Overview of Genetic Databases

Autor: Mateusz Oleszek, nr. 144608

Computer databases play a crucial role in organizing and accessing the vast amount of biological data. The Los Alamos Sequence Database, established in 1979 and later known as GenBank, served as an early repository for biological sequences. In 1982, GenBank was transferred to the National Center for Biotechnology Information (NCBI), where it currently resides. By the end of 1983, GenBank stored over 2,000 sequences with nearly 1 million base pairs. Concurrently, the International Nucleotide Sequence Database Collaboration (INSDC) was formed by NCBI, EMBL, and DDBJ to manage the growing amount of nucleotide and amino acid sequence data. The INSDC databases have witnessed exponential growth, now housing over 95 billion base pairs.

To handle the vast amount of raw sequence data, specialized databases have been created, including genome browsers, model organism databases, and process-specific databases. The number of genomics databases and tools has significantly increased, as evidenced by over 1,000 listings in a recent issue of Nucleic Acids Research. Furthermore, it is estimated that there are approximately 3,000 distinct genomic resources, tools, and databases available online.

Sequence Data Repositories:

The International Nucleotide Sequence Database Collaboration (INSDC) comprises GenBank, EMBL, and DDBJ, which accept sequence submissions from researchers worldwide. Each submission receives a unique identification number and is stored in a library. However, the exponential growth of data poses challenges in maintaining accuracy and accessibility. Duplication and incomplete or incorrect information are common issues. Efforts like RefSeq aim to curate and correct sequence data, but the vast amount of data makes it difficult to annotate and organize every entry.

Another challenge is the lack of context and annotation in repository data, making them less useful for research. Developers have repurposed GenBank data to create specialized databases, tailored to specific aesthetics, project history, and community needs. These databases provide better organization and access to genomic data. The main categories of such databases are further discussed in the following sections.

General Genome Browsers:

To address the research needs of scientists, general genome browsers have emerged as valuable tools. Examples of successful browsers include the UCSC Genome Browser, Ensembl by EBI, and MapViewer by NCBI. These browsers repackage genomic and gene annotation data from databases like GenBank, offering a comprehensive genomic context for specific genome features, such as genes or disease loci. Additionally, these browsers facilitate cross-species comparisons by displaying information in common formats, enabling easier visualization and extraction of data. For instance, users can search for a specific genomic region, like a disease gene, and access visual displays of the corresponding sequence and annotations. These displays are linked to supplementary data and databases for further investigation and also provide links back to the original data sources.

Strony wspomnianych genomowych baz danych

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.

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.

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).