Advancements in Genetic Engineering

Editorial

Protein Sequence Analysis

Hendrick john

Florida State University

Bioinformatics is the use of information technology to store, organise, and interpret large amounts of biological data, such as sequences and structures of proteins and nucleic acids (the building blocks of organisms) (the information carrier). Nucleic acid biological knowledge is available as sequences, while protein data is available as sequences. Sequences are interpreted in a single dimension, while sequences' three-dimensional data is stored in the structure. Sequences are a type of pattern. There was a lot of enthusiasm in the world of Molecular Biology when Sanger first discovered the method to sequence proteins. The need to build databases of biological sequences sparked initial interest in Bioinformatics.

Series and structure databases are the two types of biological databases. Sequence databases are useful for both nucleic acid and protein sequences, while structure databases are only useful for proteins. After the Insulin protein sequence was published in 1956, the first database was developed in a short time. Insulin was the first protein to be sequenced, by the way. Insulin's sequence is made up of only 51 residues (alphabets in a sentence) that define it. The first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was discovered in the mid-1960s. During this time, researchers looked into the three-dimensional structure of proteins, and in 1972, the Protein Data Bank was created as the first protein structure database, with only 10 entries. This has now expanded into a massive database of over 10,000 records.

Although individual laboratories kept their own databases of protein sequences, the SWISS-PROT protein sequence database was established in 1986 and now contains over 70,000 protein sequences from over 5000 model organisms, a small fraction of all known organisms. Apart from maintaining the massive database, it is critical to extract valuable knowledge from the various primary and secondary databases. For data mining and information discovery, many efficient algorithms have been created. These are computationally intensive, necessitating the use of fast and parallel computing facilities to handle multiple queries at once. It is these search tools that integrate the user and the databases. One of the widely used search program is BLAST.BLAST is a collection of similarity search programmes that search all available sequence databases, regardless of whether the query is a protein or a DNA sequence. The BLAST programmes have been developed to be quick while maintaining a high level of sensitivity.

The statistical analysis of the scores assigned in a BLAST quest makes it easier to discern actual matches from random context hits. BLAST employs a heuristic algorithm that seeks local alignments rather than global alignments, allowing it to detect relationships between sequences that share only isolated regions of similarity.