FunPDBe

Progress Report - End of Work Package 1

Mihaly Varadi

October 2018

PDBe - KB

Protein Data Bank in Europe Knowledge Base

Table of Contents

Executive Summary

FunPDBe is one of the flagship projects of the Protein Data Bank in Europe - Knowledge Base (PDBe-KB), a community-driven integrated and accessible resource of structural and functional annotations for macromolecular structure data in the Protein Data Bank (PDB). It is a collaboration between PDBe-KB and world-leading providers of structural bioinformatics data. The project promotes interoperability, comparative analysis and exchange of structural and functional annotations by implementing common data standards and infrastructure to collect these enhanced annotations. The project aims to significantly increase the impact of structural data globally by implementing a central sustainable data resource and a uniform data access mechanism (via FTP and REST API) for distribution of these valuable functional and structural annotations. The data is made accessible programmatically and via web interface by developing reusable web components.

It is a 3 year long project that has been running since October 2017, divided into three main Work Packages (WPs), with an additional, concurrent work package focused on training and dissemination. **This current report covers the first WP.**

Work packages	Focus	Co-PIs
WP1 (Oct 2017-2018)	Predicted functional sites	Christine Orengo
WP2 (Oct 2018-2019)	Known functional sites	Janet Thornton
WP3 (Oct 2019-2020)	Genetic variation	Mike Sternberg

WP1 Deliverables

Work Package 1 can be divided into 3 main deliverables (Figure 1). All three deliverables have been achieved as envisaged during this first year. These are described in detail below:

1.) Designing a data exchange format that captures the commonalities between the annotations provided by various specialist data resources, who may have very diverse types of data.

2.) Designing and implementing a data deposition system that allows the collaborating partners to deposit their annotations in the agreed data exchange format.

3.) Designing and implementing programmatic access to the deposited data, and displaying the data using various visualisation tools.

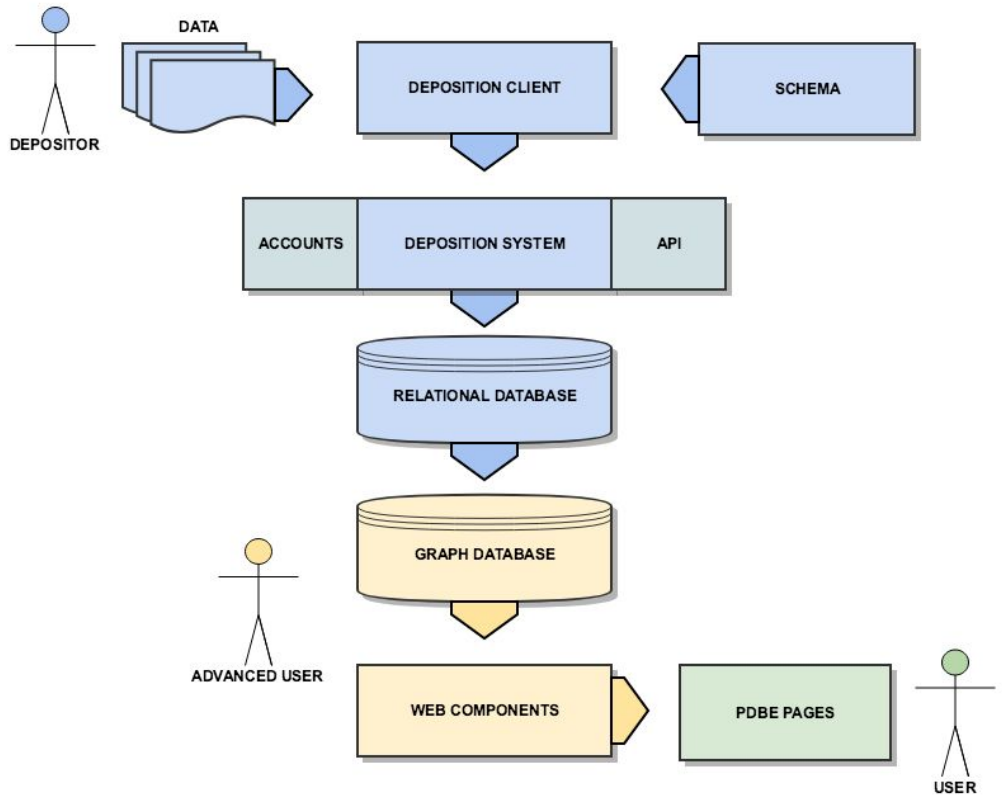

Figure 1 - Overview of the FunPDBe deliverables

The first and second deliverables are the data format and the deposition system (blue boxes), while the third deliverable is exposing the annotations using web components (yellow boxes)

Outcomes

Consortium terms of reference

During the first year of FunPDBe, there were three workshops (October and November 2017 and May 2018) where collaborating partners and other interested parties discussed the goals, progress, next steps of the project and collaboration guidelines. The guidelines (Appendix A) were circulated in August 2018 to project PIs and to all participating teams, and have been made available on the PDBe-KB web page (https://www.ebi.ac.uk/pdbe/pdbe-kb/guidelines).

Data exchange format and schema

The initial recommendation presented in the inaugural workshop in October 2018 was to design a JSON (JavaScript Object Notation) schema that can capture residue-level functional annotations keyed on PDB entries. After a number of iterations, the schema was finalised during the 2nd workshop in Nov 2017. The key data items are the residue-level predictions scores (raw scores), the confidence scores and their classification into simplified levels (null, low, medium, high), and evidence and conclusion ontology (ECO) codes and terms. Importantly, the JSON schema allows for storing URLs linking back to the original data contributor. This allows access to more comprehensive data that cannot be captured by the unified data exchange format, while also giving credit to the collaborating partners for the data they provide.

The JSON schema is available publicly, hosted on EMBL-EBI's GitLab: https://gitlab.ebi.ac.uk/pdbe-kb/funpdbe/funpdbe-schema.

Partner resources agreed on using this format when depositing functional annotations.

Deposition system

In order to allow the collaborating partners to deposit their data, we have designed and implemented a deposition system. This deposition system consists of 3 units:

1. FunPDBe Deposition Client
2. FunPDBe Deposition API
3. FunPDBe Deposition Database

The FunPDBe Deposition Client is a lightweight Python package that can be used to validate data against the FunPDBe JSON schema specifications, and for communicating with the FunPDBe Deposition API, creating, reading, updating and deleting data. The client is hosted on GitLab and is publicly available: https://gitlab.ebi.ac.uk/pdbe-kb/funpdbe/funpdbe-client.

Collaborating partners have agreed to download and use this client for depositing their JSON files.

The FunPDBe Deposition API is providing programmatic access to the deposition database. It listens to requests sent by the FunPDBe client, performs final data checks, and loads the data into the deposition database. It is hosted on virtual machines internal to EBI.

The FunPDBe Deposition Database is a MySQL database that serves as a staging database for the data deposited by the contributing partner resources. Access to this database is provided only via the deposition API.

Accessing and visualising data

Annotations deposited to the FunPDBe Deposition Database are periodically exported, and integrated with annotations from other projects and core PDBe data in a Neo4j Graph Database. Using the graph approach is especially well-suited for this type of highly interconnected data, and allows performing complex queries, effectively rendering the database into a scientific research tool. Each annotation is linked to the corresponding PDB residues, and these residues are also linked to their UniProtKB counterparts, provided by the SIFTS infrastructure. This enables the transfer of structure-based functional annotations onto UniProtKB sequences, allowing queries based on UniProtKB accessions, sequences and residues.

The Neo4j database itself (currently at 250GB) will be made publicly available, allowing the scientific community to perform complex queries, and to integrate this rich data resource with their own data and/or perform extensive data mining.

We are developing an API that allows programmatic access to this database, exposing functional annotations in the context of PDB entries. Current API endpoints answer the following questions, related to the FunPDBe project:

- Which data resources have annotations for a specific PDB entry?
- What are the annotations for every residue within a specific PDB entry from a specific resource?

- What annotations are there for a specific PDB residue?

These API endpoints are available on our development server, with production release scheduled in early 2019: https://wwwdev.ebi.ac.uk/pdbe/graph-api/pdbe_doc/

In order to visualise the functional annotations, our initial goals are to integrate the API endpoints with ProtVista and LiteMol. ProtVista is a sequence feature viewer developed by UniProt (http://ebi-uniprot.github.io/ProtVista/) and used by UniProt, InterPro and PDBe. It is well suited for displaying and comparing residue-level annotation. An example of the FunPDBe implementation is shown below (Figure 2). Integration of FunPDBe data with the reusable and portable ProtVista widget means that the annotations can be plugged in and displayed on any web service in a straightforward manner, facilitating access and providing visibility to the data of the contributing resources.

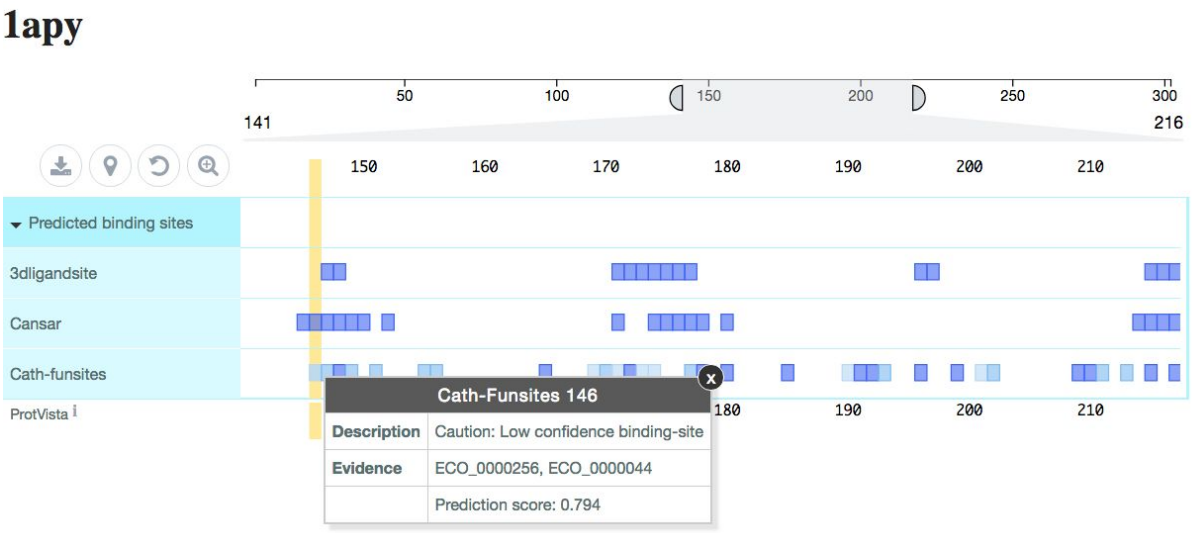

Figure 2 - FunPDBe annotations rendered in ProtVista

ProtVista is a sequence feature viewer developed by UniProt. We have integrated the new API calls with ProtVista so that FunPDBe data can be accessed, consumed and rendered by ProtVista tracks.

LiteMol is a 3D molecular viewer already used on the PDBe pages. We are currently working on establishing direct two-way communication between ProtVista and LiteMol, so that selecting a residue/annotation in ProtVista will highlight that residue on the 3D structure, and vice versa.

Networking

The original group of contributors in Work Package 1 consisted of six labs (see Appendix B). Over the course of the first year, additional groups have expressed an interest in collaborating in the FunPDBe project. To date, four more resources provided residue-level annotations, and there are discussions with six further groups.

Recommendations, next steps

Once the integration of FunPDBe annotations with the 3D molecular viewer, LiteMol, is completed, the data will be available for displaying on PDBe pages, and other web services. Moving forward to Work Package 2 (Oct 2018 - 2019), there are planned changes to the deposition system, enabling higher throughput depositions using FTP (file transfer protocol) and local, parallelised data validation and loading, in addition to the deposition API and client that are already in place.

Appendices

A - Consortium Guidelines
B - Participating Partner Resources and Statistics

Appendix A - Consortium Guidelines

Terms of collaboration

PDBe-KB

- The infrastructure for data deposition and retrieval will be maintained by PDBe-KB
- Data exchange format schema(s) will be maintained by PDBe-KB
- The schema will evolve in consultation with collaborating partners
- PDBe-KB will provide programmatic access to expose contributed annotations
- PDBe-KB will link back to the original collaborating partners resource, attributing credit for their contributions
- PDBe-KB will maintain an open-access library of reusable data visualisation components

Collaborating partners

- The data contributed to PDBe-KB by collaborating partners will be free from any restrictions on distribution and re-use
- The partners are responsible for the quality of the data they contribute
- Protocols for data generation must be published in peer-reviewed publications
- In case of predicted/calculated annotations, the contributing partner makes a commitment of depositing data at least once a year: e.g., to provide annotations for newer PDB entries and/or update the existing annotations when the underlying algorithms change significantly.
    - Manually curated annotations may be exempt from this condition on a case-by-case basis
    - Depositors can change/update/delete their entries at any time

General Data Protection Regulation (GDPR) notice

Collaborating partners have to agree to the PDBe-KB GDPR notice before registering a data deposition account. The privacy notice is available here

Appendix B - Participating Partner Resources and Statistics

Resource name	Owner	Data type	Entries contributed
CATH-FunSites	Christine Orengo	Conserved sites	23,975
canSAR	Bissan al-Lazikani	Druggable pockets	17,804
3DLigandSite	Mark Wass	Binding sites	975
NoD	Geoff Barton	Binding sites	1
14-3-3-Pred	Geoff Barton	Binding sites	1
POPSCOMP	Franca Fraternali	Solvent accessibility	0
CREDO	Tom Blundell	Binding pockets	0
COSPI-Depth	Madhusudhan	Residue depth	141,097
DynaMine	Wim Vranken	Dynamics	139,049
AKID	Manuela Citterich-Helmer	PTM sites	39,763
ProKinO	Natarajan Kannan	PTM sites	3,671