints may then be used in an outside-in approach where complementary chemical groups are placed in favourable positions relative to a selection of these sitepoints in order to assemble a molecular scaffold. Alternatively, the sitepoints may be used as a basis for an inside-out approach to 'guide' the growth of a ligand within the binding site.

Currently available knowledge-based binding site-analysis methods suffer from a significant combinatorial problem that requires manual intervention to be overcome. This problem arises from the fact that, of the multiple hydrogen-bonding sitepoints often identified within a given binding site, only a fraction are directly involved in the interactions with any one ligand. For example, the combinatorics for a binding site comprising twenty sitepoints, where five may be satisfied by the ligand, result in more than fifteen thousand possible combinations [13]. Clearly, a significant proportion of such combinations of sitepoints may be rejected on the basis of the hydrogen-bonding pattern of known co-crystallised ligands, distance criteria between the sitepoints, and the arbitrary definition of binding site sub-regions (pockets). However, if this particular approach to *de novo* drug design is to become more efficient, this problem must be overcome, as to search every possible solution would be impractical.

An initial step towards solving this problem could be to prioritise the sitepoints in some way, thus reducing the number of sitepoints to be considered to only those that carry more 'weight' for binding. The most obvious way of doing such prioritisation would be by measuring the intrinsic hydrogen-bond strength of the sitepoints in a binding site. There have been a number of methods used for scoring the hydrogen-bonding ability of an atom or chemical group. Early methods distinguished between atoms that could form hydrogen bonds and those that could not [14]. However, this particular method was unable to distinguish between groups with differing hydrogen-bonding ability. Another early method of scoring hydrogen-bonding groups within a binding site was that used by the GRID algorithm [1]. This method superimposes a regular grid onto the binding site and the energy of a probe placed at each of the vertices of the grid is calculated using an empirical force field. Different probes can be used to identify favourable positions for a particular chemical group within the binding site.

More complex methods have modelled the hydrogen-bonding ability of an atom using quantum mechanics-derived properties, such as the net atomic charge and the energy of the lowest unoccupied molecular orbital (E_{LUMO}) for the donor atom, and the net atomic charge and the energy of the highest occupied molecular orbital (E_{HOMO}) for the acceptor atom [15–17]. More recently it has been demonstrated that electron donor superdelocalisability and self-atom polarisability also correlate with hydrogen-bonding ability [18].

Thermodynamic parameters have been used to describe hydrogen-bond formation and calculate the optimum bond energy for a given hydrogen bond [19–20]. From this optimum energy the corresponding ideal bond length is calculated. The hydrogen-bond energy of a given interaction is then estimated from its deviation from the calculated ideal geometry. This method, however, requires information on both the hydrogen-bonding group within the protein and the corresponding group in the ligand.

Comprehensive donor and acceptor hydrogen-bonding scales have been derived from experimental data using 4-nitrophenol as the standard donor, N-methyl-pyrrolidinone as the standard acceptor and 1,1,1-trichloroethane as the solvent [21]. However, the data was derived for small molecules and there is insufficient data for all the hydrogen-bonding groups observed in a protein.

Our objective in this work was to develop a method for prioritising hydrogen-bonding sitepoints found in the binding site of a protein independently from any information about the ligand. We concentrated on the two features likely to contribute mostly to the hydrogen-bonding ability of a protein sitepoint: the intrinsic ability of a chemical group on the surface of a protein to form hydrogen bonds and the physical accessibility of such a group to an incoming ligand. Each feature was assigned a scoring function, and a final combined score was then introduced. We then developed a simple geometric method for determining which sitepoints are orientated towards the binding cavity. A set of tests was carried out on a number of protein-ligand complexes to validate our approach.

Table 1. Typical ranges of d_{rel} values and the corresponding hydrogen-bond strengths (taken from *IsoStar*).

d_{rel} value	Hydrogen-bond strength
2.0–8.0	Strong
1.0–2.0	Average
0.7–1.0	Weakish
0.3–0.7	Very weak

Materials and methods

Intrinsic hydrogen-bonding ability

The ISOSTAR database [22] uses experimentally-derived data from an extensive survey of the Cambridge Structural Database (CSD) for small molecules and the Brookhaven Protein Data Bank (PDB) for proteins. This database contains geometrical information on the distribution of contact groups (up to 43) around specific central groups (298 in total). These distributions can provide density maps showing preferred geometries of central and contact groups. The relative density (d_{rel}) score is a measure of the tendency for the central and contact groups to form short interactions:

$$d_{rel} = d_{short}/d_{long} \quad (1)$$

Here d_{short} is the density of contacts within the sum of van der Waals distances (vdWD) and d_{long} is the density of contacts between vdWD and $\text{vdWD} + 0.5 \text{ \AA}$. The larger the d_{rel} score is, the greater the tendency for the central and contact groups to form short interactions and, therefore, the more attractive they are likely to be [22]. Typical ranges for d_{rel} are listed in Table 1, as taken from the ISOSTAR database (Version 1.4, Cambridge Crystallographic Data Centre).

Correlation coefficients of between 0.8 and 0.9 between the logarithm of d_{rel} and calculated hydrogen-bond energies have been reported [22]. The ISOSTAR database provides d_{rel} values for a variety of functional groups found on ligands, in addition to the functional groups found within proteins. The protein-based groups are divided into three classes: terminal groups (those found in the sidechains of residues, such as hydroxyl groups), ring systems (also found in the sidechains of residues, e.g. indole) and linker groups (such as those found between residues, e.g. disulphide bridges). For each of the constituent groups of these classes a d_{rel} value is provided for an interaction with a specific probe or probe type. For

this study the data from two probe types were used. The first probe type was any polar X–H group, where X may be an oxygen, nitrogen or sulphur atom. This probe type is used to provide a d_{rel} value for the protein group acting as a hydrogen-bond acceptor. The second probe type used was any C=O group. This probe is used to provide a d_{rel} value for the protein group acting as a hydrogen-bond donor. The actual values for the protein groups with each of the two probes are shown in Table 2. For certain sitepoints, e.g. carboxylic acids and imidazole groups, the ionisation state may vary due to variations in the side chain's pK_a . For these groups, d_{rel} values have been determined for the different states.

As can be seen from Table 2, the charged carboxylic acid group scores highest as an acceptor with a very low score as a donor. It is interesting to note that, despite the absence of a hydrogen atom to be donated, the d_{rel} score for the group acting as a donor is not zero. This does not necessarily mean that a charged carboxylic acid group can act as a donor to a carbonyl group (the standard acceptor group), rather that in the crystal survey, examples were observed where the contact distance between the two groups was less than the sum of the van der Waals radii. This may be due to the contacts not having the correct geometry for a hydrogen bond or due to errors within the coordinates of the PDB / CSD file. Despite the use of filters to remove corrupt crystal structures before their inclusion into the ISOSTAR database, some inaccuracies may still remain.

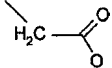
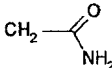
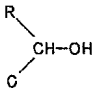
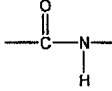
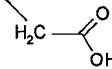
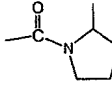
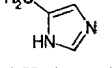
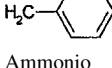
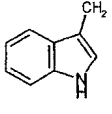
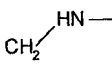
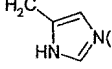
We have taken the d_{rel} values quoted in Table 2 as a representative measure of the intrinsic hydrogen-bonding ability of the various functional groups that can be found on the surface of a protein and, in particular, in its binding site. Such values represent the average strength of the hydrogen bonds that may be made with a ligand.

Hydrogen-bonding sitepoint accessibility

The accessibility of a hydrogen-bonding sitepoint within the binding pocket of a protein determines its availability to complementary hydrogen-bonding groups of a ligand. The greater the accessibility of a sitepoint, the greater the likelihood that it will be involved in the binding of a ligand.

A simple approach to measuring the accessibility of an atom within a binding site would be to compute its solvent accessible surface area (SASA) [23]. Although this method could accurately calculate the

Table 2. Mean d_{rel} values for the functional groups found in proteins with the acceptor (C=O) and donor (X-H) probes, along with the single letter code for the amino acids in which the group is found. Standard errors are shown in parentheses.

Functional group	Residue	Any polar X-H	Any C=O
Charged carboxylic acid 	DE	4.9 (0.3)	0.4 (0.0)
Carbamoyl 	NQ	2.0 (0.1)	1.7 (0.2)
Hydroxy (aliphatic) 	ST	2.0 (0.1)	1.6 (0.1)
Peptide 	All	1.4 (0.1)	0.7 (0.0)
Uncharged carboxylic acid 	DE	1.0 (0.0)	0.8 (0.0)
Prolyl down conformer 	P	0.8 (0.1)	1.0 (0.1)
Uncharged imidazole 4-yl 	H	0.8 (0.1)	N/A
4-Hydroxyphenyl 	Y	0.6 (0.1)	0.9 (0.1)
Ammonio $\text{CH}_2\text{—NH}_3^+$	K	0.4 (0.0)	4.0 (0.3)
Indol-3-yl 	W	0.3 (0.0)	0.8 (0.1)
Guanidinio 	R	0.3 (0.0)	2.0 (0.3)
Charged imidazole 	H	0.3 (0.0)	1.5 (0.3)

percentage of the van der Waals surface accessible to the solvent (or ligand atom), it would not take into account the three-dimensional geometric preferences seen in hydrogen bonding.

For example, the oxygen in a carbonyl group carries two lone pairs of electrons that have the potential to form a hydrogen bond. The inherent directionality of hydrogen-bonding interactions results in the preferential geometries seen in hydrogen bonding. This preferential geometry results in regions around the hydrogen-bonding sitepoint where it would be more favourable for the complementary atom to be positioned [24].

To demonstrate the shortcomings of the simple SASA method, consider the carbonyl example described above. If the oxygen atom was positioned within the protein binding site in such a way that 75% of its SASA was occluded by surrounding atoms, then the more simple SASA method would report the accessibility of the atom as 25%. Whilst this may be an accurate value for the solvent accessibility of the atom, it does not necessarily reflect the effect of reduced accessibility on the capability of that atom to form a hydrogen bond with a suitable donor. For example, if the 25% of the SASA that remains accessible does not include the regions that are, as determined by the positions of the lone pairs of electrons on the oxygen atom, most favourable for hydrogen bonding, then the hydrogen bonding capability of that atom will be reduced by far more than 75%. If, on the other hand, the 25% of the SASA that remains includes the regions that are most favourable for hydrogen bonding then the effect of the reduced accessibility would be far less.

To overcome this problem, a method has been designed using data from the HBMAP algorithm [25], which can predict the location of hydrogen-bonding sitepoints within a protein based on the superposition of hydrogen-bonding maps of a range of ligands for that site. HBMAP generates hydrogen-bonding maps for a ligand based upon crystal survey data. The maps display positions where complementary hydrogen-bonding atoms may be found. The densities of the maps relate to the probability of finding a protein sitepoint at that position.

We have applied the HBMAP method, where hydrogen-bonding groups are first identified and classified (e.g. carbonyls, hydroxyls, carboxylic acids) using previously described criteria [26]. The area surrounding each identified group is then divided into equal volume bins. These bins are then checked for steric clashes. A bin is discarded if it is closer to

the hydrogen-bonding atom than the closest observed contact in the crystal survey. Since they represent the positions of complementary hydrogen-bonding heavy atoms, a bin is also discarded if the distance between it and the centre of the atom is less than the sum of the van der Waals radii of the atom and a heavy atom – i.e. oxygen or nitrogen. However, complementary hydrogen-bonding atoms are allowed to approach within this distance, for example, an oxygen atom and a hydrogen atom bound to a nitrogen atom. Bins that are within the van der Waals radii of neighbouring atoms are also removed. For each of the accessible bins remaining, a probability value is assigned from the crystal survey that represents the normalised number of occurrences of corresponding groups observed within the bin. So for each group, all the observed contacts from the survey are collated by overlaying the test group and the total number of groups located in each bin is calculated. All non-zero values are then normalised by dividing each one by the maximum number of contacts observed in a single bin for that particular group. This results in a value of between zero and one for each bin. This value is directly related to the probability of finding a complementary hydrogen-bonding atom in this bin. Figure 1 shows the hydrogen-bonding map of one of the oxygens of a fully accessible charged carboxylic acid group (COO^-).

We have incorporated these hydrogen-bonding probability maps into a hydrogen-bonding accessibility method by first summing the probability values from all accessible bins for a particular group within the confines of the protein binding site. This value can then be divided by the sum of the probability values from all the bins for the group when it is fully accessible (i.e., when it is treated in isolation):

$$SAPS = \sum_{bins} P_t / \sum_{bins} P_i \quad (2)$$

Here P_t is the probability of an available bin in the test group, P_i is the probability of an available bin in the isolated group. The ratio between the two gives a normalised value of between 0 and 1, which we have termed the *solvent-accessible probability score* (SAPS). The difference between our method and the simple SASA method described above is that the orientations of the accessible regions relative to the test atom are taken into account. For example, bins located in regions of high hydrogen-bonding probability are likely to have a probability value close to one, whereas bins in regions of low hydrogen-bonding

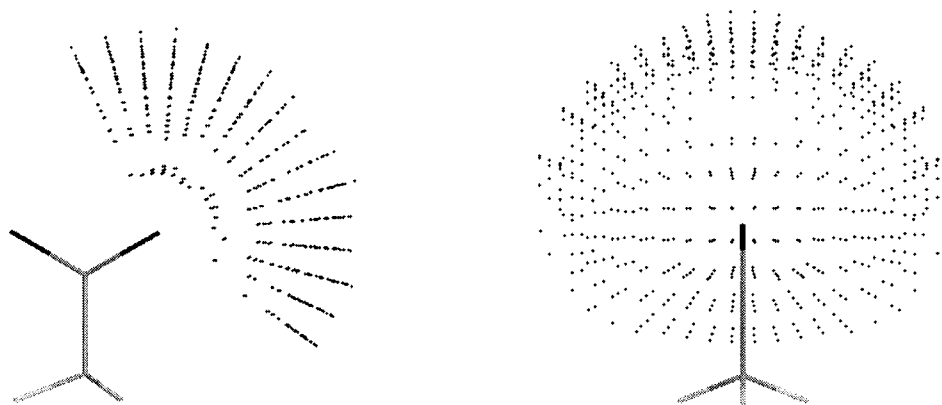


Figure 1. Hydrogen-bonding map of one of the oxygen atoms of a fully-accessible charged carboxylic acid group (COO^-). For each point, representing the centre of a bin, a value is assigned that represents the probability of finding a complementary hydrogen bonding atom within that particular bin.

probability will have a probability value closer to zero. If a hydrogen-bonding atom within the binding site is positioned relative to its neighbouring atoms such that more favourable regions for a complementary hydrogen-bonding partner are inaccessible, then bins with higher probability values will be discarded. This would result in a lower value for the sitepoint accessibility than if mainly low value bins were discarded. The final *SAPS* value for a hydrogen-bonding sitepoint reflects the total accessibility of its hydrogen-bonding probability maps to the solvent and/or an incoming ligand.

Total hydrogen-bonding score

In order to obtain a single score for the hydrogen-bonding ability of a sitepoint, the intrinsic hydrogen-bonding ability as measured by the d_{rel} value is multiplied by the hydrogen-bonding sitepoint accessibility as measured by the *SAPS* value. This produces a final score that we have termed the *strength-weighted accessible probability score (SWAPS)*:

$$SWAPS = d_{rel} \times SAPS \quad (3)$$

We can see that this procedure is equivalent to weighting the intrinsic hydrogen-bonding ability by the accessibility of a sitepoint's hydrogen-bonding probability map. If the sitepoint is fully accessible (i.e. $SAPS = 1$) then the d_{rel} score is unaltered and the final score remains the same. These final *SWAPS* values enable the ranking of all the sitepoints of a binding site so that those with higher scores are likely to be of greater weight for ligand binding than those with low scores. This arises from the fact that higher values

for d_{rel} and *SAPS* indicate the possibility of forming stronger hydrogen bonds with the ligand and the better accessibility of the sitepoint to an incoming ligand, respectively. It is important to stress that these scores are calculated in the absence of any ligand and are derived purely from information available in the protein itself.

Sitepoint orientation within the binding site

There is usually a large number of sitepoints that can be found in the binding site of a protein within a given cut-off distance of a bound ligand. Some of these sitepoints are not orientated towards the binding site and are thus geometrically unable to participate in hydrogen bonding with an incoming ligand. In order to remove these undesirable sitepoints, we developed a set of simple geometric criteria. Sitepoints are assigned to one of three classes, as can be seen in Figure 2: sitepoints with the heteroatom (donor or acceptor) bonded to a single carbon atom (Figure 2a), sitepoints with a donor heteroatom bonded to two carbons (Figure 2b), and sitepoints with an acceptor heteroatom bonded to two carbons (Figure 2c). The basis for this classification is that each requires a different approach to determine its orientation relative to the binding site.

Before the orientations of the sitepoints are determined it is first necessary to specify a number of reference points. Since the ligand is used to define the binding site, the orientations of the sitepoints were calculated relative to the ligand atoms. It would be impractical to calculate the orientations of the sitepoints relative to every ligand atom. Therefore we defined the dimensions of any ligand using a maximum of

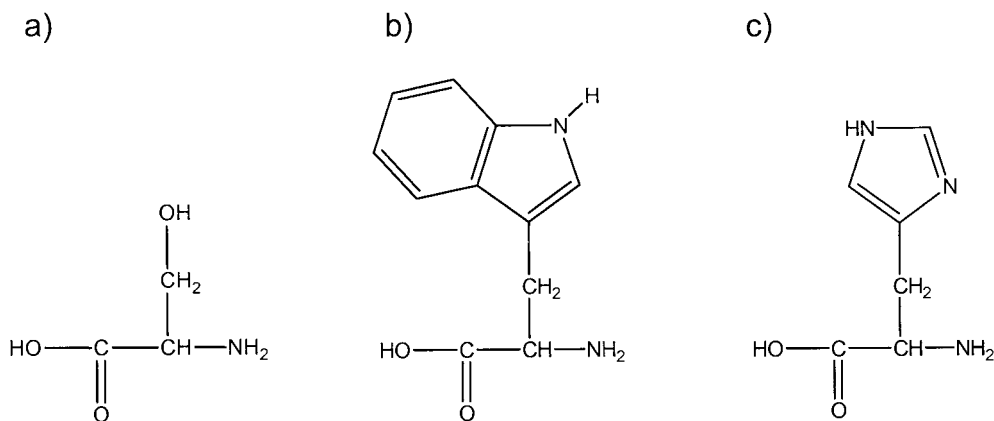


Figure 2. An example of each of the three classes of hydrogen-bonding sitepoints used to determine their orientation with respect to the binding cavity. (a) Heteroatom (donor or acceptor) bonded to a single carbon atom. (b) Donor heteroatom bonded to two carbons. (c) Acceptor heteroatom bonded to two carbons.

four points. First the geometric centre of the ligand is calculated. Then the ligand atom furthest from this point is identified, becoming the first reference point A. The second reference point B is the ligand atom furthest from the first reference point. The third reference point C is the ligand atom furthest from the previously defined reference points A and B, as long as the ratio $|(C-A) - (C-B)| / \max[(C-A), (C-B)]$ is equal to 0.3 or less. This ensures that for very elongated molecules only two reference points are needed to describe their geometry. The fourth reference point D is the ligand atom furthest from the previously defined reference points A, B and C. For this fourth reference point, to prevent it from being too close to any of the other reference points (as we observed in the case of ring systems), a minimum distance of 4.5 Å between reference point D and all the others was introduced. The orientation of each of the hydrogen-bonding sitepoints is then calculated relative to each of the previously defined reference points. Each sitepoint must then pass the orientation criteria relative to any one of the representative points to be considered as having the right orientation within the binding site.

For the first class, a heteroatom (donor or acceptor) bonded to a single carbon atom (Figure 2a), the angle between the carbon–heteroatom vector and the carbon–reference point vector is measured. If the angle is less than or equal to 90°, i.e. the heteroatom points towards the binding site, the sitepoint is retained. If the angle is greater than 90°, the sitepoint is excluded. For the second class, a donor heteroatom bonded to two carbons (Figure 2b), the angle between the heteroatom–hydrogen vector and the heteroatom–

reference point vector is used. The same rejection criterion as before is used for this angle. For the third class, an acceptor heteroatom bonded to two carbons (Figure 2c), since there is no hydrogen atom present, the angle between the vector of the heteroatom–midpoint of the two carbon atoms and the vector of the midpoint to the reference point is used. Again, the same rejection criterion as before is used for this angle.

This method required some modification for groups where the position of the hydrogen atom cannot be accurately predicted. For example, the rotatable hydroxyl group of a serine may be positioned in such a way that by using the standard criteria, the group would be rejected as having the wrong orientation, yet, the O–H group could actually point towards the reference point. As a consequence, the group should be retained, since it is the orientation of the O–H group relative to the reference point that determines its ability to act as a donor. The reason for this problem is that the C–O vector is used to represent the orientation of the sitepoint, since the exact position of the hydrogen atom is unknown. The same problem may arise with the hydroxyl groups of threonine and tyrosine, and the NH₃ group of lysine. To overcome this problem, an angle tolerance is included to allow the C–O (or C–N) vector to form an angle relative to the O–reference point vector greater than 90° and still be accepted, provided the O–H (or N–H) vector in one of its possible orientations is able to satisfy the orientation criteria.

Table 3. Protein Data Bank (PDB) code and description of the 29 crystal structures used in the validation. The LIGPLOT (Wallace et al., 1995) code for each ligand is shown in parentheses.

PDB code	Description
19GS	Glutathione S-Transferase P1-1 (BSP)
1AHT	α -Thrombin (APA)
1BIO	Human Complement Factor D (GOL)
1BOA	Methionine Aminopeptidase 2 (FUM)
1BZS	Mmp8 (BSI)
1C88	Protein Tyrosine Phosphatase 1B (OTA)
1DVT	Transthyretin (FLP)
1E1V	Cyclin Dependent Kinase 2 (CMG)
1E2E	Thymidylate kinase (TMP)
1FDS	17-Hydroxysteroid-Dehydrogenase Type 1 (EST)
1FKN	Secretase Complexed (LOL)
1FJS	Factor Xa (Z34)
1G32	Prothrombin (R11)
1G35	HIV-1 Protease (AHF)
1G53	Carbonic Anhydrase II (F6B)
1G7F	Ptp1B (INZ)
1GD0	Macrophage Migration Inhibitory Factor (CIT)
1HI3	Eosinophil-Derived Neurotoxin (A2P)
1HNE	Neutrophil elastase (MSU)
1IDA	HIV-2 protease (HBP)
1J4R	Fk506 Binding Protein (001)
1JK7	Protein Phosphatase-1 (OKA)
1KVO	Phospholipase A2 (OAP)
1QPC	Lymphocyte-specific kinase Lck (ANP)
1QTN	Caspase-8 (DTD)
2OAT	Ornithine Aminotransferase (PFM)
2SRC	Tyrosine-Protein Kinase C-Src (ANP)
2KI5	Thymidine Kinase (AC2)
3GSS	Glutathione S-Transferase P1-1 (EAA)

Validation

In order to test the ability of our methods to identify hydrogen-bonding sitepoints that play a key role in ligand binding, we took a selection of proteins where the binding mode of a ligand is known. The crystal structures of the proteins were taken from the Brookhaven Protein Data Bank (PDB). Table 3 lists all the protein structures used in this work. All the structures chosen have a resolution more accurate than 2.0 Å and contain a non-covalently bound ligand with at least 10 atoms. Metal ions and water molecules were removed from the protein. Hydrogen atoms were added to all the structures using the program BABEL [27]. Only those sitepoints within 6.0 Å of any atom of the ligand in the crystal structure were analysed. Values for

d_{rel} , $SAPS$ and $SWAPS$ were calculated as outlined before. Amphiprotic sitepoints, such as hydroxyl groups, were assigned two d_{rel} (and $SWAPS$) values, one for its donor ability and one for its acceptor ability.

A visual representation of the binding site is created (Figure 3) in which the sitepoints are coloured magenta for donors, yellow for acceptors and white for amphiprotic groups. For ease of visualisation of donors and amphiprotic groups, small cyan spheres represent the positions of their hydrogens. The sizes of the spheres representing the sitepoints are scaled proportionally to the final $SWAPS$ value awarded: the larger the sphere the higher the score. The remaining atoms within the binding site of the cavity are displayed as wireframe using CPK colouring in order to give an impression of the shape of the binding site.

Our analysis of the binding site of each of these proteins was compared with the actual hydrogen-bonding network observed between the ligand and the protein. In each of the 28 structures we used our methods to identify and rank the hydrogen-bonding sitepoints within the binding site. The hydrogen-bonding network formed between the ligand and protein was analysed using the geometric and atomic criteria for hydrogen bonding derived from a survey of the Cambridge Structural Database (CSD) [28]. These criteria consist of angular and distance ranges observed between complementary hydrogen bonding groups, so that if two complementary (i.e. donor and acceptor) hydrogen-bonding groups are positioned within the ranges specified for those particular groups, a hydrogen bond is reported.

Solvent-accessible surface areas (SASA) [17] were calculated using the NACCESS program [29]. Non-polar atoms were defined as any carbon atoms and the hydrogen atoms bound to them. This provided an estimate of the non-polar SASA of each ligand.

Results and discussion

The hydrogen-bonding analysis produced a list of hydrogen bonds between the ligand and protein. The number of hydrogen bonds formed between each ligand and protein set used in the validation is listed in Table 4. We can see that the number of hydrogen bonds found between protein and ligand varies between 0 and 11. The existence of protein-ligand complexes in which no hydrogen bonds are formed highlights the importance of other factors contributing

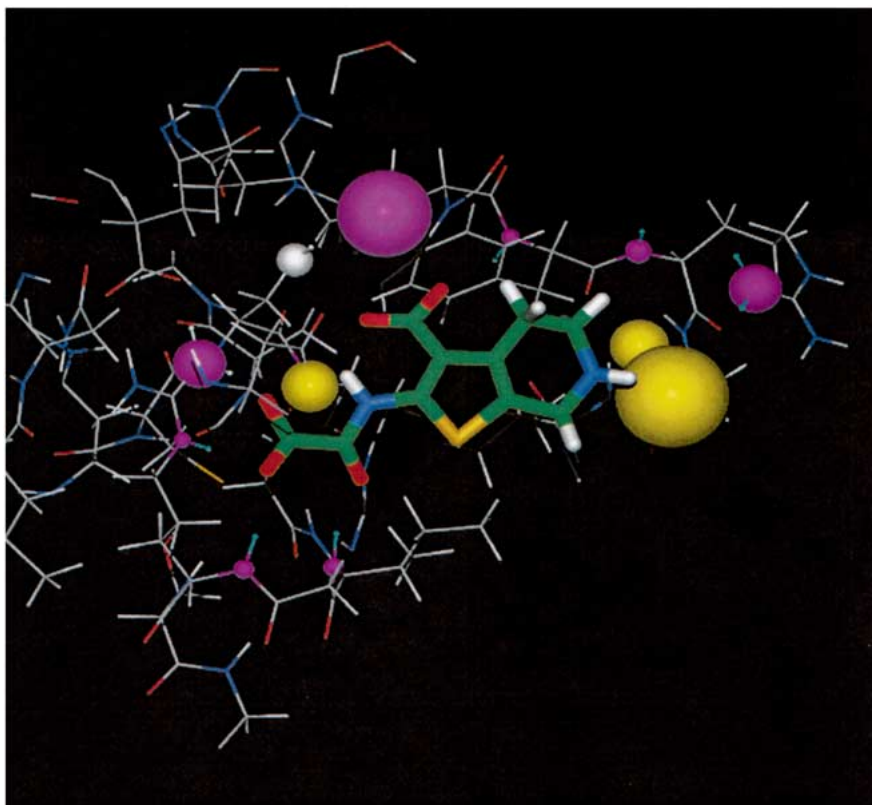


Figure 3. The graphical output for Protein Tyrosine Phosphatase 1B with ligand bound (PDB entry '1C88'). The ligand is shown with CPK colouring (except carbons are green). In our method, spheres representing donor sitepoints are coloured magenta, those for acceptor sitepoints are yellow and amphiprotic sitepoints are white. The diameters of the spheres are proportional to their *SWAPS* value. As can be seen in the figure, the larger sitepoints are within hydrogen-bonding distance of a complementary ligand group.

Table 4. The number of hydrogen bonds formed between the proteins used in the validation of PHARMASITE and their respective ligands.

PDB code	H-bonds observed	PDB code	H-bonds observed
19GS	0	1G7F	11
1AHT	6	1GD0	4
1BIO	5	1HI3	5
1BOA	1	1HNE	0
1BZS	9	1IDA	3
1C88	8	1J4R	3
1DVT	1	1JK7	3
1E1V	2	1KVO	1
1E2E	4	1QPC	3
1FDS	2	1QTN	1
1FKN	4	2OAT	9
1FJS	4	2SRC	3
1G32	3	2KI5	5
1G35	9	3GSS	9

to ligand binding, such as hydrophobic interactions and electrostatic potential matching.

The first indication of the ability of our methods to identify and prioritise hydrogen-bonding sitepoints within a binding site would be whether or not they can detect those highlighted by the hydrogen-bonding analysis. We were able to identify all hydrogen-bonding sitepoints used by the ligands in the test sets listed in Table 4, with the exception of two. The two hydrogen-bonding groups not identified were the SH groups of cysteine residue 203 in protein 1fjs and 216 in protein 1g32. Our methods rely on the HBMAP algorithm in order to generate a sitepoint accessibility score, but this method contains no data for sulphur groups. The HBMAP algorithm considers only nitrogen and oxygen atoms as hydrogen-bonding groups. The reason being that, compared to nitrogen and oxygen atoms, sulphur atoms formed very few hydrogen bonds in the crystal survey performed by Wilson and Famini [15] and therefore, representative 3D distri-

Table 5. The sum of SWAPS for all sitepoints identified by the hydrogen-bonding criteria analysis as forming a hydrogen bond with the ligand and the corresponding hydrophobicity score for each of the complexes used in the validation of our methods.

PDB code	Sum of SWAPS	Hydrophobicity score
19GS	0.0000	0.816
1AHT	4.6766	0.485
1BIO	2.0482	0.457
1BOA	0.4159	0.821
1BZY	9.9075	0.289
1C88	12.0311	0.476
1DVT	3.9076	0.744
1E1V	0.2092	0.684
1E2E	7.1188	0.445
1FDS	2.7178	0.826
1FKN	4.5203	0.741
1FJS	3.2088	0.624
1G32	3.6069	0.780
1G35	14.1114	0.801
1G7F	18.8287	0.684
1GD0	3.3853	0.334
1HI3	4.0620	0.283
1HNE	0.0000	0.761
1IDA	3.9365	0.803
1J4R	2.4862	0.812
1JK7	4.9000	0.831
1KVO	0.7000	0.840
1QPC	2.8755	0.302
1QTN	3.8896	0.382
2OAT	13.0617	0.475
2SRC	5.4934	0.259
2KI5	5.1864	0.446
3GSS	13.1163	0.548

butions could not be derived. Also, the propensity for these sulphur groups to form hydrogen bonds was much lower than for any of the other groups surveyed. Sulphur groups have been shown to form very weak and non-directional hydrogen bonds [30–31].

The next step in the validation of our methods was to verify that the sitepoints identified by the hydrogen-bonding analysis had ‘good’ *SWAPS* values. The *SWAPS* values for each of the sitepoints identified by the hydrogen-bonding criteria analysis were summed for each binding site studied, as shown in Table 5. We can see that the sum of *SWAPS* varies noticeably from 0 to almost 19. This is to be ex-

pected considering the varied importance of hydrogen bonding to ligand binding in the different protein–ligand complexes investigated, as well as the number of hydrogen-bonding groups found in the binding site of the proteins analysed (Table 4). We then assumed that the non-hydrogen bonding feature that contributes greatest to ligand binding is hydrophobicity. Hydrophobic interactions occur between non-polar atoms and, unlike hydrogen bonds, they are non-directional [32–33]. To provide a rough estimate of the degree to which hydrophobic interactions are involved in a particular protein–ligand interaction we took the non-polar SASA of the ligand (Table 5).

The relationship between the sum of *SWAPS* for those sitepoints identified by the hydrogen-bonding analysis and the non-polar SASA is shown in Figure 4. We can see that those protein–ligand complexes with a low sum of *SWAPS* (e.g. below 5) have a corresponding high value for hydrophobicity, and as the sum of *SWAPS* decreases further, the hydrophobicity score increases. Although this would be expected if the sum of *SWAPS* accurately represents the contribution of hydrogen bonding to the interaction, the converse is not necessarily true. For instance, a large ligand comprising mainly hydrophobic atoms and only one or two polar atoms may form hydrogen bonds through these polar atoms to corresponding sitepoints within the protein that score very highly according to our methods (e.g. a fully accessible oxygen atom of a charged carboxylic acid group). In this example, the sum of *SWAPS* for sitepoints identified by the hydrogen-bonding analysis would be high, and the non-polar SASA would also be high due to the large number of non-polar atoms in the ligand.

We then proceeded to investigate how the hydrogen-bonding sitepoints utilised by the ligand score relative to the other sitepoints identified by our methods. To examine this, we studied the position of the highest scoring sitepoint used by the ligand, the ‘anchor point’, relative to the highest scoring sitepoint in our analysis (Table 6).

Out of the 28 protein–ligand complexes studied, 24 of these had the highest-scoring utilised sitepoint in the top third of the hydrogen-bonding ranking list (25 were in the top half of the ranking list). Of the 4 protein–ligand complexes whose highest-scoring utilised sitepoint was not in the top third of the table, two (19gs and 1hne) formed no hydrogen bonds with the ligand and, along with the other two (1boa and 1e1v), all four have high hydrophobicity scores (0.816 for 19gs, 0.716 for 1hne, 0.821 for 1boa and 0.684

Table 6. Table listing highest SWAPS value observed for a sitepoint in each of the binding pockets along with the highest SWAPS value for a sitepoint utilised by the ligand. The ratio of these two values is also included, as is the position in the table of highest scoring utilised sitepoint (total number of sitepoints identified in brackets). The hydrophobicity score for each ligand is also shown.

PDB code	Highest in table	Highest used in table	Position in table	Highest observed/highest in table	Hydrophobicity score
19GS	1.700	0.000	N/A	0.00	0.816
1AHT	4.897	2.853	4 (26)	0.58	0.485
1BIO	4.771	0.699	5 (15)	0.15	0.457
1BOA	4.279	0.416	17 (35)	0.10	0.821
1BZY	4.538	2.321	4 (45)	0.51	0.289
1C88	4.798	4.798	1 (29)	1.00	0.476
1DVT	4.143	3.908	2 (20)	0.94	0.744
1E1V	4.000	0.115	18 (28)	0.03	0.684
1E2E	3.879	2.000	4 (35)	0.52	0.445
1FDS	4.899	1.178	3 (18)	0.24	0.826
1FKN	2.524	2.524	1 (33)	1.00	0.741
1FJS	4.896	1.700	5 (48)	0.35	0.624
1G32	3.750	2.841	3 (41)	0.76	0.780
1G35	4.846	4.261	4 (51)	0.88	0.801
1G7F	4.886	4.886	1 (41)	1.00	0.684
1GD0	1.240	1.223	2 (19)	0.99	0.334
1HI3	3.966	1.365	7 (26)	0.34	0.283
1HNE	2.000	0.000	N/A	0.00	0.761
1IDA	4.700	3.143	3 (36)	0.67	0.803
1J4R	2.516	0.915	7 (22)	0.36	0.812
1JK7	4.900	2.000	6 (39)	0.41	0.831
1KVO	3.379	0.700	9 (35)	0.21	0.840
1QPC	4.663	4.663	1 (25)	1.00	0.302
1QTN	3.890	3.890	1 (15)	1.00	0.382
2OAT	3.996	2.000	5 (52)	0.50	0.475
2SRC	4.874	4.000	3 (35)	0.82	0.259
2KI5	2.000	1.728	7 (29)	0.86	0.446
3GSS	4.639	2.531	3 (44)	0.55	0.548

for 1e1v), suggesting that hydrophobic interactions play a more significant role in binding the ligand for these cases. These results support our method's ability to prioritise key hydrogen bonding sitepoints (anchor points) in the protein's binding site.

In the orientation-based rejection component of our method, of the 84 sitepoints removed, none of them were highlighted by the hydrogen-bonding analysis as being directly involved in ligand binding, indicating that our rejection criterion is not too strict. Those sitepoints rejected in this method showed a range of SWAPS scores ranging from the very low for those buried sitepoints angled away from the pocket

into the core of the protein, to the high for the largely accessible sitepoints around the mouth of the pocket on the surface of the protein. This latter category of rejected sitepoints demonstrates that it is possible to find high scoring sitepoints outside of the main binding pocket.

During the course of the validation we observed that very high scoring sitepoints (e.g. charged carboxylic acid) may not always be used by a particular ligand. Instead, non-matched high-ranking sitepoints may be used as alternative anchor points for other ligands. To investigate this observation further, four proteins for which multiple crystal structures are

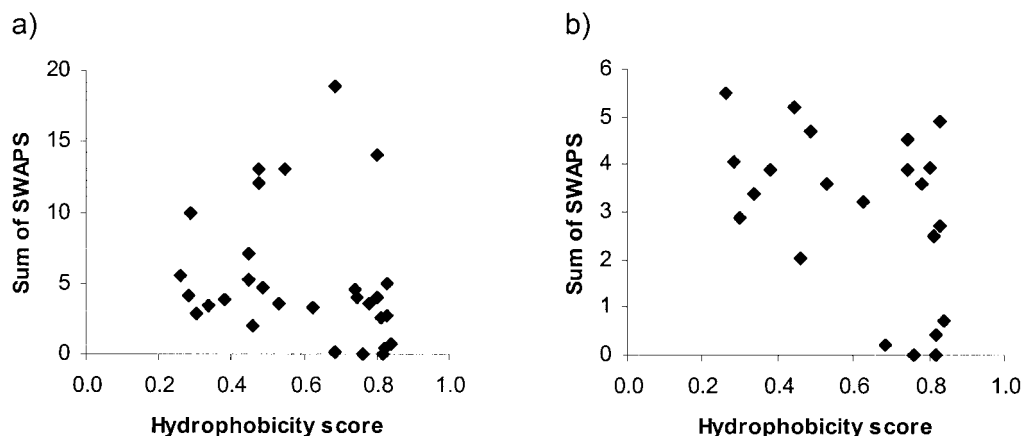


Figure 4. (a) Scatter plot revealing the relationship between the sum of *SWAPS* and the hydrophobicity score for each protein-ligand complex. (b) A closer look at the relationship between the lower-scoring sum of *SWAPS* and the hydrophobicity score. Those protein-ligand complexes with a very low sum of *SWAPS* have a corresponding high hydrophobicity score.

available with different ligands bound were identified. The proteins selected were: human carbonic anhydrase II, protein tyrosine phosphatase 1b, human cyclophilin a, and alpha thrombin (PDB codes of the structures used are shown in Table 7). By selecting proteins for which the binding modes of multiple ligands are known, a more comprehensive knowledge of useful sitepoints within a particular binding pocket can be gleaned. Our method was used to analyse each of the structures taken from the PDB and the outputs were collated. The sitepoints identified for each of the structures were divided into two populations. One population contains all sitepoints for which no incidence of hydrogen bonding to a ligand or water molecule is observed for that specific sitepoint in all structures for that protein. Sitepoints having been observed to form a hydrogen bond with a ligand or water molecule in at least one of the structures investigated for a particular protein are placed in the second population. For each of these populations, the mean *SWAPS* were calculated, and any difference between the two was tested for significance using the Student's t-test. As seen in Table 8, the mean *SWAPS* for the population containing utilised sitepoints is significantly higher (almost 4-fold) than that for the population containing unused sitepoints. This demonstrates our method's ability to award higher *SWAPS* to those sitepoints that are observed to participate in ligand and/or water binding when compared to those that are not. This analysis was then repeated using a different criterion for creating the populations. In this case, the utilised population contained only those sitepoints hydrogen bonded to a ligand or water molecule in each individual structure

investigated. The unutilised population contained all sitepoints identified by our methods for each crystal structure that are not hydrogen bonding to a ligand or water molecule even if they are used in another example for the same protein. Again, the mean *SWAPS* for the utilised population was significantly higher than that for the unused population (Table 9), although less so than the alternative subdivision of populations. This reduced difference would be expected, as sitepoints that are observed to be used in some crystal structures may be included in the unused population since there could be a structure in which that particular sitepoint is not used. This would have the effect of increasing the mean *SWAPS* for the unused population. However, one benefit of using this method for dividing the populations is that the frequency of use of a particular sitepoint in ligand and/or binding is taken into account. For example, using the first method of division, both a sitepoint observed to hydrogen bond to a ligand or water molecule only once out of ten structures investigated, and a sitepoint observed to bond on all ten occasions would contribute equally to the utilised population. The second method of population division does not suffer from this problem. It does, however, fail to utilise the more comprehensive knowledge of useful sitepoints within a particular binding pocket made available by the use of multiple ligands for a single binding site.

Given the range of *SWAPS* possible (0–4.9) a mean value in the used population of less than one may appear low (Tables 8 and 9). The reason for obtaining a value of this magnitude is that the distribution

Table 7. PDB codes for the multiple structures of the proteins: human carbonic anhydrase II, protein tyrosine phosphatase 1b, human cyclophilin a, and alpha thrombin used in the validation of our methods.

Human carbonic anhydrase II			Protein tyrosine phosphatase 1b	Human cyclophilin a	Alpha thrombin	
1A42	1CAN	1IF5	1AAX	1BCK	1A2C	1C5O
1AM6	1CAO	1IF6	1BZC	1CWF	1A3B	1G37
1AVN	1CIL	1IF7	1BZH	1CWH	1A3E	1HGT
1AZM	1CIM	1IF8	1BZJ	1CWI	1A4W	1IHT
1BCD	1CIN	1IF9	1C83	1CWK	1A61	1LHC
1BN1	1CRA	1OKL	1C84	1CWL	1AHT	1LHD
1BN3	1CRM	1OKM	1C85	1CWO	1AI8	1LHE
1BN4	1CZM	1OKN	1C86		1AIX	1LHF
1BNM	1EOU	1QQ0	1C87		1AWF	1LHG
1BNN	1G1D	1RAY	1C88		1BCU	1NRS
1BNQ	1G52	1RZB	1ECV		1BHX	1QHR
1BNT	1G53	1RZD	1EEN		1BMM	1THS
1BNU	1G54	1RZE	1EEO		1BMN	1TMB
1BNV	1I8Z	2CBC	1G7F		1C1U	1TOM
1BNW	1I90	2CBD	1G7G		1C1V	3HAT
1BV3	1I91		1KAV		1C1W	
1BZM	1IF4				1C5N	

Table 8. Mean SWAPS and corresponding standard errors for the two populations of sitepoints, utilised ($n = 2390$) and unutilised ($n = 186$). The significance of the difference between the two populations as determined by the Student's t-test is 1.87×10^{-16} .

	Mean	Standard error
Utilised population	0.856	0.021
Unused population	0.221	0.038

Table 9. Mean SWAPS and corresponding standard errors for the two populations of sitepoints, utilised ($n = 1387$) and unutilised ($n = 1189$), with the alternative division of populations. The significance of the difference between the two populations as determined by the Student's t-test is 1.11×10^{-6} .

	Mean	Standard error
Utilised population	0.899	0.027
Unused population	0.706	0.028

of *SWAPS* (not shown) has the majority of sitepoints carrying a *SWAPS* of below one.

Due to the dependence of the *SWAPS* value on the accessibility of a sitepoint, conformational variation within the binding pocket can affect the *SWAPS* ranking. For example, sitepoints that become more accessible through a rotameric side chain rearrangement will receive higher *SWAPS* values. This is likely to affect the probability of forming a hydrogen bond with an incoming ligand when the binding site exhibits a different conformation.

The validation of our methods highlighted the varying importance of hydrogen bonding to ligand binding with some interactions seemingly independent of them. By taking into consideration the contribution of other types of interaction, namely the hydrophobic one, it could be seen that the anchor points (highest scoring sitepoint used by the ligand) were ranked in the top third by our methods. Finally, our analysis demonstrated that those sitepoints observed to form hydrogen bonds are awarded higher *SWAPS* than those that are not. Clearly, this ability is key to the success of any binding-site analysis program.

Conclusions

We have developed a new method for estimating the importance of hydrogen-bonding sitepoints in the binding site of proteins as part of a structure-based design strategy. Our method identifies hydrogen-bonding sitepoints and ranks them according to both the accessibility of their hydrogen bonding regions (*solvent-accessible probability score*, 'SAPS') to incoming ligands and the strength of the hydrogen bond that they have the potential to form (d_{rel}). These components are then combined to produce a final *strength-weighted accessible probability score* (SWAPS). This ranking is performed independently of any information from the ligand.

To validate our method, a dataset of known protein-ligand interactions was used to compare the prioritisation of sitepoints identified by our methods, with those sitepoints observed to be involved in hydrogen bonding in the crystal structures. This analysis demonstrated the ability of our method to remove those sitepoints unable to bind the ligand due to low accessibility or an unfavourable orientation and, of the remaining, award significantly higher hydrogen-bonding values to those sitepoints observed to form hydrogen bonds.

In conclusion, our methods are able to identify and prioritise hydrogen-bonding sitepoints within a binding site, highlighting those that should then be targeted for drug design. However, only through the incorporation of additional factors such as hydrophobicity can such a method give a complete account of the binding properties of a ligand-binding site.

Acknowledgements

M.D.K. would like to thank *De Novo* Pharmaceuticals for the award of a Ph.D. studentship. R.L.M. is also a Research Fellow of Hughes Hall, Cambridge.

References

- Goodford, P.J., *J. Med. Chem.*, 28 (1985) 849.
- Kuntz, I.D., *Science*, 257 (1992) 1078.
- Bohm, H.J., *J. Comput.-Aided Mol. Des.*, 6 (1992) 593.
- Gillet, V.J., Johnson, A.P., Mata, P., Sik, S. and Williams, P., *J. Comput.-Aided Mol. Des.*, 7 (1993) 127.
- Miranker, A. and Karplus, M., *Proteins: Struct., Func. and Genetics*, 11 (1991) 314.
- Glen, R.C. and Payne, A.W.R., *J. Comput.-Aided Mol. Des.*, 9 (1995) 181.
- Pearlman, D.A. and Murcko, M.A., *J. Med. Chem.*, 39 (1996) 1651.
- Lewis, R.A. and Leach, A.R., *J. Comput.-Aided Mol. Des.*, 8 (1994) 467.
- Lauri, G. and Bartlett, P.A., *J. Comput.-Aided Mol. Des.*, 8 (1994) 51–66.
- Moon, J.B. and Howe, W.J., *Proteins: Struct., Func. and Genetics*, 11 (1991) 29.
- Todorov, N.P. and Dean, P.M., *J. Comput.-Aided Mol. Des.*, 11 (1997) 175.
- Jeffrey, G.A. and Saenger, W., *Hydrogen Bonding in Biological Structures*, Springer-Verlag, Berlin, 1991.
- Dean, P.M., In Dean, P.M. (Ed.), *Molecular Similarity in Drug Design*, Blackie A & P, UK, 1995, pp. 7–8.
- Fujita, T., Nishioka, T. and Nakajima, M., *J. Med. Chem.*, 20 (1977) 1071.
- Wilson, L.Y. and Famini, G.R., *J. Med. Chem.*, 34 (1991) 1668.
- Dearden, J.C. and Ghafourian, T., *J. Chem. Inf. Comput. Sci.*, 39 (1999) 231.
- Dearden, J.C. and Ghafourian, T., In Sanz, F., Giraldo, J. and Manaut, F. (Eds.), *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, Prous Science Publishers, Barcelona, 1995, pp. 117–119.
- Gancia, E., Montana, J.G. and Manallack, D.T., *J. Mol. Graph. Mod.*, 19 (2001) 349.
- Raevsky, O.A., *J. Phys. Org. Chem.*, 10 (1997) 405.
- Raevsky, O.A., *Newsletter of QSAR and Modelling Society*, 7 (1996) 16.
- Abraham, M.H., Duce, P.P., Prior, D.V., Barrat, D.G., Morris, J.J. and Taylor, P.J., *J. Chem. Soc. Perkin Trans.*, 2 (1989) 1355.
- Bruno, I.J., Cole, J.C., Lommerse, J.P.M., Rowland, R.S., Taylor, R. and Verdonk, M.L., *J. Comput.-Aided Mol. Design*, 11 (1997) 525.
- Lee, B. and Richards, F.M., *J. Mol. Biol.*, 55 (1971) 379.
- Taylor, R., Kennard, O. and Versichel, W., *J. Am. Chem. Soc.*, 105 (1983) 5761.
- Mills, J.E.J., Perkins, T.D.J. and Dean, P.M., *J. Comput.-Aided Mol. Des.*, 11 (1997) 229.
- Mills, J.E.J. and Dean, P.M., *J. Comput.-Aided Mol. Des.*, 10 (1996) 607.
- Walters, M. and Stahl, M., *BABEL version 1.6; Copyright (C) 1992, 1993, 1994*. <http://smog.com/chem/babel>.
- Mills, J.E.J. (1995) *Analysis of hydrogen-bond data applied to drug-design strategies*, Doctoral Dissertation, University of Cambridge, Cambridge, U.K., 1995.
- Hubbard, S.J. and Thornton, J.M., *NACCESS V2.1.1*, <http://wolf.bms.umist.ac.uk/naccess>.
- Allen, F.H., Bird, C.M., Rowland, R.S. and Raithby, P.R., *Acta Cryst.*, B53 (1997) 680.
- Allen, F.H., Bird, C.M., Rowland, R.S. and Raithby, P.R., *Acta Cryst.*, B53 (1997) 696.
- Casari, G. and Sippl, M.J., *J. Mol. Biol.*, 224 (1992) 725.
- Islam, S.A. and Weaver, D.L., *Proteins*, 8 (1990) 1.