FunPDBe

Progress Report - End of Work Package 2

Mihaly Varadi

October 2018

PDBe - KB

Protein Data Bank in Europe Knowledge Base

Table of Contents

Executive Summary

FunPDBe is a flagship project that contributes data to the Protein Data Bank in Europe - Knowledge Base (PDBe-KB, https://pdbe-kb.org), 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 the PDBe team 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 a web interface by developing reusable web components.

Over the course of the first two years of the project, the consortium members agreed on the collaboration guidelines (https://pdbe-kb.org/guidelines) and established a common data exchange format for functional site annotations and biophysical parameters. By the end of WP2, the contributing resources transferred over 828,000 entries using the deposition infrastructure designed and implemented over the duration of the FunPDBe project, and this

data is being exposed using novel aggregated views keyed on UniProt identifiers in addition to PDB identifiers (https://pdbe-kb.org/proteins).

In addition to the original collaborating partners listed in the grant proposal, several new resources joined the consortium, in part through the ELIXIR 3DBioInfo community, where PDBe-KB is designated as a recommended activity.

FunPDBe project is described in the PDBe-KB publication in the 2020 Database Issue of Nucleic Acids Research (*in press*).

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 second work package.

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

WP2 Deliverables

Work Package 2 builds upon the previous work, focusing on the deliverables described below. The overall workflow is depicted in Figure 1.

1.) Evaluate the suitability of the data exchange format developed in WP1 and extend it if required to accommodate annotations that are evidence-based and curated.

2.) Maintain, improve and further develop the data deposition system to allow the partners from both WP1 and WP2 to deposit their annotations using the agreed data exchange format.

3.) Design and implement additional RESTful API endpoints to expose the new annotations and ensure that the developed visualisation components support the new data.

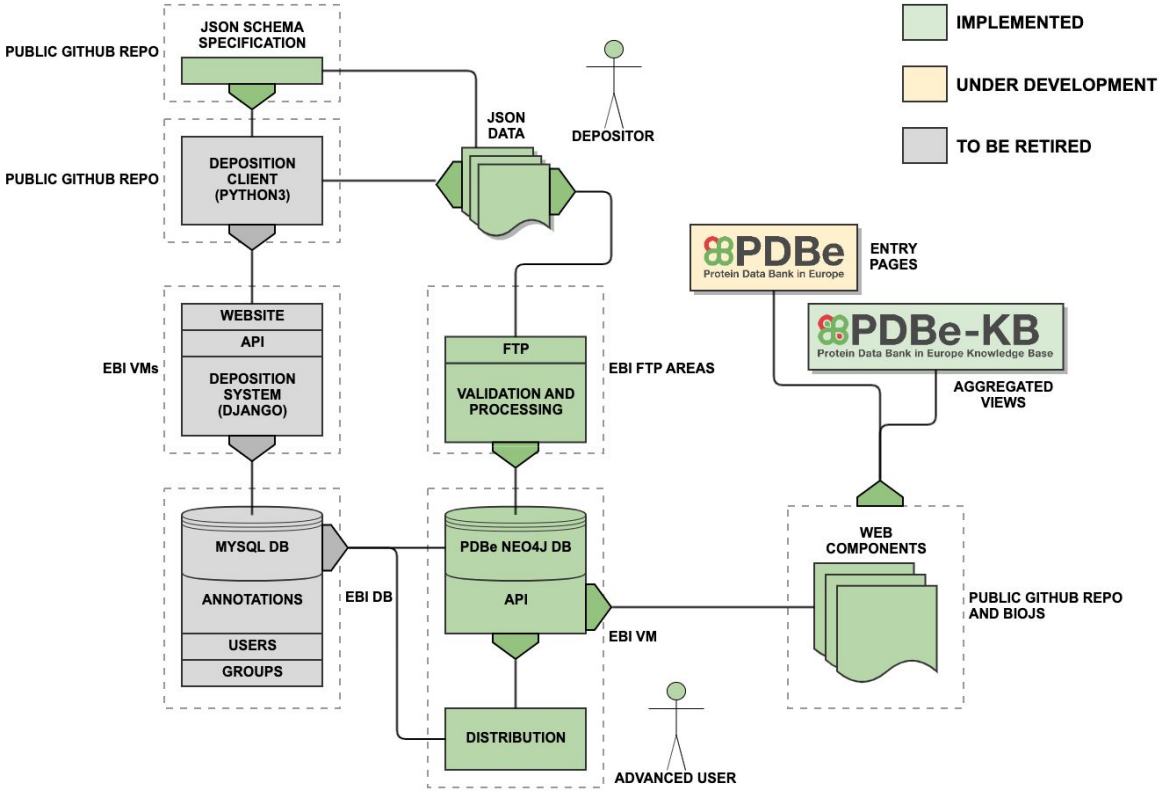

Figure 1 - Overview of the FunPDBe deliverables

The deposition system validates and processes the contributed annotations (left-side of the figure), while the data is exposed from the PDBe Neo4j graph database (middle of the figure) using web components that are reused on PDBe and PDBe-KB pages (right-side of the figure). The deposition API and MySQL deposition database is being retired as the FTP and local validation pipeline is more efficient for high-throughput depositions.

Outcomes

Consortium terms of reference

Over the second year of FunPDBe, the collaboration guidelines were presented both to the Scientific Advisory Board and to new potential PDBe-KB partners during the annual PDBe-KB meeting in June 2019. The guidelines (Appendix A) remained unchanged except for the technical appendix, where the changes to the infrastructure were included, and are available at https://www.ebi.ac.uk/pdbe/pdbe-kb/guidelines.

Data exchange format and schema

During WP1 it was agreed that the data exchange format should be defined as JSON (JavaScript Object Notation) schema that can capture residue-level functional and

biophysical annotations keyed on PDB entries. The schema is available at https://gitlab.ebi.ac.uk/pdbe-kb/funpdbe/funpdbe-schema.

Over the course of the second year of FunPDBe (Oct 2018 - Sep 2019) there have been changes to this schema in order to support the annotations of the new partner resources, while ensuring backward compatibility for existing contributors. The changes are recorded in the “changelog” file of the repository and can be found below, in chronological order:

- 24/10/2018 - "confidence_score" can be greater than 1.0, if the calculation method justifies it
- 15/01/2019
    - "confidence_score" is optional - curated annotations don't have confidence score
    - "raw_score" is optional - curated annotations don't have raw scores
    - "curated" added to "confidence_classification" enumeration list
    - allow additional annotations fields
- 22/03/2019 - "aa_variant" is an optional field in "site_data" for mutations/variants
- 10/09/2019 - "site_url" is an optional field in "sites" for linking directly to site information
- 16/09/2019 - "source_version" is an optional field for the version of data which was used to derive annotations

Deposition system

The PDBe-KB deposition system changed extensively over WP2 of FunPDBe in order to improve its scalability and efficiency. This was necessary as some of the partner resources started to provide residue-level annotations for each residue in the PDB, making the deposition via the deposition API cumbersome, often taking several days to complete the data transfer.

In order to address this issue, the following major changes were carried out:

1) The FunPDBe client is retired except for its “JSON validation” functionality. This functionality was moved to a new “FunPDBe Validator” tool, available at https://gitlab.ebi.ac.uk/pdbe-kb/funpdbe/funpdbe-validator and performs significantly more validation than the original client. In particular, it executes all the data “sanity checks”, ensures compliance with the data exchange format JSON schema, and performs residue-level checking against PDBe data to ensure the residue numbering is correct and that only non-obsoleted entries are processed.

2) The FunPDBe deposition API is also retired, and its data validation and processing functionality was moved to the “FunPDBe Validator” described above.
3) The FunPDBe SQL database will be retired, once all the partner resources move their annotations to the new deposition system that will be described below.

The new deposition system consists of private FTP (File Transfer Protocol) areas specific to each collaborating resource where they can transfer their annotations. The transferred JSON files are then validated and processed locally at EBI rather than performing the checks at the resource sites using the client and via the deposition API. The process can provide log files to the contributing partners with detailed errors descriptions if an entry failed.

The processed JSON files are then converted directly to the CSV format which is expected by the PDBe graph database loader pipeline. Previously, the CSV files were generated by exporting the data from the deposition SQL database, but this new pipeline can go straight from JSON to CSV.

The CSV files are used by the graph database loader pipeline which runs weekly, and regenerates the complete PDBe graph database.

Accessing and visualising data

Annotations deposited to PDBe-KB are provided to the PDBe archive weekly process so these can be integrated with annotations from other PDBe archive 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. In future, this can be extended to proteins that are within the same UniRef90 cluster as the directly mapped UniProtKB sequence, i.e. protein sequences with 90% or higher sequence identities where the structural coverage of the referred UniProtKB sequence is 70% or higher.

The Neo4j database itself (currently at 500GB) was made publicly available by the PDBe team during WP2, 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.

The database is available over FTP at ftp://ftp.ebi.ac.uk/pub/databases/msd/graphdb/ and the underlying data schema is made available at https://www.ebi.ac.uk/pdbe/pdbe-kb/schema)

Over the course of WP2, 70+ programmatic access endpoints were added to the PDBe graph API, exposing functional annotations in the context of PDB entries and UniProtKB proteins. The API and its documentation is available at https://www.ebi.ac.uk/pdbe/graph-api/pdbe_doc/.

The annotations data is visualised using ProtVista and LiteMol on the PDBe pages (Figure 2). This update is available on the test server at https://wwwdev.ebi.ac.uk/pdbe/entry/pdb/1cbs/protein/1 and will be made available by the end of 2019.

Aggregated Views for Proteins

In March 2019 PDBe-KB launched a new type of web pages which are keyed on UniProt accession identifiers instead of PDB entry identifiers and provide an overview of all the available structural information and FunPDBe annotations for a protein of interest. These pages aggregate data using the PDBe graph API described above, and use the web components developed as part of the FunPDBe project. In particular, the sequence feature viewer ProtVista is playing a prominent role in displaying the structural information.

One of the sections of these aggregated views is focused on the residue-level (FunPDBe) annotations provided by the PDBe-KB partners.

The aggregated views for proteins are available at https://www.ebi.ac.uk/pdbe/pdbe-kb/protein using either UniProt or PDB identifiers. For example:
https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/Q14676
https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/2etx

Networking

The original group of contributors in Work Package 2 consisted of 3 labs. Over the course of the first year, additional groups have expressed an interest in collaborating in the FunPDBe project and yet more resources joined during the second year of FunPDBe. As of October

2019, 14 resources provided residue-level annotations, with 4 more preparing to contribute their data (see Appendix B).

Publications

PDBe-KB will be published in the 2020 Database Issue of Nucleic Acids Research (*in press*).

Recommendations, next steps

The final year of FunPDBe (WP3) will focus on mutation and variation data from collaborating partner resources. Initial discussions have already been started with these groups, and the first step will be to ensure the schema of the data exchange format is suitable for supporting their new type of data.
Additionally, more aggregated views are planned for PDBe-KB which will serve as further platforms to display annotations collected as part of the FunPDBe project.

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 at https://www.ebi.ac.uk/data-protection/privacy-notice/funpdbe-deposition-service.

Appendix B - Participating Partner Resources and Statistics

Category	Resource name	Owner	Data type	Entries contributed
WP1 participants	Arpeggio	T. Blundell	Ligand interactions	117,023
	POPSCOMP	F. Fraternali	Solvent accessibility	77,578
	3DLigandSite	M. Wass	Predicted binding sites	901
	CATH-FunSites	C. Orengo	Conserved sites	23,975
	14-3-3-Pred	G. Barton	Predicted binding sites	1,887
	canSAR	B. al-Lazikani	Druggable pockets	17,804
WP2 participants	M-CSA	J. Thornton	Curated catalytic sites	919
	3DComplex	E.D. Levy	Interaction interfaces	106,100
Additional partners	COSPI-Depth	M.S. Madhusudhan	Residue depth	139,509
	AKID	M. Helmer-Citterich	Predicted kinase-targets	41,251
	P2rank	D. Hoksza	Predicted binding sites	138,544
	ChannelsDB	R. Svobodova	Molecular channels	22,305
	CamKinet	M. Kumar, T. Gibson	Curated PTM sites	1,076
	FoldX	L. Serrano	Predicted effects of mutations	6,804
	ProKinO	N. Kannan	Curated PTM sites	5,035
	DynaMine	W. Vranken	Predicted backbone flexibility	130,552