Genetic Databases: A Journey of Challenges and Promises

Autor: Mateusz Oleszek, nr. 144608

Introduction:

Genetic databases have evolved over the years to become essential tools for organizing and accessing biological data. These databases provide researchers with the nucleotide and amino acid data they need. The development of genomic databases began in the late 1970s with the establishment of the Los Alamos Sequence Database, which later became GenBank. Since then, the International Nucleotide Sequence Database Collaboration (INSDC) has been formed to collect and disseminate the growing amount of sequence data. The exponential growth of these databases has led to the creation of specialized databases and genome browsers. Despite their usefulness, these resources face challenges such as accuracy, accessibility, and organization of data. In this essay, we will explore the challenges and promises of genomic data resources, including sequence data repositories, general genome browsers, species- and taxa-specific databases, subject-specific databases, and potential solutions to the challenges they face.

Sequence Data Repositories:

The INSDC, consisting of GenBank, EMBL, and DDBJ, serves as a repository for biological sequences. Researchers submit their sequences, which are stored in a library and given unique identification numbers. The data in these repositories have exponentially grown over the years, doubling every 18 months. However, maintaining accuracy and accessibility across the databases is challenging due to the volume of data and potential errors. Efforts such as RefSeq aim to curate and correct sequence data. Despite these challenges, sequence repositories remain crucial resources for researchers.

General Genome Browsers:

General genome browsers like the UCSC Genome Browser, Ensembl, and NCBI's MapViewer repackage genome and gene annotation data from databases like GenBank. They provide a genomic context for individual genome features and facilitate cross-species comparisons. These browsers offer advanced search capabilities and allow researchers to access and analyze data in a customized fashion. They serve as valuable tools for visualizing and extracting information from genomic data.

Species- and Taxa-Specific Databases:

Species-specific or taxa-specific genome databases have been developed to provide deeper information about various genomes. These databases are publicly available and often curated. They offer accurate annotations and incorporate species-specific data types. Researchers can access these databases for a more in-depth analysis of specific genomes.

Standardized Genome Database Tools: GMOD:

GMOD is an open-source standard database and visualization toolset that promotes standardized querying, browsing, and usage of genome databases across species. Several species- and taxa-specific genome databases have adopted GMOD tools to enhance their annotation, visualization, and query options. The goal is to facilitate research and comparative studies by providing a common framework for database development and data exploration.

Subject-Specific Databases:

Subject-specific databases focus on specific biological data categories such as protein domains, protein structures, expression data, and genome-wide association studies. These databases serve as valuable resources for researchers working in specific areas of study. However, the large number of these databases can lead to redundancy and lack of integration.

Solutions to the Current Challenges of Accuracy and Curation:

Ensuring accurate data and efficient management and curation are ongoing challenges for genetic databases. Proposed solutions include education of database biocurators, standardized inclusion of sequence data and references in publications, and community curation. While some databases have embraced community curation, others like GenBank, have concerns about maintaining authoritative repositories. Efforts to incentivize researchers to contribute and improve the reliability of curation are ongoing, even if they are hampered by factors like the reliability of curation.

Conclusion:

Genomic data resources, including sequence data repositories, general genome browsers, species- and taxa-specific databases, and subject-specific databases, have revolutionized the field of bioinformatics. They provide researchers with invaluable tools to explore and analyze genetic information. However, challenges such as maintaining accuracy, accessibility, and organization of data persist. Privacy concerns have necessitated the removal of large datasets from public databases, and the analysis of community genomes requires unique tools and databases. Another obstacle is the overwhelming number of available genomic resources, making it difficult for researchers to locate and effectively utilize them. Despite these challenges, the rewards and opportunities provided by genomic data have been significant, enabling new discoveries and interdisciplinary relationships. Looking ahead, as the amount of data continues to grow, it is crucial for the scientific community to find solutions, such as advanced training, to ensure that amazing discoveries can be made.

References

The essay was based on "Genomic Data Resources: Challenges and Promises" by Warren C. Lathe III (OpenHelix), Jennifer M. Williams (OpenHelix), Mary E. Mangan (OpenHelix), and Donna Karolchik (University of California, Santa Cruz Genome Bioinformatics Group), published in Nature Education in 2008 (Lathe, Williams, Mangan, & Karolchik, 2008, p. 2).