Commentary

iMedPub Journals
www.imedpub.com

International Journal of Innovative Research in Computer and Communication Engineering

2022
Vol.9 No.9:92

Databases of Protein Sequences and Functional Domains

Daniel John*

Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung, Taiwan

***Corresponding author:** Daniel John, Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung, Taiwan, E-mail: John_d@gmail.com

Received date: October 23, 2022, Manuscript No. IJIRCCE-22-15419; **Editor assigned date:** October 26, 2022, PreQC No. IJIRCCE-22-15419 (PQ); **Reviewed date:** November 07, 2022, QC No. IJIRCCE-22-15419; **Revised date:** November 17, 2022, Manuscript No. IJIRCCE-22-15419 (R); **Published date:** November 23, 2022, DOI: 10.36648/ijircce.7.9.92.

Citation: John D (2022) Databases of Protein Sequences and Functional Domains. Int J Inn Res Compu Commun Eng Vol.7 No.9: 92.

Description

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, chemistry, physics, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data. Bioinformatics has been used for in silico analyses of biological queries using computational and statistical techniques. Bioinformatics includes biological studies that use computer programming as part of their methodology, as well as specific analysis "pipelines" that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of candidate's genes and Single Nucleotide Polymorphisms.

Human Genome Project

Often, such identification is made with the aim to better understand the genetic basis of disease, unique adaptations and desirable properties differences between populations. In a less formal way, bioinformatics also tries to understand the organizational principles within nucleic acid and protein sequences, called proteomics. Image and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. There has been a tremendous advance in speed and cost reduction since the completion of the Human Genome Project, with some labs able to sequence over 100,000 billion bases each year and a full genome can be sequenced for a thousand dollars or less. Computers became essential in molecular biology when protein sequences became available after Frederick Sanger determined the sequence of insulin in the early 1950s. Comparing multiple sequences manually turned out to be impractical.

Antibody Sequences Release

A pioneer in the field was Margaret Oakley Dayhoff. She compiled one of the first protein sequence databases, initially published as booksand pioneered methods of sequence alignment and molecular evolution. Another early contributor to bioinformatics was Elvin A. Kabat, who pioneered biological sequence analysis in 1970 with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1910. In the 1970s, new techniques for sequencing DNA were applied to bacteriophage and the extended nucleotide sequences were then parsed with informational and statistical algorithms. These studies illustrated that well known features, such as the coding segments and the triplet code, are revealed in straightforward statistical analyses and were thus proof of the concept that bioinformatics would be insightful Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data. Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce atremendous amount of information related to molecular biology processes. Common activities in bioinformatics include mapping and analysing DNA and protein sequences, aligning DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures Most DNA sequencing techniques produce short fragments of sequence that need to be assembled to obtain complete gene or genome sequences. The so-called shotgun sequencing technique which was used, for example, by The Institute for Genomic Research (TIGR) to sequence the first bacterial genome, Haemophilus influenza ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments and the resulting assembly usually contains numerous gaps that must be filled in later. Shotgun sequencing is the method of choice for virtually all genomes in the context of genomics annotation is the process of marking the genes and other biological features in a DNA sequence. This process needs to be automated because most genomes are too large to annotate by hand, not to mention the desire to annotate as many genomes as possible, as the rate of sequencing has ceased to pose a bottleneck. Annotation is made possible by the fact that genes have

recognisable start and stop regions, although the exact sequence found in these regions can vary between genes. Genome annotation can be classified into three levels: the nucleotide, protein and process levels. Gene finding is a chief aspect of nucleotide-level annotation. For complex genomes, the most successful methods use a combination of ab initio gene prediction and sequence comparison with expressed sequence databases and other organisms. Nucleotide-level annotation also allows the integration of genome sequence with other genetic and physical maps of the genome. The principal aim of protein-level annotation is to assign function to the products of the genome. Databases of protein sequences and functional domains and motifs are powerful resources for this type of annotation. Nevertheless, half of the predicted proteins in a new genome sequence tend to have no obvious function. The core of comparative genome analysis is the establishment of the correspondence between genes orthology analysis and other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution.

At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectrum of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.