A Searchable Database for Comparing Protein−Ligand Binding Sites for the Analysis of Structure−Function Relationships

Nicola D. Gold and Richard M. Jackson*

Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, U.K.

Received August 31, 2005

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 a large database of ligand binding sites extracted automatically from the Protein Data Bank. This has been combined with a method for calculating binding site similarity based on geometric hashing to create a relational database for the retrieval of site similarity and binding site superposition. It contains an all-against-all comparison of binding sites and holds known protein−ligand binding sites, which are made accessible to data mining. Here we demonstrate its utility in two structure-based applications: in determining site similarity and in aiding the derivation of a receptor-based pharmacophore model. The database is available from http://www.bioinformatics.leeds.ac.uk/sb/.

INTRODUCTION

The field of structure-based design increasingly influences the pharmaceutical lead discovery and optimization process. This process relies on using available 3D structural information to inform the design of novel small molecules to interact in a specific way with binding sites on proteins. The ability to manipulate and use protein−ligand binding site information in this design process is important for targeting specific proteins or protein families. The rapid expansion of structural information for protein−ligand binding sites is now a valuable source of information in structure-based drug design and will aid in understanding ligand cross reactivity and toxicity. With this in mind we have developed a large database (SitesBase) of ligand binding sites extracted automatically from the Protein Data Bank (PDB).[1] This has been combined with a new fast method for calculating binding site similarity based on geometric hashing.[2] A relational database (SitesBase) for the retrieval of site similarity and binding site superposition of most ligands (excluding metal ions and solvent molecules) in the structural database has been created. This contains an all-against-all comparison of binding sites that is accessible to data mining. 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 precalculated, easily accessible resource for the retrieval of site similarities for any of over 30 000 protein−ligand binding sites. Other databases also store structural information about protein−ligand binding sites (e.g. LigBase,[3] PDB-Ligand,[4] and Relibase[5]), and several methods have been developed to search a protein against a database of 3D structure patterns or surfaces (e.g. PINTS,[6] EF-site,[7] pvSOAR,[8] SiteEngine[9]) to identify the presence of significant similarities. The premise is that similarity in the shape and properties of binding and active sites can indicate common modes of molecular recognition or common biochemical function (e.g. enzyme activity or specificity) even in the absence of overall fold similarity.[10,11] SitesBase may also prove useful in function prediction; however, currently it is restricted to searching sites already present in the PDB. In general, site comparison methods may prove more powerful when used in combination with other available information, such as tools that combine multiple methods and data sources for functional assignment such as ProFunc.[12]

Since proteins have largely evolved and diverged from common ancestors, their folds are likely to be conserved even where sequence similarity is no longer detectable. For this reason it is thought that in nature there are almost certainly a limited number of protein folds. It is therefore 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 alpha helical proteins and their ligands in the PDB.[13] Nobeli et al.[14] have recently performed an extensive analysis of the diversity of protein−ligand interactions in *E. coli* and shown that only a few protein superfamilies show a great deal of substrate diversity. They also show that all the metabolites in *E. coli* can be reduced to a series of "structural keys" i.e., a modular set of molecular fragments consisting of 61 structural keys e.g. adenine, pyridoxal, ribose, and phosphate (this excludes individual metal or halogen atoms) that assemble to produce all the 900 small molecule metabolites identified (from pathways) in *E. coli*. It can be anticipated that in nature there are almost certainly a limited number of small molecule metabolites. However, this combinatorial library of 61 small molecule fragments represents a very small proportion of chemical space. The level of molecular diversity in small molecule databases of druglike compounds (e.g. the World Drug Index[15]) is several orders of magnitude larger. It is therefore interesting to speculate that there are also a limited number of different protein−ligand recognition motifs in protein space. This has implications for toxicity and the

* Corresponding author phone: +44 (0)113 343 2592; fax: +44 (0)-113 343 3167; e-mail: r.m.jackson@leeds.ac.uk.

10.1021/ci050359c CCC: $33.50
Published on Web 00/00/0000 PAGE EST: 6.9

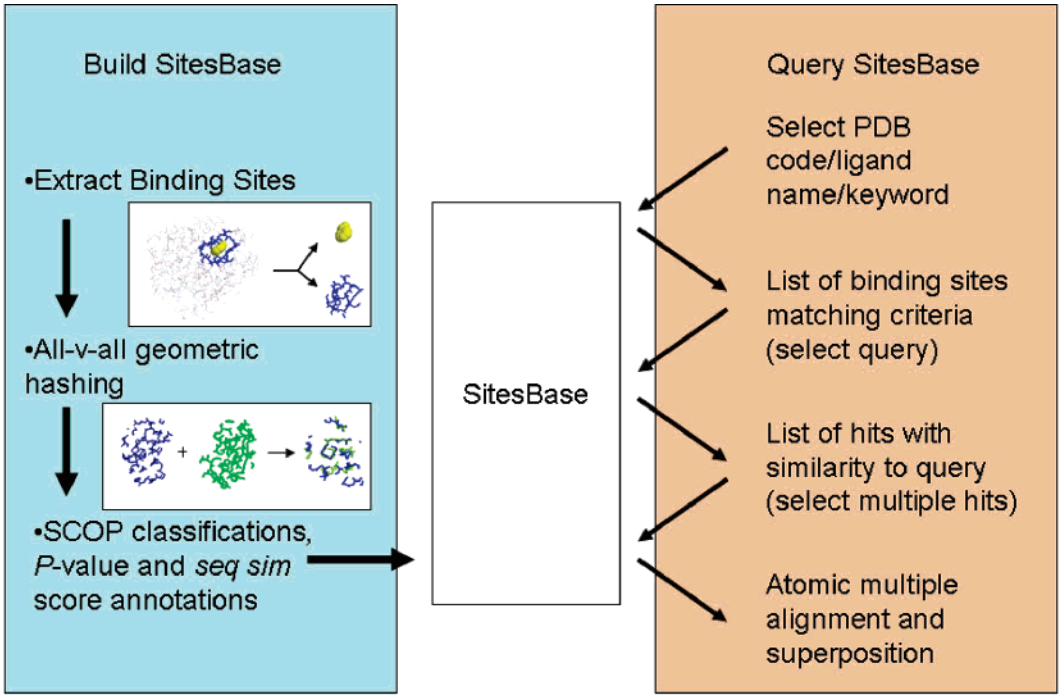

Figure 1. Scheme for the construction and querying of SitesBase.

potential for ligand cross-reactivity in structure-based drug design. Alternatively, this property could be exploited in structure-based drug design or chemical genetics approaches by employing different structural combinations of a fairly small combinatorial library of molecular fragments that are analogues of those occurring in nature.

Here we describe the database structure and content of SitesBase. We show that binding site structural similarity is common in the large class of nucleotide binding proteins. A specific example from the structurally conserved P-loop proteins is chosen that transcends different protein folds occurring in several different structural superfamilies. We also show the utility of the database for the generation of 3D multiple alignments of protein−ligand binding sites for the creation of a structure-based pharmacophore for glycogen synthase kinase GSK-3β a regulatory serine/threonine kinase. This example shows that the method is able to create an effective structure-based alignment of crystallographic ligands despite considerable structural variability of the binding site flexible loop when in complex with different ligands.

METHODS

SitesBase is a WWW accessible relational database holding known protein−ligand binding sites and the extent of structural similarity between them. The database was created by extracting binding sites from the PDB (Figure 1) using bound ligands to establish their location (the binding site is defined as all protein atoms within a 5 Å radius of any ligand atom). In each case all available data for each atom within both the protein binding site and the ligand were stored. Small molecules with fewer than 6 atoms are not considered (excluding metal ions and small solvent molecules) and are not included within SitesBase. We also discard several commonly occurring solvent molecules such as glycerol.

Atomic positions within each binding site were then compared to all other sites within SitesBase using a fast geometric hashing algorithm[2] to determine the level of structural similarity between each binding site pair (Figure 1). In this process only the atoms of the protein sites were compared; ligand atoms were not taken into account. Similarity was measured by atom−atom score (i.e. the number of atoms matching in similar spatial orientations and with similar types (carbon, nitrogen, oxygen, etc.). A conservative atom−atom score cutoff of 20 matching atoms was used to determine if the pairwise similarity should be stored within the database preventing numerous and insignificant matches being stored. Atom−atom scores below 20 were assessed to find the percentage of atoms within the smallest site covered by the identified similarity and were included within SitesBase if this was >30%. If sites were considered similar using these criteria, then a *seq sim* score (see below) and rotation/translation matrix were also stored to enable superposition of the sites.

SitesBase searches proceed by identifying a specific binding site within a protein to act as a query to retrieve all other binding sites with atomic similarity. To aid analysis of any similar hits retrieved from the database with a given query, useful annotations of the data were stored within SitesBase (Figure 1). Each binding site was annotated with relevant structural classification codes from the SCOP database[16] (i.e. by *class*, *fold*, *superfamily*, and *family*). This allows easy separation of structurally highly similar sites (often also with high sequence similarity) from more distant relatives. The SCOP database is hierarchically clustered according to sequence, structure, and function relationships of protein domains. Domains with similar structure (overall fold) appear in the same *fold* group and are further clustered according to sequence and functional similarity into *superfamilies* and *families*. Assignment of SCOP classification is not always straightforward because binding sites can occur at the interface of more than one structural domain. If this were the case, then all SCOP domains contributing atoms to the binding site were found and ranked according to the

number of atoms they contribute. The domain supplying the most atoms was termed primary with others listed as secondary and tertiary in ranked order. 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 seen in context with other relatives.

For each query site, a generation of a background random extreme value distribution (EVD) 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. Calculated probabilities for each binding site with atomic similarity to the query were then annotated within SitesBase. It should be noted that *P*-values will change as new sites are incorporated into SitesBase and that using this method to remove known relatives can sometimes result in overestimating *P*-values if the data no longer resemble a true random distribution. This can occur where the query site has significant site similarity to proteins with different folds or if it is unclassified in SCOP and all relatives are retained in the presumed random distribution.

A combination of sequence and structural similarity between any pair of binding sites was calculated using the method of Stark et al.[6] and stored. This *seq sim* score provides an assessment of how well certain atoms (Cα, Cβ, and a functional atom[17] where available) within sequence-conserved residues align. A low *seq sim* score indicates that important atoms are found in highly similar positions with high sequence similarity.

The SitesBase WWW interface is simple and requires only a PDB code, ligand name, or keyword to identify all the ligand binding sites within the specified proteins. Selection of one of these sites searches the database for all binding sites with atomic-level structural similarity to the query. These hits are provided in a list ranked by decreasing atom−atom score. The hits are colored according to their overall structural similarity to the query using the SCOP database classification. User selection of hits then generates an atomic multiple alignment of the query and selected binding sites in a format similar to a multiple sequence alignment with residues colored according to the clustalX color scheme.[18] A PDB format file containing the 3D coordinates of hits superimposed on the query is also provided. A schema showing the database and query process is given in Figure 1.

APPLICATIONS IN STRUCTURE-BASED ANALYSIS OF PROTEIN−LIGAND INTERACTIONS

Querying Binding Site Similarity. The importance of examining ligand binding sites when predicting ligand cross reactivity is demonstrated by the large family of nucleotide binding proteins. These proteins are structurally diverse but many employ similar 3D binding site motifs to bind related nucleotide ligands[19−25] (ATP, GTP, NAD, FAD, FMN, etc.). Figure 2 shows an all-against-all similarity matrix of over 5000 nucleotide binding sites. Throughout SitesBase all of the ligand atoms are used to extract the protein binding sites.

Two thresholds were identified to define different levels of binding site similarity. A threshold score of 25 is used here to indicate the lowest score for which binding sites are considered similar. It represents 25 atoms of the same atom type occurring in a similar relative spatial orientation. A second atom−atom score threshold of 40 is also used here to indicate higher confidence of functional similarity and restrict inclusion of false positives. Sites scoring 25−39 are indicated with an orange dot, and a higher degree of similarity (score $\geq$40) is indicated by a blue dot. Sites are sorted and ordered along the axes by their SCOP classification numbers causing family relatives to be clustered together along the diagonal adjacent to superfamily and then fold relatives. Fold relatives are grouped according to their structural class. The class order from left to right (and bottom to top) is as follows: all-α, all-β, α/β, $\alpha+\beta$, multidomain-($\alpha+\beta$), coiled-coil, low resolution, designed proteins. The far right and top of the matrix is populated by nucleotide binding proteins that are currently unclassified by SCOP. Some of the most highly represented superfamilies have been labeled (see Figure 2 legend for details). The two largest clusters represent the P-loop containing nucleotide triphosphates (labeled H−L) and the Rossmann-like fold proteins (labeled C−G). Matches away from the diagonal indicate binding site similarity between different folds showing instances of structural and potentially functional similarity where the overall folds are different and indicate sites of potential ligand cross-reactivity. For example, there is similarity between the P-loop binding sites and the phosphoenolpyruvate (PEP) carboxykinase-like binding sites (Figure 2, label 2) previously identified by Matte et al.[26] These proteins have different folds but bind similar ligands (GDP and ADP, respectively). This figure also demonstrates the resource is useful prior to the classification of new proteins in the SCOP database by identifying similarity of known binding sites among the unclassified proteins.

In addition to possible ligand cross-reactivity, similarity in binding sites is important to (1) corroborate function predictions based on overall sequence or fold by providing additional evidence of functional similarity, (2) predict common ligand recognition or similar function between proteins with different folds, and (3) determine possible binding partners for proteins with known superfolds (fold associated with multiple functions[27]). Here we demonstrate binding site similarity in the absence of fold similarity with the results of a database search with the GDP binding site of query protein elongation factor Tu from *E. coli* (1dg1). This protein is a member of the large G-protein family and promotes the GTP-dependent binding of aminoacyl tRNA to the A-site of the ribosome during protein biosynthesis. It has three domains: a GDP binding P-loop containing nucleoside triphosphate hydrolase domain, an EF-Tu/eEF-1α/eIF2δ C-terminal domain, and a reductase/isomerase/elongation factor common domain. The binding site of the GDP binding domain contains 83 atoms within a 5 Å radius of GDP. The hits retrieved from the database search are detailed in Figure 3a in ranked order of decreasing similarity to the query and are colored according to their relationship to the query in the SCOP database allowing distant relatives to be easily distinguished from close relatives. This figure also highlights proteins that are not yet classified in SCOP. In this example many of the unclassified proteins can be classified as functional relatives because of their high similarity scores and low Root Mean Square Deviations

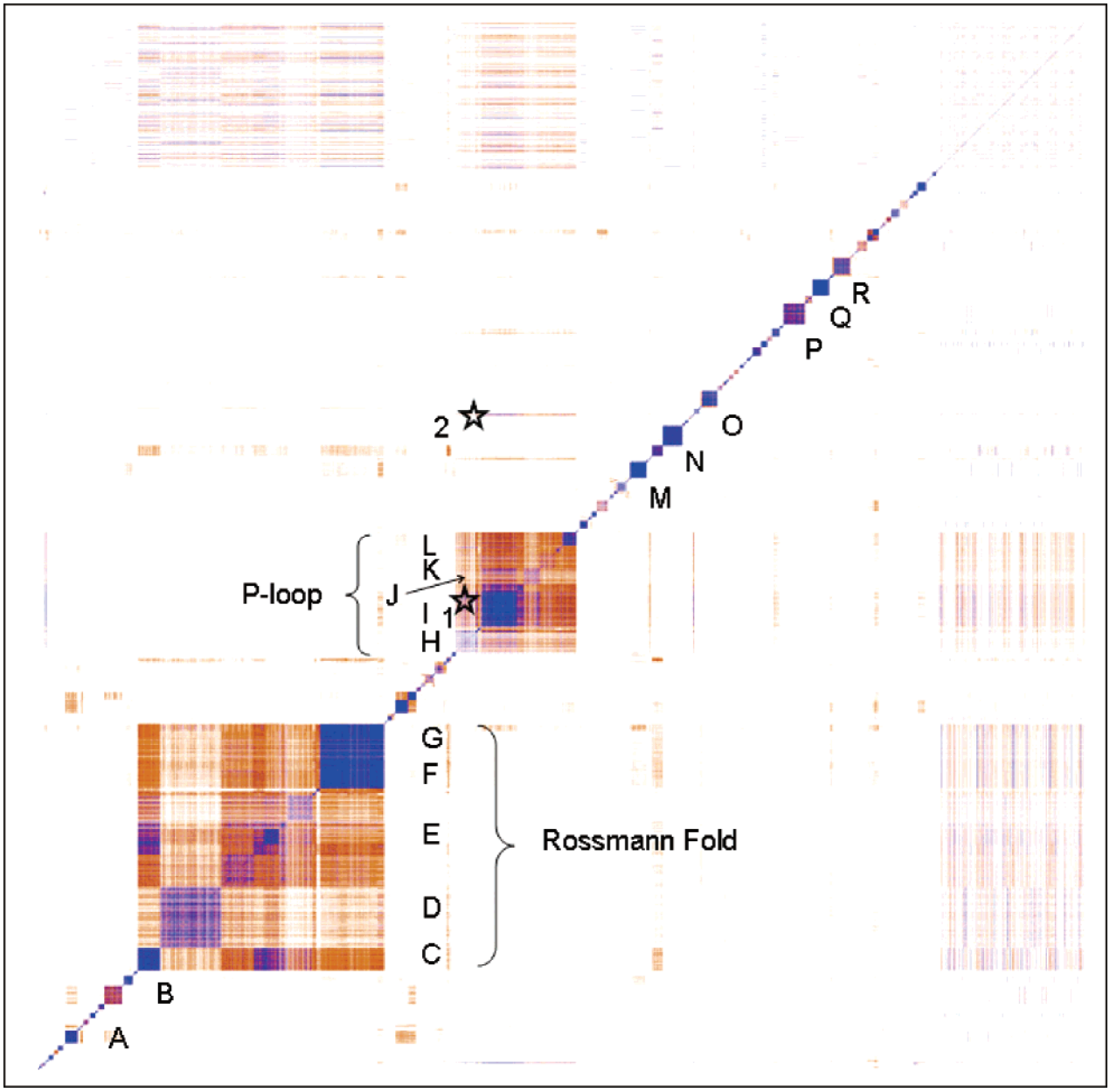

Figure 2. All-against-all similarity matrix. An atom−atom structural similarity score between sites of >40 is given in blue. A score of 25−39 is given in orange. Prominent SCOP superfamilies within the database are labeled, except in the cases of the Rossmann and P-loop superfamilies where each family is listed. (A) Riboflavin synthetase domainlike. (B) FMN-linked oxidoreductases. (C) Alcohol dehydrogenase. (D) Tyrosine-dependent oxidoreductases. (E) Glyceraldehyde-3-phosphate dehydrogenases; formate/glycerate dehydrogenases; LDH dehydrogenase-like; 6-phosphogluconate dehydrogenase-like; amino acid dehydrogenase-like; CoA binding domain; potassium channel; bifunctional dehydrogenase/ferrochelatase Met8p. (F) FAD/NAD(P) binding domain (fold). (G) Nucleotide binding domain (fold). (H) Nucleotide and nucleoside kinases; shikimate kinases; chloramphenicol phosphotransferase; adenosine-5;-phosphosulfate kinase; PAPA sulfotransferase. (I) Phosphoribulokinase/pantothenate kinase; 6-phosphofructo-2-kinase/fructose 2,6-bisphosphate kinase, G proteins; motor proteins. (J) Nitrogenase iron protein-like. (K) RecA protein-like, ABC transporter, helicase-like "domain" of reverse gyrase, gluconate kinase, YjeE-like, tandem AAA-ATPase domain. (L) Extended AAA-ATPase domain. (M) Dihydrofolate reductases. (N) Aldehyde reductases. (O) Microbial ribonucleases. (P) Thymidylate synthetase. (Q) Glutamine synthetase. (R) Glutathione synthetase. Hits found with query 1dg1 are labeled (1) 1rj9 and (2) 1k3c.

(RMSDs) prior to their classification in the structural databases. For example the highest scoring unclassified protein is 1sqk, also an elongation factor from S*ulfolobus solfataricus*. The best hits are family relatives; however, superfamily relatives such as 1rj9 (signal sequence recognition protein) (Figures 2 and 3, label 1) also score highly (atom − atom score = 57). This protein is necessary for the efficient export of extracytoplasmic proteins, and although these proteins have different functions within the cell, their molecular functions are similar as both proteins have GTP binding GTPase domains. Superposition of the binding sites of these two domains (Figure 3b) has an RMSD of 0.54 Å. Similarity can be seen around the P-loop or Walker A motif.[28] In this case a match occurs between part of the Walker A motif ($GX_4GK[T/S]$) which comprises the first strand and helix of the P-loop domain and is involved in binding the triphosphate moiety of the cofactor in both proteins. It can also be noted in Figure 3a that two hits in red (which indicate proteins in a different class to the query) rank above 1rj9. These represent hits at a domain interface where the

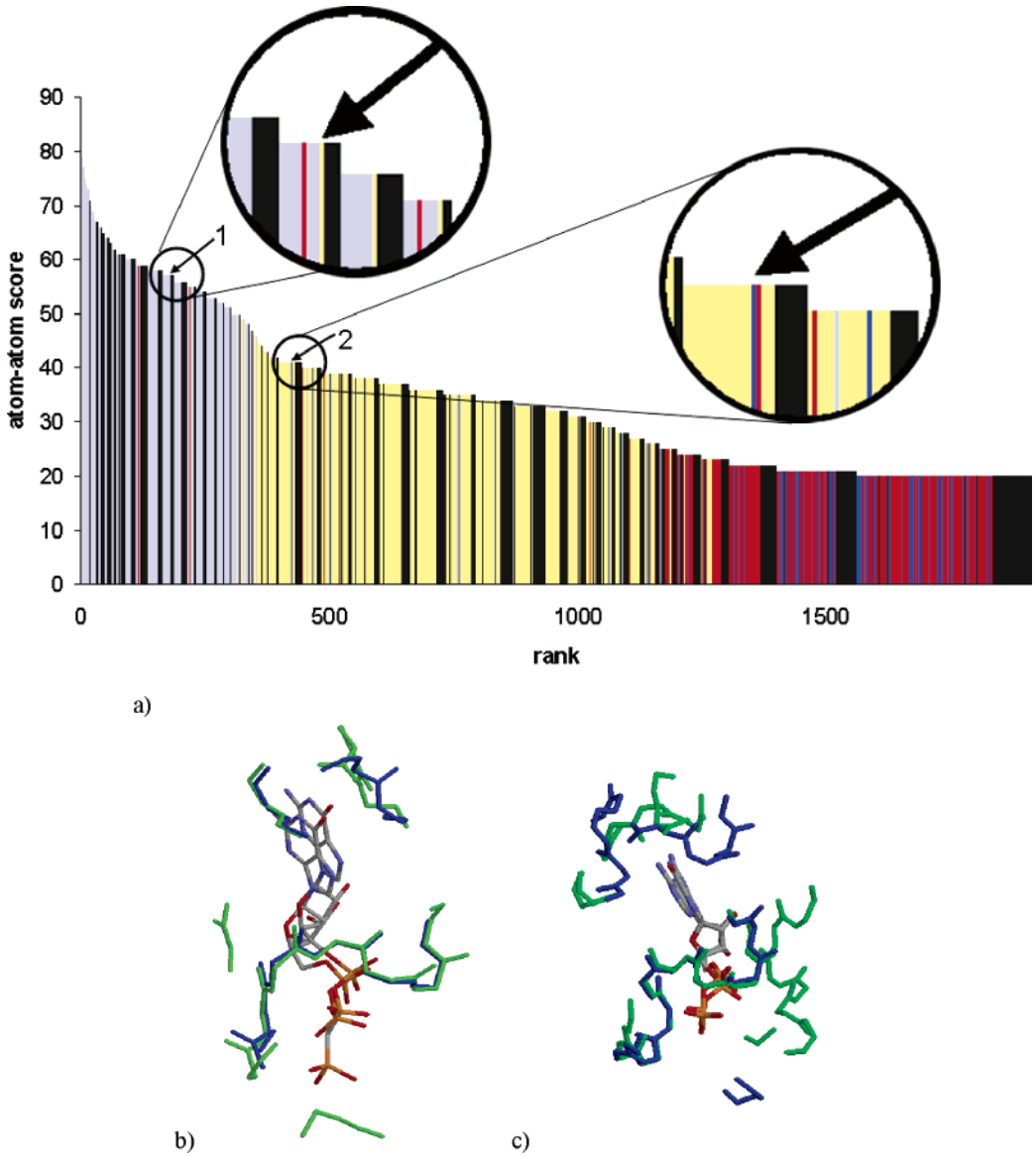

Figure 3. (a) Ranked hits with query GDP binding site from elongation factor Tu 1(dg1). The similarity atom−atom scores are plotted on the *y*-axis and the rank of the site on the *x*-axis. Hits are colored according to SCOP classification. Other members of the query's SCOP family are colored light blue. Superfamily and class relatives of the query are colored yellow and dark blue, respectively. Binding sites with unrelated SCOP classifications are shown in red, and proteins which are not yet classified in SCOP are colored black. Hits are labeled (1) 1rj9 and (2) 1k3c. (b) Superposition of the query GDP binding site of elongation factor Tu (1dg1) (blue) and GDP binding site from the highest scoring superfamily relative signal sequence recognition protein (1rj9) (green). The RMSD for the superposition is 0.54 Å. (c) Superposition of the query GDP binding site of elongation factor Tu (1dg1) (blue) and ADP binding site from the highest scoring class relative phosphoenolpyruvate carboxykinase (1k3c) (green). The RMSD for the superposition is 0.60 Å.

secondary SCOP assignment (see methods) is in fact a family relative and would have been more appropriate.

Another protein with a high atom−atom score reveals a protein with a different fold to the query. Phosphoenolpyruvate (PEP) carboxykinase (1k3c) which binds ADP within its PEP-carboxykinase-like domain is a gluconeogenic enzyme catalyzing the conversion of oxaloacetate to phosphoenolpyruvate (Figures 2 and 3, label 2). The binding site of 1k3c consists of 96 atoms and 41 atoms superimposed on the query with an RMSD of 0.53 Å (Figure 3c), and although the ligands are different, they align very closely. Interestingly this protein contains a P-loop binding motif although its fold classification is different to the query. The method is therefore able to identify binding site structural similarity where overall fold similarity is absent.

Other folds with similarities to the P-loop include flavodoxin-like (e.g. 1obv), ribokinase-like (e.g. 1p3d), Acyl-CoA N-acyltransferases (e.g. 1bo4), carbohydrate phosphatase (e.g. 1fsa), and molybdenum cofactor binding domain (e.g. 1n62). In each case their nucleotide binding sites superimpose on to known P-loop structures. Both the flavodoxin-like protein and the ribokinase-like protein have the distinctive GKT sequence motif which aligns to known P-loop folds and binds the phosphate moiety of their ligands. The acyl-CoA N-acyltransferase has a very similar backbone structure alignment for six residues (RQGIAT) with the P-loop query (1kjy) (ESGKST), and the AMP moieties of their GDP and coenzyme-A ligands align closely; however, only the glycine residue is sequence conserved. The backbone conformation of carbohydrate phosphatase is also highly

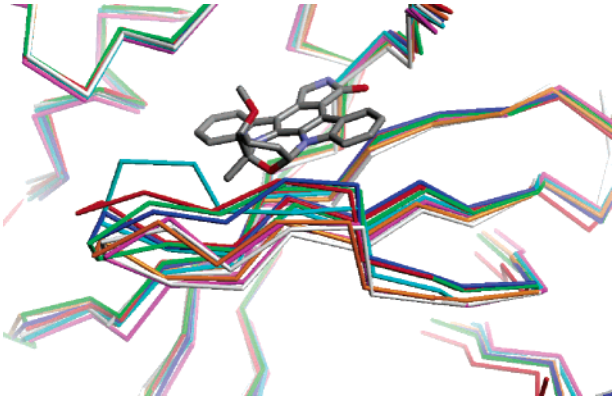

Figure 4. Superposition of twelve binding sites from GSK-3 β proteins.

similar to the P-loop structure of the query (1k5d). Here the motif sequences GTGKTT and GTGEMT align closely. The phosphate moieties of the GNP and AMP ligands are in close proximity when the structures are superimposed. A significant level of structural similarity is also observed between the GDP P-loop binding site of 1fqj and the pterin cytosine dinucleotide (MCN) binding site of 1n62. Here the aligned motifs are KST (preceded by SG) and RST (preceded by GS). Examination of the structures shows that the serine and threonine residues align closely, and the phosphate moieties of the GDP and MCN ligands align well. We have further examined these structures by superimposing their whole protein domains based on the center of mass positions of the 5 approximately aligned residues. This reveals a very similar hairpin turn between the two proteins. We comment that although these proteins are very unlikely to share common ancestry, the structure by which they recognize their nucleotide ligands is highly similar over a relatively short sequential stretch of residues.

These examples demonstrate that common ligand recognition can be predicted by binding site atomic similarity to known proteins. The P-loops of elongation factor Tu and phosphoenolpyruvate carboxykinase, for example, can be superimposed with high score and low RMSD confirming their recognition of similar molecules despite differences in their overall folds and sequences. This site comparison methodology will become increasingly important in the future as more protein structures are solved in structural genomics initiatives without knowledge of protein function and without detectable homology (i.e. sequence or fold similarity) to known proteins. It is thought that binding site similarity can aid protein characterization efforts by predicting possible binding partners or putative functions for these "unknown" proteins.

Receptor-Based Pharmacophore Modeling. Pharmacophore modeling in drug design is often used when a list of active small molecule compounds is known but the 3D structure of a receptor protein is not. However, a recent further innovation is to include information about both bound ligand(s) and the 3D protein receptor site (when this latter information is available) to more fully represent specific ligand−protein interactions. It would appear that using the protein structure combined with a ligand-based approach can result in improved performance in pharmacophore modeling.[29] If several ligands can be identified which bind with reasonable affinity, then common properties of these ligands can be exploited to determine key functional groups or pharmacophores. The pharmacophore model that collectively describes these ligands can then be used in virtual screening against small molecule libraries to discover other compounds with similar features.

Here we aim to aid the derivation of receptor-based pharmacophore models by using protein binding sites similarity to create a structural superposition of protein sites that consequently aligns the bound ligands in 3D space. Figure 4 shows an example, in which twelve glycogen synthase kinase-3β (GSK-3β) proteins complexed with different ligands were superimposed using the whole protein domain (nine with ligand bound and three in the unbound state). A single ligand molecule, staurosporine, is shown in the binding site. There is a high level of structural conservation in much of the kinase domain and in most of the ligand binding site. However, there is considerable structural variability of a binding site flexible loop that adopts different conformations when complexed with different ligands.

If one of the binding sites (1j1b) is used to query SitesBase the same eight sites are retrieved as high scoring matches, and the superposition of their protein binding sites is shown in Figure 5. The structurally static part of the binding site

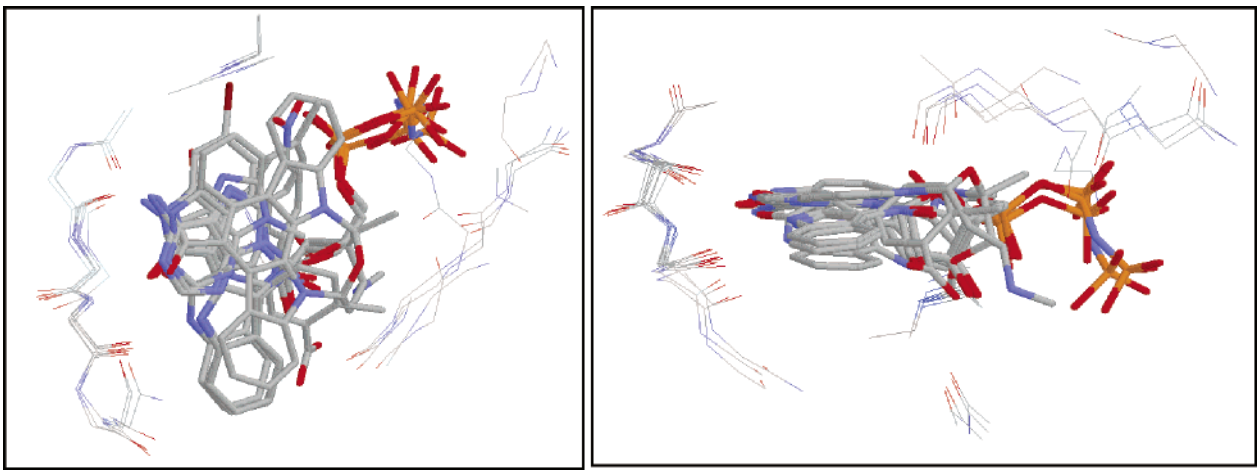

Figure 5. Superposition of eight ligand binding sites from GSK-3 β proteins (1o9u, 1q3d, 1q41, 1uv5, 1j1c, 1pyx, 1q3w and 1q4l) onto the query 1j1b. Shown from above and from the side. The structurally conserved part of the protein binding site can be seen on the left, and the variable binding site flexible loop on the right-hand side of both images.

superposes very closely; however, the structurally variable binding site flexible loop shows considerable variability. The superposition recreates the same binding site orientation of the sites and the ligands seen when using the whole kinase domains (Figure 4). This shows that SitesBase is useful even when there is considerable conformational variability in parts of a binding site since it will always find the maximum common substructure. In addition it provides a useful overlay of the bound ligands in terms of their orientation relative to each other and the protein binding site. This could form the basis for a receptor-based pharmacophore model.

CONCLUSION

SitesBase is a novel database which can be used to identify protein binding site similarity and compare the spatial location of ligands in similar binding sites. Atomic similarity can reflect similarities in the shape of the binding pocket as well as between hydrogen bonding and salt-bridge interaction patterns. This is useful in studies of functional assignment, understanding the potential for ligand cross reactivity, and creating receptor-based pharmacophore models.

Detection of binding site similarity can be achieved for highly diverged proteins where sequence homology is undetectable and even for proteins where there is no overall fold similarity such as between members of the nucleotide binding protein family. Here we show an example where the P-loop structural motif is used to bind both ADP and GDP even without conservation of the overall fold. Predictions of the potential for drug cross reactivity are important in structure-based drug design and in understanding the molecular basis for specificity and toxicity. Last, we show that SitesBase can also be used to produce pharmacophore models of ligands based on the relative position of the protein binding site.

AVAILABILITY

SitesBase is available at http://www.bioinformatics.leeds.ac.uk/sb/.

ACKNOWLEDGMENT

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council. We thank Alex Stark and Robert Russell for help in implementing their method for calculating *seq sim* scores. We thank Peter Oledzki for generating the image in Figure 4.

REFERENCES AND NOTES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−42.

(2) Brakoulias, A.; Jackson, R. M. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56* (2), 250−60.

(3) Stuart, A. C.; Ilyin, V. A.; Sali, A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* **2002**, *18* (1), 200−1.

(4) Shin, J. M.; Cho, D. H. PDB−Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res.* **2005**, *33* (Database issue), D238−41.

(5) Hendlich, M. Databases for protein−ligand complexes. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54* (Pt 6 Pt 1), 1178−82.

(6) Stark, A.; Sunyaev, S.; Russell, R. B. A model for statistical significance of local similarities in structure. *J. Mol. Biol.* **2003**, *326* (5), 1307−16.

(7) Kinoshita, K.; Furui, J.; Nakamura, H. Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics* **2002**, *2* (1), 9−22.

(8) Binkowski, T. A.; Adamian, L.; Liang, J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J. Mol. Biol.* **2003**, *332* (2), 505−26.

(9) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339* (3), 607−33.

(10) Dodson, G.; Wlodawer, A. Catalytic triads and their relatives. *Trends Biochem. Sci.* **1998**, *23* (9), 347−52.

(11) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* **2003**, *13* (3), 389−95.

(12) Laskowski, R. A.; Watson, J. D.; Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W89−93.

(13) Mitchell, J. B. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1617−22.

(14) Nobeli, I.; Spriggs, R. V.; George, R. A.; Thornton, J. M. A ligand-centric analysis of the diversity and evolution of protein−ligand relationships in E.coli. *J. Mol. Biol.* **2005**, *347* (2), 415−36.

(15) WorldDrugIndex http://www.daylight.com/products/databases/WDI.html.

(16) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247* (4), 536−40.

(17) Russell, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **1998**, *279* (5), 1211−27.

(18) Chenna, R.; Sugawara, H.; Koike, T.; Lopez, R.; Gibson, T. J.; Higgins, D. G.; Thompson, J. D. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **2003**, *31* (13), 3497−500.

(19) Carugo, O.; Argos, P. NADP-dependent enzymes. I: Conserved stereochemistry of cofactor binding. *Proteins* **1997**, *28* (1), 10−28.

(20) Carugo, O.; Argos, P. NADP-dependent enzymes. II: Evolution of the mono- and dinucleotide binding domains. *Proteins* **1997**, *28* (1), 29−40.

(21) Denessiouk, K. A.; Lehtonen, J. V.; Korpela, T.; Johnson, M. S. Two "unrelated" families of ATP-dependent enzymes share extensive structural similarities about their cofactor binding sites. *Protein Sci.* **1998**, *7* (5), 1136−46.

(22) Denessiouk, K. A.; Johnson, M. S. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins* **2000**, *38* (3), 310−26.

(23) Denessiouk, K. A.; Rantanen, V. V.; Johnson, M. S. Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins* **2001**, *44* (3), 282−91.

(24) Moodie, S. L.; Thornton, J. M. A study into the effects of protein binding on nucleotide conformation. *Nucleic Acids Res.* **1993**, *21* (6), 1369−80.

(25) Moodie, S. L.; Mitchell, J. B.; Thornton, J. M. Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol.* **1996**, *263* (3), 486-500.

(26) Matte, A.; Goldie, H.; Sweet, R. M.; Delbaere, L. T. Crystal structure of *Escherichia coli* phosphoenolpyruvate carboxykinase: a new structural family with the P-loop nucleoside triphosphate hydrolase fold. *J. Mol. Biol.* **1996**, *256* (1), 126−43.

(27) Orengo, C. A.; Jones, D. T.; Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **1994**, *372* (6507), 631−4.

(28) Walker, J. E.; Saraste, M.; Runswick, M. J.; Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1982**, *1* (8), 945−51.

(29) Steindl, T.; Langer, T. Influenza virus neuraminidase inhibitors: generation and comparison of structure-based and common feature pharmacophore hypotheses and their application in virtual screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1849−56.

CI050359C