Published online 15 November 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D267–D272***
doi:10.1093/nar/gkt1127

FireDB: a compendium of biological and pharmacologically relevant ligands

Paolo Maietta[1,*], Gonzalo Lopez[1], Angel Carro[1], Benjamin J. Pingilley[1], Leticia G. Leon[1], Alfonso Valencia[1,2] and Michael L. Tress[1,*]

[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Madrid, 28029, Spain and [2]Spanish National Bioinformatics Institute (INB-ISCIII)

Received September 13, 2013; Revised October 20, 2013; Accepted October 24, 2013

ABSTRACT

FireDB (http://firedb.bioinfo.cnio.es) is a curated inventory of catalytic and biologically relevant small ligand-binding residues culled from the protein structures in the Protein Data Bank. Here we present the important new additions since the publication of FireDB in 2007. The database now contains an extensive list of manually curated biologically relevant compounds. Biologically relevant compounds are informative because of their role in protein function, but they are only a small fraction of the entire ligand set. For the remaining ligands, the FireDB provides cross-references to the annotations from publicly available biological, chemical and pharmacological compound databases. FireDB now has external references for 95% of contacting small ligands, making FireDB a more complete database and providing the scientific community with easy access to the pharmacological annotations of PDB ligands. In addition to the manual curation of ligands, FireDB also provides insights into the biological relevance of individual binding sites. Here, biological relevance is calculated from the multiple sequence alignments of related binding sites that are generated from all-against-all comparison of each FireDB binding site. The database can be accessed by RESTful web services and is available for download via MySQL.

INTRODUCTION

The growth in protein sequence and structural databases is accelerating thanks to genome sequencing projects (1) and structural genomics initiatives (2). This rapid growth of the primary databases is generating an enormous quantity of potentially interesting data. Secondary databases that can analyse and process this information and present it in a usable form are necessary to allow us to make use of this wealth of new biological data.

Much of the untapped functional information in the main repository for protein 3D structures, the Protein Data Bank [PDB, (3)], can be found at the residue level in the form of the amino acid residues involved in ligand binding and implicated in catalysis. Functional information at the residue-level, such as the amino acid residues implicated in protein–protein interactions and in molecular function, can be of crucial importance in the elucidation of protein function. Pinpointing catalytic residues and ligand-binding sites by computational means provides vital clues for the design of targeted biochemical experiments, and could play a role in drug design and screening.

The PDB database is the largest source of these functionally important residues. FireDB (4), a database of ligand binding and catalytic residues culled from the protein structures deposited in the PDB, was developed specifically to make use of the PDB ligand-binding data.

FireDB is more than a simple repository of PDB residue–ligand contacts, it also attempts to bring some order to the protein–ligand interactions; many ligands in the deposited structures in the PDB do not have any strict biological meaning and FireDB puts a value on the biological importance of each protein–ligand interaction. The separation of biological and non-biological ligands in the PDB is a major issue when defining what a binding site is. This definition is especially difficult for small organic or inorganic molecules and ions that can be biologically important in some cases, while in others may simply be crystallized along with the protein structure as part of the buffer or solvent.

Many ligand databases attempt to divide ligands into biological and non-biologically relevant and different

*To whom correspondence should be addressed. Tel: +34 91 732 80 59; Fax: +34 91 224 69 76; Email: pmaietta@cnio.es
Correspondence may also be addressed to Michael L. Tress. Tel: +34 917 32 80 00; Fax: +34 912 246 976; Email: mtress@cnio.es
Present addresses:
Gonzalo Lopez, Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA.
Leticia G. Leon, Instituto de Tecnologías Biomédicas, Center for Biomedical Research of the Canary Islands, University of La Laguna, Tenerife, Spain.

approaches have been used to deal with this problem, generally based on the nature of the ligand. LigASite (5) uses the size of ligand and characteristic of the binding site (>10 heavy atoms and >70 inter-atomic contacts with protein) in order to filter out uninteresting binding sites. But this strict approach leaves out ions and small molecules that are known to be important for the structure and function of proteins. The most recent version of the database contains annotation for 391 non-redundant data entries and a total of 1194 unique ligands. Binding MOAD (6) uses pre-established criteria (manually curated lists) but does not take metals into account. The latest version contains 21 109 proteins in contact with 10 156 different ligands. The recently introduced BioLip (7) has a more sophisticated composite automated and manual procedure in order to avoid the loss of information. It is updated weekly and the August 2013 version contained 56 763 proteins and 11 185 unique ligands. No other database attempts to annotate biological relevance at the level of individual binding sites.

Computer predictions of functional residues have now become an integral part of the process of protein function determination. Many functional residue prediction methods have been developed in recent years (8–11) and the most effective methods involve some form of homologous transfer of ligand-binding data. FireDB has a companion web server, *firestar* (12,13) that bases its ligand binding and catalytic residue predictions on the binding sites in FireDB.

Here we present the new developments in FireDB. A number of new features have been incorporated into the database extending substantially its coverage and making improvements to the quality of the FireDB ligand annotations and to the usability of FireDB data.

CONTENTS

FireDB brings together ligands crystallized in PDB structures, the residues in contact with those ligands, and the catalytic sites annotated by hand in the Catalytic Site Atlas (14). FireDB also incorporates detailed information on each ligand along with two tools, SQUARE (15) and *firestar*. FireDB clusters are cross-linked with UniProt accession codes (16) EC enzyme numbers (17) via MSD (18) and GO terms from GOA-PDB (19). The general flowchart is available in the on-line documentation.

The ligands in FireDB are extracted from the mmCIF (20) data file. Since FireDB is oriented towards small molecule ligands, interactions with proteins, and DNA and RNA interactions are excluded, as are large ligands such as photosystems where the number of ligand atoms is two thirds or greater than the number of protein atoms. In addition many solvent molecules are filtered out at an early stage. The remaining ligands are tagged as metal or non-metal depending on the nature of the ligand. Ligands are cross-linked with the publicly available chemical databases as detailed below. FireDB defines residues in contact with ligand as those atom–atom distances <0.5 Å plus Van Der Walls radii cut-off.

In order to reduce the redundancy inherent in the PDB, ligands, binding residues and CSA catalytic residues are associated to FireDB master sequences. Master sequences are consensus sequences generated by clustering all PDB chains at 97% sequence identity using CD-HIT (21), and building multiple sequence alignments with MUSCLE (22).

The database schema has been updated in order to integrate the new features in FireDB and to discard information considered not useful. A complete graphical schema of the database structure is now available in the online documentation; full descriptions of tables are also provided.

Collapsing binding sites

Multiple binding sites in the same FireDB cluster are collapsed together into master sequence binding sites (MSS) if they overlap over at least of 60% of the binding residues. Overlapping binding sites are clustered even if the ligands in the sites are different, although overlapping metal and non-metal-binding sites are always collapsed independently. Catalytic residues from the CSA in FireDB clusters are also collapsed into MSS in the same way as the ligand-binding sites. This means that there are three types of MSS, metal, non-metal and catalytic.

The MSS is composed of the residues from all sites that make up the MSS. Residues in the MSS are given an occupancy score, calculated from the frequency with which each residue is in contact with a ligand in each of the separate sites that make up the MSS.

The reduction of multiple sites into a single MSS is a key step in the construction of the database and can provide much information on its own. First of all, the comparison between the constituent binding sites can shed light on ligand flexibility [especially for co-enzymes such as ADP/ATP, (23)] and second it allows comparison of the different residues involved in binding different ligands in the same binding site.

Functionality

Users of FireDB can retrieve the detailed annotations for each MSS via PDB code, UniProt accession code or associated keywords. Detailed ligand information can be retrieved via the mmCIF three-letter code or keywords.

FireDB in numbers

FireDB has grown with the PDB. The first stable version of FireDB (July 2006) had 76 504 protein chains. The latest version (August 2013) has 224 691 chains, with binding site annotations for 141 199, and 16 661 annotated PDB ligands. In total there are 116 514 non-redundant ligand-binding MSS and 11 416 catalytic site MSS. The most recent version of FireDB contains 26 287 master sequences with at least one MSS. A comparison between the September 2006 and August 2013 releases can be seen in Table 1.

FireDB annotates binding sites and ligands for more than a quarter of sequence space. FireDB master sequences covered 6519 out of the 14 381 PfamA families in the

Table 1. The growth of FireDB, a comparison of the September 2006 and August 2013 releases

	FireDB September 2006	FireDB August 2013
Chains	78 300	224 691
Ligand compounds	6 926	16 661
Sites	160 588	488 984
CSA catalytic sites	47 512	95 414
Chains with at least one site	50 909	141 199
Cognate ligand-binding sites	57 125	156 101
Metal ligand-binding sites	34 756	98 226
Master sequences	16 151	42 938
Number of sites per master sequence	9.94	11.39
Master sequences with at least one MSS	9060	26 046

March 2013 release of Pfam and 4246 of these families were annotated with ligands. The functional residue prediction server *firestar* that is attached to FireDB can extend predicted binding sites to a total of over 5400 PfamA families (approximately 40%, data unpublished).

NEW ADDITIONS AND IMPROVEMENTS in FireDB

Biological activity of PDB Ligands

FireDB categorizes all protein–small ligand interactions in the PDB. The PDB contains a diverse range of bound ligands, but many PDB ligands have little biological relevance. The structures deposited in the PDB often contain solvents and non-biological molecules that are part of the crystallization conditions. In addition there are many compounds that would not be found under strictly biological conditions, but that are nonetheless interesting, such as antagonists, inhibitors and other drugs. In order to deal with this range of functions all ligands in the FireDB repository are now classified in terms of their biological relevance and placed in one of three classes: 'COGNATE', 'NON-COGNATE' or 'AMBIGUOUS' based on manual curation and exhaustive literature searches.

COGNATE

The aim of FireDB is to annotate functionally important residues and for this reason the information obtained from natural occurring protein–ligand interactions are given more weight in FireDB. COGNATE compounds are those that are the natural biological ligands of the protein they are in contact with. These ligands will be metal ions, co-factors, substrates and/or products).

A great deal of effort has gone into expanding the available manual annotation of FireDB ligands. Ligands have been annotated as COGNATE by the database curators based on extensive literature searches and based on the role of each ligand in the PDB structure it is crystallized with. The August 2013 release of FireDB has a list of 655 natural biological compounds and as far as we know this is the largest similar list.

AMBIGUOUS

Biologically relevant compounds are informative because of their role in protein function, but they are a small fraction of the entire compound set. And in some cases cognate compounds can also act as non-biological ligands. Those compounds that can be biological ligands, but that are also often found as part of the crystallization conditions are defined as AMBIGUOUS. For example sucrose (Figure 1) is present as a ligand in more 190 entries in the PDB but is often used as buffer in crystallization solutions. So far there are 54 compounds in the list of AMBIGUOUS ligands.

NON-COGNATE

All other compounds. FireDB provides cross-referencing of PDB ligands to publicly available biological, chemical and pharmacological compound databases for the non-cognate ligands in the PDB. Complete searchable lists of all these ligands are available on the website.

Extended annotations for NON COGNATE ligands

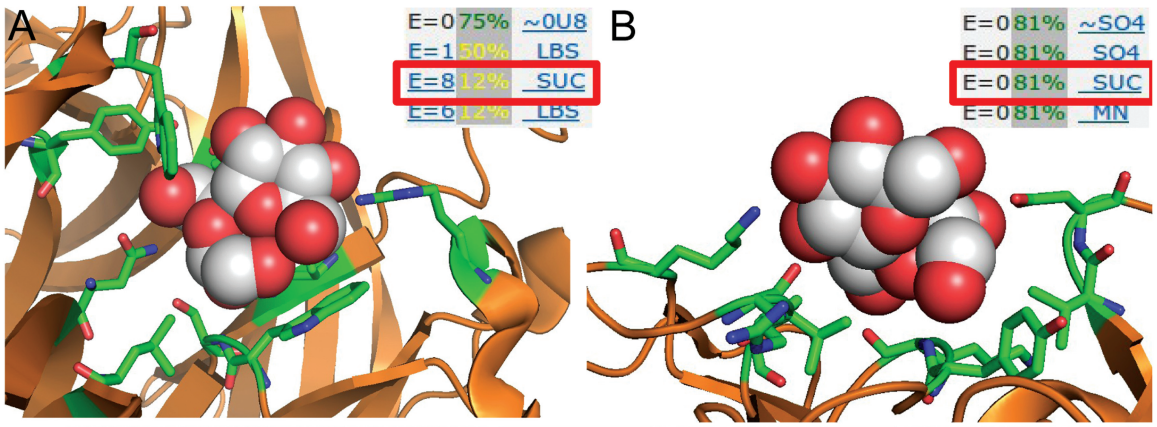
The vast majority of the PDB compounds are classified as NON COGNATE. As of August 2013, there were 14 319 compounds in contact with at least one chain in FireDB; 13 610 of these were annotated as NON-COGNATE compounds. They often have little or no easily accessible annotation.

There are many public chemical and pharmacological databases that store a diverse range of data about chemical compounds. Unfortunately it is often difficult to cross-reference this valuable information with PDB ligands. The RCSB PDB have made efforts to map compounds to some of these databases, but so far only direct associations to DrugBank (24) are shown on the web page and this only covers 33% of the compounds: for other databases a search link is provided, making the collection of information complicated.

We have developed a pipeline for the automatic cross matching of PDB ligands with information gathered from well-known publicly available compound databases. We selected the KEGG COMPOUND database (25) and MetaCyc (26) in order to cross-reference possible new COGNATE compounds. We selected ChEMBL (27), chEBI (28), DrugBank and KEGG DRUG (25) because they are known repositories of molecules with pharmacological activity. And finally we used PubChem (29,30) because it is one of the most complete databases of small molecules information. PubChem also provides biological activity annotation.

The pipeline was able to assign an external reference to approximately 95% of PDB compounds with at least one contact in the FireDB database. Small ligands in FireDB now have external references to PubChem, KEGG COMPOUND, MetaCyc, ChEMBL, chEBI, DrugBank or KEGG DRUG and FireDB also stores annotations of biological activity by cross-linking with PubChem, KEGG DRUG or DrugBank when available.

PubChem is the major contributor, followed by ChEMBL and DrugBank. Additionally we retrieved pharmacological annotations from PubChem, DrugBank and KEGG DRUG, obtaining at least one pharmacological description for 1300 compounds. For those compounds for which we had little or no information, we



C

Cluster Info	% seq Id	Residues	Size	Score		Ligand Compound
t_4ffgA clst	100	N L W S R R D E Y W	10	1.00	Expand	SUC
t_3kf5A clst	22	D Q F S - R D E Y W	13	0.85	Expand	FRU
t_1y4wA clst	25	D Q F S - R D E Y W	12	0.76	Expand	FRU
t_1uypF clst	21	D Q F S - R D E Y W	14	0.76	Expand	SUC
t_2ac1A clst	18	D Q W S - R D E Y -	15	0.57	Expand	FRU SUC
t_1st8A clst	19	D Q W S - R D E Y -	11	0.57	Expand	DQQ FRU DQR SUC
t_3ugfB clst	17	D Q F S - R D E Y -	12	0.56	Expand	DQR
t_3rwkX clst	25	E Q F T - R D E - -	7	0.55	Expand	FRU
t_3pigB clst	22	D Q F S - R D E Y -	14	0.54	Expand	FRU

Figure 1. Biological and non-biological binding of sucrose. (**A**) sucrose-binding site of a bacterial levan fructotransferase (PDB: 4FFH). This site is described in the associated paper as biological and sucrose is the substrate. FireDB annotations for the ligands in the collapsed MSS are shown in the upper right panel. 'E' represents the number of evolutionarily related sites found in FireDB and the percentage shows the occupancy, calculated from the number of chains in the MSS that bind ligands at this site (in this case 2 of 16 chains). (**B**) sucrose-binding site of a bacterial tRNA-splicing ligase (PDB: 4DWR). This site is not biological, and sucrose is cited as part of the crystallization mix in the related paper. The FireDB data for this MSS shows that the occupancy is high (9 of 11 chains) but there are no homologous sites ($E = 0$). (**C**) Clicking on the 'E = 8' icon in (A) will lead the user to the alignment of evolutionarily related sites for the levan fructotransferase-binding site (4FFH). Residues in the binding site alignment are coloured: the darker is the blue, the more conserved the position. Much of the site is conserved even in distant homologous sites and the nature of the bound ligands (Ligand Compounds) suggests that this is a biologically important sugar-binding site.

began a process of manual curation. So far over 300 non-cognate compounds have been manually curated and we have added direct scientific references, and data about activity and target organisms. An example of the new ligand web pages can be seen in Figure 2.

This annotation effort makes FireDB a more complete database, filling a need for the annotation of pharmacological information on PDB ligands, and also offering users the possibility of exploring the collapsed MSS from a pharmacological point of view. This effort is especially important because, apart from the direct information itself, it also allows homology-based function prediction methods such as *firestar* to make predictions for drug-binding sites.

Evolutionarily related sites

At the binding site level FireDB now includes evolutionary analyses of binding site residues. The evolutionary information used in the biological relevance analysis comes from running biological residue prediction server *firestar* in an all-against-all mode for each FireDB master sequence. The *firestar* searches allow FireDB to cluster together related MSS. Multiple alignments of homologous MSS are generated when the detected sites overlap for 40% of the annotated residues. The alignments form the basis of the calculation biological relevance of individual binding sites in FireDB and this information is an important aid in the prediction of functional residues by the *firestar* server.

In Figure 1 we show the information retrieved from FireDB for two sucrose-binding proteins (4FFH and 4DWR). For the biological binding site (from 4FFH) the occupancy is only 12%, but there are evolutionarily related binding sites for the sucrose. Indeed there is a core of well-conserved residues that bind sugars in the evolutionarily related binding sites even when they are remote homologues. The evolutionarily related binding sites confirm the biological role of sucrose in 4FFH.

For non-biological ligand (from 4DWR) the occupancy is 81%. However, there are no evolutionarily related binding sites for the sucrose in 4DWR. Although the sucrose in 4DWR is non-biological the high occupancy of the site in homologous proteins suggests that this site may have some ligand-binding role. It should also be

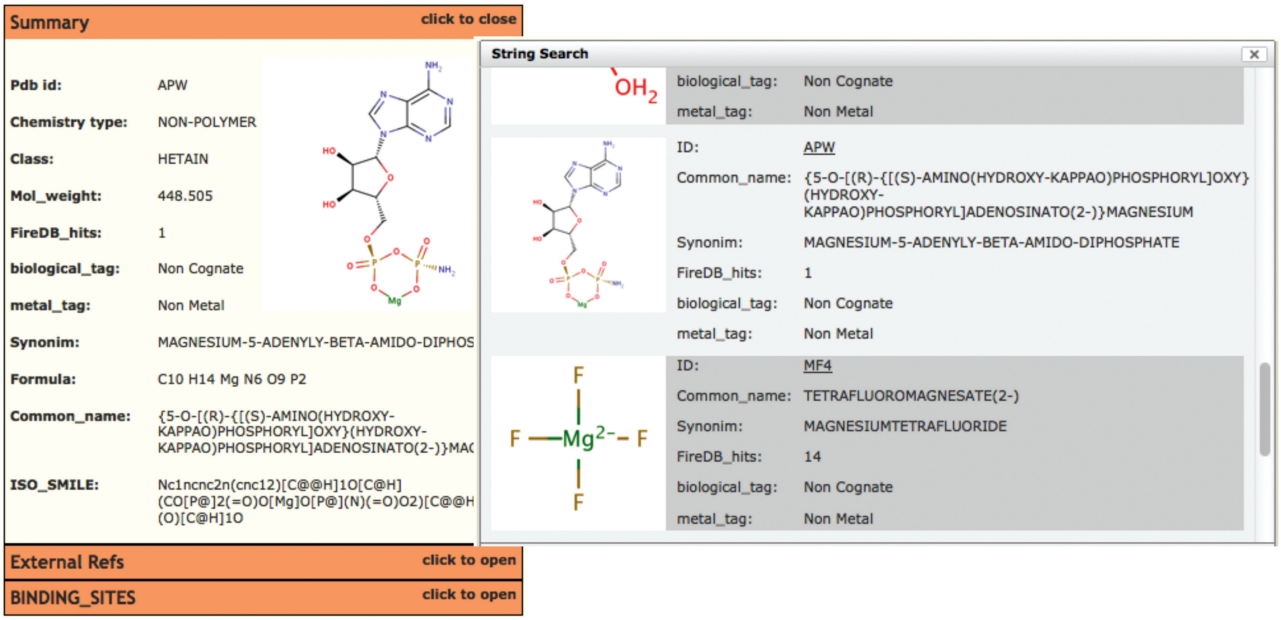


Figure 2. A screen capture from the new ligand information web pages in FireDB. The user can directly query the database using the PDB ligand three-letter code or can search using a keyword. Searches with keywords generate a window with the result of the search (right). Information is shown in pull-down tabs. General information is shown in the summary tab, and an additional three tabs are generated when information is available. The external references tab appears when a match with an external database has been found; a manual references tab is generated if manual annotation has been collected from the literature.

pointed out that due to the heterogeneous distribution of the structures deposited in the PDB, the absence of evolutionarily related binding sites does not automatically imply that a binding site is not biological.

We have automatically evaluated the biological relevance of all MSS in the current release. We used SQUARE to identify conserved residues and motifs in the *firestar*-generated multiple alignments for each MSS. MSS that bound cognate ligands and that had at least a core of conserved ligand-binding residues were considered biologically relevant. Specific amino acid composition filters were used for MSS that bound metal ligands.

Out of 116 514 ligand-binding MSS in the current release, 64 896 have at least one homologue and of these 10 320 non-metal and 6976 metal MSS were tagged as biologically relevant. Beyond this we were able to tag another 1393 MSS as 'novel' biologically relevant sites. These MSS were those that did not have homologous MSS, but where all other features (cognate ligand, residue composition) pointed to their biological activity.

These biologically relevant and novel MSS combined with the catalytic site MSS mean that FireDB contains a total of approximately 30 000 biologically relevant MSS. The entire set of biologically relevant ligand-binding MSS can be downloaded from FireDB. Further information on the decision-making process involved in determining biological relevance can be found on the web pages.

RESTFUL web services

FireDB is freely available via the web. The database is available as a MySQL dump, and we have also developed RESTFUL web services to make the resource easier and faster to access for the scientific community. All the annotations are easily retrievable. An example script is available at http://firedb.bioinfo.cnio.es/rest/FireDB_rest.pl

ROOM FOR FUTURE IMPROVEMENTS

Like all inventories that base annotations on experimental data, FireDB information and quality could be increased by adding other reliable sources of annotated functional residues beyond those in the PDB. We would like to integrate other sources of experimentally annotated functionally important residues where they exist.

At present FireDB only contains protein–small molecule ligand interactions. We are looking into ways of including protein–protein and protein–DNA interactions and also information from post-translational modifications and mutations to extend the coverage of FireDB.

We will continue to add to the manual curation of ligands and in particular extend the literature links where possible.

FUNDING

The Spanish Ministry of Economy and Competitiveness (BIO2012-40205); technical assistance was provided by the Spanish Bioinformatics Institute (INB), a platform of the ISCII; European Union [FP7-REGPOT-2012-CT2012-31637-IMBRAIN to L.G.L.]. Funding for open access charge: Spanish Ministry of Economy and Competitiveness [grant: BIO2012-40205].

Conflict of interest statement. None declared.

REFERENCES

1. Sterk,P., Kulikova,T., Kersey,P. and Apweiler,R. (2007) The EMBL nucleotide sequence and genome reviews databases. *Methods Mol. Biol.*, **406**, 1–21.
2. Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci.USA*, **104**, 3183–3188.
3. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
4. Lopez,G., Valencia,A. and Tress,M. (2007) FireDB–a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.*, **35**, D219–D223.
5. Dessailly,B.H., Lensink,M.F., Orengo,C.A. and Wodak,S.J. (2008) LigASite–a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36**, D667–D673.
6. Benson,M.L., Smith,R.D., Khazanov,N.A., Dimcheff,B., Beaver,J., Dresslar,P., Nerothin,J. and Carlson,H.A. (2005) Binding MOAD – a high quality protein-ligand database. *Nucleic Acids Res.*, **36**, D674–D678.
7. Yang,J., Roy,A. and Zhang,Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
8. Wass,M.N., Kelley,L.A. and Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
9. Fischer,J.D., Mayer,C.E. and Soding,J. (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.
10. Roy,A., Yang,J. and Zhang,Y. (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.*, **40**, W471–W477.
11. Roche,D.B., Buenavista,M.T. and McGuffin,L.J. (2013) The FunFOLD2 server for the prediction of protein-ligand interactions. *Nucleic Acids Res.*, **41**, W303–W307.
12. Lopez,G., Valencia,A. and Tress,M.L. (2007) firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.
13. Lopez,G., Maietta,P., Rodriguez,J.M., Valencia,A. and Tress,M.L. (2011) firestar–advances in the prediction of functionally important residues. *Nucleic Acids Res.*, **39**, W235–W241.
14. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
15. Tress,M.L., Grana,O. and Valencia,A. (2004) SQUARE–determining reliable regions in sequence alignments. *Bioinformatics*, **20**, 974–975.
16. The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
17. IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and Nomenclature Committee of IUBMB (NC-IUBMB). (1999) *Eur. J. Biochem.*, **264**, 607–609.
18. Velankar,S., Alhroub,Y., Best,C., Caboche,S., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Golovin,A., Gore,S.P. *et al.* (2012) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **40**, D445–D452.
19. Ponomarnko,J.V., Bourne,P.E. and Shindyalov,I.N. (2005) Assigning new GO annotations to protein data bank sequences by combining structure and sequence homology. *Proteins*, **58**, 855–865.
20. Bourne,P.E., Berman,H.M., McMahon,B., Watenpaugh,K.D., Westbrook,J. and Fitzgerald,P.M.D. (1997) The macromolecular crystallographic information file (mmCIF) *Meth. Enzymol.*, **277**, 571–590.
21. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Stockwell,G.R. and Thornton,J.M. (2006) Conformational diversity of ligands bound to proteins. *J. Mol. Biol.*, **356**, 928–944.
24. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
25. Nakaya,A., Katayama,T., Itoh,M., Hiranuka,K., Kawashima,S., Moriya,Y., Okuda,S., Tanaka,M., Tokimatsu,T., Yamanishi,Y. *et al.* (2013) KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res.*, **41**, D353–D357.
26. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–D753.
27. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
28. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.
29. Bolton,E,, Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities in Wheeler,R.A. and Spellmeyer,D.C. (eds.) Annual Reports in Computational Chemistry, Elsevier, Vol. 4, pp.217–241.
30. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J., Zhou,Z., Han,L., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.