Journal of **Concepts in Structural Biology & Bioinformatics** (JSBB)

BIO-DATABASES ARTICLES

Concept article: Minireviews

Databases in Biology

RACHEDI Abdelkrim

Laboratory of Biotoxicology, Pharmacognosy and biological valorisation of plants, Faculty of Sciences, Department of Biology, University Dr Tahar Moulay Saida, 20100 Saida, Algeria.

Correspondence: Abdelkrim RACHEDI – E.mail: abdelkrim.rachedi@univ-saida.dz

Abstract

The emergence of high-throughput technologies that generate large amounts of data and database have become essential platforms and tools in the field of biology implemented to store, retrieve, and analyze the data. Databases are used extensively in genomics, proteomics, medicine, novel drug design and virtually in all biology related field, allowing researchers to organize, search, and analyze the vast amounts of information generated by these fields.

In this paper, we will discuss the concept of databases,their importance in biology, with a focus on their applications in a number of fields including bioinformatics, genomics, proteomics, medicine, drug design. We will discuss the methods used to create and curate these databases, and highlight some of the most widely used databases and search engines in the field. Consideration is given to some of the challenges and limitations associated with these tools.

Key words

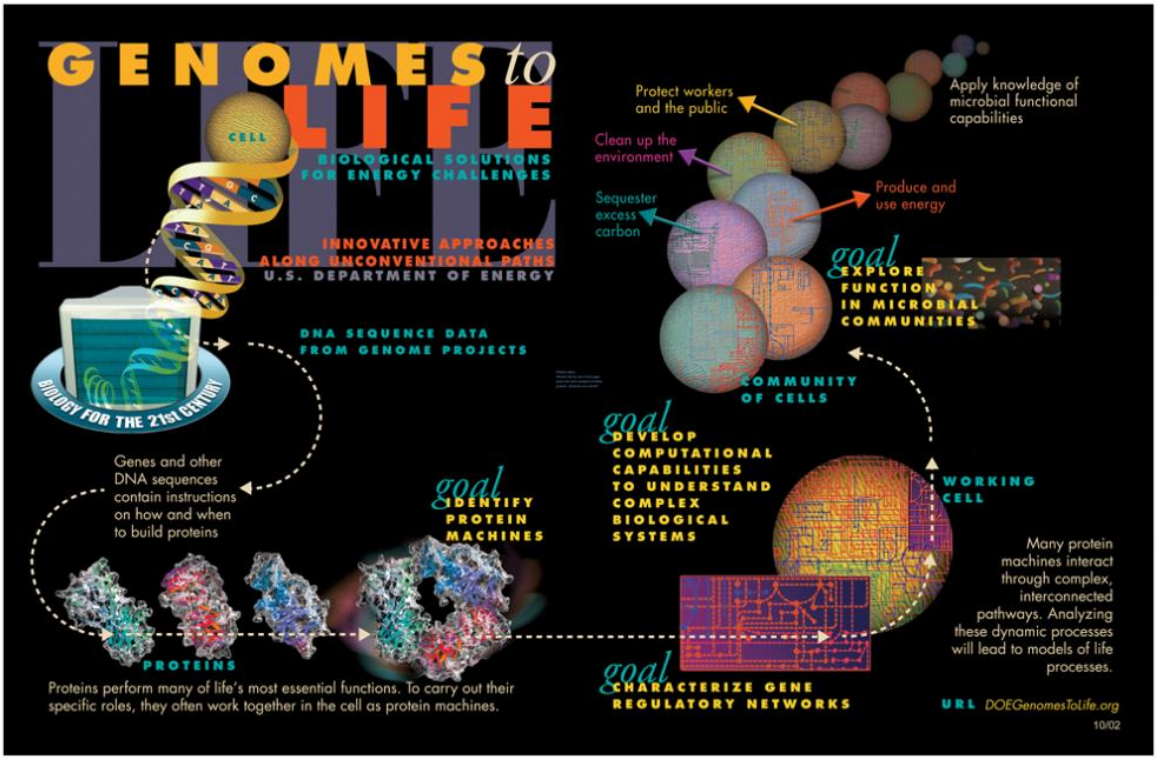
Databases, Bioinformatics, Biology, Genomes, Proteomes, Sequence data, Structural data

Introduction

A database is a collection of data that is organized in a specific way to enable efficient retrieval and manipulation of the data. In biology, databases are essential tools for storing and managing the vast amounts of data generated by high-throughput technologies. The development of these tools has been crucial for advancing our understanding of biological processes, diseases, and drug discovery.

Researchers need to be able to access and analyze this data to maske meaningful discoveries, but this can be a daunting task without the right tools. Databases and search engines have become essential tools for biologists, providing a way to organize and retrieve biological information from a wide range of sources. These tools allow researchers to discover new relationships between genes, proteins, and other biological molecules, and to develop new hypotheses about biological processes.

The study of genomics and proteomics have been revolutionized by the explosion of biological data generated in recent years. The massive amounts of data produced by high-throughput technologies have resulted in the development of a range of databases and search engines that allow researchers to store, retrieve and analyze biological information. These tools have become crucial for understanding the molecular basis of diseases and the development of new drugs.



Databases in Biology - Systems Biology

Website from the U.S. Department of Energy Genome Programs. http://genomics.energy.gov

Methods

Databases are typically created using a combination of manual curation and automated data mining techniques. Data are collected from a variety of sources, including published literature, experimental data, and publicly available databases. The data are then curated to ensure its accuracy and completeness, and are organized in a way that makes it easy to search and retrieve. Search engines, on the other hand, use algorithms to search for and retrieve data from one or more databases. These algorithms are designed to identify relevant information based on user queries, and to rank the results in order of relevance.

Database concepts and methodologies have been implemented in a biology under what's known the field of Bioinformatics and Computational Biology. In the following, some highlights are given about a number of fields of biology where database technology play major roles in their development and advancement.

Genomics:

Genomics is the study of an organism's entire genome, including its genes, regulatory sequences, and non-coding DNA. The field of genomics has been transformed by the development of high-throughput sequencing technologies, which have resulted in the generation of massive amounts of genomic data. Databases such as GenBank, the Sequence Read Archive (SRA), and the Genome Reference Consortium (GRC) have been developed to store and manage this data, making it accessible to researchers worldwide. These databases are essential for the annotation, analysis, and comparison of genomes across different organisms, providing valuable insights into evolution, genetic diversity, and disease.

Proteomics:

Proteomics is the study of an organism's entire complement of proteins, including their structures, functions, and interactions. Like genomics, the field of proteomics has been transformed by the development of high-throughput technologies such as mass spectrometry and protein microarrays, which generate large amounts of protein data. Databases such as UniProt, ProteomeXchange, and the Protein Data Bank (PDB) have been developed to store and manage this data, making it accessible to researchers worldwide. These databases are essential for the identification, characterization, and comparison of proteins across different organisms, providing valuable insights into protein function, structure, and disease.

Medicine:

Databases have become essential tools in the field of medicine, providing valuable resources for disease diagnosis, treatment, and drug discovery. The Human Gene Mutation Database (HGMD) is a comprehensive database of genetic mutations associated with human disease, providing a valuable resource for the identification of disease-causing mutations. The DrugBank database is a comprehensive resource for information on drugs and drug targets, providing valuable information for drug discovery and development.

Novel drug design:

Databases are also essential tools for novel drug design, providing valuable resources for drug target identification, lead compound selection, and optimization. The Protein Data Bank (PDB) provides valuable information on protein structures, enabling the identification of potential drug targets. The ChEMBL database is a comprehensive resource for bioactive molecules and their targets, providing valuable information for lead compound selection and optimization.

Discussion

There are many different types of databases and search engines used in biology, each with its own strengths and limitations. Some of the most widely used databases include GenBank, UniProt, and the Protein Data Bank. GenBank is a database of genetic sequences, and is used extensively in genetics and genomics research. UniProt is a database of protein sequences and functions, and is used in many areas of biology, including molecular biology, biochemistry, and drug discovery. The Protein Data Bank is a database of protein structures, and is used to study the relationship between protein structure and function.

One example of a database used extensively in genomics research is the GenBank database, which is maintained by the National Center for Biotechnology Information (NCBI) and contains over 400 million sequences. This database is used to store and retrieve nucleotide sequences, including those from whole genome sequencing projects. The database is widely used by researchers for sequence alignment, gene annotation, and phylogenetic analysis.

In proteomics, the UniProt database is a comprehensive resource for protein sequence and functional information. It contains over 200 million protein sequences and is a valuable resource for protein identification, functional analysis, and pathway mapping. UniProt also provides access to manually curated

information, including information on protein interactions, post-translational modifications, and disease associations.

Databases and search engines are also used extensively in the field of drug discovery. For example, the Protein Data Bank (PDB) provides structural information for proteins and nucleic acids, which is useful for drug design and virtual screening. The Human Gene Mutation Database (HGMD) is a valuable resource for identifying genetic variations associated with diseases, and it is used to inform drug development strategies.

In addition to these well-established databases, there are also numerous search engines that allow researchers to search across multiple databases simultaneously. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database is a widely used resource for pathway mapping and genome analysis. The KEGG search engine allows researchers to search for pathways, genes, and proteins across multiple databases and provides a valuable tool for analysing large datasets.

Search engines are also used extensively in biology research. Some of the most widely used search engines include Google Scholar, PubMed, and Web of Science. Google Scholar is a general-purpose search engine that is used to find scientific articles and papers. PubMed is a search engine that is specifically designed to search for articles in the biomedical literature, and it is used extensively in medical research. Web of Science is a search engine that is used to search for articles in a wide range of scientific disciplines, including biology.

While databases and search engines have revolutionized the way that researchers access and analyse biological data, there are also some challenges associated with these tools. For example, the data in these databases is often incomplete or out-of-date, and it can be difficult to find relevant information using search engines. In addition, there are often problems with data integration, as different databases use different formats and standards. These challenges will need to be addressed in order to continue to improve the usefulness and accuracy of these tools.

Conclusion

Databases are essential tools in the field of biology, providing a way to store, retrieve and analyse vast amounts of biological data generated by high-throughput technologies. Moreover, databases have become essential resources in the biology-related fields such genomics, proteomics, medicine and drug discovery, to mention some, providing valuable insights into

biological processes, molecular basis of disease, and drug development. The development of new databases and data management tools will continue to drive advances in these fields, enabling researchers to better understand the complexities of biological systems and develop new treatments for disease.

References

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36-42.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-242.

Bramer WM, Giustini D, Kramer BM, Anderson P. The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic reviews: a review of searches used in systematic reviews. Syst Rev. 2013;2:115.

GenBank. (2023). NCBI. https://www.ncbi.nlm.nih.gov/genbank/

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D361.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research, 28(1), 27-30.

PubMed (2023). National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/

PDB (2023). RCSB Protein Data Bank. https://www.rcsb.org/

Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD, Cooper DN. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 136(6):665-677.

UniProt (2023). EBI-UniProt. https://www.uniprot.org

UniProt Consortium (2023). UniProt: the Universal Protein Resource. Nucleic acids research, 51, D523-D531.