OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Perspective

Management and Storage of Genomic Database Resources

Stanley Batley*

Department of Human Genetics, University of Queensland, Brisbane, Australia

DESCRIPTION

A breakthrough in DNA (Deoxyribonucleic Acid) sequencing technology over the past few years has reduced the cost of DNA sequencing and enabled the sequencing of an increasing number of genomes both practical and affordable. The type of sequence data being created has also undergone a significant change, with a great deal of short reads or pairs of short reads replacing the conventional relatively lengthy reads generated by Sanger sequencing. The administration, storage, and presentation of sequence data have had to be rethought as a result of these improvements in data quality and quantity, which presents a problem for bioinformatics. A genome database may be thought of as a collection of DNA sequences from several different plant and animal species. To facilitate the understanding of the sequencing data, they are connected to supporting databases. In the digital world of one or more computers, genome databases are created and maintained using a variety of operating systems, software programmes, file transfer protocols, and user interfaces. Genome databases hold sequence information created by molecular biologists, geneticists, and others utilizing methods that may be used to identify the precise nucleotide order of a whole DNA sequence. DNA is the substance found inside the cells of all living things that enables effective cellular activity and is passed on to next generations. Web browsers may access a variety of publicly available databases from the GenBank Database server at the National Center for Biotechnology Information (NCBI).

Genome mapping and sequencing initiatives are producing a growing amount of data; as a result, dependable and effective storage methods are needed. The first organization of genome sequencing and mapping data involved manual data collection and deposit in a central location, such as a table with columns and rows. This uses resources in a laborious, time-consuming, and ineffective manner. A drawback was rapidly identified as the absence of synergy between big independent collections of different sorts of data. Informatics is the science that combines information technology with data obtained from laboratory experiments to enable the gathering, analysis, and distribution of practical sets of pertinent data. In order to minimize mistake, eliminate redundancy across comparable data sources, and make well-informed judgments regarding outcomes, informatics offers a semi-automated method of retrieving, filtering, and making comparisons and contrasts of data in an electronic format. Many databases are constructed using tables or relational databases as their foundation. Alternatives include object-oriented techniques, which provide users the flexibility to categories data and analyses outcomes based on this classification, as well as to store and retrieve data.

The main repository for DNA sequence data are the International Nucleotide Sequence Databases, which are made up of GenBank, the DNA Data Bank of Japan (DDBJ), and the European Molecular Biological Laboratory (EMBL). They hold the text sequence data in addition to basic annotation and frequently the original data from which the text sequences were created. The storage of raw data for the new technologies is challenging because of the enormous size of the pictures, even though supplying the raw trace file data for Sanger sequences is frequently required for publication. According to estimates, it will cost more to store the terabytes of raw data generated by each run of the Illumina GAII or AB SOLiD than it will to produce the data itself. The raw picture files are now often deleted after being processed to create the relatively short text sequence and high-quality data files. Although the text sequence files can be stored for a long time using current tape and disc technologies, it is more difficult to keep the data in a useful state where it can be easily accessed by users. The size of the GenBank sequence repository is growing exponentially, and it takes a long time to scan this data using conventional sequence comparison techniques. Additionally, using common programmes like BLAST requires a lot of processing resources.

The goal of the human genome sequencing project is to make available to the public the whole sequence of the human genome. The efforts of computer science and information technology have already yielded results. Few if any undertakings in the biological sciences have needed the enormous amounts of data storage, assimilation, retrieval, and communication needs of genome sequence databases. In order to analyses the genomes of new species of bacteria, plants, and animals as they are fully

Correspondence to: Stanley Batley, Department of Human Genetics, University of Queensland, Brisbane, Australia, E-mail: stanbatley@edu.au

Received: 03-Jan-2023, Manuscript No. JDMGP-23-19626; **Editor assigned:** 06-Jan-2023, JDMGP-23-19626 (PQ); **Reviewed:** 20-Jan-2023, QC No. JDMGP-23-19626; **Revised:** 27-Jan-2023, Manuscript No. JDMGP-23-19626 (R); **Published:** 03-Feb-2023, DOI: 10.4172/2153-0602.23.14.275

Citation: Batley S (2023) Management and Storage of Genomic Database Resourcesse. J Data Mining Genomics Proteomics. 14:275

sequenced, sophisticated analytical tools will continue. The volume genome mapping and sequencing projects is increasing; hence, reliable and efficient means to store these data are required. Manual collecting and depositing of data in a central repository such as a table with columns and rows had served as the initial organization of genome sequencing and mapping data. This is a tedious, time consuming and inefficient use of resources. The lack of synergy between large separate collections of these types of data was quickly recognized as a limitation.