



Calculating the knowledge-based similarity of functional groups using crystallographic data

Paul Watson^a, Peter Willett^a, Valerie J. Gillet^a & Marcel L. Verdonk^{b,*}

^aKrebs Institute for Biomolecular Research, Department of Information Studies, University of Sheffield, S10 2TN, UK; ^bCambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK

Received 20 October 2000; accepted 10 August 2001

Key words: bioisosterism, crystallographic databases, knowledge-based similarity, non-bonded interactions, structure-based design, 3D molecular similarity

Summary

A knowledge-based method for calculating the similarity of functional groups is described and validated. The method is based on experimental information derived from small molecule crystal structures. These data are used in the form of scatterplots that show the likelihood of a non-bonded interaction being formed between functional group A (the 'central group') and functional group B (the 'contact group' or 'probe'). The scatterplots are converted into three-dimensional maps that show the propensity of the probe at different positions around the central group. Here we describe how to calculate the similarity of a pair of central groups based on these maps. The similarity method is validated using bioisosteric functional group pairs identified in the BioSter database and Relibase. The BioSter database is a critical compilation of thousands of bioisosteric molecule pairs, including drugs, enzyme inhibitors and agrochemicals. Relibase is an object-oriented database containing structural data about protein-ligand interactions. The distributions of the similarities of the bioisosteric functional group pairs are compared with similarities for all the possible pairs in IsoStar, and are found to be significantly different. Enrichment factors are also calculated showing the similarity method is statistically significantly better than random in predicting bioisosteric functional group pairs.

Introduction

The concept of *bioisosterism*, as introduced by Friedman[1], was originally used to describe functional groups or molecules that have chemical and physical similarities producing broadly similar biological properties. The structural knowledge gained from protein crystallography has allowed this definition to be refined. Bioisosteres can now be described as molecules or functional groups that are structurally different but form similar intermolecular interactions [2].

Intermolecular interactions between ligands and protein functional groups are fundamental to the way in which drug molecules bind to a given active site.

Non-bonded interactions provide the driving force for ligand-active site binding as well as the specificity of the receptor site to certain ligands [3]. As such, non-bonded interactions between functional groups have been studied extensively. The principal techniques for studying non-bonded interactions between small functional groups are theoretical (quantum mechanical) and knowledge-based techniques [4], which involve the statistical analysis of crystallographic databases. Each method has its advantages and disadvantages. Theoretical methods can provide information about the strength and geometry of intermolecular interactions; however, these are usually only calculated in the gas-phase and pertain isolated complexes. Crystallographic databases offer the most detailed and reliable source of information about the frequencies and geometries of non-bonded interactions [4], but direct information about the strength of a given interaction

*To whom correspondence should be addressed. Present address: Astex Technology Ltd, 250 Cambridge Science Park, Cambridge CB4 0WE, UK. E-mail: m.verdonk@astex-technology.com

is not available. Crystallographic databases such as the Cambridge Structural Database [5] (CSD) and the Protein Data Bank [6] (PDB) have been used widely in the past to investigate non-bonded interactions [7–12]. The interest shown in utilising crystallographic databases for this purpose has led to the formation of the IsoStar database [13].

IsoStar is a library of information about intermolecular interactions, based on experimental information from the CSD and the PDB. The information is presented in the form of scatterplots showing the distribution of one functional group (the ‘contact group’) around another (the ‘central group’). The distributions of the contact groups can be converted into propensity maps for a specific contact group atom (the ‘probe’), which can be viewed as a contoured surface. The propensity map shows the probability of finding a probe atom at a point around the central group compared to that expected by chance. IsoStar V1.1 (which was used for this work) contains scatterplots for 298 central groups combined with up to 43 contact groups. There are 10,379 scatterplots derived from the CSD and 2,563 scatterplots derived from the PDB. The name IsoStar was derived from bioisostere, i.e. the intended use of the database was to find bioisosteric functional groups, in terms of their non-bonded interactions.

Similarity methods, using 3D fields such as MEPs (Molecular Electrostatic Potential), lipophilic fields and steric fields, are now widely used to identify bioisosterically similar whole molecules [14]. Whilst the fields do not explicitly define the preferential non-bonded interactions formed by the structures within the database, the fields themselves are responsible for non-bonded interactions formed and as such infer such information. These 3D approaches usually involve aligning the molecules in a position of maximum similarity, using any one of a number of similarity indices [15–17] to calculate the similarity of the putative alignment, usually based upon the extent of overlap of the 3D fields. The generated alignments are of particular use in the early stages of the drug-design process in which structures of the protein-inhibitor complexes are not available, and as such are useful for generating models of ligand-binding alignment [18]. The alignments can also be used in 3D QSAR studies and pharmacophore elucidation [19–21]. Many automated procedures have been developed for the alignment of whole molecules and the calculation of their similarities based upon 3D fields. Lemmen and Lengauer give a comprehensive review of these techniques [22].

As far as we are aware, the calculation of the similarity and alignment of *small functional groups* contained within ligand molecules has so far been overlooked. However such a technique would be of great use in identifying novel bioisosteric replacements for functional groups. Here we describe the development and validation of a technique that uses the scatterplots in the IsoStar database to identify bioisosteric functional groups. In the current version of IsoStar the user must compare scatterplots manually for bioisosteric functional groups to be detected. Here, we compute the similarity of the central groups within the database in terms of the propensity maps produced from CSD-based IsoStar scatterplots, in order to identify bioisosteric functional groups in an automated fashion.

Two types of similarity measure are described: *single probe* similarities and *diverse set* similarities. The single probe similarity between two central groups is calculated in terms of a single probe atom from a given contact group, in effect only taking one interaction type into consideration. The diverse set similarity between two central groups is calculated using four probe atoms: any N–H hydrogen, any O–H oxygen, any C=O oxygen and *methyl* carbon. In this case a range of interaction types is taken into account when calculating the similarity of the central groups. These interaction types have been shown to be highly diverse in a previous study [23]. In both cases Gaussian approximations of the propensity maps, first suggested by Good et al. [24], are used to increase the speed of the alignment process. The results of example similarity searches are shown in which a target functional group is compared with all others in IsoStar.

The method is validated using the Bioster database [25] and Relibase [26]. Both databases are not compilations of bioisosteric functional group pairs, however we describe methods for selecting bioisosteric functional group pairs from their parent molecules in the validation section. The Bioster database is a critical compilation of thousands of bioisosteric molecule pairs, including drugs, enzyme inhibitors and agrochemicals. The entries are selected from literature spanning the past 35 years. The January 1998 version of the database was used for the purposes of this validation and contains 3545 bioisosteric pairs of molecules, derived from 5499 active compounds from 4727 literature citations. Relibase is an object-oriented database containing structural data about protein-ligand interactions derived from the PDB. The February 2000 alpha test version was used for the purposes of this

study. 185 bioisosteric functional group pairs were retrieved from the BioStar database and 137 bioisosteric functional group pairs were retrieved from Relibase. The manner in which bioisosteric functional group pairs were retrieved from both databases is described in the validation section. In each case the distributions of their similarities are compared with the distribution of similarities for all the possible pairs of functional groups in IsoStar for both single probe and diverse set methods.

Methods

Overview of single probe and diverse set similarity calculations

The steps involved in calculating the similarity of a pair of central groups in terms of a given probe atom (i.e., the single probe method) and for the diverse set of probes (i.e., the diverse set method) are listed below. Each of the stages is then described in detail in the following subsections.

- (1) Generate propensity maps for the chosen probe atom(s) for both central groups.
- (2) Model the propensity maps using Gaussian functions.
- (3) Overlay the central groups based on the degree of overlap of the Gaussian functions.
- (4) Calculate the similarity based on the optimal overlay of the central groups.

Generation of propensity maps

(a) *IsoStar scatterplots.* A complete description of the techniques employed to create the IsoStar database is given by Bruno et al. [13]. The Quest package was used to search the CSD for non-bonded contacts between two functional groups, A and B, where a non-bonded contact is defined as an intermolecular contact between a pair of target atoms (defined below) in A and B shorter than $d_{\text{vdw}} + 0.5 \text{ \AA}$, where d_{vdw} is the sum of the van der Waals radii of the atoms involved. Target atoms are those which define a non-bonded contact, the remaining atoms within A and B (backbone atoms) are not considered when defining the non-bonded contacts and are used only to define the chemical connectivity of the group. The scatterplots are created by overlaying the A moieties, thus giving the scattered distribution of B (the contact group) around A (the central group). The same procedure was also carried out using the PDB. In this study scatter-

plots derived from the CSD were used from IsoStar V1.1.

(b) *Calculation of density maps.* Propensity maps can be calculated for a selected probe atom, where a probe atom is defined as an atom in the contact group, e.g., any C=O oxygen. The first stage of the propensity map generation is the determination of the optimal grid spacing, based upon the amount of data contained within the scatterplot. The grid spacing was obtained by first calculating the average number of probe atoms per grid cube, via Equation 1:

$$n_g(j) = \frac{\Delta^3 n_p(j)}{V_a(j)}, \quad (1)$$

where $n_g(j)$ is the average number of probe atoms per grid cube, $n_p(j)$ is the number of probe atoms in scatterplot j , Δ is the gridspacing and $V_a(j)$ is the volume around the central group accessible to the chosen probe. $V_a(j)$ was estimated by stochastic sampling of the volume around the central group. $n_g(j)$ is calculated for Δ from 0.5 up to and including 1.5 Å. The minimum grid spacing for which $n_g(j)$ was greater than or equal to 4.0 was used. If $n_g(j)$ was less than 4.0 for a grid spacing of 1.5 Å, the similarity calculation was not performed.

The probe density is the number of probes per unit volume. For each scatterplot j , each probe atom k contributes to the densities at the eight surrounding grid points, the quantity of which has an inverse dependence on the distance from the grid point (Equation 2).

$$w(i, j, k) = \frac{[r(i, j, k)]^{-1}}{\sum_{i'=1}^8 [r(i', j, k)]^{-1}}, \quad (2)$$

$w(i, j, k)$ is the contribution of the probe atom k to the density at grid point i . $r(i, j, k)$ is the distance of the probe atom k to the grid point i . The summation is over the eight grid points i' surrounding the probe atom k . The probe density $d(i, j)$ at each grid point i is calculated using Equation 3.

$$d(i, j) = \frac{1}{\Delta^3} \sum_{k=1}^{n_p(j)} w(i, j, k). \quad (3)$$

(c) *Normalisation of density maps.* The densities at each grid point are normalised using the 'average density' of the contact group. The average density is calculated from the occurrence of the contact group within the CSD database, and is provided for all

the scatterplots derived from the CSD. The average density of scatterplot j is given as:

$$d_{av}(j) = \sum_c \frac{N_{\text{central}}(c, j) \cdot N_{\text{contact}}(c, j)}{V_{\text{cell}}(c)} \quad (4)$$

where $d_{av}(j)$ is the average density of the contact group (i.e., probe atoms) in scatterplot j , if the contact groups and central groups were uniformly distributed in each crystal structure. $N_{\text{central}}(c, j)$ is the number of unique central groups and $N_{\text{contact}}(c, j)$ is the total number of contact groups found in the unit cell of crystal structure c that has unit volume $V_{\text{cell}}(c)$. The summation is over all crystal structures that contain both the central group and contact group which could potentially contribute to scatterplot j . The density map is normalised by dividing the probe density at each grid point by the average density, yielding the propensity $p(i, j)$ (Equation 5).

$$p(i, j) = \frac{d(i, j)}{d_{av}(j)}. \quad (5)$$

The propensity, $p(i, j)$, indicates whether the density at grid point i in scatterplot j is higher or lower than expected by chance. A propensity of 2.0 at a given grid point implies that contacts at that grid point are twice as frequent as would be expected by chance.

(d) *Multiple probe correction.* Using the average density to normalise the density maps is not necessarily correct for contact groups that contain chemically equivalent probe atoms. An example of such a multiple probe is the *methyl* hydrogen probe. In this case a possible three probes from each occurrence of the contact group within the scatterplot can be included in the propensity map. Normalisation using the average density would only be correct if one probe atom from each contact group is used. As such the average density is modified for these probe atoms (Equation 6).

$$md_{av}(j) = \frac{n_p(j) \cdot d_{av}(j)}{n_c(j)}. \quad (6)$$

Here, $n_p(j)$ is the number of probe atoms and $n_c(j)$ is number contact groups in scatterplot j used to construct the propensity map. $md_{av}(j)$ is the modified average density.

(e) *Hydrophobic correction.* It has been shown in a previous study [27] that hydrophobic interactions are more likely to occur at protein-ligand interfaces than in small molecule crystal structures. This ‘hydrophobic effect’ will not be sufficiently incorporated

into CSD-based propensity maps generated using hydrophobic probe atoms that are forming close contacts with hydrophobic central group atoms. Therefore the propensities in these maps are multiplied by a factor, $P(\text{hydro}) = 2.0$, suggested previously by Verdonk et al. [28].

(2) Calculation of peak maps

Peak map representations of the propensity maps can be generated using 3D Gaussian functions [27]. The peaks contained within the propensity map for scatterplot j are located at the grid-points i for which $p(i, j) > p(i', j)$ for all the 26 surrounding grid-points i' . One-term 3D Gaussian functions are assigned for each peak (see Equation 7).

$$F(\mathbf{v}, j, g) = h_p(j, g) \cdot \exp\left(-\frac{|\mathbf{v} - \mathbf{v}_p(j, g)|^2}{2\sigma_p^2(j, g)}\right) \quad (7)$$

where $F(\mathbf{v}, j, g)$ is the propensity at position \mathbf{v} , from peak g , used to describe propensity map j . $h_p(j, g)$, $\mathbf{v}_p(j, g)$ and $\sigma_p(j, g)$ are all estimated from the 26 surrounding grid-points i' that surround i , in propensity map j . $h_p(j, g)$ is the maximum height of the peak, $\mathbf{v}_p(j, g)$ is the position of the maximum and $\sigma_p(j, g)$ controls the rate of decay. A peak map, M , representing the propensity map in ‘Gaussian form’ is given by Equation 8.

$$M(\mathbf{v}, j) = \sum_g F(\mathbf{v}, j, g), \quad (8)$$

where $M(\mathbf{v}, j)$ is the propensity at position \mathbf{v} in the peak map used to approximate propensity map j . The summation is over all Gaussian peaks. The peak fitting process is repeated a number of times in order to detect second order peaks. From the list of peaks already identified a difference map, D , can be calculated via Equation 9 giving the difference between the propensity map and the peak map.

$$D(i, j) = p(i, j) - M(\mathbf{v}(i), j), \quad (9)$$

where $D(i, j)$ represents the difference map at grid point i and $\mathbf{v}(i)$ is the positional vector corresponding to grid point i . The local maxima can be detected in the difference map and fitted, as before, with Gaussian functions. These functions can be added to the list of peaks and the cycle can be repeated replacing $p(i, j)$ with $D(i, j)$.

Peaks are only found if the following criteria are met. The first of these is that the peaks are not allowed within a certain distance tolerance, d_{min} , of each other. The second criterion is that the height of the peak is

above a certain tolerance h_{\min} (i.e., $h_p > h_{\min}$). For this work $d_{\min} = 0.5 \text{ \AA}$ and $h_{\min} = 0.01$. The peak fitting cycle was repeated three times.

The final step of the peak fitting process optimises the parameters $h_p(j, g)$, $v_p(j, g)$ and $\sigma_p(j, g)$, by minimising the root-mean-square difference between the original propensity map and the peak map, using the SIMPLEX algorithm [29]. Using this methodology peak maps were generated for all of the propensity maps, for which there is enough data in the IsoStar database. 85% of these peak maps have a similarity greater than 70% compared to the corresponding original propensity map.

(3) Superposition of central groups

The superposition procedure involves overlaying the Gaussian peak maps representing the propensity maps in a position of maximum similarity, whilst placing a restraint upon the relative positions of the chosen connecting bonds from the two central groups. The restraint imposed on the connecting bonds is to ensure a chemically relevant superposition of the Gaussian peak maps is attained, and hence a bioisosterically relevant similarity calculation. The SIMPLEX algorithm is used to find the optimal superposition of the Gaussian peak maps, by maximising Equation 10.

$$S(A, B) = R(A, B) - Q(A, B). \quad (10)$$

Here, $S(A, B)$ is the score for the alignment of central groups A and B . $R(A, B)$ is the similarity in terms of the Gaussian peak maps (calculated using the Carbo coefficient [15]) and is given by Equation 11:

$$R(A, B) = \frac{\int_v M(v, A)M(v, B) dv}{\left(\int_v [M(v, A)]^2 dv\right)^{1/2} \left(\int_v [M(v, B)]^2 dv\right)^{1/2}} \quad (11)$$

$$0 < R(A, B) \leq 1.$$

The integrations are over the entire grid. At this point, the use of the Gaussian peak map descriptions of the original maps becomes apparent. The integral over the product of two Gaussian functions is a straightforward analytical expression. Hence Equation 11 can be evaluated very quickly, which is essential at this superposition stage. In the case of the diverse set method, $R(A, B)$ is calculated for the four diverse probes, and the average value is used in Equation 10.

$Q(A, B)$ is a penalty calculated from the difference in position of the connecting bond from each of the

central groups.

$$Q(A, B) = C \cdot ([r_1(A, B)]^2 + [r_2(A, B)]^2). \quad (12)$$

In the above equation $r_1(A, B)$ and $r_2(A, B)$ are the distances between the first and second atom in the connecting bonds in central groups A and B . C is a constant.

The purpose of such a function is to ensure that the final overlay will represent a situation of maximum overlap of the Gaussian approximations from the different scatterplots whilst keeping the connecting bonds aligned. The inclusion of the constant C in equation 12 allows for some flexibility in the amount that the connecting bonds of the central group are allowed to deviate from being perfectly overlaid. After testing, a value of $C = 1.0 \text{ \AA}^{-2}$ was used. If the central groups being compared have more than one connecting bond (e.g., linking and ring groups), similarity calculations are made for all possible combinations of connecting bonds.

(4) Calculation of final grid-based similarity

After the optimum superposition is found, a final grid-based similarity calculation is performed because it is more accurate than the similarity calculations conducted using the Gaussian peak maps. Both the Carbo [15] and Hodgkin [16] similarities are calculated via Equations 13 and 14, respectively. In the case of the diverse set method, the grid-based similarity calculation is performed for each pair of propensity maps for each probe atom. The mean similarity is then taken as the diverse set similarity.

$$R(A, B) = \frac{\sum_i p(i, A)p(i, B)}{\sqrt{\sum_i [p(i, A)]^2} \sqrt{\sum_i [p(i, B)]^2}} \quad (13)$$

$$0 \leq R(A, B) \leq 1,$$

$$H(A, B) = \frac{2 \sum_i p(i, A)p(i, B)}{\sum_i [p(i, A)]^2 + \sum_i [p(i, B)]^2} \quad (14)$$

$$0 \leq H(A, B) \leq 1.$$

Here, $p(i, A)$ is the propensity at grid point i for scatterplot A and $p(i, B)$ is the propensity at grid point i for scatterplot B . $R(A, B)$ is the Carbo coefficient and $H(A, B)$ is the Hodgkin coefficient for the scatterplots A and B . The summations are over the entire grid.

Table 1. Top ten-ranked functional group replacements for the carbamoyl group calculated using propensity maps derived from the any N-H hydrogen atom, using the Carbo, $R(A, B)$, and Hodgkin, $H(A, B)$, coefficients. The connecting bonds used to overlay the groups are highlighted

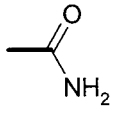
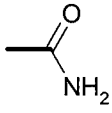
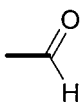
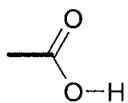
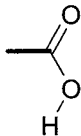
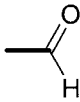
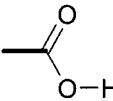
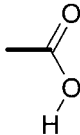
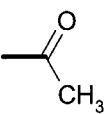
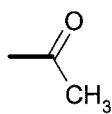
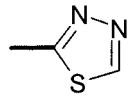
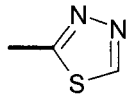
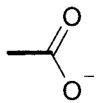
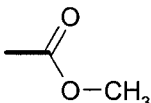
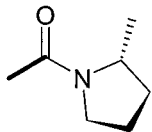

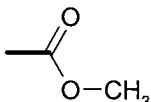
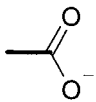

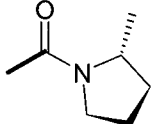
Rank	Functional group	$H(A, B)$	Functional group	$R(A, B)$
1	 carbamoyl	1.00	 carbamoyl	1.00
2	 formyl	0.91	 uncharged carboxylic acid- <i>cis</i>	0.91
3	 uncharged carboxylic acid- <i>trans</i>	0.90	 formyl	0.91
4	 uncharged carboxylic acid- <i>cis</i>	0.87	 uncharged carboxylic acid- <i>trans</i>	0.90
5	 acetyl	0.83	 acetyl	0.86
6	 1,3,4-thiadiazole	0.72	 1,3,4-thiadiazole	0.73

Table 1. (Continued)

Rank	Functional group	$H(A,B)$	Functional group	$R(A,B)$
7	 charged carboxylic acid	0.65	 methoxycarbonyl	0.68
8	 prolyl-up conformer C-C(C=O) ^a	0.64	 epoxide	0.67
9	 methoxycarbonyl	0.59	 charged carboxylic acid	0.66
10	 epoxide	0.56	 prolyl-up conformer C-C(C=O) ^a	0.65

^aC-C(C=O) denotes the connecting bond used for alignment purposes in the prolyl group.

Results

This section shows the results of some example similarity searches in which a target functional group is compared with all others in IsoStar (provided there is enough data for the chosen probe). Note that functional groups are only compared if they can replace one another, i.e. terminal functional groups are not compared with linking groups. The target functional groups used in the example searches are (a) carbamoyl, (b) planar 4-aminophenyl and (c) aliphatic-aliphatic amide.

(a) Carbamoyl

Table 1 shows ranked lists of the ten most similar functional groups in IsoStar to the carbamoyl group,

calculated using the single probe method for both the Carbo and Hodgkin coefficients. The probe used was the any N-H hydrogen probe atom. As can be seen the lists contain many functional groups that are fairly isostructural to the target group. Figure 1 shows an example alignment of the target group with that of the *cis*-carboxylic acid group ($R(A, B) = 0.91$, $H(A, B) = 0.87$). In this case the functional groups and the propensity maps are shown displaced from the optimum alignment so that the similarity in the propensity maps can be seen. The 1,3,4-thiadiazole ($R(A, B) = 0.73$, $H(A, B) = 0.72$) group is perhaps a less obvious bioisosteric replacements for the carbamoyl group; however if the propensity maps are observed for these groups then the relatively high similarity scores are justified. Figure 2 shows the

Table 2. Top ten-ranked functional group replacements for the planar 4-aminophenyl group calculated using propensity maps derived from the any C=O oxygen atom, using the Carbo, $R(A, B)$, and Hodgkin, $H(A, B)$, coefficients. The connecting bonds used to overlay the groups are highlighted

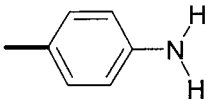
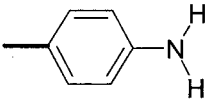
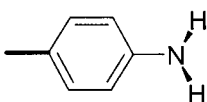
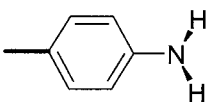
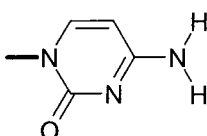
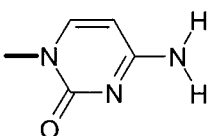
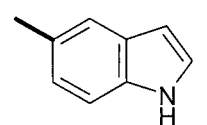
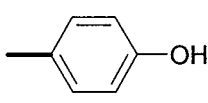
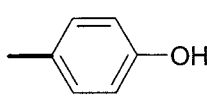
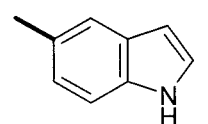
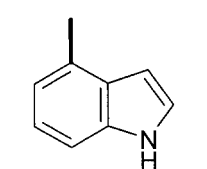
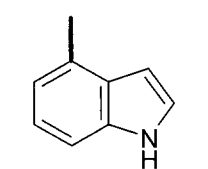
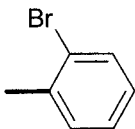
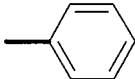
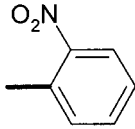
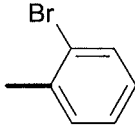
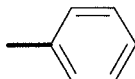
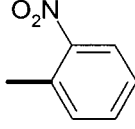
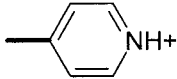
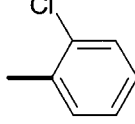
Rank	Functional group	$H(A, B)$	Functional group	$R(A, B)$
1	 planar 4-aminophenyl	1.00	 planar 4-aminophenyl	1.00
2	 pyramidal 4-aminophenyl	0.87	 pyramidal 4-aminophenyl	0.87
3	 uncharged cytosine	0.82	 uncharged cytosine	0.83
4	 indole C5 ^a	0.69	 4-hydroxyphenyl	0.71
5	 4-hydroxyphenyl	0.68	 indole C5 ^a	0.70
6	 indole C4 ^a	0.64	 indole C4 ^a	0.65

Table 2. (Continued)

Rank	Functional group	$H(A,B)$	Functional group	$R(A,B)$
7	 2-bromophenyl	0.56	 phenyl	0.60
8	 2-nitrophenyl	0.50	 2-bromophenyl	0.57
9	 phenyl	0.45	 2-nitrophenyl	0.52
10	 charged pyridine C4 ^a	0.44	 2-chlorophenyl	0.51

^aC4, C5 denote the carbon atoms in the indole and charged pyridine groups, that the connecting bond is attached to, used for alignment purposes.

propensity maps for the carbamoyl group and the 1,3,4-thiadiazole group, showing the similarity in the propensity maps thus explaining the high similarity score.

The difference in the Carbo and Hodgkin coefficients is illustrated if the similarities and rank numbers of the methylene group are observed. In the case of the Carbo coefficient the similarity is 0.57 and the rank number is 13, whereas the Hodgkin coefficient gives a similarity of 0.21 and ranks the epoxide group 66th. The reason for the difference is that the Hodgkin coefficient takes into account the difference in magnitudes of the propensity maps being compared as well as the extent of the overlap, while the Carbo coefficient only takes into account the overlap of the propensity maps. The magnitude of the propensity maps for the carbamoyl and methylene groups, calculated using the

any N–H hydrogen probe atom, are quite different. As expected, forming hydrogen bonds with the carbonyl group of the carbamoyl functional group is more favourable than with the methylene group. As such, the Hodgkin coefficient ranks the methylene group much lower than in the case of the Carbo coefficient.

(b) Planar 4-aminophenyl

The ranked lists for the ten most similar functional groups for this group are shown in Table 2, again calculated using the Carbo and Hodgkin coefficients, but this time the any C=O oxygen atom is used as the probe. Predictably the pyramidal conformer of the 4-aminophenyl group is most similar ($R(A, B) = 0.87$, $H(A, B) = 0.87$); however the uncharged cytosine, ranked third, is perhaps a less obvious bioisosteric substitution ($R(A, B) = 0.83$, $H(A, B) = 0.82$), until

Table 3. Top ten-ranked functional group replacements for the aliphatic-aliphatic amide group calculated using the diverse set method, using the Carbo, $R(A, B)$, and Hodgkin, $H(A, B)$, coefficients. The connecting bonds used to overlay the groups are highlighted

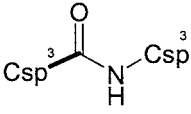
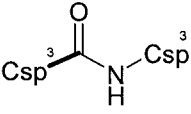
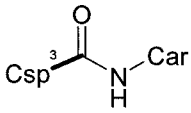
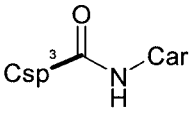
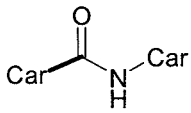
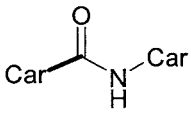
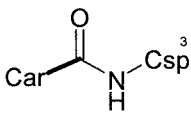
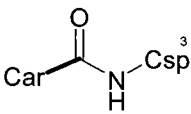
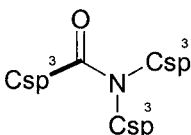
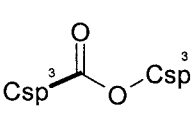
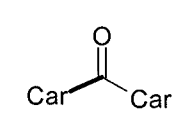
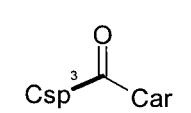
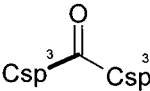
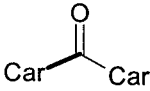
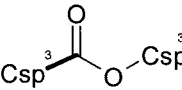
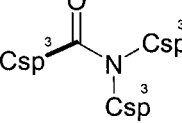
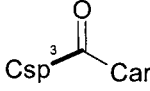
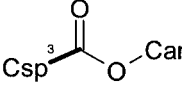
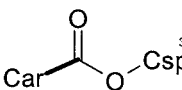
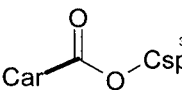
Rank	Functional group	$H(A, B)$	Functional group	$R(A, B)$
				
1	aliphatic-aliphatic amide $C(sp^3)-C(=O)^a$	1.00	aliphatic-aliphatic amide $C(sp^3)-C(=O)^a$	1.00
				
2	aromatic-aliphatic amide $C(sp^3)-C(=O)^a$	0.92	aromatic-aliphatic amide $C(sp^3)-C(=O)^a$	0.95
				
3	aromatic-aromatic amide $C(ar)-C(=O)^a$	0.81	aromatic-aromatic amide $C(ar)-C(=O)^a$	0.86
				
4	aliphatic-aromatic amide $C(ar)-C(=O)^a$	0.79	aliphatic-aromatic amide $C(ar)-C(=O)^a$	0.84
				
5	amide, N,N-disubstituted $C-C(=O)^a$	0.64	aliphatic-aliphatic ester $C(sp^3)-C(=O)^a$	0.70
				
6	aromatic-aromatic ketone	0.63	aliphatic-aromatic ketone $C(sp^3)-C(=O)^a$	0.67

Table 3. (Continued)

Rank	Functional group	$H(A,B)$	Functional group	$R(A,B)$
			$C(=O)^a$	
				
7	aliphatic-aliphatic ketone	0.61	aromatic-aromatic ketone	0.67
				
8	aliphatic-aliphatic ester $C(sp^3)-C(=O)^a$	0.59	amide, N,N-disubstituted $C(=O)^a$	0.66
				
9	aliphatic-aromatic ketone $C(sp^3)-C(=O)^a$	0.57	aromatic-aliphatic ester ($C(sp^3)-C(=O)^a$)	0.66
				
10	aliphatic-aromatic ester $Car-C(=O)^a$	0.57	aliphatic-aromatic ester $C(ar)-C(=O)^a$	0.64

^a $C(hyb)-C(=O)$, where $hyb = sp^3$ or ar , indicates the connecting bond used for superposition purposes.

the propensity maps for the two groups are observed (shown in Figure 3).

(c) Aliphatic-aliphatic amide group

In this instance the diverse set method was used to calculate the similarity of the target group with that of the others in the IsoStar database. The groups were aligned with the $C(sp^3)-C(=O)$ connecting bond of the amide group. The top ten bioisosteric groups are shown in Table 3, for both the Carbo and Hodgkin coefficients. As can be seen the ten groups comprise mainly amides, ketones and esters from different bonded environments. The high similarity of the aliphatic-aliphatic amide to the aliphatic-aromatic amide ($R(A, B) = 0.84$, $H(A, B) = 0.79$, aligned

using the $C(sp^3)-C(=O)$ connecting bond from the aliphatic-aromatic amide group), is fairly obvious given that the groups are isostructural aside from the difference in their bonded environment. Figure 4 shows the alignment of the target molecule with the aliphatic-aliphatic ester group ($R(A, B) = 0.70$, $H(A, B) = 0.59$), aligned using the $C(sp^3)-C(=O)$ connecting bond from the aliphatic-aliphatic ester group). Each alignment in Figure 4 shows the propensity maps for one of the probes used to calculate the diverse set similarity.

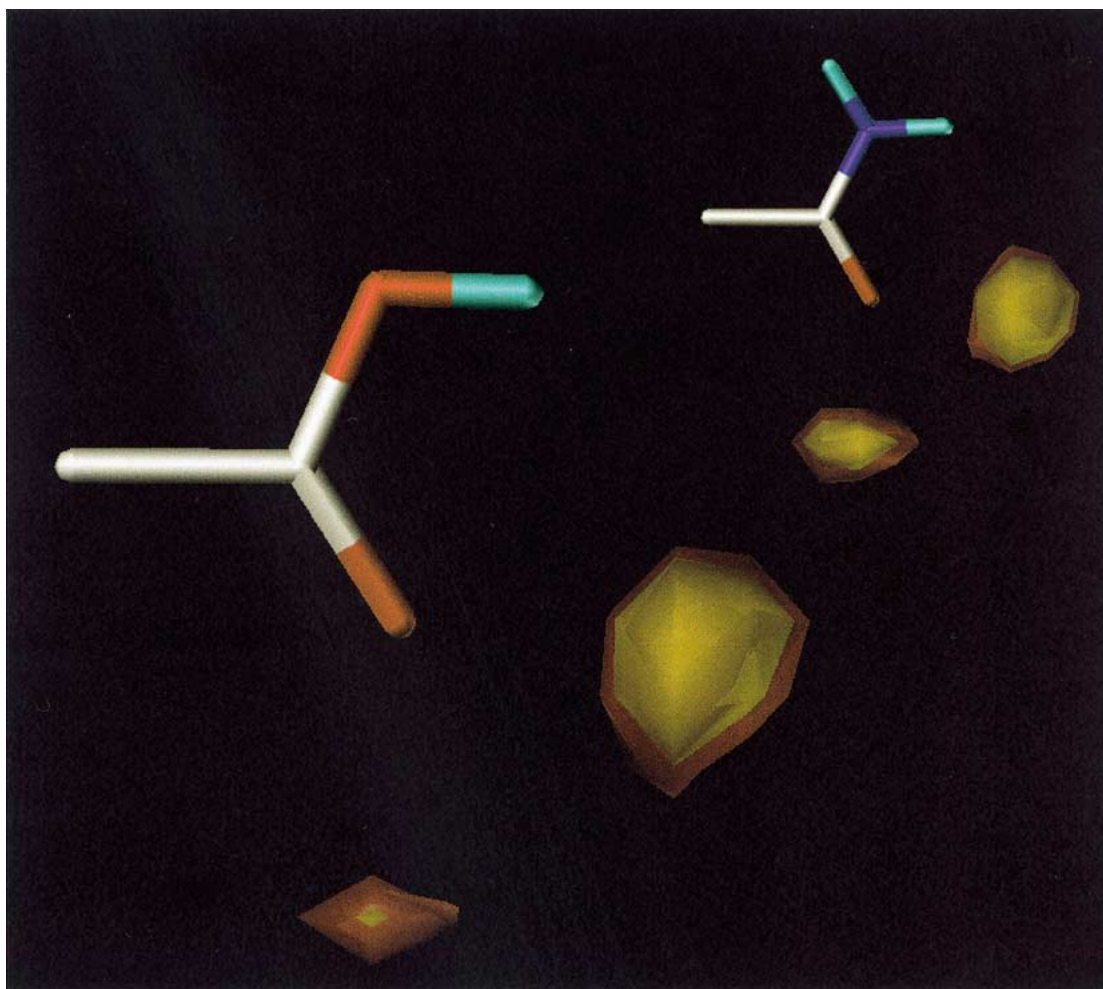


Figure 1. Propensity maps for the carbamoyl central group (rear) and the *cis*-uncharged carboxylic acid group (front). Calculated using the any N–H hydrogen probe atom. The propensity maps are shown at levels of 4.0 (red) and 6.0 (yellow).

Validation

The examples discussed thus far suggest that our similarity procedure is able to identify resemblances between functional groups that may not be obvious simply by considering the groups' topology. We now further validate the methodology with known bioisosteric pairs extracted from Bioster and Relibase. The way in which the bioisosteric pairs were extracted is described below.

Bioster bioisosteric functional group pair extraction

Each pair of bioisosteric molecules in the Bioster database was visually inspected in order to identify bioisosteric functional group pairs. If two bioisosteric molecules in the database are identical save one func-

tional group replacement, those two functional groups were regarded bioisosteric. Only those pairs for which both functional groups exist in IsoStar were extracted. Duplicate pairs were removed. Using this approach 185 bioisosteric functional group pairs were found (available as supplementary material).

The similarities of each of the bioisosteric functional group pairs were calculated using the methods described above. The single probe similarities were calculated for each of the pairs, provided that the scatterplots contain enough data, for the contact group probe atoms shown in Tables 4–6. The probes chosen represent a non-redundant unique set determined from a previous study [23]. The diverse set similarities were calculated using the probes previously defined.

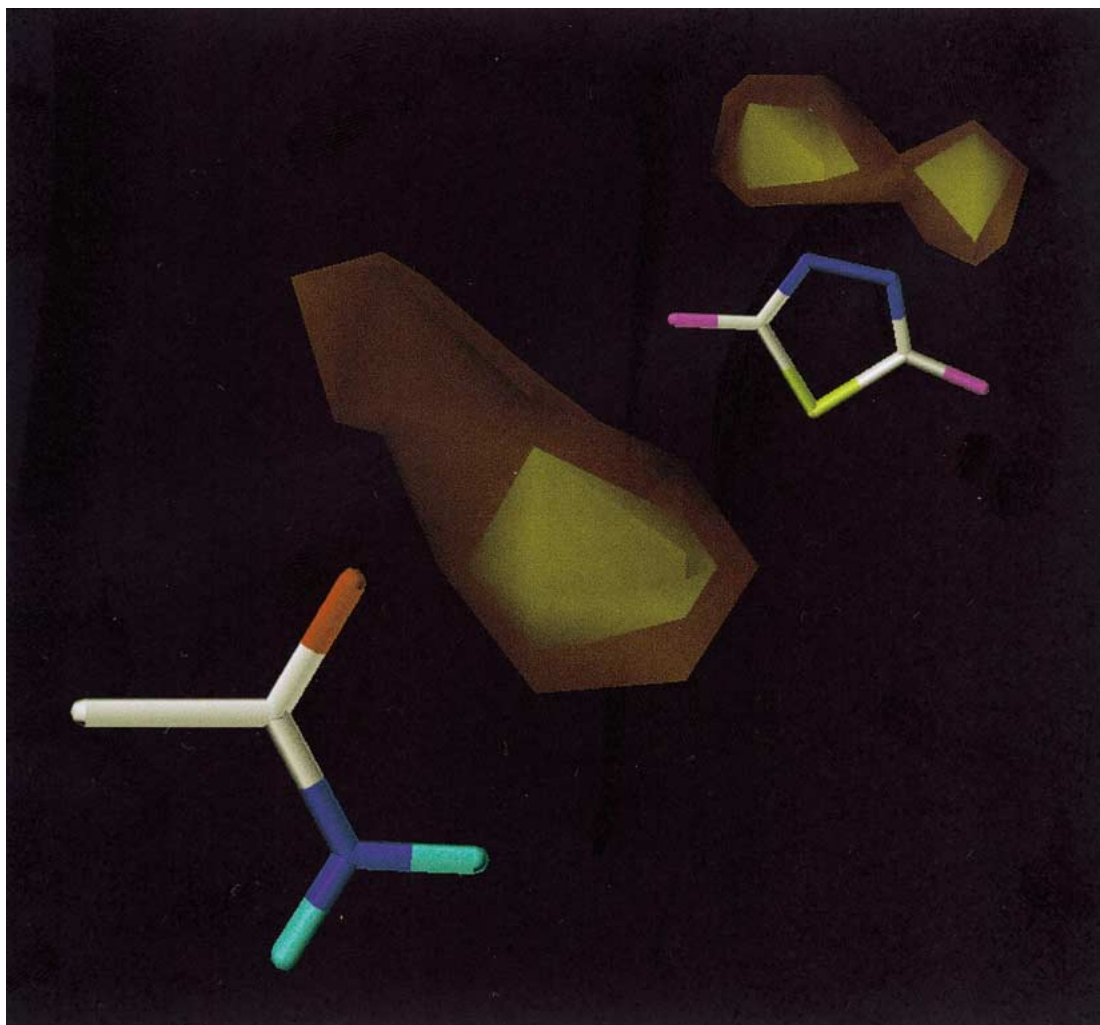


Figure 2. Propensity maps for the carbamoyl central group (front) and the 1,3,4-thiadiazole group (rear). Calculated using the any N–H hydrogen probe atom. The propensity maps are shown at levels of 2.0 (red) and 3.0 (yellow).

The number of bioisosteric functional group pairs obtained from Bioster for which there was enough data to conduct the similarity calculations, n_{bp} , are shown in Table 4 for each type of similarity calculation and, where appropriate, for each probe atom.

Relibase bioisosteric functional group pair extraction

For each functional group in IsoStar, ligands in the PDB were identified that contain the functional group of interest using the 2D/3D-search facility provided in Relibase. The interaction of the ligand with the binding site was visually inspected to check that the functional group concerned is interacting with binding site residues. If this is so, similar binding sites and the associated ligands were retrieved in order to find

different ligands bound to the same (or highly similar) binding site. Similar binding sites were retrieved with a sequence identity greater than 95% to that of the target binding site. Each of the retrieved binding sites and associated ligands were then superposed in turn onto the target ligand/binding sites. Functional group pairs were extracted from the pairs of ligands provided that the functional groups interact with the same part of the binding site.

It is important to note that the extraction of bioisosteric pairs from Relibase is not a comprehensive process as it would be impossible to screen every ligand contained within the database using the tools provided by Relibase. However the process of bioisos-

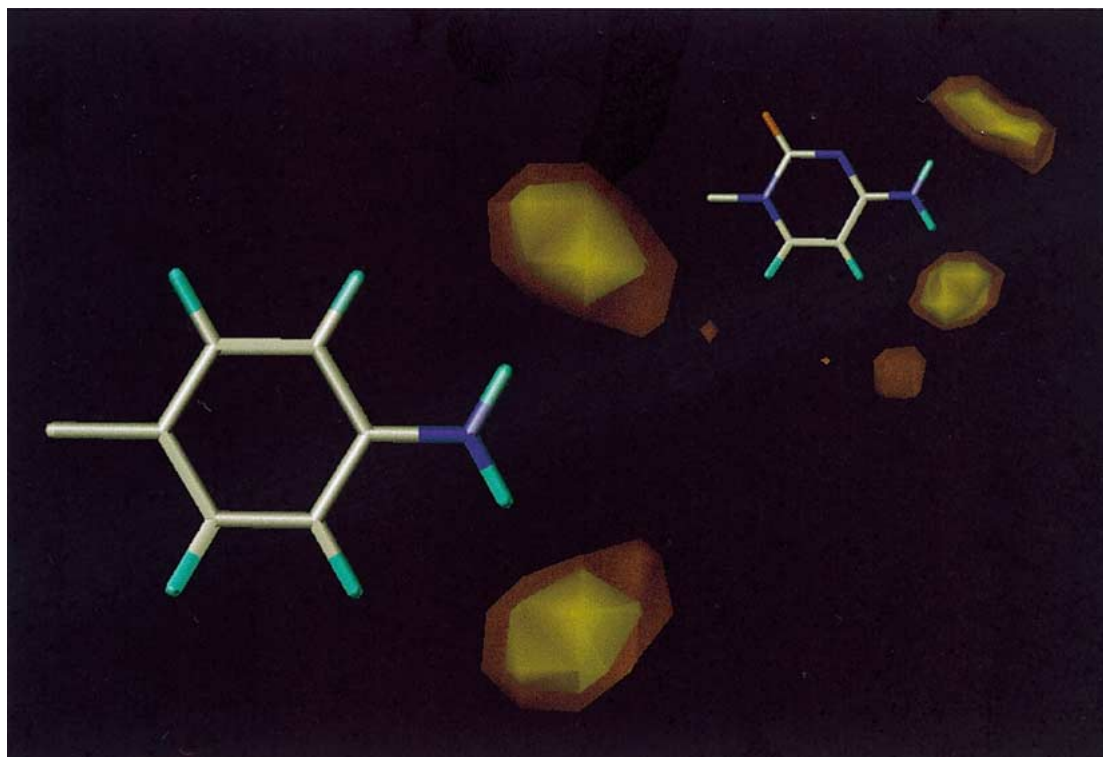


Figure 3. Propensity maps for the planar 4-aminophenyl group (front) and uncharged cytosine group (rear). Calculated using the any C=O oxygen probe atom. The propensity maps are shown at levels of 4.0 (red) and 6.0 (yellow).

teric pair retrieval was designed to make it as objective as possible.

An example of the ligand and binding site superposition is shown in Figure 5. In this case the functional group searched for using the Relibase 2D/3D-substructure search was the methoxy group. The target protein-ligand complex used in the similar binding site search has the PDB code 1bqm [30]. The binding site and ligand from the PDB entry 1rt6 [31] is shown superposed on the target complex. As can be seen the methoxy group from the 1bqm ligand is interacting with the same protein residues as the methyl group from the ligand in the 1rt6 complex, therefore the two functional groups are considered 'bioisosteres'.

Using this method it was possible to identify 137 unique bioisosteric functional group pairs (available as supplementary material). As before, the similarities of each of the bioisosteric functional group pairs were calculated using the single probe and diverse set methods. The number of bioisosteric functional group pairs extracted from Relibase for which there was enough data in IsoStar to conduct the similarity calculations, n_{bp} , are shown in Table 4 for each type of similar-

ity calculation and, where appropriate, for each probe atom.

'All pairs' functional group extraction

To observe whether the similarities obtained using these methods are of use in predicting bioisosteres, similarities of the bioisosteric pairs are compared with those calculated for all the pairs of functional groups within IsoStar provided there is enough data for the given probe. The number of functional group pairs for which similarity calculations could be made, n_{ap} , are shown in Table 4, for each method and, where appropriate, for each probe. The distributions of the bioisosteric functional group pair similarities are then compared with the corresponding 'all pairs' sets.

Single probe data

Figures 6 and 7 show two histograms in which the similarities have been calculated using the any aromatic C-H hydrogen probe atom for both the Carbo and Hodgkin coefficients. Plotting these histograms clearly shows a marked difference in the distributions

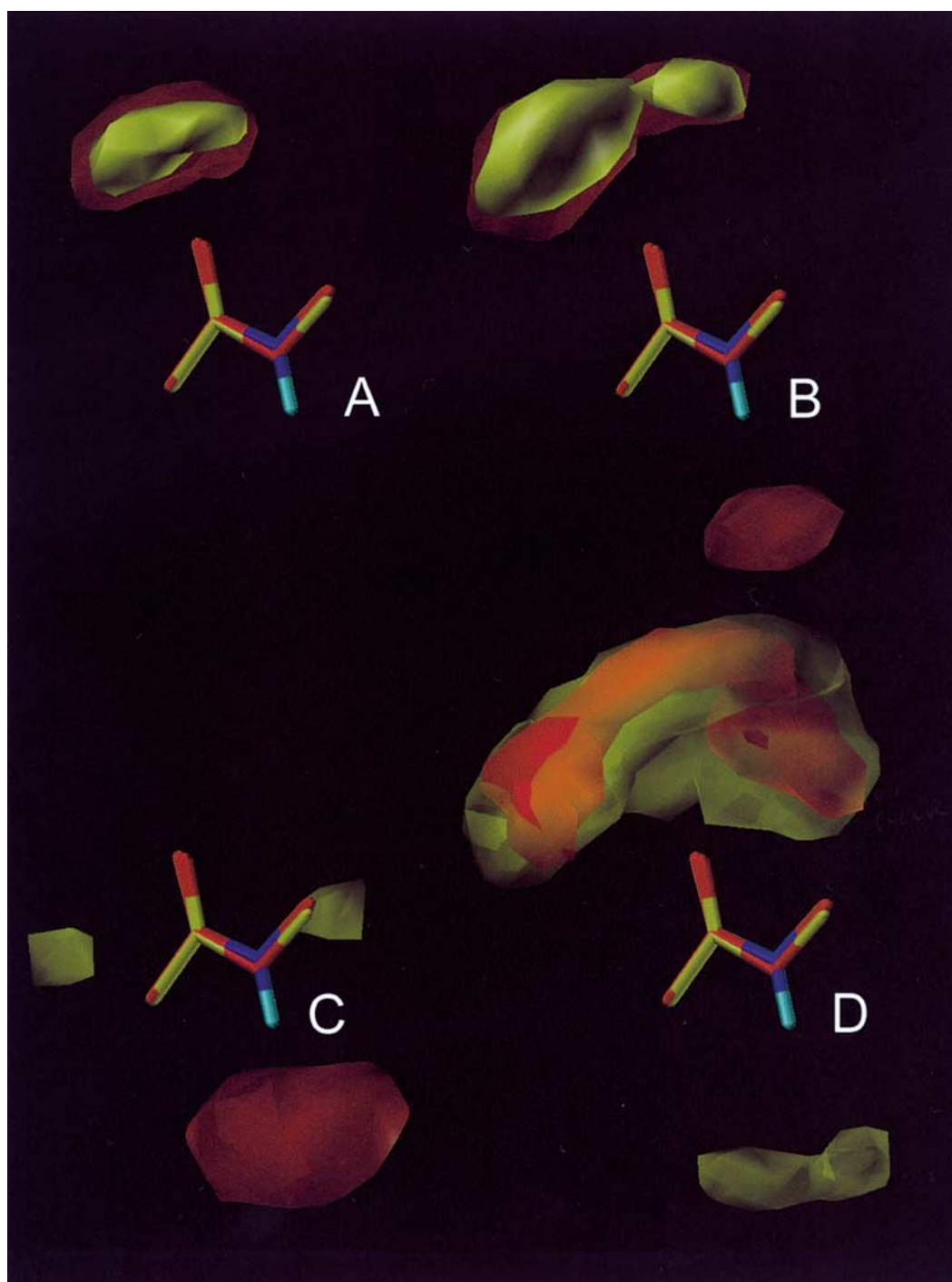


Figure 4. Superposition of the aliphatic-aliphatic amide group with the aliphatic-aliphatic ester group. The groups are superposed using the $C(sp^3)-C(=O)$ and $C(sp^3)-C(=O)$ bonds, respectively. Calculated using the diverse set method. Propensity maps are shown in each picture for the four probes used in red (the aliphatic-aliphatic amide) and yellow (aliphatic-aliphatic ester). (A) shows the propensity maps for the any N-H hydrogen probe, at a level of 3.0, (B) shows propensity maps for the any O-H oxygen probe, at a level of 3.0, (C) shows propensity maps for the any C=O oxygen probe, at a level of 3.0 and 1.0 for the aliphatic-aliphatic amide and aliphatic-aliphatic ester, respectively, and (D) shows propensity maps for the *methyl* carbon probe, at a level of 1.0. The carbon atoms of the are also coloured accordingly (i.e., red for the aliphatic-aliphatic amide group and yellow for the aliphatic-aliphatic ester group).

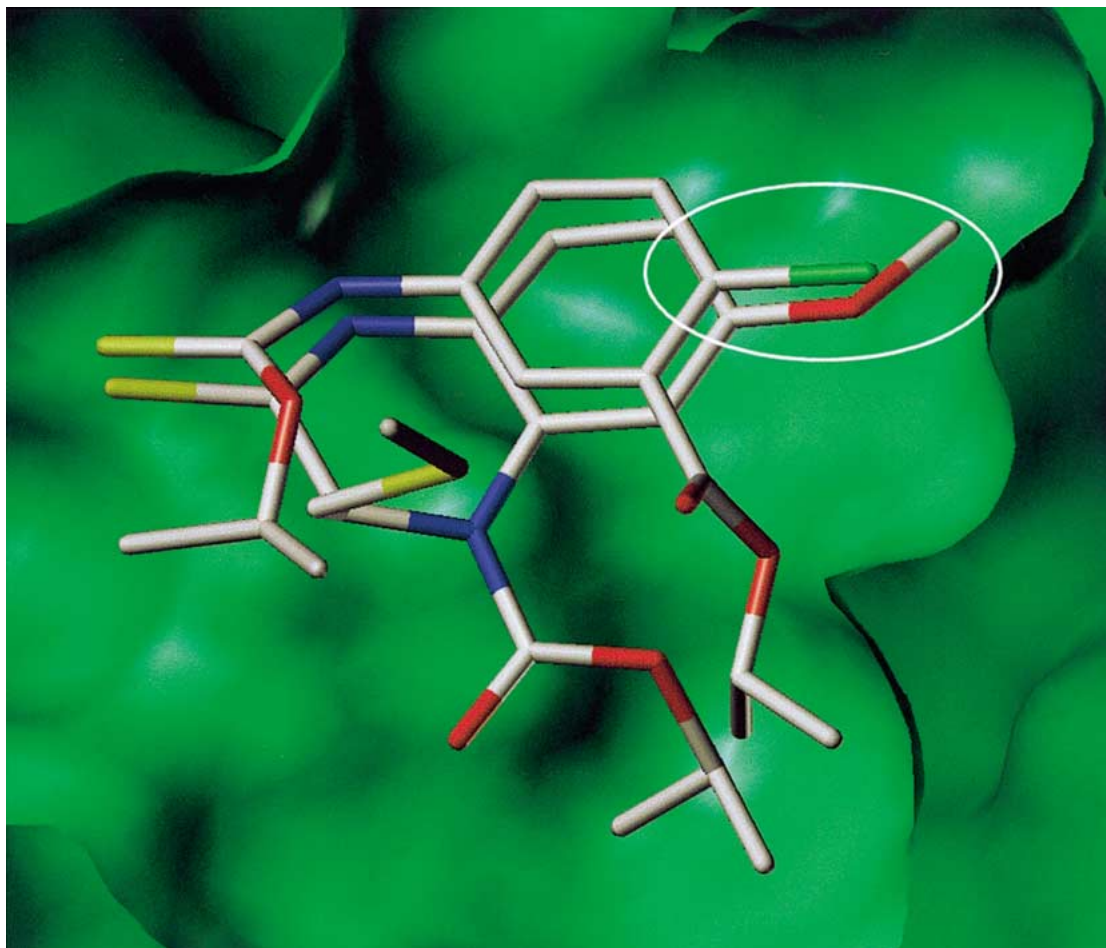


Figure 5. Binding site superposition of the 1bqm complex with the 1rt6 complex. The chosen bioisosteric functional groups are circled. The Connolly surface of the active site of the 1bqm complex is shown in green.

of the similarities for the bioisosteric pairs and the all pairs set. All of the histograms for the other single probe experiments and the diverse set method have distributions very similar to the two examples shown in Figures 6 and 7. The bioisosteric pair similarities for both the Bioster and Relibase sets are generally distributed at higher similarity values than the all pairs sets, showing that the calculated similarities are successful at predicting likely bioisosteres. However it is noted that the distribution of the similarities of the bioisosteric pairs is quite broad in some cases. In the case of the histograms derived from the Hodgkin similarities, the peaks of the histograms lie at lower similarity values compared to their counterparts in the Carbo histograms.

A χ^2 test was used to determine whether the distributions of the similarities for the bioisosteric pairs are significantly different to those of the all pairs sets.

The χ^2 values for the Bioster pairs are given in Table 5, the χ^2 values for the Relibase pairs are given in Table 6, for each of the probe atoms for both sets of bioisosteric groups and for both the Carbo and Hodgkin coefficients.

Generally the χ^2 values are quite large indicating that the distributions compared are significantly different. In all but five examples for the Bioster pairs and eight examples for the Relibase pairs, the level of confidence with which we can assume the distributions of the all pairs sets are different from distributions of the bioisosteric sets is at the 0.005 level of statistical significance (assuming that we have 10 degrees of freedom, $\chi^2 \geq 25.12$). Examples where the χ^2 values are lower than this threshold value are due to the lack of data for the particular probe atom and hence the similarity will have only been calculated for a small number of bioisosteric pairs. This is demon-

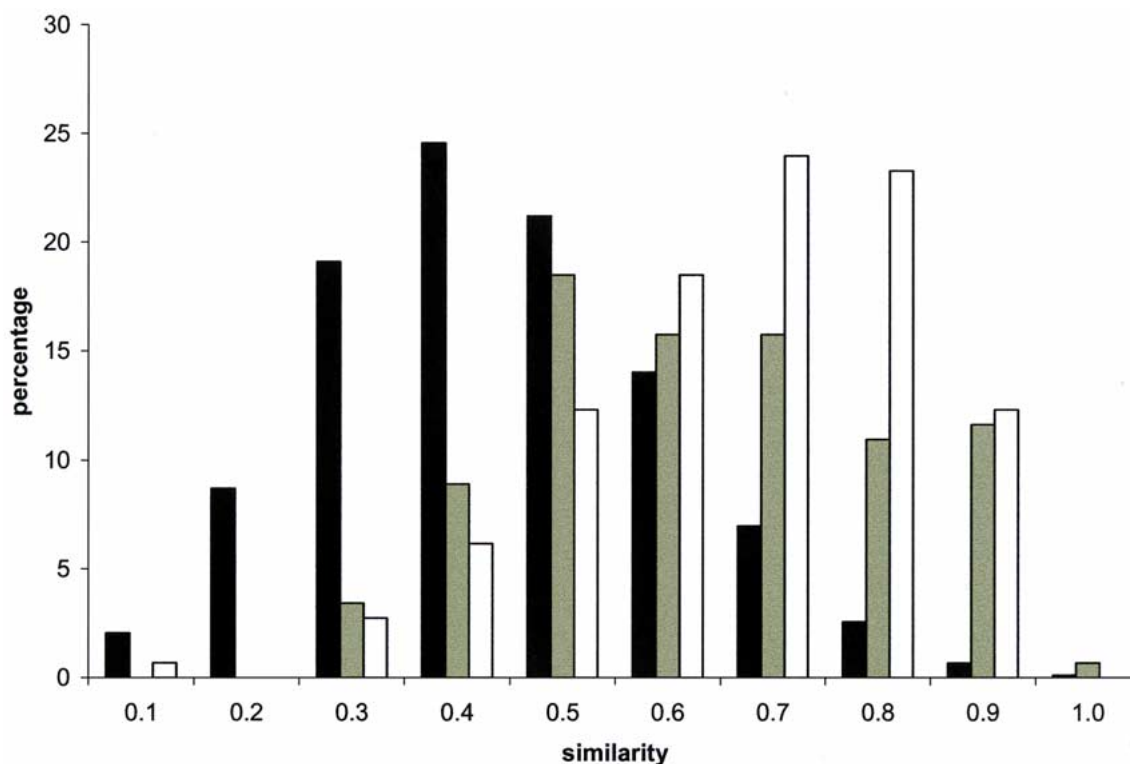


Figure 6. Histogram showing the relative distribution of the similarities for the bioisosteric functional group pairs from Biooster and Relibase, and the all pairs set. Calculated using the single probe method in terms of the aromatic C–H hydrogen probe atom using the Carbo coefficient. The all pairs similarities are shown in black, the bioisosteric pairs from Biooster are shown in white and the pairs from Relibase are shown in grey.

strated by comparing the entries in Table 4 with the corresponding χ^2 values in Tables 5 and 6, where the non-significant entries are shown in italics.

Enrichment factors for the various similarity calculations can also be calculated. The enrichment factors will show whether using the IsoStar similarity method to predict the bioisosteres is better than just picking functional group replacements at random.

The following method was used to calculate the enrichment factor for each probe atom for both the Carbo and Hodgkin coefficients. The number of bioisosteric pairs, n_{exp} , expected in the top 10% of a ranked list (ranked according to the similarity of each pair) of the all pairs set, provided that the bioisosteric pairs are distributed randomly is given by Equation 15.

$$n_{exp} = 0.1 \times n_{bp} \quad (15)$$

Using the number of bioisosteric pairs that are actually contained within the top 10% of the ranked all pairs list, n_{obs} , it is possible to calculate the enrichment

factor, E , via Equation 16.

$$E = \frac{n_{obs}}{n_{exp}} \quad (16)$$

. Tables 5 and 6 show the enrichment factors calculated for both the Carbo and Hodgkin equations for all the probe atoms considered for the Biooster and Relibase data sets respectively. The estimated standard deviations have been calculated assuming a Poisson distribution.

If the enrichment factors for the Biooster pairs are analysed it is seen that 19 of the 26 probe atoms have enrichment factors that are significantly larger than 1.0 for both the Carbo and Hodgkin coefficients, taking into account the estimated standard deviations (i.e., $\pm 2\sigma$). Therefore using this method of similarity calculation to identify bioisosteres is better than choosing them at random for the Biooster dataset for these probe atoms. Of the remaining seven, six have enrichment factors greater than 1.0, but all have large estimated standard deviations. The probe atoms that fall into this category are: *cationic RNH₃* nitrogen/hydrogen, *car-*

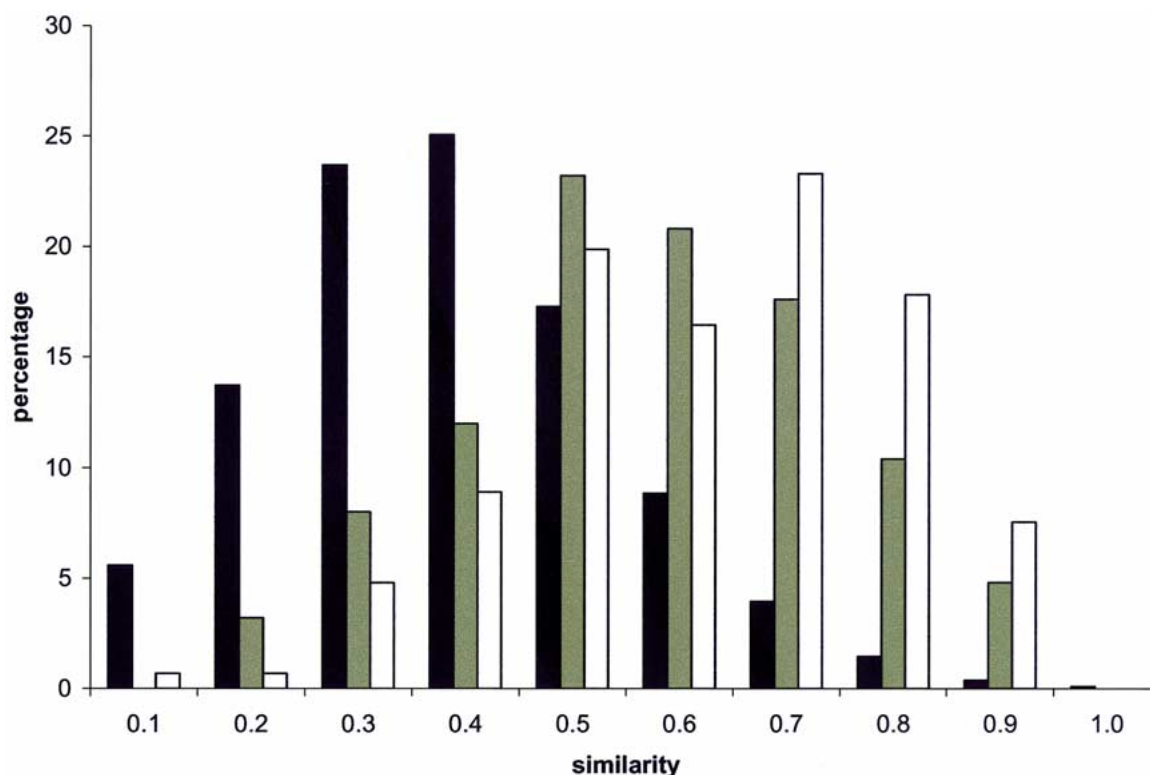


Figure 7. Histogram showing the relative distribution of the similarities for the bioisosteric functional group pairs from Biooster and Relibase, and the all pairs set. Calculated using the single probe method in terms of the aromatic C–H hydrogen probe atom using the Hodgkin coefficient. The all pairs similarities are shown in black, the bioisosteric pairs from Biooster are shown in white and the pairs from Relibase are shown in grey.

boxylate carbon/oxygen, *cyano* nitrogen and *chloride ion* chlorine probes. The remaining probe atom is the *aliphatic* ether oxygen, which is the only probe atom to have an enrichment factor lower than 1.0. The estimated standard deviation for this probe is relatively low, and we hence conclude that this probe is not very useful for identifying likely bioisosteres.

If the enrichment factors for the Relibase pairs are observed it is seen that 17 of the 26 probe atoms have enrichment factors that are larger than 1.0 for both the Carbo and Hodgkin coefficients, taking into account the estimated standard deviations. Using this method of similarity calculation to identify likely bioisosteric functional group pairs is better than choosing them at random for the Relibase dataset for these probe atoms. Of the remaining nine, eight have enrichment factors greater than 1.0, but all have large estimated standard deviations that do not preclude the possibility that the enrichment factors may be artificially high, and as such not truly greater than 1.0. The probe atoms that fall into this category are: any cationic N–H ni-

trogen/hydrogen, cationic RNH₃ nitrogen/hydrogen, carboxylate carbon/oxygen, nitro nitrogen, and chloride ion chlorine probes. Again the weakest probe atom is the *aliphatic ether* oxygen atom, which is the only probe atom to have an enrichment factor lower than 1.0.

There is not much difference between the enrichment factors obtained using the Carbo and Hodgkin coefficients. In the case of the Biooster pairs, for 15 of the 19 probe atoms that have enrichment factors greater than 1.0, the Carbo coefficient gives marginally higher enrichment factors than the Hodgkin coefficient. In the case of the Relibase pairs, eight of the 17 probe atoms with enrichment factors greater than 1.0 give higher enrichment factors for the Hodgkin coefficient.

The enrichment factors of the Biooster dataset are generally larger than those obtained from the Relibase dataset. For the Carbo coefficient it is seen that out of the possible 16 probe atoms that can be compared, 13 of the enrichment factors from the Biooster set are

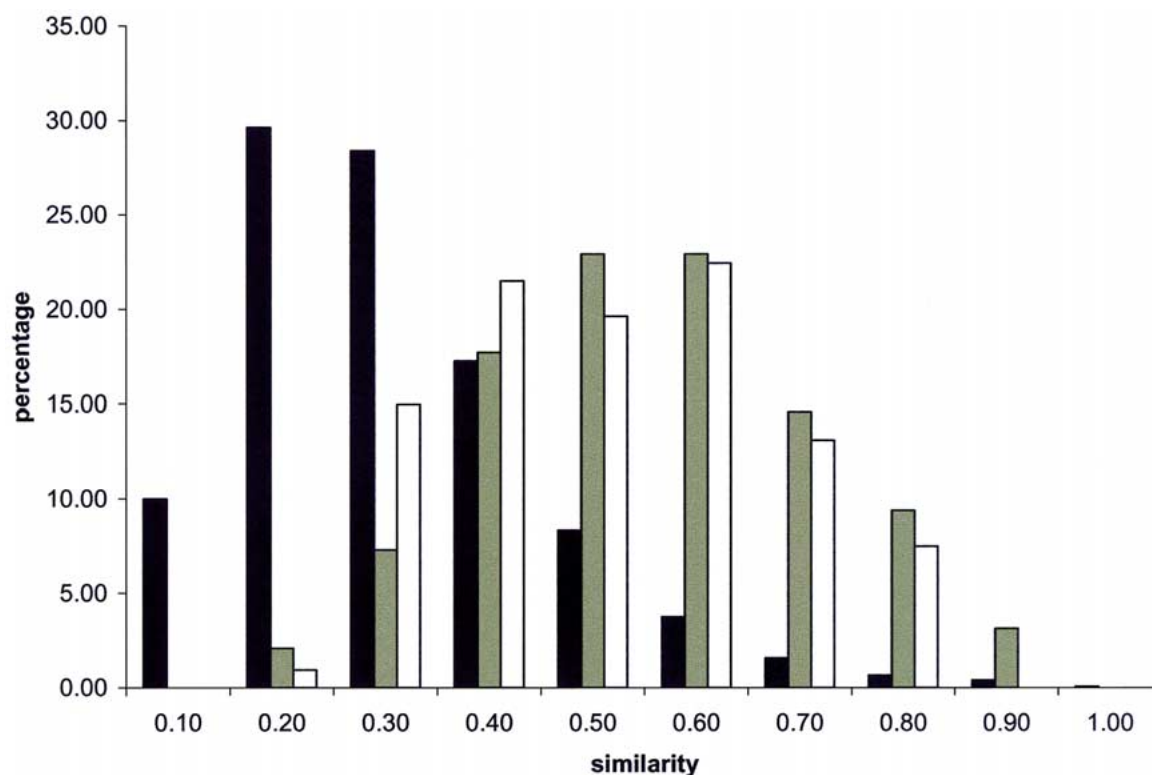


Figure 8. Histogram showing the relative distribution of the similarities for the bioisosteric functional group pairs from Bioster and Relibase, and the all pairs set. Calculated using the diverse set method using the Carbo coefficient. The all pairs similarities are shown in black, the bioisosteric pairs from Bioster are shown in white and the pairs from Relibase are shown in grey.

larger than those from the Relibase set. In terms of the Hodgkin coefficient, 12 of the enrichment factors are larger for the Bioster set. A possible explanation for this lies in shortcomings of the methodology for the validation. Although great care was taken in the selection of the Relibase and Bioster pairs of bioisosteric groups, we have no information on the relative affinities of the groups in each pair, or the ligands from which they came. Therefore the difference in the relative binding affinities of the groups within each bioisosteric pair may be quite large. As such 'weak' functional group replacements could have been made, particularly in the case of the Relibase set, where the ligands from which the functional groups were extracted are often radically different. The low binding affinity of the replacement functional group could be compensated for by stronger interactions formed by other parts of the ligand, hence retaining the binding mode.

The largest enrichment factors for both the Bioster and Relibase pairs are obtained using probe atoms from the hydrophobic contact groups (e.g.

methyl carbon/hydrogen and any aromatic C–H carbon/hydrogen), especially in the case of the Carbo coefficient. This may be explained as follows. Hydrophobic probes, apart from taking into account the hydrophobic effect, generally do not form very directional interactions. Hence, they provide a good descriptor of the shape of a functional group. If the 'hydrophobic' propensity maps for two central groups are similar, this implies they fit in a similar pocket. As stated above, the relative binding strengths of two groups that constitute a Bioster or Relibase pair may vary considerably, which implies that, in this validation experiment, shape similarity may be the most important factor that determines enrichment factors.

Diverse set data

The same analysis was carried out for the diverse set method. The similarities of the bioisosteric functional group pairs and the all pairs sets show very similar distributions to the histograms shown above for the any aromatic C–H hydrogen single probe experiment. Figure 8 shows the distributions of the bioisosteric

Table 4. Number of bioisosteric functional groups pairs, n_{bp} , from both Bioster and Relibase and the number of pairs of functional groups pairs in the IsoStar database, n_{ap} , used in the validation for which similarity calculations can be made.

Contact group	Probe atom	n_{bp} (Bioster)	n_{bp} (Relibase)	n_{ap}
Any N–H	N	150	106	25759
	H	143	111	22204
Any cationic N–H	N	44	53	2180
	H	35	44	1730
Cationic RNH ₃	N	16	14	244
	H	24	29	775
Any O–H	O	150	125	26058
	H	136	117	23258
Water	O	71	79	5188
	H	80	84	8082
Any C=O	C	158	121	30678
	O	154	124	31652
Carboxylate	C	12	35	1116
	O	26	51	2174
Aliphatic ether	O	26	31	770
Nitro	N	34	35	1865
	O	76	84	9210
Cyano	C	36	54	1890
	N	29	37	2945
Methyl	C	143	119	33862
	H	161	131	50039
Any aromatic C–H	C	164	131	66876
	H	146	125	65714
Any C–F	F	54	75	3745
Any C–Cl	Cl	68	75	5226
Chloride ion	Cl	36	44	1632
Diverse set		107	96	13113

functional group pairs from both Bioster and Relibase with the distribution of the all pairs set using the diverse set method in terms of the Carbo coefficient.

As before, χ^2 values can be calculated illustrating the difference in the distribution between the bioisosteric functional group pairs and the all pairs set. The χ^2 values are shown in Table 5 for the Bioster set and Table 6 for the Relibase set. As before they are high indicating that the distribution for the bioisosteric pairs and the all pairs sets are significantly different.

The enrichment factors for the diverse set method are also shown in Tables 5 and 6 for both the Bioster and Relibase pairs respectively. The enrichment factors are above 1.0, indicating using this method to predict bioisosteric pairs is better than choosing them

at random from IsoStar. The enrichment factors are relatively high in comparison with the enrichment factors obtained for the single probe method indicating that calculating the similarity of the functional groups in terms of a range of interaction types is perhaps more successful than using one interaction type given the limitations of the validation.

Discussion and conclusions

A method for calculating the similarity of functional groups in terms of their non-bonded interactions has been described and validated using Bioster and Relibase. The single probe and diverse set methods have both been shown to be able to predict the existence of bioisosteric pairs of functional groups better than would have been expected by chance. It is important to note that it is impossible to validate every pair of functional groups with a high similarity score (i.e., a bioisosteric pair) obtained using this method. If it were, then IsoStar would not be offering any novel pairs of bioisosteric functional groups. Rather the validation illustrates that enough of the pairs derived using this method are valid and as such the results for the other pairs can be considered valid too.

This methodology has several advantages over existing techniques to identify bioisosteres. Whilst a user of IsoStar could identify the similarities in the propensity maps for a pair of central groups visually, this method has the advantage of automating the procedure. It is also possible to identify bioisosteric functional groups pairs using the Bioster and Relibase databases as described above, however using IsoStar similarities it is possible to identify bioisosteric functional group pairs that are not contained within these databases. For example, using Bioster four bioisosteres were found for the carbamoyl group; uncharged carboxylic acid, thiazole, acetyl and aliphatic cyano. Many other groups are listed in the top ten lists shown in Table 1 that are not present in Bioster, such as formyl ($R(A, B) = 0.91$, $H(A, B) = 0.91$), 1,3,4-thiadiazole ($R(A, B) = 0.73$, $H(A, B) = 0.72$), methoxycarbonyl ($R(A, B) = 0.68$, $H(A, B) = 0.59$), epoxide ($R(A, B) = 0.67$, $H(A, B) = 0.56$) and prolyl ($R(A, B) = 0.65$, $H(A, B) = 0.64$). As such we feel that using IsoStar similarities it is possible to identify novel bioisosteres that are not present in the existing databases.

Whilst the IsoStar similarity method appears successful there are some limitations which should be

Table 5. χ^2 values and enrichment factors for both single probe and diverse set methods for the validation using bioisosteric pairs extracted from BioStar

Contact group	Probe atom	χ_c^2	χ_h^2	E_c	esd	E_h	esd
Any N–H	N	214.84	183.32	3.33	0.54	3.47	0.56
	H	151.95	169.23	3.64	0.59	3.85	0.39
Any cationic N–H	N	37.64	31.44	3.41	1.02	2.73	0.89
	H	45.54	83.07	3.71	1.21	3.71	1.21
Cationic RNH ₃	N	25.29	21.37	1.88	1.18	4.38	1.98
	H	24.61	18.94	3.33	1.36	3.33	1.36
Any O–H	O	190.62	187.02	4.00	0.37	3.73	0.34
	H	122.47	142.91	3.16	0.56	2.57	0.49
Water	O	121.11	114.98	3.80	0.86	3.38	0.80
	H	103.77	96.04	3.75	0.80	3.63	0.79
Any C=O	C	262.53	293.43	4.49	0.41	4.37	0.40
	O	197.53	207.76	3.83	0.59	3.77	0.58
Carboxylate	C	9.26	9.79	1.67	1.27	2.50	1.61
	O	21.92	57.27	1.92	0.94	4.62	1.61
Aliphatic ether	O	11.98	19.94	0.77	0.56	1.54	0.83
Nitro	N	54.40	64.20	3.82	1.25	4.12	1.31
	O	84.84	83.05	3.42	0.78	3.29	0.76
Cyano	C	61.27	50.73	3.61	1.17	3.33	1.11
	N	32.81	25.80	2.41	1.02	3.10	1.18
Methyl	C	365.58	347.06	4.55	0.68	3.85	0.61
	H	422.09	345.64	5.03	0.69	4.41	0.63
Any aromatic C–H	C	357.18	299.20	5.18	0.69	3.72	0.56
	H	658.30	680.00	6.51	0.86	5.07	0.72
Any C–F	F	59.78	65.27	3.33	0.91	3.15	0.88
Any C–Cl	Cl	104.37	69.21	4.41	0.97	3.82	0.88
Chloride ion	Cl	28.59	33.32	3.06	1.05	2.78	0.99
Diverse set	na	227.50	309.25	5.14	0.85	4.86	0.82

discussed before possible applications are suggested. The single probe method, although appearing useful for the determination of bioisosteric functional groups pairs, should, on occasion be treated with some caution. The reason for this is that only one type of interaction is taken into account when using this method and as such will sometimes give misleading results. An example of this is found if the similarity of the charged amino group is calculated with all the other central groups in the IsoStar database using the any aromatic C–H hydrogen probe atom. The top ranked functional group in terms of the Carbo coefficient is the methyl group. Obviously this is not a sensible bioisosteric replacement for the charged amino group, and illustrates care must be taken when interpreting the single probe data. However, the enrichment factors and distributions clearly indicate the success of using

the single probe approach method. An explanation for this could be that, for a given protein-ligand complex, the ‘right’ bioisosteric substitution depends on which protein residues the chosen functional group is interacting with. If the chosen group is only accepting hydrogen bonds from the binding pocket then using the single probe approach (e.g., using the any N–H hydrogen probe) is probably more appropriate than the diverse set method.

The superposition process does not take into account the bonded environment of the functional groups. For example if the aromatic bromo and aromatic chloro groups are superposed, the ring atoms (included in the scatterplots) are not necessarily aligned in the superposition process. In this case it is possible to do this as the ring atoms are defined, however in most cases the atoms defining the bonded

Table 6. χ^2 values and enrichments factors for both single probe and diverse set methods for the validation using bioisosteric pairs extracted from Relibase

Contact group	Probe atom	χ_c^2	χ_h^2	E_c	esd	E_h	esd
Any N–H	N	175.05	147.71	3.68	0.69	3.49	0.67
	H	165.13	174.16	3.15	0.61	3.78	0.47
Any cationic N–H	N	17.22	10.55	1.70	0.61	1.89	0.65
	H	19.27	33.65	1.82	0.70	2.27	0.80
Cationic RNH ₃	N	23.12	19.66	2.14	1.36	2.14	1.36
	H	21.09	19.50	1.72	0.83	3.45	1.26
Any O–H	O	163.49	200.69	3.84	0.43	4.40	0.51
	H	127.18	142.59	4.79	0.78	3.59	0.65
Water	O	34.51	39.67	2.44	0.62	2.69	0.66
	H	59.24	57.26	2.38	0.59	2.50	0.61
Any C=O	C	70.63	89.20	2.81	0.30	3.06	0.35
	O	69.22	91.22	3.23	0.59	2.66	0.52
Carboxylate	C	6.22	15.05	1.14	0.60	1.71	0.67
	O	35.10	50.51	2.35	0.75	3.33	0.93
Aliphatic ether	O	4.83	12.64	0.65	0.47	1.29	0.69
Nitro	N	20.23	26.54	2.00	0.83	1.71	0.76
	O	47.74	53.98	2.38	0.59	2.86	0.66
Cyano	C	53.60	82.31	2.59	0.78	2.78	0.81
	N	38.76	38.45	2.43	0.70	3.78	0.89
Methyl	C	261.74	159.54	4.45	0.74	3.28	0.71
	H	312.13	180.52	4.50	0.71	3.05	0.55
Any aromatic C–H	C	201.13	169.42	3.89	0.64	2.37	0.47
	H	426.67	255.04	5.12	0.79	3.84	0.65
Any C–F	F	93.36	125.01	3.87	0.85	4.00	0.86
Any C–Cl	Cl	72.16	75.73	3.60	0.81	3.33	0.70
Chloride ion	Cl	22.01	28.20	1.56	0.65	1.56	0.65
Diverse set	na	403.79	369.66	6.25	1.03	6.56	1.06

environment of the functional group are not present in the IsoStar definition of the central group. Therefore these atoms were not taken into account. As such, very occasionally, the alignments for a pair of functional groups are not the alignments that would occur if the functional groups were interchanged.

The superpositions of a given pair of functional groups may vary for different probes. Ideally alignments would have been generated using the diverse set method and used for the single probe calculations for each probe; however, as stated above, this would only be possible for 141 groups.

Perhaps another limitation of the methodology is that it is difficult to ascribe what level of similarity constitutes a bioisosteric relationship between one functional group and another. It would be difficult to apply a hard cut-off defining the level of similarity

required to define a bioisosteric pair. Rather, there is continuum between safe and obvious bioisosteres which are likely to work but are not likely to be novel, and more speculative groups which are perhaps less likely to work but are far more novel and which may explore new areas of chemistry.

Numerous applications of this form of similarity calculation are possible. The most obvious of these is in the lead optimisation stage of the rational drug design process in which functional groups in a putative ligand could be replaced by similar groups determined using this methodology. The similarity of the functional groups in IsoStar (for which we have calculated all the possible pairwise similarities) can also be used for *de novo* ligand design. If an active ligand is known it could be broken into IsoStar fragments. Similar functional groups can be retrieved for each of these

and linked in order to create new ligands. The similarity of entire ligands can be calculated using the propensity maps derived from IsoStar data. We are currently exploring the possibility of using functional group similarities in *de novo* ligand design and the subsequent similarity calculations of entire ligands using the propensity maps described.

Another approach for predicting bioisosteric functional groups using data derived from IsoStar would be to use complementarity instead of similarity. In this case we could identify a section of the active site surrounding a selected functional group within a ligand (not necessarily included in IsoStar). Propensity maps for all the functional groups in IsoStar could be generated in terms of the surrounding probe atoms from the active site moieties and the functional groups can be 'docked' into the selected area based on the overlap of the propensity maps with the active site moieties. The score of each docking would give another measure of the ability of one functional group to act as a bioisosteric replacement for another.

Acknowledgements

This work is funded by the CCDC. The author wishes to thank Synopsys Scientific Systems for the use of the Bioster database. The author also wishes to thank Jason Cole, Steve Maginn, Greg Shields and Robin Taylor for their contributions to this work.

References

1. Friedman, H.L., *Influence of Isosteric Replacements upon Biological Activity*, National Academy of Sciences-National Research Council Publication No. 206, Washington DC, 1951, pp. 295–395.
2. Thornber, C.W., *Quart. Rev. Chem. Soc.*, 8 (1979) 563.
3. Gilbert, A.S. and Champness, J.N., In Beddell, C.R. (Ed.), *The Design of Drugs to Macromolecular Targets*, John Wiley and Sons, Chichester, 1992, pp. 25–47.
4. Pitchford, N.A. and Taylor, R., In Martin, Y.C. and Willett, P. (Eds.), *Designing Bioactive Molecules*, American Chemical Society, Washington DC, 1998, pp. 19–46.
5. Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M. and Watson, G.G., *J. Chem. Inf. Comput. Sci.*, 31 (1991) 187.
6. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., *J. Mol. Biol.*, 112 (1977) 535.
7. Taylor, R., Kennard, O. and Vershiel, W., *J. Am. Chem. Soc.*, 105 (1983) 5761.
8. Hunter, C.A., Singh, J. and Thornton, J.M., *J. Mol. Biol.*, 218 (1991) 837.
9. Klebe, G., *J. Mol. Biol.*, 237 (1994) 212.
10. Mitchell, J.B.O., Nandi, C.L., McDonald, I.K., Thornton, J.M. and Price, S.L., *J. Mol. Biol.*, 239 (1994) 315.
11. Flanagan, K., Walshaw, J., Price, S.L. and Goodfellow, J.M., *Protein Eng.*, 8 (1995) 109.
12. Lommerse, J.P.M., Stone, A.J., Taylor, R., and Allen, F.H., *J. Am. Chem. Soc.*, 118 (1996) 3108.
13. Bruno, I.J., Cole, J.C., Lommerse, J.P.M., Rowland, R.S., Taylor, R. and Verdonk, M.L., *J. Comput. Aided Mol. Des.*, 11 (1997), 525.
14. Dean, P.M. and Perkins, T.D., In Martin, Y. C. and Willett, P. (Eds.), *Designing Bioactive Molecules*, American Chemical Society, Washington DC, 1998, pp. 199–218.
15. Carbo, R., Leyda, L. and Arnau, M., *Int. J. Quant. Chem.*, 17 (1980) 1185.
16. Hodgkin, E.E. and Richards, W.G., *Int. J. Quant. Chem.*, 14 (1987) 105.
17. Reynolds, C.A., Burt, C. and Richards, W.G., *Quant. Struct. Act. Relat.*, 11 (1992) 34.
18. Klebe, G., In Kubinyi, H. (Ed.), *3D QSAR in Drug Design*, ESCOM, Leiden, 1993, pp. 173–225.
19. Martin, Y.C., Bures, M.G., Danaher, E.A., Delazzer, J., Lico, I. and Pavlik, P.A., *J. Comput. Aid Mol. Des.*, 7 (1993) 83.
20. Pickett, S.D., Mason, J.S. and McLay, I.M., *J. Chem. Inf. Comput. Sci.*, 36 (1996) 1214.
21. Van Drie, J.H., *J. Comput. Aid Mol. Des.* 11 (1997) 39.
22. Lemmen, C. and Lengauer, T., *J. Comput. Aid Mol. Des.* 141 (2000) 215.
23. Verdonk, M.L., Cole, J.C., Watson, P., Gillet, V.J., Willett, P., *J. Mol. Biol.*, accepted.
24. Good, A.C., Hodgkin, E.E. and Richards, W.G., *J. Chem. Inf. Comput. Sci.*, 32 (1992) 188.
25. Bioster (version 98.1), Synopsys, 5 North Hill Road, Leeds, LS6 2EN, UK.
26. Relibase+ (alpha version), CCDC, 12 Union Road, Cambridge, CB2 1EZ, UK.
27. Verdonk, M.L., Cole, J.C. and Taylor, R., *J. Mol. Biol.*, 289 (1999) 1093.
28. Boer, D.R., Kroon, J., Cole, J.C., Smith, B., Verdonk, M.L., unpublished work.
29. Nelder, J.A. and Mead, R., *Comput. J.*, 7 (1965) 308.
30. Hsiou, Y., Das, K., Ding, J., Clark, J.D., Kleim, J.P., Rosner, M., Winkler, I., Riess, G., Hughes, S.H. and Arnold, E., *J. Mol. Biol.*, 284 (1998) 313.
31. Ren, J., Esnouf, R.M., Hopkins, A.L., Warren, J., Balzarini, J., Stuart, D.I. and Stammers, D.K., *Biochem.*, 37 (1998) 14394.