BIOINFORMATICS

Vol. 24 ECCB 2008, pages i105–i111
doi:10.1093/bioinformatics/btn263

Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites

Rafael Najmanovich[1,*], Natalja Kurbatova[2] and Janet Thornton[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK and [2]Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, LV-1459, Latvia

ABSTRACT

Motivation: Current computational methods for the prediction of function from structure are restricted to the detection of similarities and subsequent transfer of functional annotation. In a significant minority of cases, global sequence or structural (fold) similarities do not provide clues about protein function. In these cases, one alternative is to detect local binding site similarities. These may still reflect more distant evolutionary relationships as well as unique physico-chemical constraints necessary for binding similar ligands, thus helping pinpoint the function. In the present work, we ask the following question: is it possible to discriminate within a dataset of non-homologous proteins those that bind similar ligands based on their binding site similarities?

Methods: We implement a graph-matching-based method for the detection of 3D atomic similarities introducing some simplifications that allow us to extend its applicability to the analysis of large all-atom binding site models. This method, called IsoCleft, does not require atoms to be connected either in sequence or space. We apply the method to a cognate-ligand bound dataset of non-homologous proteins. We define a family of binding site models with decreasing knowledge about the identity of the ligand-interacting atoms to uncouple the questions of predicting the location of the binding site and detecting binding site similarities. Furthermore, we calculate the individual contributions of binding site size, chemical composition and geometry to prediction performance.

Results: We find that it is possible to discriminate between different ligand-binding sites. In other words, there is a certain uniqueness in the set of atoms that are in contact to specific ligand scaffolds. This uniqueness is restricted to the atoms in close proximity of the ligand in which case, size and chemical composition alone are sufficient to discriminate binding sites. Discrimination ability decreases with decreasing knowledge about the identity of the ligand-interacting binding site atoms. The decrease is quite abrupt when considering size and chemical composition alone, but much slower when including geometry. We also observe that certain ligands are easier to discriminate. Interestingly, the subset of binding site atoms belonging to highly conserved residues is not sufficient to discriminate binding sites, implying that convergently evolved binding sites arrived at dissimilar solutions.

Availability: IsoCleft can be obtained from the authors.

Contact: rafael.najmanovich@ebi.ac.uk

1 INTRODUCTION

A central goal in the field of structural biology is to understand how protein structure determines and affects protein function. Predicting the function of a protein from its 3D structure is a major intellectual and practical challenge. Despite the details inherent in the structure, extracting knowledge about what a protein does biologically and how it does it is not straightforward.

In the absence of reliable tools that would permit *ab initio* prediction of function from structure, we are currently restricted to the transfer of functional annotation based on similarities. The detection of local 3D atomic similarities may provide useful clues when sequence similarity and overall structural similarity (fold) are insufficient. In particular, knowing what small molecules bind to a protein, may provide valuable functional information.

Independent of our ability to detect such binding site 3D atomic similarities, one question still remains: is it possible, on the basis of binding site 3D atomic similarities, to discriminate binding sites that bind the same ligand[1] from binding sites that bind different ligands? The answer to this question is not at all obvious and only in the case of an affirmative answer there can be a hope of being able to predict what ligands may bind to a given protein binding site and also to extract unique binding site characteristics that may help us understand what are the atomic requirements necessary for binding specific small molecules.

While in 83% of proteins, the ligand-binding site can be found within the largest cleft (Laskowski *et al.*, 1996), the accurate detection of binding site atoms within clefts is still an open question. In the present work, we are not interested in the question of predicting the binding site, we know in advance which atoms belong to the binding site, i.e. which are the ligand-interacting cleft atoms and therefore can avoid the question of locating the binding site atoms. However, we are interested in understanding how important is it to know the location of the binding site.

Mutations on positions in which specific residues are necessary from a structural or functional point of view are negatively selected. Thus, residue conservation is used to detect important residues. However, it is worth noting that conserved residues may be important for a variety of reasons: to maintain the structure, to control dynamic aspect of the structure (conferring or restricting flexibility), in the interaction with small molecules or other proteins, etc. In the context of the present work, we are interested in determining the extent which conservation reflect similarities between non-homologous proteins in binding similar ligands. In other words, to what extent conservation reflects the need to maintain specific atoms in specific positions in space relative to the ligand, presumably in order to satisfy physico-chemical constraints.

Methods for the detection of local structural similarities vary primarily in the type of representation (generally simplified) and search method, usually via the detection of sub-graph isomorphisms

*To whom correspondence should be addressed.

[1]Or equivalent parts in different ligands.

(Kresher and Stinson, 1998) or geometric hashing (Rigoutsos and Wolfson, 1997). However, due to the time-consuming nature of the detection of sub-graph isomorphisms, methods have mostly resorted to simplifications in the form of pseudo-atoms (Schmitt *et al*., 2002; Weskamp *et al*., 2004). Shulman-Peleg *et al*. (2004, 2005) combined the simplified representation, graph-matching-based method of Schmitt *et al*. with a geometric hashing pre-screening step. Methods that make use of full atomic representation, i.e. utilising the coordinates of all non-hydrogen atoms, are few and only applicable in limited cases. The method of Kobayashi *et al*. (1997) requires the superimposition of bound ligands. Brakoulias *et al*. (2004) use a geometric method to compare a large dataset of molecular environments of PO_4 groups that also require the pre-definition of the molecular environments from the position of the phosphate groups. A more thorough review of methods for the detection of local structural similarities can be found in Najmanovich *et al*. (2005) and references therein.

Here we present IsoCleft, a graph-matching-based method for the detection of pairwise local 3D atomic similarities. IsoCleft is suited to compare large sets of atoms using full atomic representation and does not require any bonding or sequence alignment information. We analyse its usefulness to discriminate different ligands' binding sites based on 3D atomic similarities. That is, rather than finding similarity alone, we are interested in being able to discriminate similar binding sites that are functionally related from others. We define a family of binding site models with decreasing knowledge about the identity of the ligand-interacting binding site atoms to uncouple the questions of detecting the binding site atoms and predicting binding site similarities. Furthermore, we calculate the individual contributions of binding site size (in terms of number of atoms), chemical composition and geometry.

2 METHODS

2.1 Dataset

Our dataset is a subset of the dataset of Kahraman *et al*. (2007) comprising a total of 72 examples of structures of non-homologous proteins each bound to a cognate ligand. In the interest of space, we present here only the PDB codes of the entries in each class. Further information can be found in the original publication. There are nine different ligands represented in the dataset: Adenosine monophosphate (AMP): *12as, 1amu, 1c0a, 1ct9, 1jp4, 1qb8, 1tb7, 8gpb*; Androsteneolone (AND): *1e3r, 1j99*; Adenosine triphosphate (ATP): *1a0i, 1a49, 1ayl, 1dv2, 1dy3, 1e8x, 1esq, 1kvk, 1rdq, 1tid, 3r1r*; 17β-Estradiol (EST): *1fds, 1lhu, 1qkt*; Flavin-adenine dinucleotide (FAD): *1cqx, 1e8g, 1hsk, 1iqr, 1jqi, 1jr8, 1k87, 1pox*; Flavin mononucleotide (FMN): *1dnl, 1f5v, 1ja1, 1mvl, 1p4c, 1p4m*; Glucose (GLC): *1bdg, 1cq1, 1k1w, 1nf5, 2gbp*; Heme (HEM): *1d0c, 1d7c, 1dk0, 1eqg, 1ew0, 1gwe, 1icq, 1naz, 1np4, 1po5, 1pp9, 1qhu, 1qla, 1qpa, 1sox, 2cpo*; Nicotinamide-adenine-dinucleotide (NAD): *1ib0, 1jq5, 1mew, 1mi3, 1o04, 1og3, 1qax, 1rlz, 1s7g, 1tox, 1zpt, 2a5f, 2npx*.

2.2 Definition of clefts

In the present work, we know in advance which cleft atoms define the binding site as all proteins in the dataset are present in the holo form. We use this information to define different models of the binding site atoms. These models simulate varying amounts of knowledge of the identity of the binding site atoms within the cleft.

Clefts atoms are determined using the Surfnet algorithm (Laskowski, 1995). A cleft is defined as a set of overlapping spheres. Each sphere is defined in the mid-point between any two protein atoms as long as its radius lies within the range of 1.5 and 4 Å and does not overlap the van der Waals radius of any atom in the protein. The upper and lower bounds for surfnet spheres are empirical and designed to prevent the formation of one single cleft across the whole protein via the union of different clefts by means of very small or very large volumes that are not biologically relevant. While the surfnet algorithm defines a cleft volume, here we are interested in the atoms in the cleft surface. We define a cleft as the atoms that generate the surfnet spheres as well as the corresponding residues' C_α atoms.

Clefts identified through the presence of the bound cognate ligand are defined using the Surfnet algorithm as described above. These are referred as original model (OM). The cleft produced by the strictly geometric Surfnet algorithm is a rough over predicted idealization of the binding site often containing much more than the atoms within interacting distance to the bound ligand. In the absence of any more specific information about the location of the binding site in a protein of unknown function, the OM is a suitable cleft model for function prediction.

Quite often, the function of a protein may not be known but given the wealth of sequence information amassed in current databases, it is easy to find related proteins and thus detect which residues within a cleft are highly conserved. As discussed in Section 1, residue conservation is often taken as a sign of importance but this importance may not necessarily be related to physico-chemical determinants of ligand binding. Despite the intrinsic uncertainty behind residue conservations, we are interested here to determine to what extent the use of such information may improve our ability to hint at what ligand may bind to a protein. In the present work, we use the phylogenetic residue conservation scores from the ConSurf-HSSP database (Armon *et al*., 2001; Glaser *et al*., 2005) to define the conserved cleft model (CM) as the subset of atoms from the original cleft model that belong to residues with the highest consurf-hssp conservation scores (score$\geq$8).

It is important to note that in principle the atoms in the conserved cleft model could interact with the bound ligand, but may not contain all atoms that interact with a ligand, as not all ligand-interacting atoms need to be highly conserved (Glaser *et al*., 2006). Likewise, conserved atoms within a cleft can be found at considerable distances from the ligand (Fig. 1).

Finally, we define the interaction models (IM) as means to analyse the relationship between binding site similarity and ligand binding without the added confounding effect of the uncertainty in defining the binding site. Atoms in the interaction models contain the subset of original cleft model atoms within d Å of the bound ligand ($5 \leq d \leq 15$). Thus, as d increases so does the uncertainty in our knowledge about the identity of the ligand interacting atoms. The database of IM models with $d = 5.0$ Å referred as IM5 for short, is of particular interest as it is the database to which all other cleft models are compared against. Figure 1 shows the various cleft models used in the present work for one particular protein.

It is entirely possible that even the IM5 subset of atoms still contains atoms that are not necessary to bind the ligand in question but this is exactly one of the questions that we are trying to answer with the current work. Namely, how unique the set of atoms in contact to a given type of ligand need to be?

2.3 Detection of 3D atomic similarities

Given the sets of atoms defining the clefts under comparison, the question that needs to be answered is: What is the largest subset of atoms in both clefts in direct correspondence with each other geometrically as well as chemically? This is a combinatorial optimization problem where, in principle, each possible set of atom correspondences might be a solution and the largest such set is the global solution. Graph theory offers a means to solve this problem via the detection of the maximal clique in an association graph.

Viewing each cleft as a graph, the task is to find the largest common sub-graph isomorphism. This is done through the construction of an association or correspondence graph. In the present work, an association graph is a graph with nodes representing pairs of atoms, one from each cleft that satisfy a condition of chemical similarity. Edges in the association graph are drawn based on a condition of geometrical similarity between the two pairs of atoms, one pair from each cleft composing the nodes of the association graph.

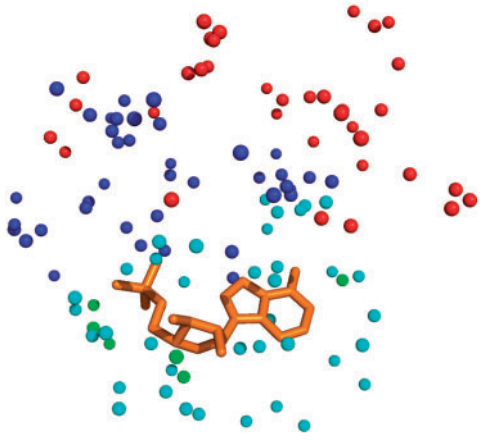

Fig. 1. Cleft models. Cleft atoms in *E. coli* Asparagine synthetase (PDB code 12as) shown as spheres. The ligand (AMP) is shown in orange. Cleft atoms within 5.0 Å of the ligand are shown in cyan (conserved atoms) and green (non-conserved). Remaining atoms are shown in blue (conserved) and red (non-conserved). Different cleft models are composed of combinations of these atoms. The IM5 model corresponds to cyan and green atoms. The CM model corresponds to cyan and blue atoms. The OM model corresponds to all atoms. Different IM models correspond to subsets of atoms at varying distances from the ligand. The cyan atoms are referred in the text as CIM5, the subset of conserved atoms within interacting distance from the ligand. The IM or CIM5 cleft models are not available when trying to predict the function of a protein but CM or OM can be used (see text for more details).

The condition of chemical similarity is implemented through the use of atom types. We utilize the atom type scheme of Sobolev *et al.* (1996). Atoms are classified into eight atom types: hydrophilic, hydrogen bond (HB) acceptor, HB donor, hydrophobic, aromatic, neutral, neutral-donor and neutral-acceptor. Each node in the association graph defines a possible correspondence between a pair of atoms of identical atom types, one from each cleft under comparison. This condition assures that the final subset of atoms in common between the two clefts corresponds pairwise to the same atom types.

The condition of geometrical similarity used when creating edges in the association graph is such that a clique corresponds to a subset of atoms in each cleft in which all pairwise distances between atoms in one cleft are satisfied by the corresponding atoms in the other cleft.

The net result of the graph matching is the detection of the largest subset of atoms of identical atom types in equivalent spatial position, thus making it possible to superimpose the two clefts based on these atoms.

The combinatorial nature of association graphs can lead to exponentially large graphs, both in terms of number of nodes as well as density of edges. This is a major drawback when employing association graphs to detect common sub-graph isomorphism since the computational cost of clique detection algorithms increases very rapidly with the size of the association graph. In the present work, we introduce two innovations that allows us to overcome this common problem associated with graph matching.

The first innovation is to perform the graph matching in two stages. In the first stage, an initial superimposition is performed via the detection of the largest clique in an association graph constructed using only C_α atoms of equivalent residues in the two clefts. A user-defined value is used to set the level of allowed residue similarity based on the rank order of each residues' JTT substitution matrix (Jones *et al.*, 1992) average probabilities (rank threshold, r). Once the largest C_α clique is obtained its transformation matrix and translation vectors are used to superimpose all atoms in the two clefts using the least square method of Arun *et al.* (1987). We utilize the Bron & Kerbosch algorithm (Bron and Kerbosch, 1973) to detect the largest clique in the association graphs on both stages of the graph matching process.

In the second graph matching stage, all non-hydrogen atoms are used. Association graph nodes are created with the requirement that two atoms, one from each cleft, be of the same atom type as well as that their spatial distance after the first stage superimposition be within a certain value. This distance threshold (cleft node radius threshold, n) is used to decrease the size of the association graph and is the reason why the initial graph matching stage is performed. The C_α atoms artificially included in the set of cleft atoms for the first stage are not utilized in the second stage and thus do not contribute directly to the detection or measurement of similarity.

The second innovation we introduce here is that we exploit the fact, noted by Bron and Kerbosch (1973), that the algorithm has the tendency to produce the larger cliques first in order to implement what we call Approximate Bron & Kerbosch. In the Approximate Bron & Kerbosch, the first clique is selected as the solution (and the search procedure stopped) rather than detecting all cliques in order to find the largest. Utilization of the approximate Bron & Kerbosch allows us to obtain an optimal or nearly optimal solution in a fraction of the time that would be needed with the original algorithm without any noticeable effect on the results (data not shown).

Four parameters are required for the execution of the algorithm. These are:

- *Rank threshold*, $r = 5$ — Used to define how much residue dissimilarity is permitted when building the C_α association graph. Values range from one (only identical residues allowed) to twenty (all substitutions allowed). This parameter represents the rank order of each residues' JTT substitution matrix (Jones *et al.*, 1992) average probabilities and makes it possible to take in account evolutionary information.
- *C_α distance difference threshold*, $c = 3.5$ Å — Used to define edges in first stage C_α association graph. A value of zero would mean that the resulting superimposed C_α atoms would have RMSD = 0 Å, this parameter places an upper bound to the resulting C_α RMSD.
- *Cleft node radius threshold*, $n = 4.0$ Å — Used to restrict the creation of nodes in the second stage association graph to those atoms of the same atom type and from different clefts that lie within this distance threshold. This parameter is used to prevent a combinatorial explosion in the number of nodes in the second stage all-atom association graph and is the main reason for the first stage graph matching.
- *Cleft distance difference threshold*, $d = 4.0$ Å — Used to define edges in the second stage, all atom association graph. This parameter is equivalent to c in nature but is applied to all atoms. This parameter places an upper bound on the final RMSD of the detected cleft atoms in common. This parameter can be used to implicitly account for the effect of flexibility to some extent.

In the present work, we utilize three measures of similarity to measure the individual contributions of binding site size, chemical composition and geometry towards prediction accuracy. All three measures are calculated as Tanimoto scores of similarity of the form:

$$\text{Similarity} = \frac{N_c}{N_A + N_B - N_c} \tag{1}$$

where N_A, N_B are the total number of atoms in clefts A,B and N_c corresponds to the number of atoms in common.

The size of the largest detected clique in the second stage association graph corresponds in effect to a measure of similarity that takes in account binding site chemical composition and geometry, as well as, implicitly, binding site size:

$$N_c = \text{Clique Size} \tag{2}$$

One can calculate the number of atoms in common between two clefts irrespective of their positions in space. For example, if the two clefts contain p and q hydrophobic atoms respectively, the maximum number of common atoms of this atom type will be $\min(p, q)$. Thus, the number

of common atoms considering size and chemical composition alone is given by:

$$N_c = \sum_{\text{atom types},i} \min(N_A^i, N_B^i) \quad (3)$$

where N_A^i, N_B^i represent the number of atoms of atom type i in each cleft.

Finally, one can ignore the atom types altogether and define a measure of similarity that reflects exclusively binding site size:

$$N_c = \min(N_A, N_B) \quad (4)$$

While predictions can be ranked according to their similarities as defined above, an independent mechanism needs to be defined to determine whether a prediction is successful or not. Here, we define a true positive (correct prediction), as a prediction that offers correct clues about the nature of the ligand that binds a given binding site.

Ligands present in the dataset in several instances share identical, sometimes large, common scaffolds such as the AMP moiety between {AMP, ATP, FAD, NAD}, the FMN moiety between {FMN, FAD}, as well as the EST moiety among {AND, EST}. If we were to find as common between an AMP and a FAD binding sites those atoms that are in contact to these ligands' equivalent atoms, this prediction should be considered successful.

We define a detected equivalent ligand atom (L_{eq}^d) as an atom in ligand A that is equivalent to an atom in ligand B[2] and each one of these atoms in turn is in close spatial proximity to cleft atoms found to be equivalent through the graph matching process. In other words, we count the number of equivalent ligand atoms detected specifically via common binding site atoms. When at least 70% of the equivalent ligand atoms are detected in this manner, the prediction is considered successful:

$$F_{eq} = \begin{cases} 1 & L_{eq}^d / L_{eq} > 0.7 \\ 0 & \text{Otherwise} \end{cases} \quad (5)$$

where L_{eq} is the number of equivalent atoms between the two ligand classes being compared. In other words, if $F_{eq} = 1$, at least 70% of the equivalent ligand atoms will be within close proximity when the two clefts are superimposed using the common cleft atoms. Figure 2 shows the result of a successful comparison.

It is worth noting that this measure of prediction success is quite stringent in the sense that it is not sufficient to find common binding site atoms, a large number of these atoms need to be functionally similar as they must interact with equivalent atoms in the ligands.

Finally, two clefts may have a large number of atoms in common and while all these are used when calculating N_c, only a fraction of these may be necessary to deem the prediction successful, as only a fraction of these will in fact be in the vicinity of the ligand.

When analysing the prediction success of measures that ignore geometry, we utilise the ligand classes as measure of success. That is, a true positive is one where the query and target proteins belong to the same ligand class as defined in Section 2.1.

The accuracy of a prediction was calculated as the average area under the receiver operator characteristic (ROC) curve, AUC. A ROC curve measures the fraction of true positive predictions as a function of the fraction of true negative predictions. Thus, an AUC value of 0.5 corresponds to the accuracy of a random predictor while, a value of 1.0 corresponds to a perfect prediction. An actual test will fall between these bounds. Particular AUC values are then averaged within ligand classes as well as over the whole dataset.

[2]For example, atom P/2569/AMP/2/X (corresponding to atom name 'P', atom number 2569, residue name AMP, residue number 2, chain identifier X) in PDB entry 12as is equivalent to atom AP/3203/FAD/405/A in PDB entry 1cqx.

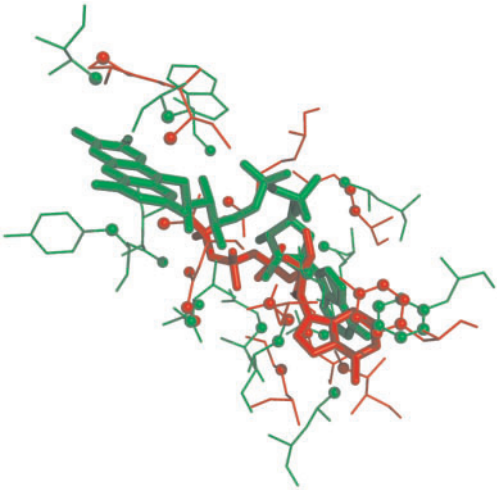

Fig. 2. Detection of similarities. Pairwise comparison of ATP-dependent DNA ligase from bacteriophage T7 bound to ATP (PDB code 1a0i) used as query (red) and rat short chain acyl-coa dehydrogenase (PDB code 1jqi) bound to FAD used as target (green). These two proteins are non-homologous and share 21% sequence identity. Comparison of the OM model for 1a0i to the IM5 model of 1jqi detectes 20 binding site atoms in common (spheres) belonging to different residues (lines). These atoms correctly identify 74% of equivalent ligand atoms, corresponding to the AMP core common to ATP and FAD. Superimposition of the binding sites based on the detected binding site similarities places the corresponding ligand atoms at a root mean square distance (RMSD) of 3.3 Å. Note that the OM model used as query is much larger than the atoms found in common (red spheres) yet the atoms in common correctly identify the common AMP core. Furthermore, if the same atoms were found in common but the equivalent ligands would not have been found, the prediction would be unsuccessful.

3 RESULTS

Default values for the algorithm parameters, were obtained through an extensive heuristic search in parameter space maximizing the average AUC. For this task, we utilized a slightly different dataset comprising the subset of the dataset with the smallest five IM models with $d = 5$ Å from each ligand class from the original dataset of Kahraman *et al.* (2007) as well as five examples of proteins bound the phosphate. Each point in parameter space involved 1250 pairwise comparisons and subsequent calculation of average AUC. The parameters obtained were utilised for all other comparisons.

3.1 Effect of uncertainty and contribution of size, chemical composition and geometry

The effects of lack of knowledge about the identity of the binding site atoms, i.e. the ligand-interacting cleft atoms, and the contributions of size, chemical composition and geometry are interconnected.

To determine these effects and contributions, the various cleft models: IM (with $5.0 \leq d \leq 15$), CM and OM, were compared against the IM5 database of models. The predictions were ranked according to the similarities calculated using Equation 1 together with the appropriate measure of N_c (Equations 2, 3 and 4). True positive predictions were marked using the appropriate method and used

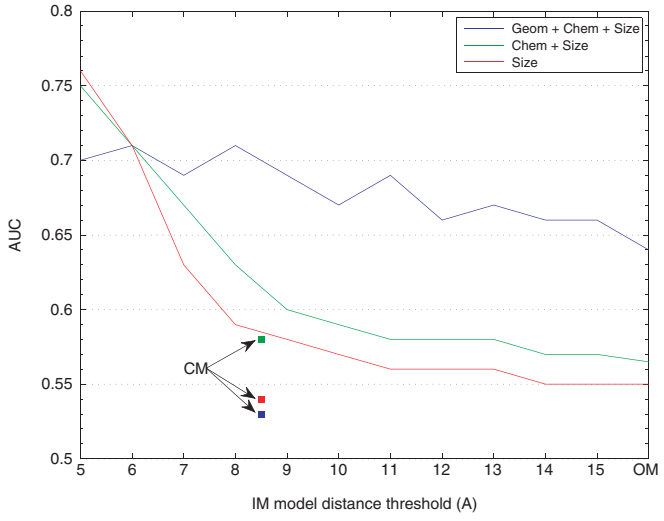

Fig. 3. Contribution of geometry, chemical composition and size to prediction accuracy as a function of uncertainty. It is possible to discriminate binding sites in the absence of uncertainty in the knowledge of the identity of ligand-interacting cleft atoms. As uncertainty in the knowledge of the binding site grows, chemical composition or binding site size alone are not sufficient to discriminate binding sites. While uncertainty also has an effect on the contribution of geometry to prediction accuracy, this effect is less drastic. A sufficiently large value of d would encompass all cleft atoms and therefore corresponds to the OM model. The last point in the graph shows the values for the OM models but it should be understood that this last point is not in scale with the rest of the axis. Conserved atoms (CM) are not informative for the discrimination of binding sites that bind similar ligands.

to calculate average AUC values. These results are presented in Figure 3.

In the ideal situation, where there is full knowledge of the binding site atoms, $d = 5.0$ Å in Figure 3, it is possible to discriminate binding sites. The set of atoms in close proximity to the ligand are more similar for proteins that bind similar ligands that those that bind dissimilar ligands. These similarities are found at all levels, binding site size, chemical composition as well as geometry. While this situation is somewhat unrealistic from the point of view of function prediction, it suggests that binding site atoms are rather conserved even in the absence of detectable homology and thus it appears that binding similar ligands introduces physico-chemical constraints on the position and nature of binding site atoms.

As uncertainty on the knowledge of the identity of the binding site grows, $d > 5.0$ Å in Figure 3, we observe that prediction accuracy decreases rather abruptly when only chemical composition or size are utilized. A decrease is also observed when geometry is used but this decrease is more moderate. As a matter of fact, in the absence of any information about the position of the binding site within the cleft, i.e. when utilizing the OM models, the prediction ability of size and chemical composition is no better than random, while that of geometry is far from ideal but considerably better than random. Furthermore, as shown in Figure 4, different ligands are easier to discriminate than others. In this dataset, ATP, EST, FAD and HEM can be discriminated with higher accuracy than others. However, given the small sampling, the error in these averages is higher. This explains the AUC value below 0.5 for GLC.

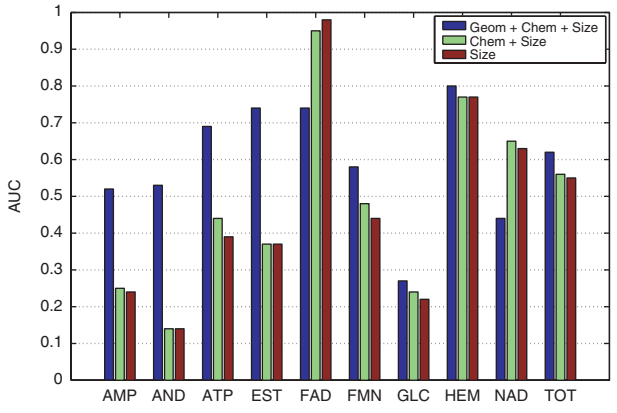

Fig. 4. Average AUC values for different ligand classes. The average AUC values from the comparison of OM cleft models against IM5 models shows that different ligands, such as ATP, EST, FAD and HEM, are easier to discriminate while others (AMP, AND, FMN and NAD) cannot be easily discriminated.

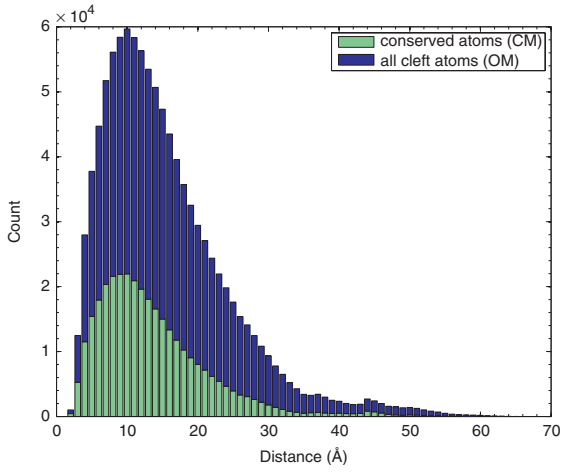

Fig. 5. Distribution of ligand-conserved cleft atom distances. Conserved atoms can be found at all distances from the ligand, they constitute a fraction of the total number of atoms at all distances and are approximately equally distributed.

3.2 Conservation does not improve predictions

Figure 3 also shows the average AUC values obtained from the comparison of conserved models (CM) against the database of IM5 models. These values are placed within the figure in the position that corresponds to the equivalent IM models based on the average number of atoms present in CM models. The results show that while size and chemical composition are not too far off from the values obtained with the corresponding IM models, the average AUC including the contribution of geometry is quite poor compared to that of the corresponding IM models.

The distribution of ligand-cleft atoms for all atoms and conserved atoms (Fig. 5) shows that conserved atoms are similarly distributed and make up a fraction of all cleft atoms at all distances. Clearly, conserved atoms at distances larger than those within which these

atoms could interact with the ligand, while functionally relevant as reflected by their conservation, cannot be relevant to ligand–protein interactions. We would like to determine if the inclusion of these atoms in the conserved models is the reason why these models perform so poorly.

We define a new cleft model composed of highly conserved atoms within 5.0 Å of the ligand (CIM5). Pairwise comparison of the CIM5 database against itself, produces average AUC values of 0.62 when considering geometry, chemical composition and size; 0.62 for chemical composition and size alone; and lastly, 0.58 for binding site size. Considering that the CIM5 clefts contain less atoms than the IM5 models, and therefore less chances for spurious similarities, we would expect the average AUC values to be closer to those for the corresponding IM5 comparison (the left most points in the curves in Fig. 3) which are 0.70, 0.75 and 0.76, respectively. This result suggest that, physico-chemical constraints cannot be the sole reason for the conservation of atoms around the ligand. If that was the case, we would expect average AUC values closer, if not higher, than those for the IM5 clefts which contain non-conserved ligand-interacting atoms in addition to those present in the CIM5 model (Fig. 1).

4 CONCLUSIONS

We describe IsoCleft in this work, a graph-matching method for the detection of local 3D atomic similarities. The method detects nearly-optimal approximate solutions for the graph-matching problem thus making it possible to compare large sets of atoms such as those obtained from naive geometric definitions of the binding site. The set of atoms need not be connected, the only information necessary are the coordinates, chemical identity and residue identity of the atoms. Furthermore, as we do not utilize any sequence alignment information, IsoCleft is also applicable to the comparison of non-homologous proteins.

We asked in this work whether it is possible to discriminate within a dataset, those proteins that bind similar ligands based on local 3D atomic similarities. The answer to this question is positive. However, the discrimination ability depends on how accurately the ligand-interacting atoms are known. While this set of atoms may not be known in advance when trying to predict the function of a protein, this result shows that the set of ligand interacting atoms are somewhat unique and thus lends further support to the use of docking techniques as well as the definition of binding templates.

Uncertainty on the knowledge of which atoms within a cleft interact with the ligand decreases our discrimination ability but also points out the power of our method in detecting functionally relevant similarities as compared to chemical composition and binding site size alone. Our results point to the need of combining our approach with information that help pinpoint which atoms within a cleft may interact with the ligand. Such information may come from computational methods as well as experimental data. Other methods can be used to predict small molecule protein binding, among these, we are currently studying how the method presented here compares to docking.

The most striking result in the present work is the poor discriminating ability when using cleft atoms belonging to highly conserved residues, particularly the subset of conserved atoms within interacting distance from the ligand. One can understand that the conservation of distant atoms in different families may not be related to the need to satisfy physico-chemical constraints posed by the ligand. However, the poor discriminatory ability of conserved atoms within interacting distance from the ligand is surprising. The fact that one needs to use all atoms around a ligand rather than just those that are conserved suggest that non-conserved atoms are functionally important.

It is still unclear what is the reason for different patterns of conservation among binding sites that evolved independently to bind similar ligands. One possibility is that a small number of crucial atoms are needed to hold a ligand in place and different protein families either utilise different such subsets or these subsets are too small to be picked out and we are unable to detect them. The remaining conserved ligand-interacting atoms might be present to satisfy other constraints resulting from the different spatial and cellular contexts in which different proteins evolve. One such constraint is the need to prevent the competitive binding of similar small molecules, which could potentially interfere with the action of the given protein. We recently found some correlation between binding site similarities and functional similarities within the Human cytosolic sulfotransferase family (Allali-Hassani *et al*., 2007; Najmanovich *et al*., 2007). We also observed that a small number of binding site differences may in some cases, but not others, have a drastic effect on protein function.

A single framework may help rationalize the observations presented in the current work for convergent evolution as well as those for the case of divergent evolution in human cytosolic sulfotransferase family members. In both cases, binding site differences may be in place to affect the free energy of binding for competing small molecules without necessarily affecting the binding of relevant small molecules. We therefore put forward the idea that molecular recognition, as the observed result of an evolutionary process, cannot be fully understood outside the spatial and temporal cellular context.

ACKNOWLEDGEMENTS

We would like to thank Abdullah Kahraman for providing the dataset on which our dataset is based. The authors would like to thank Richard Morris for discussions and comments on this article.

Funding: R.N. would like to acknowledge funding from the Structural Genomics Consortium (SGC). The SGC is a public private charitable partnership that receives funds from Canada Foundation for Innovation, Canadian Institutes for Health Research, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, the Ontario Innovation Trust, the Ontario Research and Development Challenge Fund and the Wellcome Trust. N.K. would like to acknowledge the Marie Curie 'Biostar' grant MEST-CT-2004-513973, which made her visit to EBI possible.

Conflict of Interest: none declared.

REFERENCES

Allali-Hassani,A. *et al*. (2007) Structural and chemical profiling of the human cytosolic sulfotransferases. *PLoS Biol*., **5**, e97.

Armon,A. *et al*. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol*., **307**, 447–463.

Arun,K.S. *et al*. (1987) Least-Squares Fitting Of 2 3-D Point Sets. *IEEE Trans. Pattern Anal. Mach. Intell*., **9**, 699–700.

Brakoulias,A. and Jackson,R.M. (2004) Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins*, **56**, 250–260.

Bron,C. and Kerbosch,J. (1973) Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, **16**, 575–577.

Glaser,F. *et al*. (2006) A method for localizing ligand binding pockets in protein structures. *Proteins*, **62**, 479–488.

Glaser,F. *et al*. (2005) The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, **58**, 610–617.

Jones,D.T. *et al*. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci*., **8**, 275–282.

Kahraman,A. *et al*. (2007) Shape variation in protein binding pockets and their ligands. *J. Mol. Biol*., **368**, 283–301.

Kobayashi,N. and Go,N. (1997) A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. *Eur. Biophys. J*., **26**, 135–144.

Kresher,D. and Stinson,D. (1998) *Combinatorial Algorithms: Generation, Enumeration and Search*. CRC Press, Boca Raton, Florida.

Laskowski,R.A. *et al*. (1995) Surfnet – a program for visualizing molecular-surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*., **13**, 323–330.

Laskowski,R.A. *et al*. (1996) Protein clefts in molecular recognition and function. *Protein Sci*., **5**, 2438–2452.

Najmanovich,R.J *et al*. (2005) Prediction of protein function from structure: insights from methods for the detection of local structural similarities. *Biotechniques*, **38**, 847, 849, 851.

Najmanovich,R.J. *et al*. (2007) Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics*, **23**, e104–e109.

Rigoutsos,I. and Wolfson,H.J. (1997) Geometric hashing. *IEEE Comput. Sci. Eng*., **4**, 9–9.

Schmitt,S. *et al*. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol*., **323**, 387–406.

Sobolev,V. *et al*. (1996) Molecular docking using surface complementarity. *Proteins Struct. Funct. Genet*., **25**, 120–129.

Shulman-Peleg,A. *et al*. (2004) Recognition of functional sites in protein structures. *J. Mol. Biol*., **339**, 607–633.

Shulman-Peleg,A. *et al*. (2005) SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res*., **33**, W337–W341.

Weskamp,N. *et al*. (2004) Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics*, **20**, 1522–1526.