ORIGINAL RESEARCH PAPER
Volume - 13 | Issue - 07 | July - 2024 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

AN OVERVIEW IN BIOLOGICAL DATABASES

Botany

Babita Kumari* Department of Botany, North-Eastern Hill University, Meghalaya-793022, India. *Corresponding Author

ABSTRACT

The rapid progress in biological research has resulted in a vast amount of data, which is necessary for the creation and application of biological databases. The present review offers a thorough exposition of the current state of biological databases, emphasizing their importance in the storage, organization, and facilitation of biological information access. Important databases are addressed, highlighting their special attributes, data kinds, and uses. Two of the most prominent categories are protein databases and nucleotide databases. Significant databases are examined with special emphasis on their features and applications; these databases include those devoted to proteomics, metabolomics, and genomics. Additionally, it examines how biological databases support innovative research in scientific fields. With an emphasis on enhancing user accessibility, data quality, and computational tools for data analysis, future directions for the creation and improvement of biological databases are suggested. With its insights into the present situation and potential futures of biological databases, this overview hopes to be a useful tool for biologists and other researchers.

KEYWORDS

Database, Protein, Nucleotide

INTRODUCTION

Biological databases are pivotal in the field of life sciences especially "bio-informatics", offering repositories that organize and make accessible the vast amounts of data generated by modern biological research (Jones & Pevzner 2004). It has following three components:

i) The development of new algorithms and statistics for assessing the relationship among the large sets of biological data, e.g. DNA sequence data.

ii) Application of these tools for the analysis and interpretation of the various biological data, including nucleotide sequence, etc.

iii) The development of database for efficient storage, access and management of large body information.

These databases are critical for the storage, retrieval, and analysis of biological information, enabling significant advancements in fields of genomics, proteomics, molecular biology, and bioinformatics. This access facilitates data sharing, enhances reproducibility, and accelerates scientific discoveries (Stein 2003). This literature survey provides an overview of two primary categories of biological databases: protein and nucleotide databases, highlighting their significance, key features, and recent developments.

Importance of Biological Databases

Biological databases enable the systematic storage of biological information, which is critical for data retrieval, integration, and analysis (Baxevanis & Bateman 2015). They help researchers in:

a) Store and Manage Data: Biological databases store diverse types of data, including nucleotide sequences, protein sequences, structural data, gene expression profiles, and metabolic pathways.

b) Data Retrieval and Sharing: Researchers can easily access and share data through these databases, promoting collaboration and transparency in scientific research.

c) Annotation and Curation: Databases often provide curated and annotated data, which includes functional information, bibliographic references, and links to other related data.

d) Data Integration and Analysis: It facilitate the integration of different types of data, enabling comprehensive analysis and the generation of new hypotheses.

Types of Biological Databases

Biological databases can be broadly categorized into several types based on the type of data they store (Rayl & Gaasterland 1994):

PROTEIN DATABASES

Protein databases store extensive information about the amino acid sequences of proteins and often include additional data such as protein structures, functions, and interactions. These databases are crucial for understanding the molecular functions of proteins, evolutionary relationships, and their roles in various biological processes.

Main Protein Databases:

UniProt: UniProt (Universal Protein Resource) is a comprehensive resource for protein sequence and functional information. It is a collaborative effort among the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR) (UniProt Consortium 2007).

- Features: UniProtKB: The UniProt Knowledgebase provides high-quality, manually curated information. UniRef: The UniProt Reference Clusters, which combine closely related sequences to reduce redundancy. UniParc: The UniParc, a comprehensive repository of protein sequences (UniProt Consortium. (2019).
- Recent Advances: Recent updates to UniProt include enhanced integration with other databases, improved annotation of protein function, and the incorporation of data from high-throughput proteomics experiments (UniProt Consortium. (2019).
- Protein Data Bank (PDB): PDB is a database of three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. Managed by the Worldwide Protein Data Bank (wwPDB) consortium, PDB is a critical resource for structural biology ("Protein Data Bank 2019; Berman et al 2002).
- Features: PDB entries include detailed atomic coordinates, structural classifications, and information about the biological source of the molecules. It is widely used for understanding protein function, drug design, and biomolecular interactions.
- Recent Advances: The PDB has expanded its data to include cryo-electron microscopy (cryo-EM) structures, reflecting the growing importance of this technique in structural biology. Enhanced visualization tools and better integration with functional annotation databases are also recent improvements (Lawson 2024).

InterPro: InterPro is a resource that integrates diverse protein signature databases into a single comprehensive resource for protein classification (Hunter, et al (2009).

- Features: It uses predictive models to classify proteins into families and to predict the presence of domains and important sites. This integration helps in annotating proteins and understanding their evolutionary relationships.
- Recent Advances: InterPro has recently incorporated more predictive models and has improved its interface for easier

navigation and data retrieval (Ferguson, A. A., & Rossi, H. L. (2024).

NUCLEOTIDE DATABASES

It store sequences of nucleotides (DNA and RNA) and provide a wealth of information about genetic sequences, their functions, and their variations. These databases are foundational for genomics and genetic research.

MAIN NUCLEOTIDE DATABASES

GenBank (Genetic Sequence Databank): GenBank is a comprehensive known public database of nucleotide sequences and supporting bibliographic and biological annotation. It is maintained by the National Centre for Biotechnology Information (NCBI) (Woodsmall & Benson 1993).

- Features: GenBank includes sequences from various sources, including large-scale sequencing projects, individual researchers, and patent sequences. It is closely integrated with other NCBI resources, facilitating extensive cross-referencing. The first line of a GenBank entry is designated LOCUS (1) (Benson et al 1996).

- Recent Advances: GenBank continues to expand with the increasing volume of sequencing data from next-generation sequencing technologies. Enhanced data submission tools and more sophisticated annotation algorithms have been recent improvements. (Sayers et al 2024 and 2021).

EMBL-EBI: The European Nucleotide Archive (ENA) at EMBL-DB provides a comprehensive repository for nucleotide sequence data which is primary repository for genetic sequences. It contains information scanned from literature and submitted directly. Many journals now require sequences which is always 5' to 3', with the sequence numbering starts at one with the 5' base to have been submitted to the EMBL-DB before publication (Rayl & Gaasterland1994).

- Features: ENA offers access to raw sequencing data, assembled sequences, and functional annotation. It supports a wide range of data submission types, from single sequences to complex meta-genomic studies.

- Recent Advances: ENA has improved its data submission workflows and expanded its metadata support to better capture the context of sequencing data, enhancing its utility for researchers (Bocs 2004).

DNA Data Bank of Japan (DDBJ): DDBJ is a nucleotide sequence database that collaborates with GenBank and ENA as part of the International Nucleotide Sequence Database Collaboration (INSDC) (Tateno et al 2002).

- Features: DDBJ provides access to a wide variety of nucleotide sequences and associated annotation. It supports submissions from global researchers and offers tools for sequence analysis.

- Recent Advances: DDBJ has enhanced its bioinformatics tools for better data analysis and visualization, and it continues to integrate new sequencing technologies into its database (Okido et al., 2022).

CATEGORIES/TYPES OF BIOLOGICAL DATABASES:

Bibliographic databases - Literature
Taxonomic databases - Classification
Nucleic acid databases - DNA information
Genomic databases - Gene level information
Protein databases - Protein information
Protein families, domains and functional sites - Classification of proteins and identifying domains
Enzymes/ metabolic pathways - Metabolic pathways

WHY BIOLOGY NEEDS DATABASES?

- Before analyzing data one need to assemble them into central, shareable resources
- Convenient means to handle and share large volumes of biological data
- Can support large-scale analysis efforts
- Makes data access easy and updated
- Links knowledge obtained from various fields of biology, chemistry and medicine.

PROBLEMS FACED WITH DATABASE USAGE ARE:

- Incomplete Information & spread of data over Multiple databases
- Redundant information
- Various errors, sometimes the links are incorrect
- Database standards, nomenclature conventions etc not clearly defined
- Formulating queries is a serious issue

BIOLOGICAL DATABASE RESOURCES

Site	Resources	Sequences
GenBank/DDJB/EMBL	www.ncbi.nlm.nih.gov	Nucleotide sequences
UCSC	www.genome.ucsc.edu	Genome Browser
Ensembl	www.ensembl.org	Human/mouse genome
PubMed	www.ncbi.nlm.nih.gov	Literature references
NR	www.ncbi.nlm.nih.gov	Protein sequences
Swiss-Prot	www.expasy.org	Protein sequences
InterPro	www.ebi.ac.uk	Protein domains
OMIM	www.ncbi.nlm.nih.gov	Genetic diseases
Enzymes	www.expasy.org	Enzymes
PDB	www.rcsb.org/pdb/	Protein structures
KEGG	www.genome.ad.jp	Metabolic pathways

RESULTS

Scientists can better comprehend biological phenomena such as protein structures and genetic diseases by using biological databases, which offer extensive information on proteins, DNA, and metabolic activities. They connect details from different domains, improve data access, and support large-scale analysis. Sequences, 2D gels, 3D structural images, flat files, and relational databases are the different types of databases. A web interface for searching data is present in most databases. These databases are helpful for researching species interactions, creating medications that can save lives, and fighting diseases.

CONCLUSIONS

Protein and nucleotide databases are fundamental for biological research, providing essential data for understanding molecular mechanisms, evolutionary relationships, and genetic diversity. Recent advances in these databases, driven by technological progress and increased data generation, have significantly enhanced their utility. These resources continue to evolve, incorporating more sophisticated analytical tools and integrating diverse types of biological data to support the scientific community's needs.

Acknowledgement
Funding: No funding was received.
Conflict of Interest: No.

Author's contribution: Conceptualization, original manuscript writing and final revision: Babita Kumari.

REFERENCES

1. Baxevanis, A. D., & Bateman, A. (2015). The importance of biological databases in biological discovery. Current protocols in bioinformatics, 50(1), 1-1.
2. Benson, D. A., Boguski, M., Lipman, D. J., & Ostell, J. (1996). GenBank. Nucleic Acids Research, 24(1), 1-5.
3. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., ... & Zardecki, C. (2002). The protein data bank. Acta Crystallographica Section D: Biological Crystallography, 58(6), 899-907.
4. Bocs, S. (2004). (Ré) annotation de génomes procaryotes complets-Exploration de groupes de gènes chez les bactéries (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).
5. Ferguson, A. A., & Rossi, H. L. (2024). The secretome of adult murine hookworms is shaped by host expression of STAT6.
6. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., ... & Yeats, C. (2009). InterPro: the integrative protein signature database. Nucleic acids research, 37(suppl_1), D211-D215.
7. Jones, N. C., & Pevzner, P. A. (2004). An introduction to bioinformatics algorithms. MIT press.
8. Lawson, C. L., Berman, H. M., Chen, L., Vallat, B., & Zirbel, C. L. (2024). The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids.

Nucleic Acids Research, 52(D1), D245-D254.
9. Okido, T., Kodama, Y., Mashima, J., Kosuge, T., Fujisawa, T., & Ogasawara, O. (2022). DNA Data Bank of Japan (DDBJ) update report 2021. Nucleic acids research, 50(D1), D102-D105.
10. Protein Data Bank: the single global archive for 3D macromolecular structure data." Nucleic acids research 47, no. D1 (2019): D520-D528.
11. Rayl, K. D., & Gaasterland, T. (1994). Overview of selected molecular biological databases (No. ANL/MCS-TM-200). Argonne National Lab.(ANL), Argonne, IL (United States).
12. Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2021). GenBank. Nucleic acids research, 49(D1), D92.
13. Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., & Karsch-Mizrachi, I. (2024). GenBank 2024 update. Nucleic Acids Research, 52(D1), D134-D137.
14. Stein, L. D. (2003). Integrating biological databases. Nature Reviews Genetics, 4(5), 337-345.
15. Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., & Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic acids research, 30(1), 27-30.
16. UniProt Consortium. (2007). The universal protein resource (UniProt). Nucleic acids research, 36(suppl_1), D190-D195.
17. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic acids research, 47(D1), D506-D515.
18. Woodsmall, R. M., & Benson, D. A. (1993). Information resources at the National Center for Biotechnology Information. Bulletin of the Medical Library Association, 81(3), 282.