Protein Databases- Definition, Types, Examples, Uses

August 3, 2023 by Sanju Tamang
Edited By: Sagar Aryal

Protein databases are a type of biological database that are collections of information about proteins.
The information contained in protein databases includes the amino acid sequence, the domain structure, the biological function of the protein, its three-dimensional structure, and its interactions with other proteins.

Figure: Protein Databases. Image Source: Respective database websites.
Several protein databases are publicly available. Based on the type of information stored, protein databases can be classified into several categories. Some of the most common categories of protein databases are as follows:

Table of Contents

Protein Sequence Databases

The protein sequence database contains amino acid sequences of proteins and related information. The amino acid sequence of a protein is important because it determines the protein's three-dimensional structure and function, as well as its identity.

Some of the most popular protein sequence databases are:

PIR

- PIR (Protein Information Resource) is a popular protein sequence database that provides information on functionally annotated protein sequences.
- PIR maintains three databases, the Protein Sequence Database (PSD), the Non-redundant Reference (NREF) sequence database, and the integrated Protein Classification (iProClass) database, which contains annotated protein sequences, classification information, and protein family, function, and structure information.

SWISS-PROT

- SWISS-PROT is a protein sequence database that provides high levels of annotations, including information on the protein's function, domain structure, post-translational modifications, and variants.
- Swiss-Prot is jointly managed by the SIB (Swiss Institute of Bioinformatics) and the EBI (European Bioinformatics Institute).

- The database distinguishes itself from other protein sequence databases by three criteria: (i) annotations, which cover a broad range of information, (ii) minimal redundancy, which ensures that each sequence is represented only once, and (iii) integration with other databases, which enables cross-referencing and retrieval of information from related databases.

TrEMBL

- TrEMBL is a computer-annotated supplement of Swiss-Prot. TrEMBL entries follow the Swiss-Prot format.
- It contains all the translations of EMBL (European Molecular Biology Laboratory) nucleotide sequence entries that have not yet been integrated into Swiss-Prot.

Protein Structure Databases

Protein structure databases are collections of information related to the three-dimensional structure and secondary structure of proteins.

There are several examples of protein structure databases. Some are:

PDB

- PDB (Protein Data Bank) is a worldwide repository of 3D structure data on large molecules such as proteins, nucleic acids, and other biological macromolecules.
- It stores three-dimensional structural models of macromolecules obtained through three frequently used experimental methods: X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and electron microscopy (3DEM).

SCOP

- SCOP (Structural Classification of Proteins) is a protein structure database that organizes proteins based on their secondary structure properties.
- SCOP categorizes proteins into different levels based on their evolutionary relationships and structural similarities.
- Proteins with high sequence identity or similar structure and function are grouped into families, and families with similar structures but low sequence identity are placed into superfamilies.

- Proteins with the same major secondary structures in the same arrangement are placed into the same fold category, and folds are further grouped into five structural classes.

CATH

- CATH is a database that categorizes protein domains into hierarchical levels based on their folding patterns.
- Protein domains are classified into the CATH hierarchy, which consists of four levels of increasing specificity: Class, Architecture, Topology, and Homologous Superfamily. Domains that have similar folding patterns are grouped together at higher levels of the hierarchy.

Protein-Protein Interaction Databases

Protein-protein interaction databases are collections of information on the interactions between proteins. These databases provide valuable information on the relationships between different proteins and their functions in biological systems.

Examples of protein-protein interaction databases include:

BIND

- BIND (Biomolecular Interaction Network Database) is a database that stores detailed descriptions of interactions, molecular complexes, and pathways between various biomolecules, including proteins, nucleic acids, and small molecules.
- The database is designed to be used for data mining and can be used to study networks of interactions and map pathways across different species. The database can also provide information for kinetic simulations.

DIP

- DIP (Database of Interacting Proteins) is a database that contains protein-protein interaction information that has been compiled through both manual curations and computational methods.
- It is useful for understanding protein functions, and their relationships with other proteins. It can also be used to study the properties of networks of interacting proteins, evaluate predictions of protein-protein interactions, and explore the evolution of these interactions.

MINT

- MINT (Molecular Interaction) is a database that stores information on functional interactions between biological molecules such as proteins, RNA, and DNA.
- It also stores information on enzymatic modifications of partner molecules.
- The database primarily focuses on experimentally verified protein-protein interactions and considers both direct and indirect relationships.

Protein Pattern and Profile Databases

Protein pattern and profile databases contain information on motifs found in sequences. Sequence motifs correspond to structural or functional features in proteins. So, the use of protein sequence patterns or profiles is a valuable tool in determining the function of proteins.

InterPro

- InterPro is a database that contains information on protein families, domains, and functional sites.
- It was created by combining several major protein signature databases, including PROSITE, Pfam, PRINTS, ProDom, and SMART into a single comprehensive resource.

PROSITE

- PROSITE is a collection of signatures that identify patterns or profiles in proteins, which can provide information on their biological functions.
- The signatures in the database are linked to annotation documents that provide information on the protein family or domain detected, including its name, function, 3D structure, and references.

Metabolic Pathway Databases

Metabolic pathway databases contain information about enzymes, biochemical reactions, and metabolic pathways.

ENZYME

- ENZYME is a database that stores information on enzyme nomenclature.

- It is used as the nomenclature source for enzyme names and reactions by most metabolic databases as well as by other biomolecular databases.

KEGG

- KEGG (Kyoto Encyclopedia of Genes and Genomes) is a comprehensive database that maps out molecular and cellular pathways involving interactions between genes and molecules.
- It is composed of pathway maps, molecule tables, gene tables, and genome maps, and is used to build functional maps of metabolic and regulatory pathways.

Applications of protein databases

Protein databases have numerous applications. Some of the applications are:

- Protein databases can be used in sequence analysis to identify homologous sequences and predict protein functions based on sequence similarity.
- Protein databases can also be used for predicting protein structure by comparing the amino acid sequence of a protein with known structures in the database.
- Protein databases also include tools to study protein-protein interactions.
- Protein pattern and profile databases can be used for protein family identification by identifying conserved motifs.
- Protein databases such as metabolic pathway databases can be used in drug discovery and disease research by studying the metabolic pathways involved in diseases.

References

1. Apweiler, R., Bairoch, A., & Wu, C. H. (2004). Protein sequence databases. Current Opinion in Chemical Biology, 8(1), 76–80. doi:10.1016/j.cbpa.2003.12.004
2. Bader, G. D., Donaldson, I., Wolting, C., Francis Ouellette, B. F., Pawson, T., & Hogue, W. V. (2001). BIND—The Biomolecular Interaction

Network Database. Nucleic Acids Research, 29(1), 242-245. https://doi.org/10.1093/nar/29.1.242
3. Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research, 28(1), 45-48. https://doi.org/10.1093/nar/28.1.45
4. Conte, L. L., Ailey, B., P. Hubbard, T. J., Brenner, S. E., Murzin, A. G., & Chothia, C. (2000). SCOP: A Structural Classification of Proteins database. Nucleic Acids Research, 28(1), 257-259. https://doi.org/10.1093/nar/28.1.257
5. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Castro, E. D., Langendijk-Genevaux, P. S., Pagni, M., & A. Sigrist, C. J. (2006). The PROSITE database. Nucleic Acids Research, 34(Database issue), D227. https://doi.org/10.1093/nar/gkj063
6. Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27-30. https://doi.org/10.1093/nar/28.1.27
7. Knudsen, M., & Wiuf, C. (2010). The CATH database. Human Genomics, 4(3), 207-212. https://doi.org/10.1186/1479-7364-4-3-207
8. Kwon, M., Cho, S.Y., Paik, Y. (2005). Protein Databases. In: Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-29623-9_3520
9. Wu, C. H., Yeh, S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R. S., Suzek, B. E., Vinayaka, C. R., Zhang, J., & Barker, W. C. (2003). The Protein Information Resource. Nucleic Acids Research, 31(1), 345-347. https://doi.org/10.1093/nar/gkg040