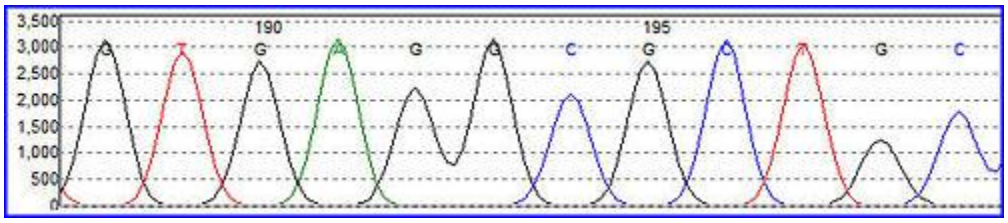
DNA Sequencing link: http://en.wikipedia.org/wiki/DNA_sequencing

DNA sequencing

From Wikipedia, the free encyclopedia

Jump to: navigation, search

The term **DNA sequencing** encompasses biochemical methods for determining the order of the nucleotide bases, adenine, guanine, cytosine, and thymine, in a DNA oligonucleotide. The sequence of DNA constitutes the heritable genetic information in nuclei, plasmids, mitochondria, and chloroplasts that forms the basis for the developmental programs of all living organisms. Determining the DNA sequence is therefore useful in basic research studying fundamental biological processes, as well as in applied fields such as diagnostic or forensic research. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the large-scale sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.



DNA Sequence Trace

Contents

[hide]

[edit] Early methods

For thirty years, a large proportion of DNA sequencing has been carried out with the chain-termination method developed by Frederick Sanger and coworkers in 1975.[1][2] Prior to the development of rapid DNA sequencing methods in the early 1970s by Sanger in England and Walter Gilbert and Allan Maxam at Harvard,[3][4] a number of laborious methods were used. For instance, in 1973[5] Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

RNA sequencing, which for technical reasons is easier to perform than DNA sequencing, was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing, dating from the pre-recombinant DNA era, is the sequence of the first complete gene and then the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), published between 1972[6] and 1976.[7]

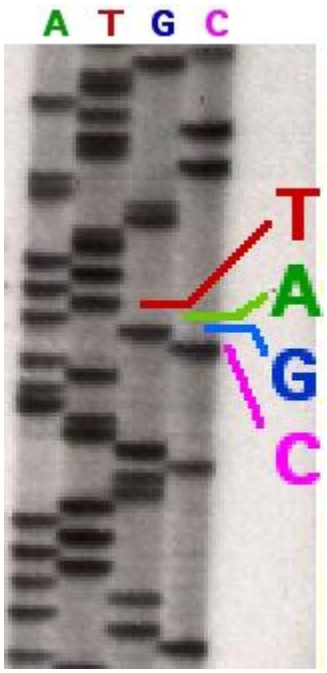
[edit] Maxam-Gilbert sequencing

In 1976-1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases [1]. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing,[8][9] Maxam-Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the development and improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity, extensive use of hazardous chemicals, and difficulties with scale-up. In addition, unlike the chain-termination method, chemicals used in the Maxam-Gilbert method cannot easily be customized for use in a standard molecular biology kit.

In brief, the method requires radioactive labelling at one end and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). Thus a series of labelled fragments is generated, from the radiolabelled end to the first 'cut' site in each molecule. The fragments are then size-separated by gel electrophoresis, with the four reactions arranged side by side. To visualize the fragments generated in each reaction, the gel is exposed to X-ray film for autoradiography, yielding an image of a series of dark 'bands' corresponding to the radiolabelled DNA fragments, from which the sequence may be inferred.

Also sometimes known as 'chemical sequencing', this method originated in the study of DNA-protein interactions (footprinting), nucleic acid structure and epigenetic modifications to DNA, and within these it still has important applications.

[edit] Chain-termination methods



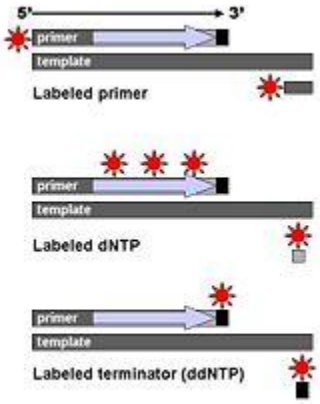
Part of a radioactively labelled sequencing gel

While the chemical sequencing method of Maxam and Gilbert, and the plus-minus method of Sanger and Coulson were orders of magnitude faster than previous methods, the chain-terminator method developed by Sanger was even more efficient, and rapidly became the method of choice. The Maxam-Gilbert technique requires the use of highly toxic chemicals, and large amounts of radiolabeled DNA, whereas the chain-terminator method uses fewer toxic chemicals and lower amounts of radioactivity. The key principle of the Sanger method was the use of dideoxynucleotides triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination or Sanger method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, radioactively or fluorescently labeled nucleotides, and modified nucleotides that terminate DNA strand elongation. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP). These dideoxynucleotides are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides during DNA strand elongation. Incorporation of a dideoxynucleotide into the nascent (elongating) DNA strand therefore terminates DNA strand

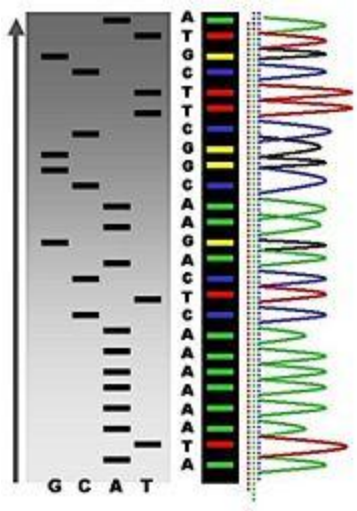
extension, resulting in various DNA fragments of varying length. The dideoxynucleotides are added at lower concentration than the standard deoxynucleotides to allow strand elongation sufficient for sequence analysis.

The newly synthesized and labeled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel. Each of the four DNA synthesis reactions is run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The terminal nucleotide base can be identified according to which dideoxynucleotide was added in the reaction giving that band. The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence as indicated.



DNA fragments can be labeled by using a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP. (click to expand)

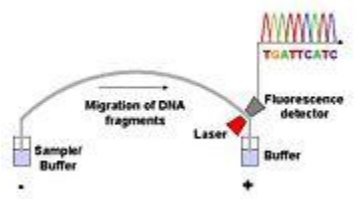
There are some technical variations of chain-termination sequencing. In one method, the DNA fragments are tagged with nucleotides containing radioactive phosphorus for radiolabelling. Alternatively, a primer labeled at the 5' end with a fluorescent dye is used for the tagging. Four separate reactions are still required, but DNA fragments with dye labels can be read using an optical system, facilitating faster and more economical analysis and automation. This approach is known as 'dye-primer sequencing'. The later development by L Hood and coworkers[10][11] of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks (click to expand)

The different chain-termination methods have greatly simplified the amount of work and planning needed for DNA sequencing. For example, the chain-termination-based "Sequenase" kit from USB Biochemicals contains most of the reagents needed for sequencing, prealiquoted and ready to use. Some sequencing problems can occur with the Sanger method, such as non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence. In addition, secondary structures within the DNA template, or contaminating RNA randomly priming at the DNA template can also affect the fidelity of the obtained sequence. Other contaminants affecting the reaction may consist of extraneous DNA or inhibitors of the DNA polymerase.

[edit] Dye-terminator sequencing



Capillary electrophoresis (click to expand)

An alternative to primer labelling is labelling of the chain terminators, a method commonly called 'dye-terminator sequencing'. The major advantage of this method is that the sequencing can be performed in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with a different fluorescent dye, each fluorescing at a different wavelength. This method is attractive because of its greater expediency and speed and is now the mainstay in automated

sequencing with computer-controlled sequence analyzers (see below). Its potential limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis (see figure to the right). This problem has largely been overcome with the introduction of new DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs", caused by certain chemical characteristics of the dyes that can result in artifacts in DNA sequence traces. The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects, as it is both easier to perform and lower in cost than most previous sequencing methods.

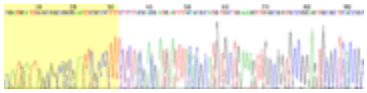
[edit] Challenges

Modern sequencing typically produces a sequence that has poor quality in the first 15-40 bases, a high quality region of 700-900 bases, and then quickly deteriorating quality. Base calling software typically outputs an estimate of quality along with the sequence to aid in quality trimming.

Before the DNA can be sequenced, linker sequences are attached to its ends, and it is inserted into a cloning vector. The resulting sequence can therefore often contain parts of the vector or the linker sequences, which must be filtered out prior to analysis. In contrast, emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors.

During PCR amplification, unrelated sequences can hybridize, and the resulting clone can be a chimaeric sequence, containing fragments from both sequences. Another problem is polymerase stuttering, where the polymerase repeatedly outputs the same fragments, giving an artificially long low-complexity part of the sequence.

[edit] Automation and sample preparation



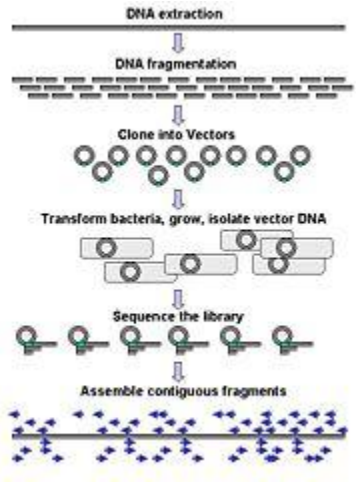
View of the start of an example dye-terminator read (click to expand)

Modern automated DNA sequencing instruments (DNA sequencers) can sequence up to 384 fluorescently labelled samples in a single batch (run) and perform as many as 24 runs a day. However, automated DNA sequencers carry out only DNA-size-based separation (by capillary electrophoresis), detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. In the past, an operator had to trim the low quality ends (see image in the right) of every sequence manually in order to remove the sequencing errors. However, today, software like Chromatogram Explorer or DNA Baser can automatically trim the ends at batch. Such programs score the peaks of each base

for quality and remove low-quality base peaks(generally located at the ends of the sequence). If too many bases have a quality score value below a certain threshold, those bases will be automatically deleted. The accuracy of such algorithms is currently below visual examination by a human operator, but is high enough for processing of sequence data sets that are too large for manual examination.

[edit] Large-scale sequencing strategies

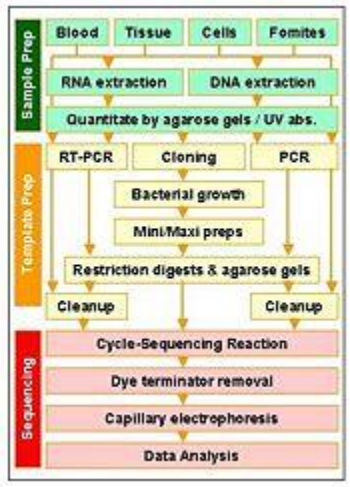
Current methods can directly sequence only relatively short (300-1000 nucleotides long) DNA fragments in a single reaction. [2]. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. Limitations on ddNTP incorporation were largely solved by Tabor at Harvard Medical, Carl Fuller at USB biochemicals, and their coworkers.[12]



Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping regions.(click to expand)

Large-scale sequencing aims at sequencing very long DNA fragments. Even relatively small bacterial genomes contain millions of nucleotides, and the human chromosome 1 alone contains about 246 million bases. Therefore, some approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, usually a bacterial artificial chromosome (BAC), and amplified in *Escherichia coli*. The amplified DNA can then be purified from the bacterial cells (a disadvantage of bacterial clones for sequencing is that some DNA sequences may be inherently *un-clonable* in some or all available bacterial strains, due to deleterious effect of the cloned sequence on the host bacterium or other effects). These short DNA fragments purified from individual bacterial colonies are then individually and completely sequenced and assembled electronically into one long, contiguous sequence by identifying 100%-identical overlapping sequences between them (shotgun sequencing). This method does

not require any pre-existing information about the sequence of the DNA and is often referred to as *de novo* sequencing. Gaps in the assembled sequence may be filled by Primer walking, often with sub-cloning steps (or transposon-based sequencing depending on the size of the remaining region to be sequenced). These strategies all involve taking many small *reads* of the DNA by one of the above methods and subsequently assembling them into a contiguous sequence. The different strategies have different tradeoffs in speed and accuracy; the shotgun method is the most practical for sequencing large genomes, but its assembly process is complex and potentially error-prone - particularly in the presence of sequence repeats. Because of this, the assembly of the human genome is not literally complete — the repetitive sequences of the centromeres, telomeres, and some other parts of chromosomes result in gaps in the genome assembly. Despite having only 93% of the full genome assembled, the Human Genome Project was declared complete because their definition of human genome sequencing was limited to euchromatic sequence (99% complete at the time), excluding these intractable repetitive regions.[13]



Resequencing steps. Sample prep: Extraction of nucleic acid. Template prep: Amplification and preparation of a small region of the target region. Sequencing steps. (click to expand)

The human genome is about 3 billion (3,000,000,000) bp long;[14] if the average fragment length is 500 bases, it would take a minimum of six million (3 billion/500) to sequence the human genome (not allowing for overlap = 1-fold coverage). Keeping track of such a high number of sequences presents significant challenges, only held down by developing and coordinating several procedural and computational algorithms, such as efficient database development and management.

Resequencing or *targeted sequencing* is utilized for determining a change in DNA sequence from a "reference" sequence. It is often performed using PCR to amplify the region of interest (pre-existing DNA sequence is required to design the PCR primers). Resequencing uses three steps, extraction of DNA or RNA from biological tissue; amplification of the RNA or DNA (often by PCR); followed by sequencing. The resultant sequence is compared to a reference or a normal sample to detect mutations.

[edit] New sequencing methods

[edit] High-throughput sequencing

The high demand for low cost sequencing has given rise to a number of high-throughput sequencing technologies.[15][16] These efforts have been funded by public and private institutions as well as privately researched and commercialized by biotechnology companies. High-throughput sequencing technologies are intended to lower the cost of sequencing DNA libraries beyond what is possible with the current dye-terminator method based on DNA separation by capillary electrophoresis. Many of the new high-throughput methods use methods that parallelize the sequencing process, producing thousands or millions of sequences at once.

In vitro clonal amplification

As molecular detection methods are often not sensitive enough for single molecule sequencing, most approaches use an *in vitro* cloning step to generate many copies of each individual molecule. Emulsion PCR is one method, isolating individual DNA molecules along with primer-coated beads in aqueous bubbles within an oil phase. A polymerase chain reaction (PCR) then coats each bead with clonal copies of the isolated library molecule and these beads are subsequently immobilized for later sequencing. Emulsion PCR is used in the methods published by Marguilis et al. (commercialized by 454 Life Sciences, acquired by Roche), Shendure and Porreca et al. (also known as "polony sequencing") and SOLiD sequencing, (developed by Agencourt and acquired by Applied Biosystems).[17][18][19] Another method for *in vitro* clonal amplification is "bridge PCR", where fragments are amplified upon primers attached to a solid surface, developed and used by Solexa (now owned by Illumina). These methods both produce many physically isolated locations which each contain many copies of a single fragment. The single-molecule method developed by Stephen Quake's laboratory (later commercialized by Helicos) skips this amplification step, directly fixing DNA molecules to a surface.[20]

Parallelized sequencing

Once clonal DNA sequences are physically localized to separate positions on a surface, various sequencing approaches may be used to determine the DNA sequences of all locations, in parallel. "Sequencing by synthesis", like the popular dye-termination electrophoretic sequencing, uses the process of DNA synthesis by DNA polymerase to identify the bases present in the complementary DNA molecule. Reversible terminator methods (used by Illumina and Helicos) use reversible versions of dye-terminators, adding one nucleotide at a time, detecting fluorescence corresponding to that position, then removing the blocking group to allow the polymerization of another nucleotide. Pyrosequencing (used by 454) also uses DNA polymerization to add nucleotides, adding one type of nucleotide at a time, then detecting and quantifying the number of nucleotides added to a given location through the light emitted by the release of attached pyrophosphates.[17][21]

"Sequencing by ligation" is another enzymatic method of sequencing, using a DNA ligase enzyme rather than polymerase to identify the target sequence.[22][18][19] Used in the polony

method and in the SOLiD technology offered by Applied Biosystems, this method uses a pool of all possible oligonucleotides of a fixed length, labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal corresponding to the complementary sequence at that position.

[edit] Other sequencing technologies

Other methods of DNA sequencing may have advantages in terms of efficiency or accuracy. Like traditional dye-terminator sequencing, they are limited to sequencing single isolated DNA fragments. "Sequencing by hybridization" is a non-enzymatic method that uses a DNA microarray. In this method, a single pool of unknown DNA is fluorescently labeled and hybridized to an array of known sequences. If the unknown DNA hybridizes strongly to a given spot on the array, causing it to "light up", then that sequence is inferred to exist within the unknown DNA being sequenced.[23] Mass spectrometry can also be used to sequence DNA molecules; conventional chain-termination reactions produce DNA molecules of different lengths and the length of these fragments is then determined by the mass differences between them (rather than using gel separation).[24]

There are new proposals for DNA sequencing, which are in development, but remain to be proven. These include labeling the DNA polymerase,[25] reading the sequence as a DNA strand transits through nanopores,[26] and microscopy-based techniques, such as AFM or electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.[27] In October 2006 the NIH issued a news release describing novel sequencing techniques and announcing several grant awards.[28]

In October 2006, the X Prize Foundation established the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."[29]

[edit] Major landmarks in DNA sequencing

- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.
- 1975 The first complete DNA genome to be sequenced is that of bacteriophage ϕX174
- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation" [3]. Fred Sanger, independently, publishes "DNA sequencing by enzymatic synthesis".
- 1980 Fred Sanger and Wally Gilbert receive the Nobel Prize in Chemistry

    - EMBL-bank, the first nucleotide sequence repository, is started at the European Molecular Biology Laboratory

- 1982 Genbank starts as a public repository of DNA sequences.
    - Andre Marion and Sam Eletr from Hewlett Packard start Applied Biosystems in May, which comes to dominate automated sequencing.
    - Akiyoshi Wada proposes automated sequencing and gets support to build robots with help from Hitachi.

- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

- 1985 Kary Mullis and colleagues develop the polymerase chain reaction, a technique to replicate small fragments of DNA

- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.

- 1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.
    - Walter Gilbert leaves the U.S. National Research Council genome panel to start Genome Corp., with the goal of sequencing and commercializing the data.

- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at 75 cents (US)/base).
    - Barry Karger (January[30]), Lloyd Smith (August[31]), and Norman Dovichi (September[32]) publish on capillary electrophoresis.

- 1991 Craig Venter develops strategy to find expressed genes with ESTs (Expressed Sequence Tags).
    - Uberbacher develops GRAIL, a gene-prediction program.

- 1992 Craig Venter leaves NIH to set up The Institute for Genomic Research (TIGR).
    - William Haseltine heads Human Genome Sciences, to commercialize TIGR products.
    - Wellcome Trust begins participation in the Human Genome Project.
    - Simon et al. develop BACs (Bacterial Artificial Chromosomes) for cloning.
    - First chromosome physical maps published:
        - Page et al. - Y chromosome[33];
        - Cohen et al. chromosome 21[34].
        - Lander - complete mouse genetic map[35];
        - Weissenbach - complete human genetic map[36].

- 1993 Wellcome Trust and MRC open Sanger Centre, near Cambridge, UK.
    - The GenBank database migrates from Los Alamos (DOE) to NCBI (NIH).

- 1995 Venter, Fraser and Smith publish first sequence of free-living organism, *Haemophilus influenzae* (genome size of 1.8 Mb).

    - Richard Mathies et al. publish on sequencing dyes (PNAS, May)[37].
    - Michael Reeve and Carl Fuller, thermostable polymerase for sequencing[12].

- 1996 International HGP partners agree to release sequence data into public databases within 24 hours.
    - International consortium releases genome sequence of yeast *S. cerevisiae* (genome size of 12.1 Mb).
    - Yoshihide Hayashizaki's at RIKEN completes the first set of full-length mouse cDNAs.
    - ABI introduces a capillary electrophoresis system, the ABI310 sequence analyzer.

- 1997 Blattner, Plunkett et al. publish the sequence of E. coli (genome size of 5 Mb)[38]

- 1998 Phil Green and Brent Ewing of Washington University publish "`phred`" for interpreting sequencer data (in use since '95)[39].
    - Venter starts new company "Celera"; "will sequence HG in 3 yrs for $300m."
    - Applied Biosystems introduces the 3700 capillary sequencing machine.
    - Wellcome Trust doubles support for the HGP to $330 million for 1/3 of the sequencing.
    - NIH & DOE goal: "working draft" of the human genome by 2001.
    - Sulston, Waterston et al finish sequence of *C. elegans* (genome size of 97Mb)[40].

- 1999 NIH moves up completion date for rough draft, to spring 2000.
    - NIH launches the mouse genome sequencing project.
    - First sequence of human chromosome 22 published[41].

- 2000 Celera and collaborators sequence fruit fly *Drosophila melanogaster* (genome size of 180Mb) - validation of Venter's shotgun method. HGP and Celera debate issues related to data release.
    - HGP consortium publishes sequence of chromosome 21.[42]
    - HGP & Celera jointly announce working drafts of HG sequence, promise joint publication.
    - Estimates for the number of genes in the human genome range from 35,000 to 120,000. International consortium completes first plant sequence, *Arabidopsis thaliana* (genome size of 125 Mb).

- 2001 HGP consortium publishes Human Genome Sequence draft in Nature (15 Feb)[43].
    - Celera publishes the Human Genome sequence[44].

- 2005 420,000 VariantSEQr human resequencing primer sequences published on new NCBI Probe database.

- 2007 For the first time, a set of closely related species (12 Drosophilidae) are sequenced, launching the era of phylogenomics.
    - Craig Venter publishes his full diploid genome: the first human genome to be sequenced completely.

- 2008 An international consortium launches The 1000 Genomes Project, aimed to study human genetic variability.

- 2008 Leiden University Medical Center scientists decipher the first complete DNA sequence of a woman.[45]

[edit] See also

- Sequencing
- Genome project - how entire genomes are assembled from these short sequences.
- Applied Biosystems - provided most of the chemistry and equipment for the genome projects. Next-generation technology for very high data generation rates.
- 454 Life Sciences - company specializing in high-throughput DNA sequencing using a sequencing-by-synthesis approach.
- Illumina (company) - Advancing genetic analysis one billion bases at a time; whole genome sequencing.
- Joint Genome Institute - sequencing center from the US Department of Energy whose mission is to provide integrated high-throughput sequencing and computational analysis to enable genomic-scale/systems-based scientific approaches to DOE-relevant challenges in energy and the environment.
- DNA field-effect transistor

[edit] Citations

1. ^ Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975 May 25;94(3):441–448
2. ^ F. Sanger, S. Nicklen, and A. R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc Natl Acad Sci U S A. 1977 December; 74(12): 5463–5467
3. ^ Maxam AM, Gilbert W., A new method for sequencing DNA, Proc Natl Acad Sci U S A. 1977 Feb;74(2):560-4
4. ^ Unknown
5. ^ Proc Natl Acad Sci U S A. 1973 December; 70(12 Pt 1-2): 3581–3584. The Nucleotide Sequence of the lac Operator, Walter Gilbert and Allan Maxam
6. ^ Min Jou W, Haegeman G, Ysebaert M, Fiers W., Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein, Nature. 1972 May 12;237(5350):82-8
7. ^ Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*. 1976 Apr 8;260(5551):500-7.
8. ^ Sanger, F. & Coulson, A. R. (1975) J. Mol. Biol. 94, 441-448
9. ^ Unknown
10. ^ Nature. 1986 Jun 12-18;321(6071):674-9. Fluorescence detection in automated DNA sequence analysis. We have developed a method for the partial automation of DNA sequence analysis. Fluorescence detection of the DNA fragments is accomplished by means of a fluorophore covalently attached to the oligonucleotide primer used in enzymatic DNA sequence analysis. A different coloured fluorophore is used for each of the reactions specific for the bases A, C, G and T. The reaction mixtures are combined and co-electrophoresed down a single polyacrylamide gel tube, the separated fluorescent bands of DNA are detected near the bottom of the tube, and the sequence information is acquired directly by computer.
11. ^ Nucleic Acids Res. 1985 Apr 11;13(7):2399-412. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Note that Oxford University Press, the publishers of the journal Nucleic Acids

Research, make the full contents of this journal available online for free - you can download a copy of this paper for yourself !!
12. ^ a b Reeve, Michael A.; Reeve MA, Fuller CW (1995-08-31). "A novel thermostable polymerase for DNA sequencing". *Nature* **376** (6543): 796–797. doi:10.1038/376796a0. PMID 7651542. Retrieved on 2007-10-08.
13. ^ International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome". *Nature* **431** (7011): 931–45. doi:10.1038/nature03001. PMID 15496913. paper available online
14. ^ Human Genome Project Information
15. ^ Neil Hall (2007). "Advanced sequencing technologies and their wider impact in microbiology". *The Journal of Experimental Biology* **209**: 1518–1525.
16. ^ G.M. Church (2006). "Genomes for ALL". *Scientific American* **294** (1): 47–54. PMID 16468433.
17. ^ a b M. Margulies, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors". *Nature* **437**: 376–380.
18. ^ a b J. Shendure, G.J. Porreca, N.B. Reppas, X. Lin, J.Pe McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra and G.M. Church (2005). "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome". *Science* **309** (5741): 1728–1732. doi:10.1126/science.1117389. PMID 16081699.
19. ^ a b http://solid.appliedbiosystems.com/ - Applied Biosystems' SOLiD technology
20. ^ Braslavsky, I., Hebert, H., Kartalov, E. and Quake, S.R. (2003). "Sequence information can be obtained from single DNA molecules". *Proceedings of the National Academy of Sciences of the United States of America* **100**: 3960–3964. doi:10.1073/pnas.0230489100. PMID 12651960.full text available online
21. ^ M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren (1996). "Real-time DNA sequencing using detection of pyrophosphate release". *Analytical Biochemistry* **242**: 84=89. doi:10.1006/abio.1996.0432.
22. ^ S. C. Macevicz, US Patent 5750341, filed 1995
23. ^ G.J. Hanna, V.A. Johnson, D.R. Kuritzkes, D.D. Richman, J. Martinez-Picado, L. Sutton, J.D. Hazelwood, R.T. D'Aquila (2000). "Comparison of sequencing by hybridization and cycle sequencing for genotyping of human immunodeficiency virus type 1 reverse transcriptase". *Journal of Clinical Microbiology* **38** (7): 2715. PMID 10878069.
24. ^ J.R. Edwards, H.Ruparel, and J. Ju. "Mass-spectrometry DNA sequencing". *Mutation Research* **573** (1-2): 3–12.
25. ^ VisiGen Biotechnologies Inc. - Technology Overview
26. ^ The Harvard Nanopore Group
27. ^ USPTO application # 20060029957 assigned to ZS genetics http://www.freepatentsonline.com/20060029957.html
28. ^ NHGRI Aims to Make DNA Sequencing Faster, More Cost Effective, NIH News Release, 4 October 2006
29. ^ "PRIZE Overview: Archon X PRIZE for Genomics"
30. ^ Karger, Barry L.; A. Guttman, A. S. Cohen, D. N. Heiger (1990-01-15). "Analytical and micropreparative ultrahigh resolution of oligonucleotides by polyacrylamide gel high-performance capillary electrophoresis". *Analytical Chemistry* **62** (2): 137–141. doi:10.1021/ac00201a010.|url=|format=|accessdate=2007-10-08
31. ^ Smith, Lloyd M.; Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB (1990-08-11). "High speed DNA sequencing by capillary electrophoresis". *Nucleic Acids Research* **18**: 4417–4421. doi:10.1093/nar/18.15.4417. PMID 2388826 doi:10.1093/nar/18.15.4417. Retrieved on 2007-10-08.

32. ^ Dovichi, Norman J.; H.P. Swerdlow, S. Wu , H.R. Harke (1990-09-07). "Capillary gel electrophoresis for DNA sequencing: laser-induced fluorescence detection with the sheath flow cuvette". *Journal of Chromatography* **516**: 61–67. PMID 2286629. Retrieved on 2007-10-08.
33. ^ Page, DC; Foote S, Vollrath D, Hilton A (1992-10-02). "The human Y chromosome: overlapping DNA clones spanning the euchromatic region". *Science* **258** (5079): 60–66. doi:10.1126/science.1359640. PMID 1359640. Retrieved on 2007-10-08.
34. ^ Cohen, Daniel; Ilya Chumakov, Philippe Rigault, Sophie Guillou, Pierre Ougen, Alain Billaut, Ghislaine Guasconi, Patricia Gervy, Isabelle LeGall, Pascal Soularue, Laurent Grinas, Lydie Bougueleret, Christine Bellanné-Chantelot, Bruno Lacroix, Emmanuel Barillot, Philippe Gesnouin, Stuart Pook, Guy Vaysseix, Gerard Frelat, Annette Schmitz, Jean-Luc Sambucy, Assumpcio Bosch, Xavier Estivill, Jean Weissenbachparallel, Alain Vignal, Harold Riethman, David Cox, David Patterson, Kathleen Gardiner, Masahira Hattori, Yoshiyuki Sakaki, Hitoshi Ichikawa, Misao Ohki, Denis Le Paslier, Roland Heilig, Stylianos Antonarakis (1992-10-01). "Continuum of overlapping clones spanning the entire human chromosome 21q". *Nature* **359** (6394): 380–387. doi:10.1038/359380a0. PMID 1406950 doi:10.1038/359380a0. Retrieved on 2007-10-08.
35. ^ Lander, E. S.; Dietrich W, Katz H, Lincoln SE, Shin HS, Friedman J, Dracopoli NC (1992-06). "A Genetic Map of the Mouse Suitable for Typing Intraspecific Crosses". *Genetics* **131**: 423–447. PMID 1353738. Retrieved on 2007-10-08.
36. ^ Weissenbach, Jean; Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M. (1992-10-29). "A second-generation linkage map of the human genome". *Nature* **359**: 794–801. doi:10.1038/359794a0. PMID 1436057 doi:10.1038/359794a0. Retrieved on 2007-10-08.
37. ^ Mathies, R. A.; Ju J, Ruan C, Fuller CW, Glazer AN (1995-05-09). "Fluorescence Energy Transfer Dye-Labeled Primers for DNA Sequencing and Analysis". *PNAS* **92**: 4347–4351. doi:10.1073/pnas.92.10.4347. PMID 7753809. Retrieved on 2007-10-08.
38. ^ Blattner, F. R.; Frederick R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, Ying Shao (1997-09-05). "The Complete Genome Sequence of Escherichia coli K-12". *Science* **277** (5331): 1453–1462. doi:10.1126/science.277.5331.1453. PMID 9278503 doi:10.1126/science.277.5331.1453. Retrieved on 2007-10-08.
39. ^ "Base-calling of automated sequencer traces using phred. II. Error probabilities" (1998-03). *Genome Reserarch* **8** (3): 186–94. PMID 9521922. Retrieved on 2007-10-08.
40. ^ The C. Elegans Sequencing Consortium,; The C. elegans Sequencing Consortium (1998-12-11). "Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology". *Science* **282** (5396): 2012–2018. doi:10.1126/science.282.5396.2012. PMID 9851916 doi:10.1126/science.282.5396.2012. Retrieved on 2007-10-08.
41. ^ Dunham, I.; Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, Ainscough R, Almeida JP, Babbage A, Bagguley C, Bailey J, Barlow K, Bates KN, Beasley O, Bird CP, Blakey S, Bridgeman AM, Buck D, Burgess J, Burrill WD, O'Brien KP, et al. (1999-12-02). "The DNA sequence of human chromosome 22". *Nature* **402** (6761): 489–495. doi:10.1038/990031. PMID 10591208. Retrieved on 2007-10-08.
42. ^ Hattori, M., A. Fujiyama, T. D. Taylor, H. Watanabe, T. Yada, H.-S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.-K. Choi, E. Soeda, M. Ohki, T. Takagi, Y. Sakaki; S. Taudienk, K. Blechschmidtk, A. Polleyk, U. Menzelk, J. Delabar, K. Kumpfk, R. Lehmannk, D. Patterson, K. Reichwaldk, A. Rumpk, M. Schillhabelk, A. Schudyk, W. Zimmermannk, A. Rosenthalk; J. KudohI, K. ShibuyaI, K. KawasakiI, S. AsakawaI, A. ShintaniI, T. SasakiI, K. NagamineI, S. MitsuyamaI, S. E. AntonarakisI, S. MinoshimaI, N. ShimizuI, G. Nordsiek, K. Hornischer, P. Brandt, M. Scharfe, O. SchoÈn, A. Desario, J. Reichelt, G. Kauer, H. Bloecker; J. Ramser, A. Beck, S. Klages, S. Hennig, L.

Riesselmann, E. Dagand, T. Haaf, S. Wehrmeyer, K. Borzym, K. Gardiner, D. Nizetickk, F. Francis, H. Lehrach, R. Reinhardt, and M.-L. Yaspo, (2000). The DNA sequence of human chromosome 21. Nature 405: 311-319.
43. ^ Lander, Eric S.; HGP consortium (2001-02-15). "Initial sequencing and analysis of the human genome". *Nature* **409** (6822): 860–921. doi:10.1038/35057062. PMID 11237011. Retrieved on 2007-10-08.
44. ^ Venter, J. C.; et al (2001-02-16). "The sequence of the human genome". *Science* **291** (5507): 1304–51. doi:10.1126/science.1058040. PMID 11181995 doi:10.1126/science.1058040. Retrieved on 2007-10-08.
45. ^ "Scientists encode first woman's genome", *msnbc.com*, Microsoft (2008-05-27). Retrieved on 2008-05-27.

[edit] External links

- DNA Sequencing: Dye Terminator Animation
- Archon Genomics X PRIZE - $10 million competition for fast and inexpensive sequencing technology

Shotgun Sequencing: http://en.wikipedia.org/wiki/Shotgun_sequencing

Shotgun sequencing

From Wikipedia, the free encyclopedia

Jump to: navigation, search

In genetics, **shotgun sequencing**, also known as **shotgun cloning**, is a method used for sequencing long DNA strands. It is named by analogy with the rapidly-expanding, quasi-random firing pattern of a shotgun.

Since the chain termination method of DNA sequencing can only be used for fairly short strands (100 to 1000 basepairs), longer sequences must be subdivided into smaller fragments, and subsequently re-assembled to give the overall sequence. Two principal methods are used for this: chromosome walking, which progresses through the entire strand, piece by piece, and shotgun sequencing, which is a faster but more complex process, and uses random fragments.

In shotgun sequencing [1] [2], DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain *reads*. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a contiguous sequence [1].

For example, consider the following two rounds of shotgun reads:

Strand	Sequence
Original	AGCATGCTGCAGTCATGCTTAGGCTA
First shotgun sequence	AGCATGCTGCAGTCATGCT------- ------------------TAGGCTA
Second shotgun sequence	AGCATG-------------------- ------CTGCAGTCATGCTTAGGCTA
Reconstruction	AGCATGCTGCAGTCATGCTTAGGCTA

In this extremely simplified example, none of the reads cover the full length of the original sequence, however, the four reads can be assembled into the original sequence using the overlap of their ends to align and order them. In reality, this process uses enormous amounts of information that are rife with ambiguities and sequencing errors. Assembly of complex genomes is additionally complicated by the great abundance of repetitive sequence, meaning similar short reads could come from completely different parts of the sequence.

Many overlapping reads for each segment of the original DNA are necessary to overcome these difficulties and accurately assemble the sequence. For example, to complete the Human Genome

Project, most of the human genome was sequenced at 12X or greater *coverage*; that is, each base in the final sequence was present, on average, in 12 reads. Even so, current methods have failed to isolate or assemble reliable sequence for approximately 1% of the (euchromatic) human genome.

[edit] Whole genome shotgun sequencing

Whole genome shotgun sequencing for small (4000 to 7000 basepair) genomes was already in use in 1979 [1] broader application benefited from pairwise end sequencing, known colloquially as *double-barrel shotgun sequencing*. As sequencing projects began to take on longer and more complicated DNAs, multiple groups began to realize that useful information could be obtained by sequencing both ends of a fragment of DNA. Although sequencing both ends of the same fragment and keeping track of the paired data was more cumbersome than sequencing a single end of two distinct fragments, the knowledge that the two sequences were oriented in opposite directions and were about the length of a fragment apart from each other was valuable in reconstructing the sequence of the original target fragment. The first published description of the use of paired ends was in 1990 [3] as part of the sequencing of the human HPRT locus, although the use of paired ends was limited to closing gaps after the application of a traditional shotgun sequencing approach. The first theoretical description of a pure pairwise end sequencing strategy, assuming fragments of constant length, was in 1991[4]. At the time, there was community consensus that the optimal fragment length for pairwise end sequencing would be three times the sequence read length. In 1995 Roach et al.[5] introduced the innovation of using fragments of varying sizes, and demonstrated that a pure pairwise end-sequencing strategy would be possible on large targets. The strategy was subsequently adopted by The Institute for Genomic Research (TIGR) to sequence the genome of the bacterium *Haemophilus influenzae* in 1995 [6] , and then by Celera Genomics to sequence the fruit fly genome in 2000 [7], and subsequently the human genome.

To apply the strategy, high-molecular-weight DNA is sheared into random fragments, size-selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends using the chain termination method yielding two short sequences. Each sequence is called an *end-read* or *read* and two reads from the same clone are referred to as *mate pairs*. Since the chain termination method usually can only produce reads between 500 and 1000 bases long, in all but the smallest clones, mate pairs will rarely overlap.

The original sequence is reconstructed from the reads using sequence assembly software. First, overlapping reads are collected into longer composite sequences known as *contigs*. Contigs can be linked together into *scaffolds* by following connections between mate pairs. The distance between contigs can be inferred from the mate pair positions if the average fragment length of the library is known and has a narrow window of deviation.

Coverage is the average number of reads representing a given nucleotide in the reconstructed sequence. It can be calculated from the length of the original genome (G), the number of reads(N), and the average read length(L) as $\frac{NL}{G}$. For example, a hypothetical genome with

2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2x coverage.

Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than more traditional approaches. Detractors argue that although the technique quickly sequences large regions of DNA, its ability to correctly link these regions is suspect, particularly for genomes with repeating regions. As sequence assembly programs become more sophisticated and computing power becomes cheaper, it may be possible to overcome this limitation[*citation needed*].

[edit] References

1. ^ [a] [b] [c] Staden, R (1979). "A strategy of DNA sequencing employing computer programs.". *Nucleic Acids Research* **6** (7): 2601-10. doi:10.1093/nar/9.13.3015. PMID 461197.
2. ^ Anderson, S (1981). "Shotgun DNA sequencing using cloned DNase I-generated fragments". *Nucleic Acids Research* **9** (13): 3015–27. doi:10.1093/nar/9.13.3015. PMID 6269069.
3. ^ Edwards, A; Caskey, T (1991). "Closure strategies for random DNA sequencing". *Methods: A Companion to Methods in Enzymology* **3** (1): 41–47. doi:10.1016/S1046-2023(05)80162-8.
4. ^ Edwards, A; Voss, H.; Rice, P.; Civitello, A.; Stegemann, J.; Schwager, C.; Zimmerman, J.; Erfle, H.; Caskey, T.; Ansorge, W. (1990). "Automated DNA sequencing of the human HPRT locus". *Genomics* **6**: 593–608. doi:10.1016/0888-7543(90)90493-E. PMID 2341149.
5. ^ Roach, JC; Boysen, C; Wang, K; Hood, L (1995). "Pairwise end sequencing: a unified approach to genomic mapping and sequencing". *Genomics* **26**: 345–353. doi:10.1016/0888-7543(95)80219-C. PMID 7601461.
6. ^ Fleischmann, RD; et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.". *Science* **269** (5223): 496–512. doi:10.1126/science.7542800. PMID 7542800.
7. ^ Adams, MD; et al. (2000). "The genome sequence of Drosophila melanogaster". *Science* **287** (5461): 2185–95. doi:10.1126/science.287.5461.2185. PMID 10731132.

- "Shotgun sequencing comes of age". *The Scientist*. Retrieved on December 31, 2002.
- "Shotgun sequencing finds nanoorganisms - Probe of acid mine drainage turns up unsuspected virus-sized Archaea". *SpaceRef.com*. Retrieved on December 23, 2006.

[edit] External links

This article contains material text from the NCBI Handbook published by the National Center for Biotechnology Information, which, as a US government publication, is in the public domain.

Hidden categories: All articles with unsourced statements | Articles with unsourced statements since February 2007

Views

- Article
- Discussion
- Edit this page
- History

Personal tools

- Log in / create account

Navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

Search

Go Search

Interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

Toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

Languages

- Deutsch
- 日本語
- 中文