Research Article

LDBPR: Latest Database of Protein Research

Shahid Ullah2*, Tianshun Gao1#, Wajeeha Rahman2, Farhan Ullah2, Riffat Jahan2, Anees Ullah3, Gulzar Ahmad2, Muhammad Ijaz2, Yihang Pan1*

1Research Center, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, Guangzhou, China
2S Khan Lab Mardan, Khyber Pakhtunkhwa, Pakistan
3Kyrgyz State Medical University, Bishkek, Kyrgyzstan

*Corresponding author: Yihang Pan, Research Center, The Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, Guangzhou, China; Shahid Ullah, S Khan Lab Mardan, Khyber Pakhtunkhwa, Pakistan.

Received: 03 February 2022; Accepted: 15 February 2022; Published: 21 February 2022


Abstract

With the vast and rapid growth of protein research data, a large number of databases are produced to annotate proteins. How to use these databases is becoming a crucial part of modern biology. Database research is usually the first step in the analysis of a new protein. The combined utilization of multiple databases could help researchers to understand the evolution, structure, and function of proteins. Therefore, a well comprehensive and large-scale resource integrated with most of databases is urgently desirable for systematic and precise studies of proteins. Here we designed a platform LDBPR with a collection of 564 latest scientific protein databases. It fully covered physical, chemical, and biological information of Protein sequence, structure, and model, domain, function, and protein-protein interactions. Furthermore, The LDBPR can be explored by three ways: (i) single database can be browsed by typing the name in the given search bar; (ii) all protein categories can be browsed by clicking on the name of the category; (iii) the image icon, could give all categorized protein databases on single click. Moreover, the programming languages including PHP, HTML, CSS, and MySQL were used to construct LDBPR for the protein scientific community that can be freely searched by clicking http://www.habdsk.org/ldbpr.php and will be updated timely.

Keywords: HABDSK; LDBPR; MySQL; Protein; Protein-protein interactions; Sequence
1. Introduction

To deposit the precious protein information for easy retrieving, a handful of databases such as “SCOP” [1], “HAMAP” [2], “Plprof” [3], “AHD” [4], “STRING” [5], and “PRIDE” [6] have been designed. These databases were mainly focused on structure, sequence, model, pathway, Protein-Protein and other Interaction (PP&OI), and expression respectively and provided comprehensive information of the proteins for the protein research community. Meanwhile, there are also a lot of well-known animal and plant databases [7] including “HMDB” [8], “P3DB” [9], “PhytAMP” [10], “Nextprot” [11], “TSTMP” [12], and “dbPAF” [13], which focused on special species. Especially, a number of articles are published based on a collection of a small number of databases simply listed in a table and didn’t construct an online website to display the compensative features for the research object (Table 1). These studies showed low coverage of category, and some of them are specific for special organisms (e.g., mouse, human, or plant). Thus, a well comprehensive and large-scale database is needed for further studies of proteins. Here we integrate a collection of the latest scientific protein data raised from physical, chemical and biological information of Protein sequence, structure, Modal, domain, function, and protein-protein interactions. These data cannot be managed without computational databases [14, 15], which become a crucial part of modern biology. Some widely known protein databases are far from being fully used by the protein scientific community. Therefore, we provided a starting point to explore the potential of all protein databases on the internet by presenting a friendly and easy searching platform. We will also update the protein information with the passage of time.

<table>
<thead>
<tr>
<th>PMID</th>
<th>YEAR</th>
<th>CATEGORY</th>
<th>FORM OF</th>
<th>DB. NO</th>
<th>JOURNAL NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDBPR</td>
<td>2022</td>
<td>Protein</td>
<td>DB+Table</td>
<td>564</td>
<td>Genomic, Proteomic Bioinformatics (GPB)</td>
</tr>
<tr>
<td>25712261</td>
<td>2015</td>
<td>Human</td>
<td>Table</td>
<td>74</td>
<td>Current Protocols in Molecular Biology</td>
</tr>
<tr>
<td>18265344</td>
<td>2012</td>
<td>Protein</td>
<td>Table</td>
<td>121</td>
<td>Nucleic Acids Research (NAR)</td>
</tr>
<tr>
<td>16381921</td>
<td>2006</td>
<td>Pathway</td>
<td>Database</td>
<td>190</td>
<td>Nucleic Acids Research (NAR)</td>
</tr>
<tr>
<td>7764641</td>
<td>1994</td>
<td>DNA+ Protein</td>
<td>Table</td>
<td>50</td>
<td>Current Opinion in Biotechnology</td>
</tr>
<tr>
<td>31906604</td>
<td>2020</td>
<td>Nucleic acid</td>
<td>Table</td>
<td>70</td>
<td>Nucleic Acids Research (NAR)</td>
</tr>
</tbody>
</table>

Table 1: Comparison table of the LDBPR with other published work.

2. Material and Method

2.1. Database construction and content

2.1.1. Construction of LDBPR

We integrated the data from four well-known resources including PubMed, Google, Google Scholar, and Web of Science. Multiple keywords such as “Protein database”, “Protein databases”, “protein database list”, “database of protein”, “databases of protein”, and “list of protein databases” were searched to retrieve published protein related databases with PubMed ids (http://www.ncbi.nlm.nih.gov/pubmed). To circumvent missing data, we have manually collected the latest protein databases from Nucleic Acids Research journal (NAR) (https://academic.oup.com/nar), and journal of Genomics, Proteomics & Bioinformatics (GPB) (https://www.journals.elsevier.com/genomics-proteomics-and-bioinformatics), which are the leading edge research journals on database issue. We only collected...
all available protein databases and removed all broken links. Programming languages such as PHP, MYSQL, HTML, CSS, and JavaScript were used to construct LDBPR. Finally, our database is easy for operation and updating (Figure 1).

2.2. Content of the LDBPR

2.2.1. Proteins databases classification

several projects [16-18] made their own special classifications of protein databases on the base of the function, application, some technical features, or a special organism such as human, mouse [17], or drosophila [19] and so on. According to the classifications in these projects, we classified all the protein databases into 6 categories, which are protein model database, protein structure database, protein sequence database, protein-protein interaction database, protein expression database, and protein pathway database in LDBPR. We have previously provided some databases of different research area like DBHR: database for human research [20] (Figure 2A), Co-19 PDB: About COVID-19 [21] (Figure 2B), and DBPR: database of plant research [7] (Figure 2C).
2.2.2. Protein model databases

Protein model databases provided the protein three-dimensional structure on the base of predication from its amino acids or primary structure [22], which could help discover the most important targets sought by bioinformatics and theoretical chemistry [23]. In addition, the protein model is of great significance in the field of medicine (e.g., drug design), while it is the development of novel enzymes in the field of biotechnology [24]. A lot of well-known databases were built in this field, such as “PMDB” [25], and “MODELLER” [26]. We have collected totally 27 protein model databases.

**Figure 2**: The screenshots of some relevant databases. (A) Database relevant to human research. (B) Covid-19 relevant database. (C) Plant related database.
2.2.3. Protein structure databases
In this classification, the databases contained a large number of experimental determinations for protein structures, and aimed to organize and annotate useful protein information [16] including unit cell dimensions and angles for structures determined by x-ray crystallography, and structure-based drug design that is the deep study about the function of the proteins [27], e.g., “PDB” [28], “PDBTM” [29], “P3DB” [9], etc.

2.2.4. Protein sequence databases
Protein sequence databases were developed for a large collection of mass-spectrometry based proteomics data [30] including protein sequences [31], post-translation modification [32], and sequence alignment [33]. They are not only the simple sequence databases but also provide a rich annotation from other known research results for proteins. As far as we know, different databases (e.g., “Proteome db” [34], “Uniprot” [35], “dbPSP” [36]) annotated sequences of proteins with different levels [16].

2.2.5. Protein–protein interaction (PPI) databases
Protein-Protein Interactions (PPIs) usually involve two or more protein molecules and could be considered as high-specific physical contacts induced by a result of biochemical events including electrostatic forces, hydrogen bonding, and hydrophobic influence occurring in a cell or a living organism [37]. Since PPIs annotate proteins in a large-scale level, much more specialized databases in this classification are designed to provide complete interactomes [38]. Some typical examples like “DIP” [39], “Biogrid” [40], and “STRING-db” [41] have been widely used as reliable references for PPI analysis.

2.2.6. Protein Expression databases
The protein expression databases integrated data of protein expression from microarray and allowed users to search proteins by gene name, splice variant, protein attribute, disease, treatment, or organism part that is a form of metadata manually curated and analyzed through standard analysis pipelines [42]. For example, the database Expression Atlas provides information about protein expression in animal and plant samples of different cell types, organism parts, developmental stages, diseases and other conditions [43].

2.2.7. Protein pathway database
Pathway diagrams are the roadmaps for molecular biology and could illustrate the connections between genes, proteins, and metabolites [44]. A well-illustrated pathway database should provide a biological context to complex molecular processes in an easily understood and highly visual manner. In this regard, the pathway databases we collected provide remarkably useful information for scientists to share, integrate, interpret, and visualize “omics data” and “omics measurements”. In this category, two databases “PathBank” [44] and “KEEG” [45] have been widely used in analysis of biological pathways.

3. Result and Discussion
3.1. Database statistics
In the current work, we have provided almost all protein databases (Table S1) and shown the category-wise, chronological order, and percentile development in LDBPR. Figure 3A showed the percentile of the proteins databases. Figure 3B displayed the chronological order of the category, while Figure 3C presented the category-wise growth distribution of protein database, indicating the tremendous growth and achievement for the protein scientific community. Furthermore, we have deleted all the broken and non-accessible database links and provided a new and updated protein database in the form of a database named LDBPR as well as a table (Table S1).

![Figure 3](image)

**Figure 3:** The statistics data of LDBPR. (A) Distribution of the database category. (B) Chronological order of the LDBPR. (C) Category-wise growth of the DBHR.

### 3.2. Usage of the LDBPR database

The (LDBPR) is developed in an easy and friendly searching way. For easier and faster search, three options are provided for accessing protein databases. First, users can browse LDBPR by clicking on the name of the category (Figure 4A), or image expression (Figure 4B) linking to the category list page (Figure 4C), or a brief overview with the original link. Users can access the database of interest by simply clicking the database name. Furthermore, to
advance specific database search, users can also type the name of the database in the search bar (Figure 4D). Here, we used the “PathBank” database as an example from the Disease Databases to display the search process.

Figure 4: The browse options of the LDBPR. (A) Browsing by clicking the name. (B) Browsing by image expression. (C) Browsing database name in the search bar. (D) A browsing example of the final result.
4. Conclusion
A useful biological database should provide facilities for storing, organizing, and retrieving biological data such as DNR, RNA, carbohydrate, protein, and so on. It should also be easily viewed, managed, and modified. Although hundreds of databases have been constructed in protein research field, and have their own classifications of protein features like sequence, function, structure and pathway, there is still a lack of research community to effectively manage these resources and give a comprehensive annotation for all proteins. Hence, we collected 564 protein related databases and divided them into 6 categories based on protein model, structure, sequence, protein-protein interaction, expression, and pathway. Furthermore, we added a short introduction for nearly each protein database and kept updating for them. Our database can be searched in an easy and friendly way by clicking on category name, image expression, or database name in the given search bar.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
These data will be available under the journal rule and regulation

Competing interests
The authors don’t have any compete of interests.

Funding
This project is supported by National Natural Science Foundation of China (32100434) and Research Start-up Fund of the Seventh Affiliated Hospital of Sun Yat-sen University (ZSQYBRJH0020).

Authors’ contributions
Dr. Shahid Ullah designed and supervised the project with Prof. Yihang Pan's assistance. Dr. Tianshun Gao worked as a co-first author performed data analysis. Farhan Ullah, Wajeeha Rahman, Muhammad Ijaz Gulzar Ahmad, Riffat Jahan and Dr. Anees Ullah contributed to data analysis. Shahid Ullah wrote the manuscript. All authors reviewed the manuscript.

Acknowledgement
To avoid future conflict and plagiarism issue, LDBPR database is uploaded on https://habdsk.org/ldbpr.phpso that we have provided some contents in this article.
**Supplementary Information**

Download the supplementary information from the below link

https://www.fortunejournals.com/supply/JBSB_5004s.pdf

**References**