New opportunities for protease ligand-binding site comparisons using SitesBase

N.D. Gold, K. Deville and R.M. Jackson¹
Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, U.K.

Abstract
The rapid expansion of structural information for protein ligand-binding sites is potentially an important source of information in structure-based drug design and in understanding ligand cross-reactivity and toxicity. We have developed SitesBase, a comprehensive database of ligand-binding sites extracted automatically from the Macromolecular Structure Database. SitesBase is an easily accessible database which is simple to use and holds pre-calculated information about structural similarities between known ligand-binding sites. These similarities are presented to the wider community enabling binding-site comparisons for therapeutically interesting protein families, such as the proteases and for new proteins to enable the discovery of interesting new structure–function relationships. The database is available from http://www.modelling.leeds.ac.uk/sb/.

Introduction
Using chemical similarity searching in small-molecule databases is a well-established method for identifying potential new ligands for a protein target. A single ligand or list of known active compounds (that have been structurally superimposed) can be used to generate either a two-dimensional fingerprint or three-dimensional pharmacophore, which can in turn be searched against small-molecule databases. The method can be used to find new active compounds for an existing target or, alternatively, an active compound for a new protein target based on sequence and assumed structural similarity to a known target with known active compounds [i.e. establishing the existence of a SAR (structure–activity relationship)]. A pharmacophore hypothesis so derived may be applicable to the whole family of proteins, allowing sequence similarity alone to allow inference of a potential SAR. However, this relationship may be taken further by considering protein structural information. Molecular recognition of small molecules occurs in binding sites or cavities on the protein surface. Therefore, ultimately, it is the properties of the binding sites that determine the ability of two independent proteins to recognize a given ligand. Studies have established that proteins of unrelated sequence and even unrelated structure (as defined by fold similarity at the domain level) can bind a ligand in the same way in the binding site (e.g. [1]). Therefore binding-site similarity analysis can complement methods based on sequence or even three-dimensional fold information for functional annotation. Alternatively, the existence of subtle differences in binding sites of proteins in the same family may represent exploitable opportunities for drug design, differences that are difficult to detect by sequence similarity alone. In addition, the importance of conformational change between apo- and ligand-bound conformational states, as well as between different ligand-bound states, may be better understood in the ligand design process.

The potential applications of binding-site similarity are evident for a number of important applications such as: (i) structure-based ligand design, including the extent to which different binding sites maintain selectivity for a particular ligand(s); (ii) predicting ligand cross-reactivity, i.e. a ligand recognizing two different proteins which have binding-site similarity; (iii) functional annotation of structurally determined proteins of unknown function; and (iv) detection of previously unrecognized functional sites on structurally determined proteins of known function (the last two applications must be used in combination with binding-site prediction methods [2,3]). With this in mind, and given the rapid expansion of structural information for protein ligand-binding sites, we have developed a large relational database of ligand-binding sites extracted automatically from the PDB [4]. This has been combined with a new fast method for calculating binding site similarity based on geometric hashing [5], to create SitesBase [6]. It contains an all-against-all comparison of binding sites of most ligands, but excludes metal ions and solvent molecules. Currently, SitesBase is restricted to protein-ligand binding sites that have a bound ligand present. SitesBase is so far unique in that it provides a pre-calculated easily accessible resource for the retrieval of site similarities for any of over 30,000 protein ligand-binding sites. SitesBase may prove useful in function prediction, particularly when used in combination with other available information, such as tools that combine multiple methods and data sources for functional assignment such as ProFunc [7].

Key words: ligand-binding site, protease, SitesBase, small-molecule database, structure-based drug design, subtilisin.

Abbreviations used: SAR, structure–activity relationship.

¹To whom correspondence should be addressed (email r.m.jackson@leeds.ac.uk).
Overview of SitesBase

SitesBase is an Internet-accessible relational database holding known protein ligand-binding sites. It contains the extent of structural similarity between any two sites and information that allows their rapid structural superposition. The database is created from the PDB by extracting binding sites using the bound ligands to establish their location [the binding site is defined as all protein atoms within a 5 Å (1 Å = 0.1 nm) radius of any ligand atom]. In each case, all available data for each atom within both the binding site and the ligand is stored. Small molecules with fewer than six atoms are not considered (this excludes metal ions and small solvent molecules) and are not included within SitesBase. Several commonly occurring solvent molecules such as glycerol are also discarded. Atomic positions within each binding site are compared with all other sites involving an all-against-all comparison using a fast geometric hashing algorithm [5] to determine the level of structural similarity between each pair. In this process, only the atoms of the protein sites are compared, ligand atoms are not taken into account. Similarity is measured by atom–atom score, i.e. the number of atoms matching in similar spatial orientations and the same element (carbon, nitrogen, oxygen, etc.). A conservative atom–atom score cut-off of 20 matching atoms is used to determine whether the pairwise similarity is retained preventing numerous and insignificant matches being stored within SitesBase. Atom–atom scores below 20 are assessed to find the percentage of atoms within the smallest site covered by the identified similarity and are included within SitesBase if this is >30%. If sites are considered to be similar using these criteria, then a seq sim score (see below) and the rotation/translation matrix are also stored to enable structural superposition of the sites.

In order to aid analysis of any similar hits retrieved from the database with a given query, useful annotations of the data are also stored within SitesBase. Each binding site is annotated with relevant structural classification codes from the SCOP database [8] (i.e. by class, fold, superfamily and family). This allows highly similar sites (often also with high sequence similarity) to be easily identified with respect to more distant relatives. Assignment of SCOP classification is not always straightforward because binding sites can occur at the interface of more than one structural domain. Where this is the case, then all SCOP domains contributing atoms to the binding site are found and listed according to the number of atoms that they contribute to the site. In most cases, the primary assignment is consistent over all other family members; however, sometimes one of the other assignments may be more appropriate when a site is viewed in context with other relatives.

For each query site, generation of a background random extreme value distribution of all atom–atom scores from SitesBase where all known relatives (family and superfamily) have been removed, was used to calculate the probability (P value) of obtaining a given atom–atom score by chance [1]. A combination of sequence and structural similarity between any pair of binding sites is calculated and stored using the method of Stark et al. [9]. This seq sim score provides an assessment of how conserved residues align by using certain atoms (Ca, Cβ and a functional atom where available) from the structure-based alignment. A low seq sim (E value) score indicates that important atoms are found in spatially similar positions with high sequence similarity.

Using SitesBase

The SitesBase interface (http://www.modelling.leeds.ac.uk/sb/) is simple and requires a PDB code, ligand name or keyword to identify all the ligand binding sites within the specified protein(s). Selection of one of these sites searches the database for all binding sites with atomic-level structural similarity to the query. The hits are provided in a list ranked by decreasing atom–atom score, and are coloured according to their overall structural similarity to the query using the SCOP domain classification. The user can select specific hits for multiple structural alignment with the query site. In addition to a PDB format file containing the three-dimensional co-ordinates of the superimposed sites and ligands, a graphical multiple structural alignment of equivalenced atoms with residues coloured according to the ClustalX [10] colour scheme is also produced.

Opportunities for protease binding-site comparison

Trypsin and subtilisin binding-site similarity

Several examples of so-called ‘convergent evolution’ exist in the structural database [11]. Most notable of these is the functional and structural similarity between two distinct serine protease fold families of trypsin and subtilisin [11,12]. Although they catalyse the same chemical reaction, they are unrelated by sequence or fold similarity. Indeed, SitesBase is able to capture this significant site similarity using either trypsin or subtilisin as the query site. For example searching (PDB code 1BH6) Subtilisin DY (SCOP c.41.1.1; EC 3.4.21.62) in complex with a synthetic inhibitor (N-benzyloxycarbonyl-Ala-Pro-Phe-chloromethane) retrieves the trypsin-like (PDB code 1SGC) proteinase A (SCOP b.47.1.1; EC 3.4.21.80) in complex with chymostatin as the first non-subtilisin-like hit. They share an atom–atom similarity score of 35, with the overlap of the critically important serine and histidine residues of the Ser-His-Asp catalytic triad in addition to conserved β-strand main-chain conformation lining the ligand-binding site (Figure 1). Indeed, the spatial proximity and chemical similarity of the functional groups comprising the two inhibitors is striking (despite the fact they are not used in the structural superposition).

Protease binding-site similarity

Proteins have largely evolved and diverged from common ancestors, so their folds are likely to be conserved even where sequence similarity is no longer detectable. Therefore
it is highly likely that related proteins bind related small-molecule substrates and metabolites. In fact, there is strong evidence to support this in terms of structural analysis of the α-helical proteins and their ligands [13]. Nobeli et al. [14] have recently performed an extensive analysis of the diversity of protein–ligand interactions in *Escherichia coli* and have shown that only a few protein superfamilies show a great deal of substrate diversity. It is interesting to speculate that there are also a limited number of different protein ligand-recognition motifs in protein space [15]. This could have important implications for toxicity owing to the potential for ligand cross-reactivity in structure-based drug design.

**Protease druggable human genome**

Recent studies suggest that the known druggable protein targets in the human genome are smaller than originally thought [16]. Further analysis shows 161 human protein families (Pfam domains) comprise the known druggable human genome [17], suggesting that this relates to 5930 unique druggable domain sequences (at 95% sequence identity). These Pfam families can be grouped into 22 target groupings based on broad functional definitions (i.e. numerous smaller target families are grouped together according to class). Overall structural coverage (as represented by PDB structures) of the unique domain sequences is still quite low at 12.5% and progress in achieving occupancy of the different enzyme families (EC levels) is slow [18]. However, total coverage increases to 31.5% if comparative modelling of targets with >50% sequence identity to proteins of known structure is taken into account (generally 50% sequence identity is considered to be sufficient quality for structure-based drug design [19]), or even 80.5% if >20% sequence identity is considered. In particular, in relation to proteases, there are three major target groupings [17]: (i) trypsin-like proteases with 26% PDB coverage (56 of 214 sequences), and 98% coverage when combining comparative models (with >20% sequence identity); (ii) metalloproteases with 19% PDB coverage (31 of 166 sequences), and 91% when combining comparative models; and (iii) other hydrolases with 18% PDB coverage (105 of 572 sequences) and 85% when combining comparative models. Coverage is high for proteases as a group (particularly trypsin-like proteases); however, major questions remain as to how useful comparative modelling is for sequence identities of lower than 50% which represents a significant proportion of the structural coverage.

**Protein kinase binding site comparison**

Like the human proteases, the protein kinases are an important target family, showing considerable diversity in pathway regulation and substrate specificities. In contrast with the protease target groupings, they only have 9% PDB coverage (67 of 738 sequences) [17]; however, have been extensively studied and modelled by comparative modelling methods (see, e.g., [20]) because of their therapeutic importance. Preliminary analysis of the protein kinase family has been undertaken using SitesBase. The dataset was generated from using the EC code as a search tool. Protein kinases belong to 39 different and well-characterized EC families (EC 2.7.10.x, EC 2.7.11.x, EC 2.7.12.x, EC 2.7.13.x and EC 2.7.99.x) as well as an additional 23 EC families from EC 2.7.1.x. In total, 62 different search criteria were used, and partially annotated proteins were also inspected (an example is EC 2.1.x.xx). After removal of wrongly assigned proteins and protein kinases crystallized without their catalytic domain, the final PDB codes collection contained 542 entries (many proteins have multiple entries containing different ligands) representing 19 distinct putative functional families. Each binding site was checked carefully to make sure that the ligand was (i) an ATP analogue, and (ii) bound in the catalytic site (ligands lying in secondary sites were removed).

It should be noted that, of the 39 distinct EC numbers corresponding to the protein kinase family, only 15 (38%) are represented (i.e. have structural coverage) within the dataset. The chosen value of 15 clusters (Figure 2) corresponds to the number of functionally different protein kinases in terms of EC number. The first cluster shows a very poor inter-cluster similarity, whereas the intra-cluster similarity is quite high. This cluster contains histidine kinase subfamily members and various atypical protein kinases, such as anti-σ factor, BCK (branched-chain α-ketoacid dehydrogenase kinase) (mitochondrial protein kinase) and PDK (pyruvate dehydrogenase kinase). The large cluster at the bottom contains some receptor and non-receptor protein tyrosine kinases, non-specific serine/threonine protein kinases, receptor serine/threonine, receptor protein kinase and MAPK (mitogen-activated protein kinase). They all belong to the same SCOP classification: d.144.1.7. Remarkably, SitesBase manages to separate the dataset into the following families: tau protein kinase, cAMP-dependent protein kinase and cyclin-dependent kinases. The other smaller families are clustered
with other small families. We also manage to separate protein kinases belonging to the same subfamily but having different folds into distinct clusters. However (as can be seen from the density of grey coloration), the overall site similarity between diverse members of the protein kinase family is high.

**Concluding remarks**

Understanding protein binding-site similarity and the influence of bound ligands on protein conformation is an emerging field. Nonetheless, it is important for understanding how different protein binding sites maintain selectivity for particular ligands and predicting ligand cross-reactivity, in addition to the structure-based functional annotation of new and existing proteins. The SitesBase database can be used to identify these similarities and compare the spatial location of ligands in similar binding sites. This has already proved useful in studies of functional assignment [1] and in creating receptor-based pharmacophore models [21]. As the number of experimentally determined protein structures increases, we can expect increasing coverage of protein binding-site space, as well as increasing depth of coverage of particular therapeutically interesting protein families, such as the trypsin-like proteases and metalloproteases. This will allow the greater exploitation of binding-site structural information for any new protein, as well as add to the overall structural coverage of binding-site space, which will be of benefit in identifying off-target liabilities or opportunities for new and existing protein families.

**References**


Received 1 February 2007