ProdoDB: a sequence-matched protein domain database with REST API service

Byung Ryul Jeon

Department of Laboratory Medicine and Genetics, Soonchunhyang University Bucheon Hospital, Soonchunhyang University College of Medicine
170 Jomaruro, Wonmi-gu, Bucheon, Korea

Abstract *— With ever-increasing amounts of genomic data being generated, most analysis of next generation sequencing data is performed with extensive use of bioinformatics tools. For proper analysis of large experimental datasets, linking to proper resources with adequate unique identifiers (IDs) is critical. However, for protein databases, although numerous genetic and protein databases provide associated unique IDs, due to the polymorphisms and isoforms of proteins used in research, protein sequences can differ among associated databases. As functional domain information is a key element for interpretation of genetic sequence variants, an easily accessible integrated protein domain database is needed. Here we present ProdoDB, a protein domain database providing sequence-matched Swiss-Prot and National Center for Biotechnology Information(NCBI) protein reference sequence unique ID mapping, as well as corresponding sequence information, including gene and domain information, as a REST API service following OpenAPI standards.*

Keywords — *ProdoDB, Protein Sequence Mapping, Protein Database, Protein Domain and Site*.

I. INTRODUCTION

With the increasing number of genomic annotation resources, researchers must make significant efforts to learn the different interfaces and data schemes. For integration of data resources, various strategies such as data warehousing, data federation, or third-party annotation database services are used[1]. Data warehousing uses flat database files to parse and load to local databases. Data federation uses remote data resources directly through provided web services such as representational state transfer (REST) application programming interface (API), or simple object access protocol (SOAP) services[2]. There are also third-party annotation databases such as MyGene.info and MyVariant.info, which are centralized repositories for aggregating and serving dispersed annotation data that offer cloud-based web services[3]. However, aggregating these data sources relies on the unique identifiers (IDs) included in each database, such as the gene IDs from the National Center for Biotechnology Information (NCBI) Gene database that are included in various gene-associated annotation databases. However, widely-used unique IDs do not always have an obvious meaning. For reference sequences, although genome sequence databases provide the unique reference sequence ID and the location of the sequence, the derived sequence from the genome build can differ from the actual sequence in the reference sequence database and this difference can cause confusion during data interpretation[4, 5].

For protein databases, although numerous genetic and protein databases provide associated unique IDs, such as gene IDs, reference sequence IDs, and Swiss-Prot IDs, due to the polymorphisms and isoforms of proteins used in research, protein sequences can differ from those in the associated databases. As functional domain information is a key element of the American College of Medical Genetics and Genomics (ACMG) interpretation guidelines for sequence variations[6], an easily accessible protein domain database is needed.

Here we present prodoDB (http://prododb.schlab.org)- a protein domain database providing sequence-matched unique IDs as well as gene and domain information as a REST API service.

II. METHODS

A. Data integration

ProdoDB makes use of gene information listed in the ClinVar database. MyGene.info was used to retrieve reference sequence information using the ClinVar gene ID[3].

The FASTA file uniprot_sprot_varsplic.fasta.gz was used for Swiss-Prot protein sequences and uniprot_sprot.fasta.gz was used for Swiss-Prot annotation. These files were downloaded from the Universal Protein Resource (UniProt) database (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete)[7].

For ClinVar data integration, clinvar.vcf.gz was downloaded from The NCBI FTP server (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/)[8]. The MyGene.info service (https://mygene.info/) was used for gene-specific information.

For Protein Data Bank (PDB) sequences, the pdb_seqres.txt.gz file was downloaded from the World Wide PDB FTP archive (ftp://ftp.wwpdb.org/pub/pdb/derived_data/)[9].

NCBI protein reference sequence IDs were acquired using the UniProt Retrieve/ID mapping tool (https://www.uniprot.org/uploadlists/) with UniProt IDs and domain information retrieved using the European Bioinformatics Institute (EBI) Proteins API (https://www.ebi.ac.uk/proteins/api/)[10].

MyGene.info, UniProt Retrieve/ID mapping, and the EBI Proteins API were accessed using a Python script.

B. Creating the sequence database

All NCBI protein reference sequences and PDB protein sequences were uploaded to a PostgreSQL database. For gene IDs listed in ClinVar, the Swiss-Prot unique ID and reference sequence were extracted. For each sequence, the NCBI protein reference sequence and PDB IDs of exact matches were acquired using a PostgreSQL query and a Python script.

C. Software architecture

The ProdoDB REST API service uses PostgreSQL (https://www.postgresql.org/) as the database backend and PostgREST(http://postgrest.org) as the API web server. Nginx (https://www.nginx.com/) was used for reverse proxy and the Swagger User Interface (UI) (https://swagger.io/tools/swagger-ui/) was used for visualization and interaction with the API.

III. RESULTS

A. Included data

In total, 50,173 entries for gene information (geneinfo) and 82,777 entries for domains were included in the database. These entries consisted of 27,810 genes and 40,806 unique Swiss-Prot IDs, including isoform IDs (e.g., Q2QGD7, Q2QGD7-2. Q2QGD7-3). When excluding the isoform ID, there were 13,929 unique Swiss-Prot IDs. Among these IDs, 56 (0.4%) had no available NCBI protein reference sequence from the UniProt Retrieve/ID mapping service and 1005 (7.2%) had no exactly matching NCBI protein reference sequence.

B. How to use the ProdoDB REST API server

As ProdoDB uses PostgREST as the REST API server, filtering and fetching of records follows the PostgREST API convention and the Swagger UI is provided for user interaction (Fig. 1). These widely adopted interfaces can minimize the necessity of learning an additional new interface. Full-text searching, and horizontal and vertical filtering are also supported.

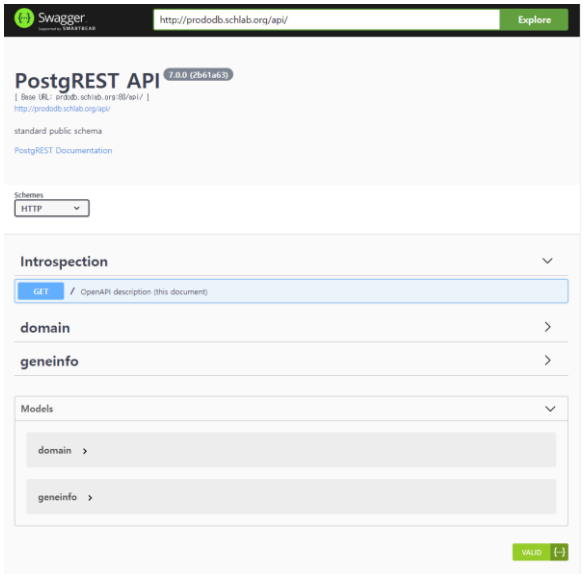

Fig. 1 The ProdoDB home page with the Swagger UI showing the domain and geneinfo REST API model.

C. An example of using the ProdoDB REST API server

An example ProdoDB entry can be found at http://prododb.schlab.org/api/domain?id=eq.7911, showing that the BRCA1 gene ZN_FING domain starts at the 24th amino acid of the Swiss-Prot sequence P38398 and ends at the 65th amino acid. The unique ID P38398 is associated with NCBI reference sequences NP_009225.1, NP_009228.2, NP_009229.2, NP_009230.2, and NP_009231.2 according to the UniProt Retrieve/ID mapping service. However, only NP_009225.1 has the exact same protein sequence (Fig. 2).

Fig. 2 Example of data fetched from ProdoDB showing the BRCA1 ZN_FING domain.

IV. DISCUSSION

Advances in human genomic studies have spurred the development of new methods for interpreting genetic variations. The ACMG guidelines are the de facto standard for interpretation of sequence

variations, and protein functional domain information is one criterion for determining pathogenicity[6]. Specifically, if a sequence variation is in a protein domain that is known to be critical for protein function, and all missense variants identified in that domain have been shown to be pathogenic, then that variant has moderate evidence for pathogenicity.

However, to collect genetic information across many protein databases, using precise database record keys is mandatory. The GeneBank and RefSeq databases are managed by NCBI (https://www.ncbi.nlm.nih.gov/) and use reference sequence numbers[11, 12]. UniProt (www.uniprot.org/) is managed by the UniProt Consortium and provides two database: UniProt Knowledgebase (UniProtKB)/TrEMBL annotations are derived from the European Nucleotide Archive (https://www.ebi.ac.uk/ena), while UniProtKB/Swiss-Prot annotations are manually curated and use Swiss-Prot IDs. Due to the difficulty in distinguishing between orthologues and paralogues in protein superfamilies, Swiss-Prot annotation is considered more reliable than annotation derived from sequence homology[7, 13].

The PDB was established as an archive for biological macromolecular crystal structures and, with the contributions of varying expertise in the techniques of X-ray crystal structure determination, nuclear magnetic resonance, cryoelectron microscopy, and theoretical modeling, it plays a central role in structural genomics and understanding biological functions of proteins. The PDB uses PDB IDs[9].

For interoperation with other databases, either a unique key mapping service, such as the UniProt Retrieve/ID mapping service, or an associated unique key included in the database are used.

As bioinformatics tools advance and increasing amounts of genomic data are generated, most analysis of genomic data is performed with the aid of information technology[14]. For proper analysis of large experimental datasets, linking appropriate resources is critical. Although genetic databases provide unique key elements for other genetic databases, considering the respective status of reference sequences and the multiple isoforms of proteins, the exact sequence relationship between the databases is not always obvious.

In this database, we collected the protein sequences from each database and extracted exact matching sequences. By doing this, ProdoDB can provide unique sequence ID mapping to exact matches. With this information, the exact domain location and sequence information can be used safely.

ProdoDB provides integrated data with REST API and the Swagger UI. For bioinformatics analysis of genomic data, programmatical access to the database is a crucial factor for use in the analysis pipeline. ProdoDB uses PostgREST as a REST based API. PostgREST is a standalone web server that makes a REST API from a PostgreSQL database and focuses on the data-centric create, read, update, and delete operations. It uses PostgreSQL's functionality extensively rather than custom programming and no object-relational mapping is involved. By delegating the functionality to the database with PostgREST, a database administrator can create an API with minimal programming. With these technologies, ProdoDB provides fast, OpenAPI-compatible services.

ACKNOWLEDGMENT

This work was supported by the Soonchunhyang University Research Fund.

REFERENCES

[1] C. Wu, I. Macleod, and A. I. Su, "*BioGPS and MyGene.info: organizing online, gene-centric information*," Nucleic Acids Res, vol. 41, no. Database issue, pp. D561-5, Jan 2013.

[2] Y. M. Park, S. Squizzato, N. Buso, T. Gur, and R. Lopez, *"The EBI search engine: EBI search as a service-making biological data accessible for all*," Nucleic Acids Res, vol. 45, no. W1, pp. W545-W549, Jul 3 2017.

[3] J. Xin et al., "*High-performance web services for querying gene and variant annotation,"* Genome Biol, vol. 17, no. 1, p. 91, May 6 2016.

[4] R. Dalgleish et al., "*Locus Reference Genomic sequences: an improved basis for describing human DNA variants*," Genome Med, vol. 2, no. 4, p. 24, Apr 15 2010.

[5] J. A. MacArthur et al., "*Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants*," Nucleic Acids Res, vol. 42, no. Database issue, pp. D873-8, Jan 2014.

[6] S. Richards et al., "*Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*," Genet Med, vol. 17, no. 5, pp. 405-24, May 2015.

[7] T. UniProt Consortium, "*UniProt: the universal protein knowledgebase*," Nucleic Acids Res, vol. 46, no. 5, p. 2699, Mar 16 2018.

[8] M. J. Landrum et al., "*ClinVar: public archive of relationships among sequence variation and human phenotype,"* Nucleic Acids Res, vol. 42, no. Database issue, pp. D980-5, Jan 2014.

[9] H. M. Berman et al., "*The Protein Data Bank*," Nucleic Acids Res, vol. 28, no. 1, pp. 235-42, Jan 1 2000.

[10] A. Nightingale et al., "*The Proteins API: accessing key integrated protein and genome information*," Nucleic Acids Res, vol. 45, no. W1, pp. W539-W544, Jul 3 2017.

[11] N. A. O'Leary et al., "*Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation*," Nucleic Acids Res, vol. 44, no. D1, pp. D733-45, Jan 4 2016.

[12] K. D. Pruitt, T. Tatusova, and D. R. Maglott, *"NCBI Reference Sequence project: update and current status,"* Nucleic Acids Res, vol. 31, no. 1, pp. 34-7, Jan 1 2003.

[13] S. Pundir, M. Magrane, M. J. Martin, C. O'Donovan, and C. UniProt, *"Searching and Navigating UniProt Databases*," Curr Protoc Bioinformatics, vol. 50, pp. 1 27 1-10, Jun 19 2015.

[14] G. A. Van der Auwera et al., "*From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline*," Curr Protoc Bioinformatics, vol. 43, pp. 11 10 1-11 10 33, 2013.