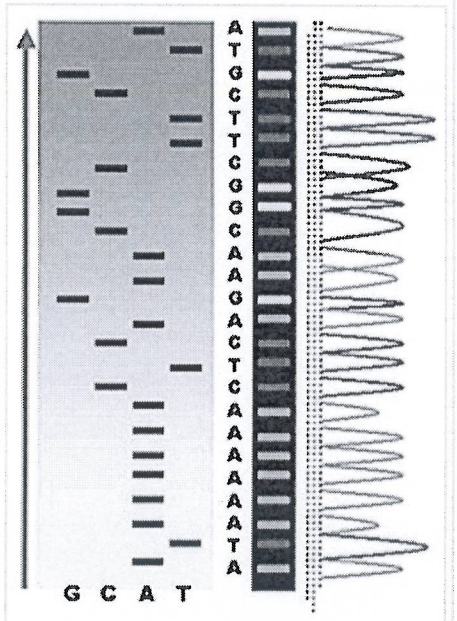
DNA sequencing

From Wikipedia, the free encyclopedia

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as medical diagnosis, biotechnology, forensic biology, virology and biological systematics. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of fluorescence-based sequencing methods with automated analysis,[1] DNA sequencing has become easier and orders of magnitude faster.[2]



An example of the results of automated chain-termination DNA sequencing.

Contents

Use of sequencing

DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e. clusters of genes or operons), full chromosomes or entire genomes. Sequencing provides the order of individual nucleotides in DNA or RNA (commonly represented as A, C, G, T, and U) isolated from cells of animals, plants, bacteria, archaea, or virtually any other source of genetic information. This is useful for:

- Molecular biology – studying the genome itself, how proteins are made, what proteins are made, identifying new genes and associations with diseases and phenotypes, and identifying potential drug targets
- Evolutionary biology – studying how different organisms are related and how they evolved
- Metagenomics – Identifying species present in a body of water, sewage, dirt, debris filtered from the air, or swab samples of organisms. Helpful in ecology, epidemiology, microbiome research, and other fields.

Less-precise information is produced by non-sequencing techniques like DNA fingerprinting. This information may be easier to obtain and is useful for:

- Detecting the presence of known genes for medical purposes (see genetic testing)
- Forensic identification
- Parental testing

The four canonical bases

The canonical structure of DNA has four bases: Thymine (T), Adenine (A), Cytosine (C), and Guanine (G). DNA sequencing is the determination of the physical order of these bases in a molecule of DNA. However, there are many other bases that may be present in a molecule. In some viruses (specifically, bacteriophage), cytosine may be replaced by hydroxy methyl or hydroxy methyl glucose cytosine.[3] In mammalian DNA, variant bases with methyl groups or phosphosulfate may be found.[4][5] Depending on the sequencing technique, a particular modification may or may not be detected, e.g., the 5mC (5 methyl cytosine) common in humans may or may not be detected.[6]

History

RNA sequencing

Though the structure of DNA was established as a double helix in 1953,[7] several decades would pass before fragments of DNA could be reliably analyzed for their sequence in the laboratory. RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), in 1972[8] and 1976.[9]

Early DNA sequencing methods

The first method for determining DNA sequences involved a location-specific primer extension strategy established by Ray Wu at Cornell University in 1970.[10] DNA polymerase catalysis and specific nucleotide labeling, both of which figure prominently in current sequencing schemes, were used to

sequence the cohesive ends of lambda phage DNA[11][12][13] Between 1970 and 1973, Wu, R Padmanabhan and colleagues demonstrated that this method can be employed to determine any DNA sequence using synthetic location-specific primers.[14][15][16] Frederick Sanger then adopted this primer-extension strategy to develop more rapid DNA sequencing methods at the MRC Centre, Cambridge, UK and published a method for "DNA sequencing with chain-terminating inhibitors" in 1977.[17] Walter Gilbert and Allan Maxam at Harvard also developed sequencing methods, including one for "DNA sequencing by chemical degradation".[18][19] In 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.[20] Advancements in sequencing were aided by the concurrent development of recombinant DNA technology, allowing DNA samples to be isolated from sources other than viruses.

Sequencing of full genomes

The first full DNA genome to be sequenced was that of bacteriophage φX174 in 1977.[21] Medical Research Council scientists deciphered the complete DNA sequence of the Epstein-Barr virus in 1984, finding it to be 170 thousand base-pairs long.

A non-radioactive method for transferring the DNA molecules of sequencing reaction mixtures onto an immobilizing matrix during electrophoresis was developed by Pohl and co-workers in the early 80's.[22][23] Followed by the commercialization of the DNA sequencer "Direct-Blotting-Electrophoresis-System GATC 1500" by GATC Biotech, which was intensively used in the framework of the EU genome-sequencing programme, the complete DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome II.[24] Leroy E. Hood's laboratory at the California Institute of Technology announced the first semi-automated DNA sequencing machine in 1986.[25] This was followed by Applied Biosystems' marketing of the first fully automated sequencing machine, the ABI 370, in 1987 and by Dupont's Genesis 2000[26] which used a novel fluorescent labeling technique enabling all four dideoxynucleotides to be identified in a single lane. By 1990, the U.S. National Institutes of Health (NIH) had begun large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* at a cost of US$0.75 per base. Meanwhile, sequencing of human cDNA sequences called expressed sequence tags began in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.[27] In 1995, Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) published the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*. The circular chromosome contains 1,830,137 bases and its publication in the journal Science[28] marked the first published use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

By 2001, shotgun sequencing methods had been used to produce a draft sequence of the human genome.[29][30]

Next-generation sequencing methods

Several new methods for DNA sequencing were developed in the mid to late 1990s and were implemented in commercial DNA sequencers by the year 2000.

On October 26, 1990, Roger Tsien, Pepi Ross, Margaret Fahnestock and Allan J Johnston filed a patent