Published online 29 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D483–D489***
doi:10.1093/nar/gks1258

SIFTS: Structure Integration with Function, Taxonomy and Sequences resource

Sameer Velankar[1,*], José M. Dana[1], Julius Jacobsen[2], Glen van Ginkel[1], Paul J. Gane[2], Jie Luo[2], Thomas J. Oldfield[1], Claire O'Donovan[2], Maria-Jesus Martin[2] and Gerard J. Kleywegt[1,*]

[1]Protein Data Bank in Europe, EMBL-EBI and [2]UniProt, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 17, 2012; Revised October 26, 2012; Accepted November 4, 2012

ABSTRACT

The Structure Integration with Function, Taxonomy and Sequences resource (SIFTS; http://pdbe.org/sifts) is a close collaboration between the Protein Data Bank in Europe (PDBe) and UniProt. The two teams have developed a semi-automated process for maintaining up-to-date cross-reference information to UniProt entries, for all protein chains in the PDB entries present in the UniProt database. This process is carried out for every weekly PDB release and the information is stored in the SIFTS database. The SIFTS process includes cross-references to other biological resources such as Pfam, SCOP, CATH, GO, InterPro and the NCBI taxonomy database. The information is exported in XML format, one file for each PDB entry, and is made available by FTP. Many bioinformatics resources use SIFTS data to obtain cross-references between the PDB and other biological databases so as to provide their users with up-to-date information.

INTRODUCTION

The explosion of biological data in recent decades has stimulated the development of archival resources to store, annotate, distribute and manage those data. The NAR database collection of 2012 (1) listed nearly 1400 databases that either archive data or provide niche annotations. Integrating the knowledge captured in all these data resources will facilitate the knowledge–discovery process in biomedical research. Institutes such as the European Bioinformatics Institute (EMBL-EBI) (2) professionally manage (often in collaboration with similar institutes in other countries) many biomedical databases, including primary data archives such as the European Nucleotide Archive (3), the UniProt Knowledgebase (UniProtKB) (4) and the Protein Data Bank (PDB) (5).

The PDB in Europe (PDBe; http://pdbe.org) (6) is a major resource at the EBI and a founding member of the Worldwide Protein Data Bank (wwPDB; http://wwpdb.org) (5), the international organization that manages the PDB, the single global archive of experimentally determined biomacromolecular structure data. The detailed information in the PDB on protein folds, protein–protein interactions and ligand-binding sites can help elucidate the biological and functional context of the increasing number of sequences with unknown function (7,8). Enriching structural data in the PDB with annotations from other biological resources adds the necessary biological context to the macromolecular structures leading to better use of PDB data. When a new structure is deposited in the PDB, the wwPDB annotation staff add appropriate cross-references to other resources such as PubMed (9), UniProtKB, the NCBI taxonomy database (10), NORINE (11) and EMDB (12,13), to capture the biological, chemical and structural context of the entry. Data held in the external resources may change over time and the cross-references to them are therefore not always immutable. The challenge of keeping the cross-reference information up-to-date is addressed by the 'Structure Integration with Function, Taxonomy and Sequences' (SIFTS) resource, maintained by the UniProt and PDBe teams at the EBI since 2002 (14). The two teams have developed the necessary infrastructure and semi-automated processes for the exchange of data between their databases, thereby dramatically improving the quality of annotation in both resources.

The original SIFTS procedure focused on standardization of taxonomy information in the PDB based on the NCBI taxonomy database, and on adding cross-references to UniProtKB for all the protein sequences in the PDB that are present in the UniProt database. The

*To whom correspondence should be addressed.Tel: +44 1223 494646; Fax: +44 1223 494468; Email: sameer@ebi.ac.uk
Correspondence may also be addressed to Gerard J. Kleywegt. Tel: +44 1223 492 698; Fax: +44 1223 494 468; Email: gerard@ebi.ac.uk

improved cross-references were fed back into the PDB archival files and these consistent data were then made available as part of the first PDB archive remediation (15). The wwPDB annotation procedures were also modified and now use the SIFTS methodology and rules to assign taxonomy and UniProtKB cross-references for newly deposited PDB entries. The wwPDB partners agreed to recognize SIFTS as the authoritative resource tasked with keeping this information up-to-date once PDB entries have been released. In addition, the SIFTS pipeline provides up-to-date cross-references to other biological resources such as IntEnz (16), GO (17), InterPro (18), CATH (19), PubMed (9), SCOP (20) and Pfam (21).

In the past 2 years, the SIFTS pipeline has been improved substantially. In thisarticle, we describe the details of the methods and the pipeline that are used by the PDBe and UniProt teams to manage the SIFTS resource. We also describe how SIFTS data can be accessed and provide a few examples of how they are used to support external bioinformatics resources and allow for the creation of advanced tools to access, integrate, correlate and analyse biomacromolecular structure data.

METHODOLOGY

The SIFTS pipeline has two main components—the semi-automated process that identifies the correct and up-to-date UniProtKB cross-reference for protein chains in the PDB and the automated pipeline that generates residue-level correspondences between proteins in the PDB and the corresponding UniProtKB sequence. The automated process also adds cross-reference information to other biological data resources and keeps this information up-to-date. Figure 1 shows a schematic overview of the SIFTS procedure. The following sections describe details of both processes.

Semi-automated mapping of proteins in PDB entries to UniProtKB entries

When a new structure is deposited into the PDB, the wwPDB annotation process adds cross-references to the UniProtKB and NCBI taxonomy databases to the PDB data file. At present, the annotation software is not identical at all wwPDB partner sites and there are some differences in how the UniProtKB cross-references are assigned. The wwPDB partners are developing a new common deposition and annotation system that will apply all the SIFTS assignment rules to identify the correct UniProtKB cross-reference. Between deposition and release of a PDB entry up to a year may pass, and the cross-reference information may no longer be up-to-date. Therefore, every week prior to the public release of new PDB entries, their protein sequences and taxonomic classifications have to be verified. This task is part of the SIFTS process and results in reassignment of the UniProtKB cross-reference for 10–20% of the PDB entries. The process first checks that the taxonomy identifier of the organism name present in the PDB data file matches the taxonomy identifier (TaxID) assigned in the PDB entry. As there may have been changes in the NCBI taxonomy database after processing of the PDB entry, the organism name (including the strain information) is submitted to the UniProt taxonomy service. This service carries out a simple similarity search of the submitted name, and the TaxID with the greatest similarity to it is used in subsequent processing of the sequence. The taxonomic lineage is then retrieved from the NCBI taxonomy database for the given TaxID up to the level of genus.

The protein sequences of the PDB entries that are about to be released are submitted to the UniProt BLAST service to search against UniProtKB (using the BLOSUM80 matrix). Any matches with >85% sequence identity are then assigned a taxonomy lineage using the same procedure as for the PDB proteins. The additional taxonomy evaluation is carried out because protein structure is more conserved during evolution than protein sequence. Therefore, proteins from different subspecies with a high level of sequence identity will have very similar structures and we can relax the rule for matching the taxonomy identity. The scoring system identifies the correct UniProtKB cross-reference from the list of accessions returned by BLAST and uses the following criteria:

(i) Is there a taxonomy match (exact, species level or none)?
(ii) Is the match to a UniProtKB/TrEMBL (i.e. automatically annotated) or a UniProtKB/Swiss-Prot entry (i.e. manually annotated)?
(iii) Is the match the longest matching sequence?
(iv) Does the match belong to a complete or reference proteome set?
(v) How many other PDB cross-references are linked to that UniProtKB entry?

Each of these criteria has an assigned score according to its importance for identifying the correct UniProtKB accession. The scoring system adds the additional score for each criterion to the '% identity' score obtained from the BLAST results to ensure that the correct UniProtKB accession is identified as a top hit. Hence, the most important consideration is the '% identity' and all accessions with >85% sequence identity are considered to ensure that any engineered mutations, tags or isoforms do not result in missing the correct identification. The process gives the highest score (an additive value of 2, i.e. it adds 2 to the percent identity of the appropriate UniProtKB accession from the BLAST results) if the taxonomy matches exactly. A score of 2 is also given if the UniProtKB entry has 'reviewed' status or if the entry is in Swiss-Prot to ensure that a well annotated UniProtKB entry is selected as a cross-reference where possible. If the match is the longest sequence or if it is from an organism for which a complete proteome is available, the score is incremented by 1 in each case. If the UniProtKB entry is from a reference organism, an additional score of 0.5 is added. This is to ensure that sequences from 'complete proteomes' and especially 'reference proteomes' are annotated ahead of other sequences. For each PDB cross-reference in the UniProtKB entry the score is incremented by 0.1 to ensure that a UniProtKB entry containing cross-references to PDB is selected given all other

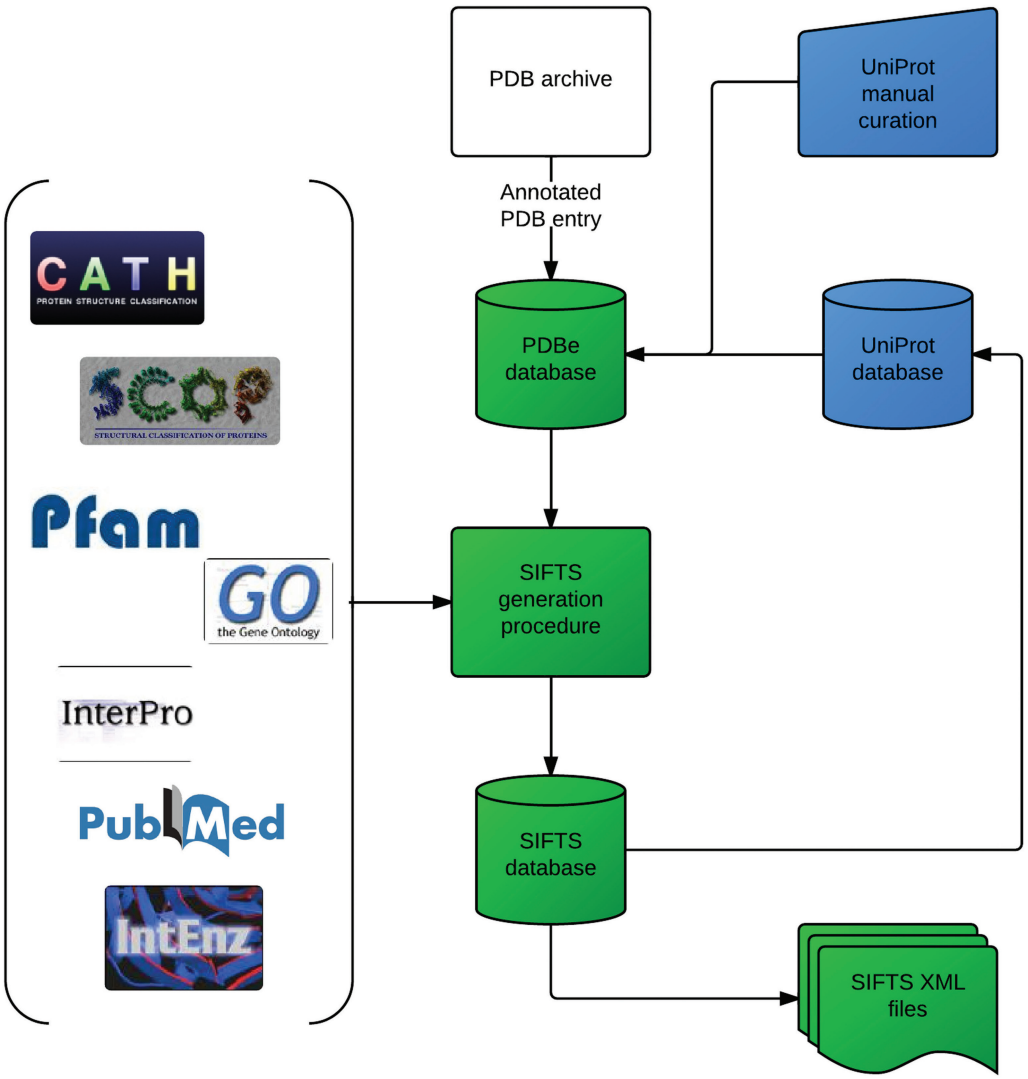

Figure 1. The SIFTS pipeline combines manual and automated processes to produce up-to-date residue-level mappings between proteins in the PDB and their corresponding UniProtKB entry. The pipeline also enriches the annotations of proteins in the PDB by adding data from other biological resources. The SIFTS data are distributed in XML format.

conditions are satisfied. Once these rules have been applied for every UniProtKB accession in the result list, the accession with the highest score is considered the best match.

In summary, the rules that determine the correct cross-reference between a protein in a PDB entry and its corresponding UniProtKB entry are:

(i) They must have a high level of sequence identity (ideally 100% but not below 90%);
(ii) The source organism must be identical or must have a common ancestor within one or two levels up to species level in the taxonomy tree.

The process results in automatic identification of the correct UniProtKB cross-reference for 80–90% of the PDB entries. In a number of cases, entries are inspected manually to make sure that the cross-references are assigned correctly. These entries include:

(i) Short peptides (<7 aa);
(ii) Synthetic constructs;
(iii) *De novo* designed polymers;
(iv) Heavily modified polymers (e.g. antibiotics);
(v) Polymers containing D-amino acids;
(vi) Polymers with unknown sequence ('UNK' used instead of the correct amino-acid residues).

The SIFTS curators also check the expression tags assigned in the PDB entries to make sure these are correct. In addition, sequences of immunoglobulins are not archived in UniProtKB so these entries are marked for manual curation based on the annotations in the UniProtKB entry corresponding to InterPro entry 'IPR013151—Immunoglobulin' which points to a presence of immunoglobulin like sequence domain.

Once the best match has been assigned, the process identifies any discrepancies with the UniProtKB accession number originally assigned by the wwPDB annotation staff. Differences may be due to minor variations in the start or end of the residue range in the UniProtKB entry, assignment of a different UniProtKB accession number or

mismatches in the taxonomy information. Protein sample sequences from the PDB entries without a BLAST hit are marked as such. A UniProt curator examines these special cases individually and makes a decision about whether and how to map the protein.

Sequences of biological origin not contained in the UniProtKB database are flagged for inclusion into UniProtKB. In such cases, a new UniProtKB entry is created from the PDB sequence, taking into account any post-translational modifications, mutations (engineered or otherwise) and expression-tag information available in the PDB entry.

Over the last 2 years, new curation interfaces have been developed to help the annotators and improve the efficiency of the SIFTS pipeline. The new interface makes it possible to see information from both the PDB and UniProtKB entries alongside the BLAST results. The interface also shows the results from the scoring system as an aggregate score and the individual scores to help annotators decide on the correct UniProtKB cross-reference. The resulting chain-level mappings are loaded into the SIFTS database. They are used as starting points to generate the residue-level correspondences between the PDB and UniProtKB sequences for each mapped protein chain in the PDB.

Residue-level mapping and cross-references

The UniProtKB cross-references from each weekly PDB release are added to the SIFTS database. The UniProt curation staff check the UniProtKB cross-reference information in the PDB and any updates from this process (as described in the previous section) are added to the SIFTS database. The process takes into account any engineered or natural variations of the sequence in a PDB entry when compared with the sequence from the corresponding UniProtKB entry, and appropriate annotation is added for all such residues. For wild-type proteins, the entire mapping procedure is quite straightforward, but to identify sequence variants, the automatic procedure uses a sequence identity cut-off of 90%. The procedure also takes into account the fact that many structures in the PDB have regions of unobserved residues in chemically continuous polypeptide chains. Such discontinuities arise when it is impossible to reliably construct a model for regions of structure that are poorly defined by the experimental data, such as flexible loops. These 'gaps' in the sequence are not properly taken into account by standard sequence-alignment algorithms, which therefore often yield incorrect alignments for regions flanking the unobserved residues. To circumvent this problem, connected segments (from N-terminal to C-terminal) of a polypeptide chain from the PDB entry are aligned individually to the sequence from the UniProtKB entry. The separate alignments are then assembled into a complete alignment between the sequence of the observed residues from the PDB entry and the complete sequence of the protein that was used in the experiment. This complex procedure also enables annotation of differences, such as variants, isoforms, modified residues, microheterogeneity or engineered mutations, between the sample sequence and the UniProtKB sequence. Annotation for any unobserved residues and N- or C-terminal tags is added automatically. Regions of the UniProtKB sequence that were not part of the sample sequence are also annotated. Furthermore, for chimeric proteins (engineered proteins where different segments of a single polypeptide are derived from different proteins or different organisms), SIFTS provides accurate cross-reference information.

Once the correct UniProtKB entry (or entries, in the case of a chimera) has been identified, further annotation is obtained from the IntEnz, Pfam and InterPro databases and cross-reference information from the structure family databases CATH and SCOP is integrated whenever new versions of these resources are released (Figure 1). The data from these resources are obtained in various ways, including direct database access and file downloads from FTP archives, and it is still a challenge to keep track of changes and updates to all these resources. The improvements to the SIFTS process have included contacting various resources to improve the data-exchange mechanisms (for instance, by identifying the latest releases on the FTP site in a directory called 'latest'). This has made the SIFTS pipeline more robust with respect to obtaining updates from other data resources. Additionally, we have improved the process of assigning cross-reference information. Until recently, the mapping of GO terms was based on the UniProtKB accession number rather than the sample sequence present in the PDB entry (which may only be a part of the complete protein, for instance the DNA-binding domain of a repressor protein). Together with the InterPro team, an improved procedure has been established. It uses InterProScan (22) on the sample sequence from the PDB and if it finds that the sequence contains <90% of the residues of the corresponding UniProtKB sequence, it identifies only those GO terms that apply to the part of the protein that was present in the sample. Similarly, InterPro assignments are now also based on the actual sample sequence from the PDB. For enzymes, in the old SIFTS process, the Enzyme Classification (EC) numbers were assigned based only on the annotation available from IntEnz, which provides EC cross-references for UniProtKB entries. To address cases where the PDB entries are not represented in UniProtKB or where the depositors of the PDB entry provide the EC

Table 1. Number of PDB entries with cross-reference information in SIFTS to other data resources (as of 24 October 2012)

Total PDB entries processed	85 582
Entries with UniProtKB cross-reference	81 029
Entries with residue-level mapping	83 143
Entries with no possible UniProtKB cross-reference	4336
Entries awaiting mapping	217
Entries with NCBI taxonomy identifier	80 608
Entries with cross-reference to InterPro	79 886
Entries with Pfam family annotation	78 401
Entries with cross-reference to Gene Ontology terms	71 227
Entries with primary citation PubMed identifier	69 417
Entries with assigned CATH identifier	50 110
Entries with SCOP cross-reference	38 054
Entries with assigned EC classification	43 730

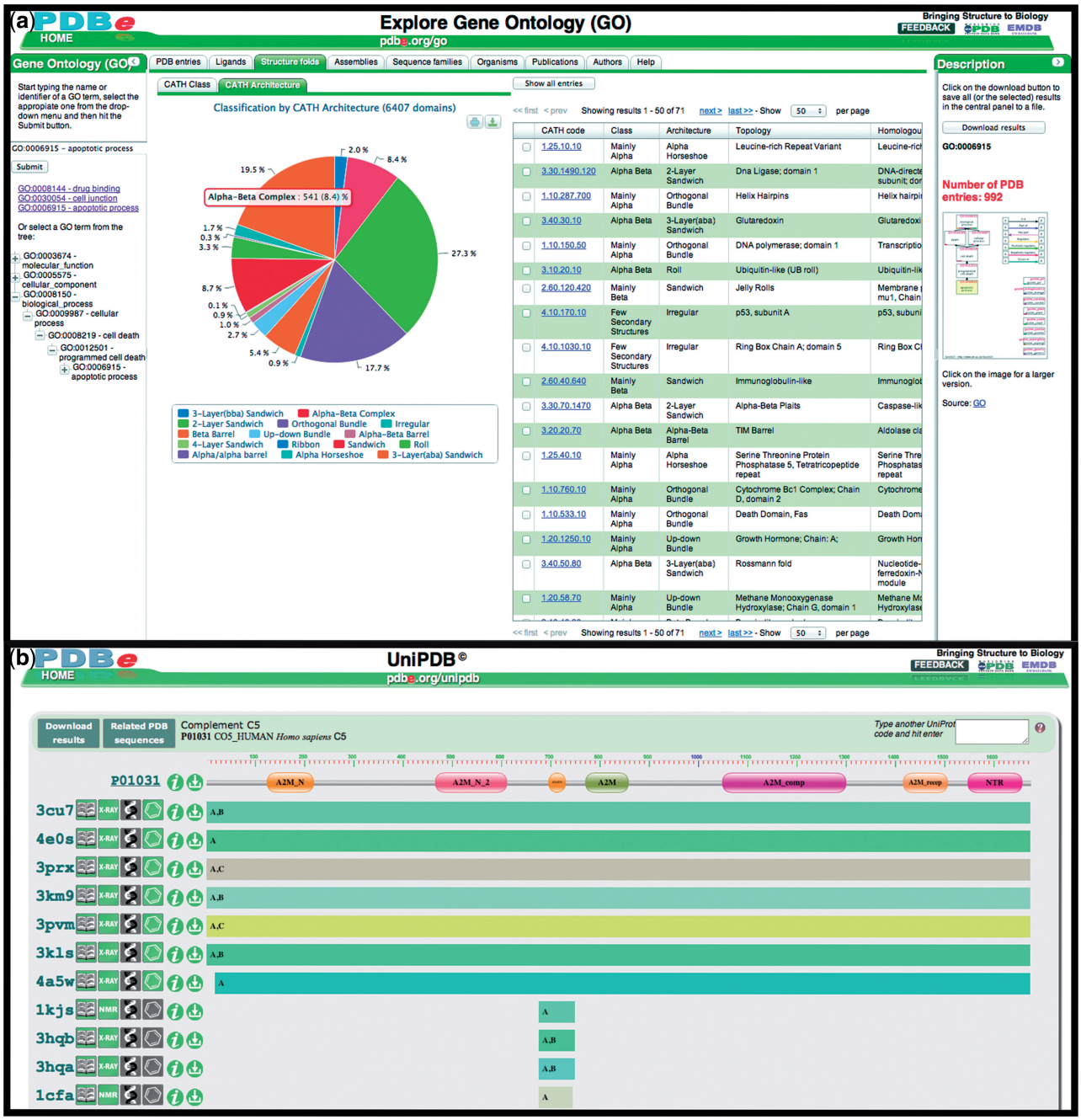

Figure 2. The PDBeXplore [6] and UniPDB [6] tools were made possible by the availability of SIFTS data. (**a**) PDBeXplore (http://pdbe.org/browse) is a browser that enables analysis of the PDB archive based on chemical and biological ontology and classification systems. The figure shows a pie chart of the distribution of 'CATH architecture' data for entries that have been annotated with the selected GO term ('apoptotic process'; GO:0006915). (**b**) UniPDB (http://pdbe.org/unipdb) provides a graphical display of the availability and extent of 3D structural coverage for a given UniProtKB entry in the PDB. The figure shows the number of PDB entries and the extent of coverage for the human complement C5 protein (UniProt accession P01031), making it easy to identify PDB entries containing the structure of the complete protein or a part of it (e.g. PDB entry 1kjs contains the structure of a small part of the sequence that includes the anaphylotoxin-like Pfam domain, PF01821).

information but the IntEnz database does not have an EC assignment for the corresponding UniProtKB entry, the new SIFTS process takes into account any information available in the PDB entry itself.

The SIFTS mapping information is kept up-to-date with each PDB release by monitoring changes to UniProtKB and other data resources using an automated procedure. In cases where the original UniProtKB reference has changed and a new UniProtKB reference cannot be identified automatically, the UniProt curation staff use the semi-automated mapping procedure to update the information manually. The updated cross-references are then used to generate up-to-date SIFTS files. The residue-level mapping is also made available as database

tables to the UniProt team at the EBI where it forms the basis of automatic annotation pipeline for UniProtKB entries with structural data.

Data distribution

The mapping and cross-reference data in SIFTS are produced semi-automatically and curated manually for each weekly PDB release and maintained and distributed by PDBe. The SIFTS data are made available in various formats through the website http://pdbe.org/sifts and the EBI FTP site at ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts. The files are now versioned making it possible to obtain SIFTS information from an old release. We have also added information to the FTP distribution that lists the new and updated files making it easy for users to identify any changes to the SIFTS archive. Residue-level annotations, including secondary structure information and cross-references to other databases are exported in XML format for each PDB entry separately. These files also have some entry-level and chain-level annotations such as the literature citation and taxonomy information. The description of the XML schema is available from the SIFTS website. Data for individual PDB entries can be found at ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/xml/1xyz.xml.gz (where '1xyz' is the PDB identifier) and at ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/splitxml/xy/1xyz.xml.gz (where 'xy' are the second and third characters of the PDB identifier). The XML files also contain residue-level mappings to other resources such as CATH, SCOP, Pfam, InterPro and GO.

The protein-level cross-reference data for the entire PDB archive are also provided as tab-delimited files at http://pdbe.org/sifts/quick.html and are part of the FTP archive at ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/. There is one tab-delimited file for each resource, i.e. UniProtKB, NCBI taxonomy, EC enzyme classification, InterPro, GO, CATH and Pfam. In addition, a file containing PubMed identifiers for primary and secondary literature references from all PDB entries is provided in the same format. The 'mapquick' (http://pdbe.org/mapping) service at PDBe provides a quick access to the SIFTS data for every chain in PDB entries.The SIFTS data are also included in the PDBe search database which can be queried via a web-based user interface using SQL statements (http://pdbe.org/database). Efforts are underway to implement a REST API to make SIFTS data available programmatically. Finally, the SIFTS data are made available through DAS servers at RCSB (http://www.pdb.org/pdb/rest/das/ based on http://biojava.org/wiki/Dazzle) and EBI (http://www.ebi.ac.uk/das-srv/proteindas/das/pdbe_summary). Table 1 shows a summary of SIFTS annotation statistics for the PDB archive as of 24 October 2012.

APPLICATIONS

The up-to-date annotation data in SIFTS make it possible to provide non-expert users with structural information in terms of familiar biological information and classification systems such as genes, proteins, pathways, enzyme nomenclature, sequence-family information (Pfam) and GO annotations. SIFTS therefore, provides critical information that helps transform the PDB from an historic archive into a valuable resource for biomedicine (23).

Based on the information available from SIFTS, PDBe has developed a number of tools and services (Figure 2). For example, PDBeXplore allows browsing and analysis of the PDB archive on the basis of known biological and chemical classification systems such as GO, Pfam, EC and taxonomy (6,24,25). Another tool, UniPDB (6), allows users to assess the coverage of any UniProtKB protein in the PDB using a graphical interface.

SIFTS data are also used by major bioinformatics resources such as UniProt, Pfam, CATH, SCOP, InterPro, RCSB (26), PDBj (27) and DAS-clients such as Spice (28) use these data. A number of resources provided by academic research groups also make direct or indirect use of SIFTS data, including PDBsum (29) and PDBfam (30). PDBfam has developed a process to improve on the Pfam assignments available in SIFTS assignments. RCSB has also developed a process based on the HMMER (31,32) web service. The latter resource takes the PDB-Pfam mappings from SIFTS and adds additional mappings to them. Xu and Dunbrack (30) also analysed the differences between three different approaches to obtain these mappings and discuss them in detail.

ACKNOWLEDGEMENTS

We wish to thank all collaborators and partners in the EBI, EMBL, wwPDB and other collaborative efforts, as well as the structural bioinformatics community for using SIFTS data and providing feedback. We would also like to thank Robert Slowley and Younes Alhroub at PDBe for their help with the SIFTS database. We acknowledge help from Pauline Haslam and Matthew Conroy during manuscript preparation. We are grateful to the InterPro team and David Binns, Craig McAnulla and Phil Jones in particular, for help with the InterPro mapping process.

FUNDING

The Wellcome Trust [088944 to PDBe]; the National Institutes of Health (NIH) [1U41HG006104-03 to UniProt]; the European Molecular Biology Laboratory (EMBL). Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Galperin,M.Y. and Fernández-Suárez,X.M. (2012) The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res.*, **40**, D1–D8.
2. Brooksbank,C., Cameron,G. and Thornton,J. (2010) The European Bioinformatics Institutes data resources. *Nucleic Acids Res.*, **38**, D17–D25.
3. Amid,C., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Gibson,R., Goodgame,N., Hunter,C. *et al.* (2012) Major submissions tool developments at the European Nucleotide Archive. *Nucleic Acids Res.*, **40**, D43–D47.

4. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
5. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
6. 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.
7. Yeats,C., Lees,J., Carter,P., Sillitoe,I. and Orengo,C. (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic Acids Res.*, **39**, W546–W550.
8. de Lima Morais,D.A., Fang,H., Rackham,O.J., Wilson,D., Pethica,R., Chothia,C. and Gough,J. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.*, **39**, D427–D434.
9. Ebbert,J.O., Dupras,D.M. and Erwin,P.J. (2003) Searching the medical literature using PubMed: a tutorial. *Mayo Clin. Proc.*, **78**, 87–91.
10. Federhen,S. (2012) The NCBI Taxonomy Database. *Nucleic Acids Res.*, **40**, D136–D143.
11. Caboche,S., Pupin,M., Leclere,V., Fontaine,A., Jacques,P. and Kucherov,G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326–D331.
12. Tagari,M., Newman,R., Chagoyen,M., Carazo,J. and Henrick,K. (2002) New electron microscopy database and deposition system. *Trends Biochem. Sci.*, **27**, 589.
13. Lawson,C.L., Baker,M.L., Best,C., Bi,C., Dougherty,M., Feng,P., van Ginkel,G., Devkota,B., Lagerstedt,I., Ludtke,S.J. *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res.*, **39**, D456–D464.
14. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
15. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2008) Remediation of the Protein Data Bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
16. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.
17. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Auchincloss,A., Axelsen,K., Blatter,M.-C., Boutet,E. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
18. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
19. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. *et al.* (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
20. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
21. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D30.
22. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) Interproscan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
23. Velankar,S. and Kleywegt,G.J. (2011) The Protein Data Bank in Europe (PDBe): Bringing structure to biology. *Acta Crystallogr.*, **D67**, 324–330.
24. Velankar,S., Best,C., Beuth,B., Boutselakis,C.H., Cobley,N., Sousa Da Silva,A.W., Dimitropoulos,D., Golovin,A., Hirshberg,M., John,M. *et al.* (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **38**, D308–D317.
25. Velankar,S., Alhroub,Y., Alili,A., Best,C., Boutselakis,H.C., Caboche,S., Conroy,M.J., Dana,J.M., van Ginkel,G., Golovin,A. *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **39**, D402–D410.
26. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P. and Berman,H. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.
27. Kinjo,A.R., Suzuki,H., Yamashita,R., Ikegawa,Y., Kudou,T., Igarashi,R., Kengaku,Y., Cho,H., Standley,D.M., Nakagawa,A. *et al.* (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.*, **40**, D453–D460.
28. Prlic,A., Down,T.A. and Hubbard,T.J.P. (2005) Adding some SPICE to DAS. *Bioinformatics*, **21**, 40–41.
29. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.
30. Xu,Q. and Dunbrack,R.L. Jr (2012) Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics*, **28**, 2763–2772.
31. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
32. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comp. Biol.*, **7**, e1002195.