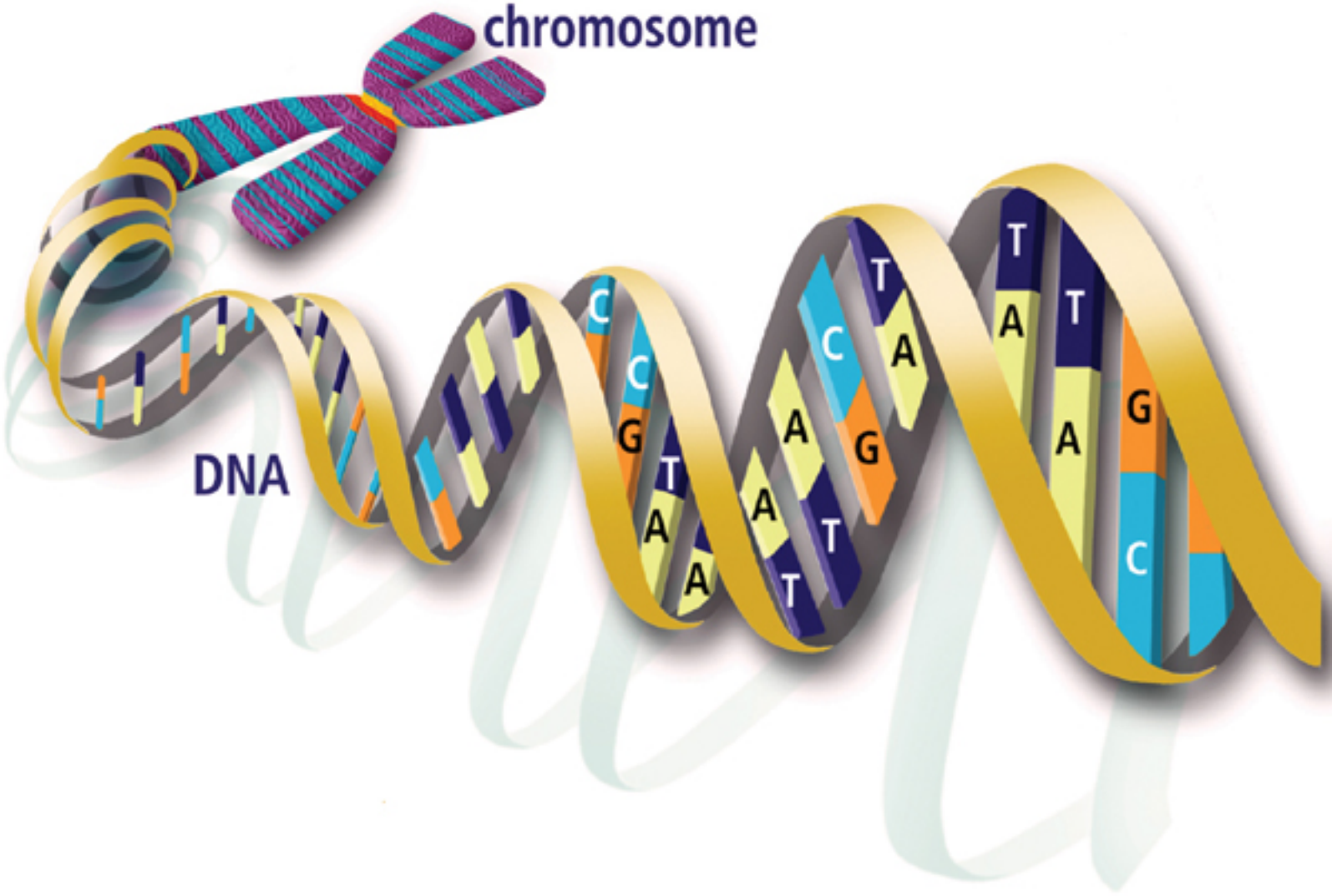
DNA Sequencing



Lucy Balderas

First Edition, 2012

ISBN 978-81-323-3272-5

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

DNA Sequencing

The term **DNA sequencing** refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics. 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 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.

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 dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster.

History

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), between 1972 and 1976.

Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.

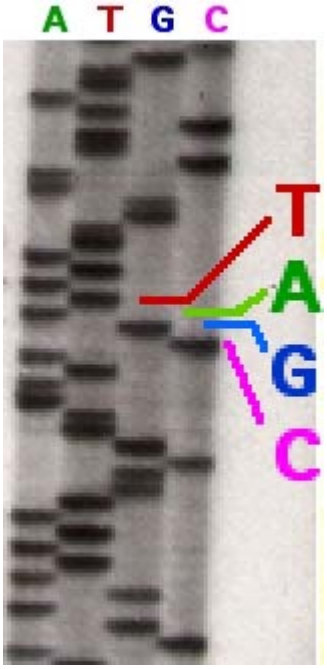
Maxam–Gilbert sequencing

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. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, 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 improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-^{32}P ATP) 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). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.

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

Chain-termination methods



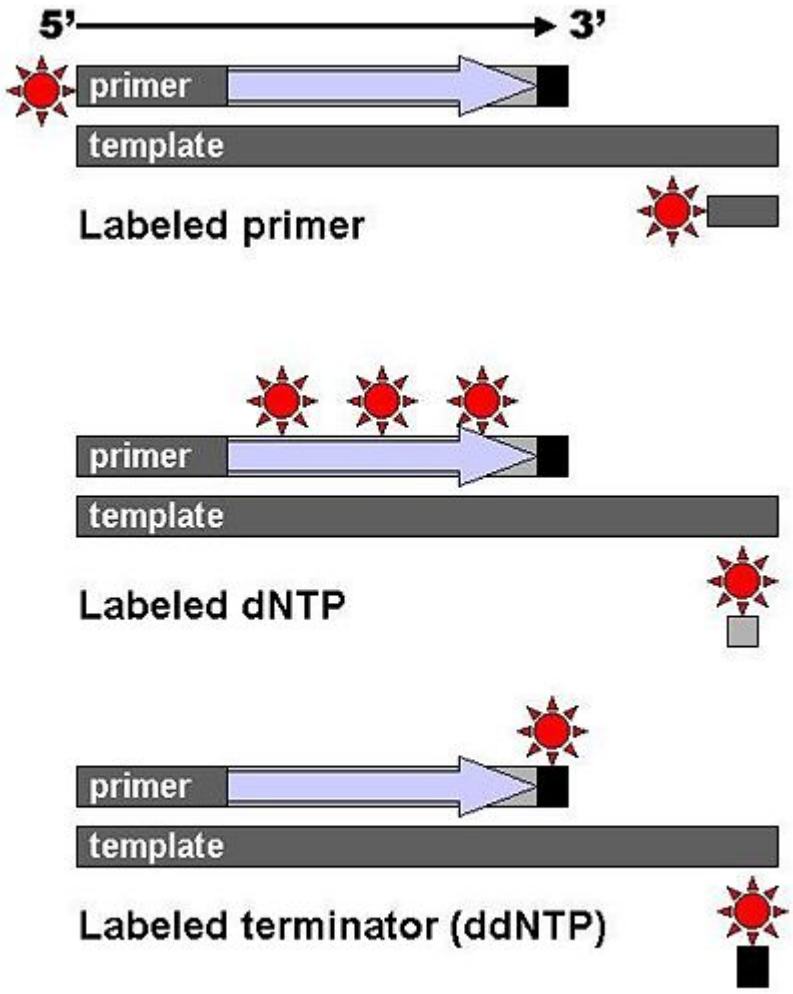
Part of a radioactively labelled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidephosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines. 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) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

The newly synthesized and labelled 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 with each of the four reactions 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

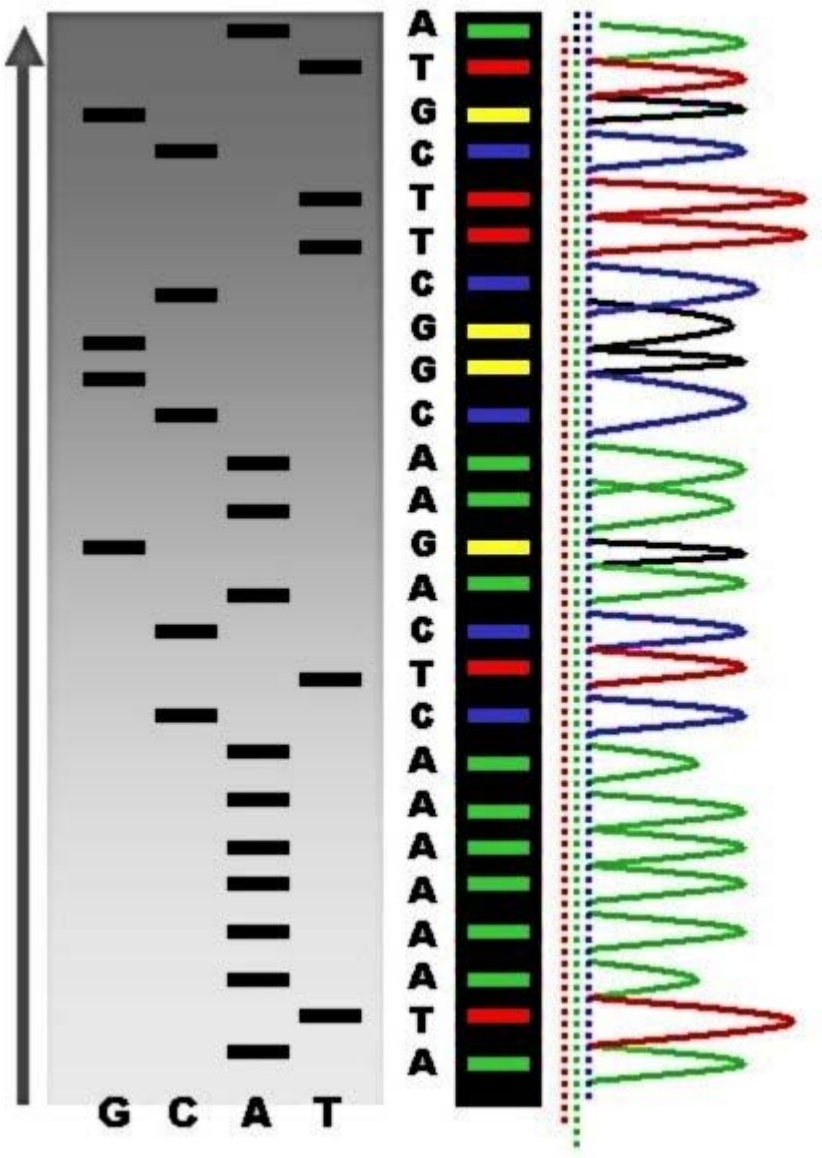
DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development

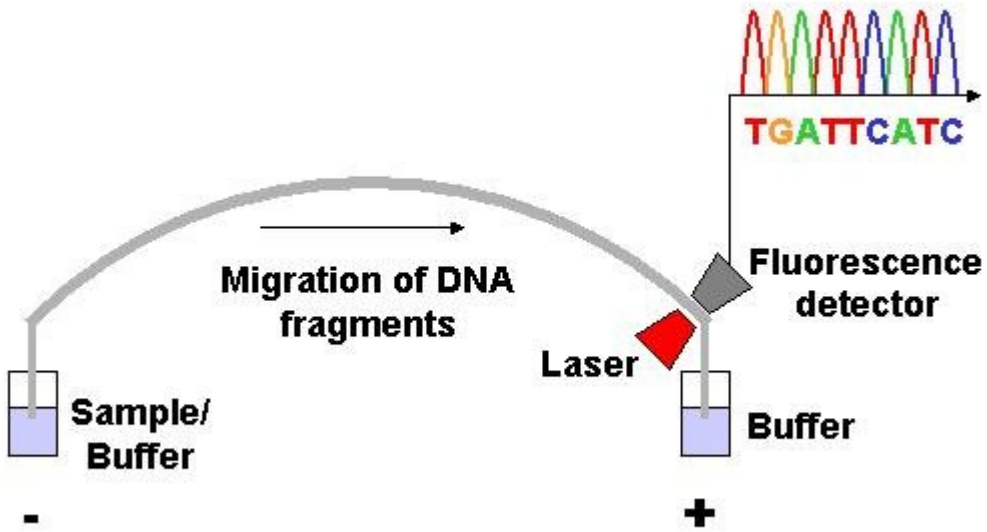
by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing 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 fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its 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.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

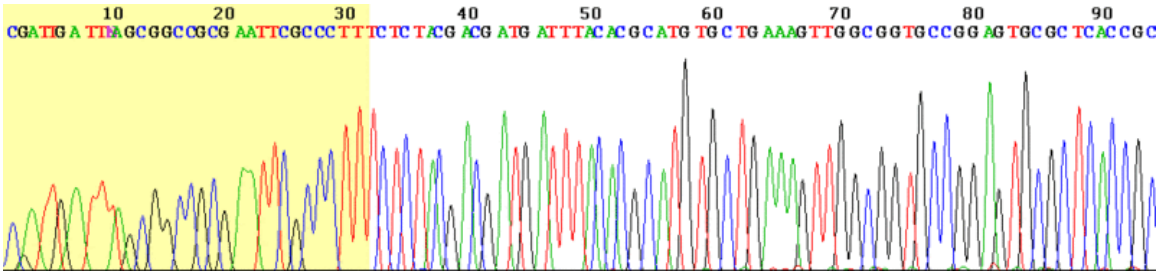
Challenges

Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. 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. In all cases the use of a primer with a free 5' end is essential.

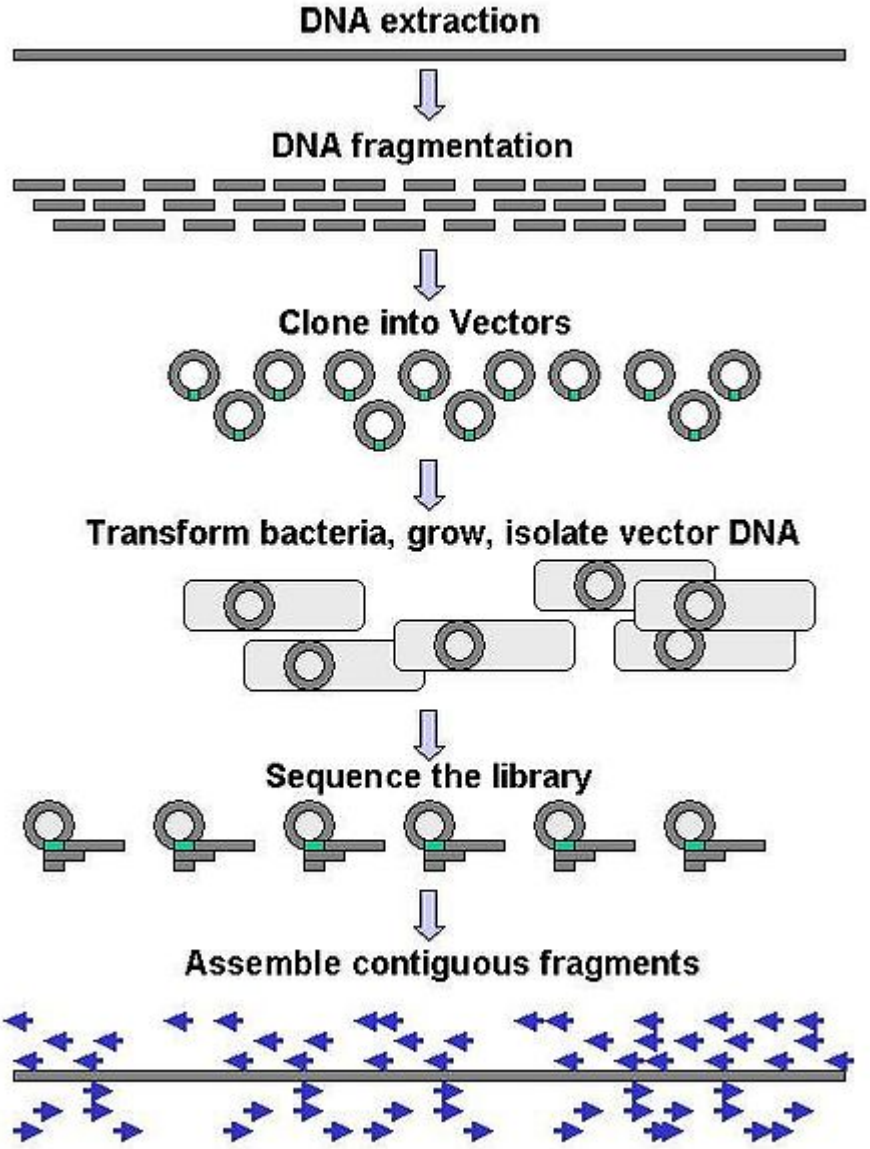
Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size separation, 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. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Amplification and clonal selection



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 DNA regions.

Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common 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, and amplified in *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence.

This method does not require any pre-existing information about the sequence of the DNA and is referred to as *de novo* sequencing. Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; *shotgun methods* are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. Polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, now Applied Biosystems).

Another method for *in vitro* clonal amplification is *bridge PCR*, where fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. The single-molecule method developed by Stephen Quake's laboratory (later commercialized by Helicos) is an exception: it uses bright fluorophores and laser excitation to detect pyrosequencing events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.

High-throughput sequencing

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)

The first of the "next-generation" sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992 by Sidney Brenner and Sam Eletr. MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides; this method made it susceptible to sequence-specific bias or loss of specific sequences. Because the technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no machines were sold; when the merger with Solexa later lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete. However, the essential properties of the MPSS output were typical of later "next-gen" data types, including hundreds of thousands of short DNA sequences. In the case of MPSS, these were typically used for sequencing

cDNA for measurements of gene expression levels. Lynx Therapeutics merged with Solexa in 2004, and this company was later purchased by Illumina.

Polony Sequencing

Polony sequencing, developed in George Church's lab at Harvard, was among the first next-generation sequencing systems used to sequence a full genome in 2005. It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an E. coli genome at an accuracy of > 99.9999% and a cost approximately 1/10th that of Sanger sequencing. The technology was licensed to Agencourt Biosciences, subsequently spun out into Agencourt Personal Genomics, and ultimately incorporated into the Applied Biosystems SOLiD platform.

454 pyrosequencing

A parallelized version of pyrosequencing was developed by 454 Life Sciences. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs. This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other. 454 Life Sciences has since been acquired by Roche Diagnostics.

Illumina (Solexa) sequencing

Solexa, now part of Illumina developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (bridge amplification). Four types of ddNTPs are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time. A camera takes images of the fluorescently labeled nucleotides then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing a next cycle.

SOLiD sequencing

Applied Biosystems' SOLiD technology employs sequencing by ligation. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

Future methods

Sequencing by hybridization is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, 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.

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost. In some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective.

In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called 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."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies

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.
- 1977 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". Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors".

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

- 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.

- 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 US$0.75/base).

- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish 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 marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

- 1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing

- 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.

- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

- 2001 A draft sequence of the human genome is published.

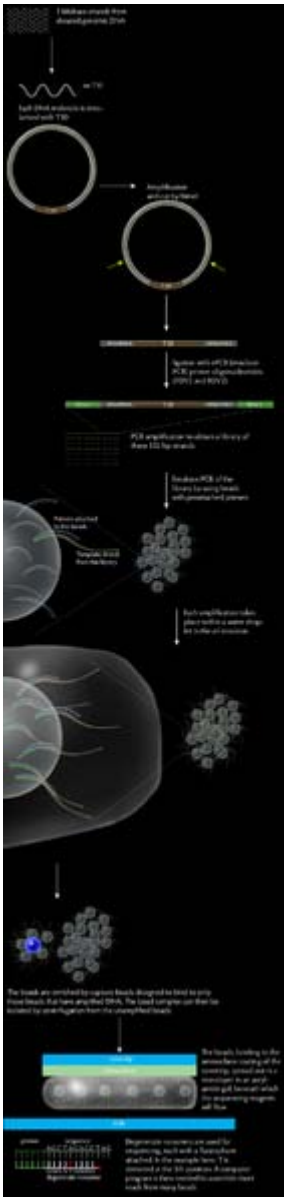
- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.

Chapter- 2

Polony Sequencing

Polony Sequencing is an inexpensive but highly accurate multiplex sequencing technique that can be used to “read” millions of immobilized DNA sequences in parallel. This technique was first developed by Dr. George Church group in Harvard Medical School. Unlike other sequencing technique, **Polony sequencing** technology is an open platform with freely downloadable, open source software and protocols. Also, the hardware of this technique can be easily set up with a commonly available epifluorescence microscope and a computer-controlled flowcell/ fluidics system. Polony sequencing is generally performed on paired-end Tags library that each molecule of DNA template is of 135bp in length with two 17-18bp paired genomic tags separated and flanked by common sequences. The current read length of this technique is 26 bases per amplicon and 13 bases per tag, leaving a 4-5 bases gap in each tag.

Workflow



An illustrated procedure for Polony sequencing.

The protocol of Polony sequencing can be break into three main parts which are the paired end-tag library construction, template amplification and DNA sequencing.

Paired end-tag library construction

This protocol begins by randomly shearing the tested genomic DNA into a tight size distribution. The sheared DNA molecules are then subjected for the end repair and A-tailed treatment. The end repair treatment converts any damaged or incompatible protruding ends of DNA to 5'-phosphorylated and blunt-ended DNA, enabling immediate blunt-end ligation. While the A-tailing treatment adds an A to the 3' end of the sheared DNA. DNA molecules with a length of 1kb are selected by loading on the 6% TBE

PAGE gel. Next step, the DNA molecules are circularized with T-tailed 30bp long synthetic oligonucleotides (T30), which contains two outward-facing MmeI recognition sites, and the resulting circularized DNA undergoes rolling circle replication. The amplified circularized DNA molecules are then digested with MmeI (type IIs restriction endonuclease) which will cuts at a distance from its recognition site, releasing the T30 fragment flanked by 17-18 bp tags (~70 bp in length). The paired-tag molecules need to be end-repaired prior to the ligation of ePCR (emulsion PCR) primer oligonucleotides (FDV2 and RDV2) to their both ends. The resulting 135 bp library molecules are size-selected and nick translated. Lastly, amplify the 135bp paired end-tag library molecules with PCR to increase the amount of library material and eliminate extraneous ligation products in a single step. The resulted DNA template consists a 44bp FDV sequence, a 17-18 bp proximal tag, the T30 sequence, a 17-18 bp distal tag, and a 25 bp RDV sequence.

Template amplification

• **Emulsion PCR** The Monosized, paramagnetic streptavidin –coated beads are pre-loaded with dual biotin forward primer. Streptavidin has a very strong affinity for biotin, thus the forward primer will bind firmly on the surface of the beads. Next, an aqueous phase is prepared with the pre-loaded beads, PCR mixture, forward and reverse primers, and the paired end-tag library. This is mixed and vortexed with an oil phase to create the emulsion. Ideally, each droplet of water in the oil emulsion has one bead and one molecule of template DNA, permitting millions of non-interacting amplification within a milliliter-scale volume by performing PCR.

• **Emulsion breaking** After amplification, the emulsion from preceding step is broken using isopropanol and detergent buffer (10mM Tris pH 7.5, 1mM EDTA pH 8.0, 100mM NaCl, 1% (v/v) Triton X‐100, 1% (w/v) SDS), following with a series of vortexing, centrifuging, and magnetic separation. The resulted solution is a suspension of empty, clonal and non-clonal beads, which arise from emulsion droplets that initially have zero, one or multiple DNA template molecules, respectively. The amplified bead could be enriched in the following step.

• **Bead enrichment** The enrichment of amplified beads is achieved through hybridization to a larger, low density, non-magnetic polystyrene beads that pre-loaded with a biotinylated capture oligonucleotides (DNA sequence that complementary to ePCR amplicon sequence). The mixture is then centrifuged to separate the amplified and capture beads complex from the unamplified beads. The amplified, capture bead complex has a lower density and thus will remain in the supernatant while the unamplified beads form a pellet. The supernatant is recovered and treated with NaOH which will break the complex. The paramagnetic amplified beads are separated from the non-magnetic capture beads by magnetic separation. This enrichment protocol is capable in enriching five times of amplified beads.

• **Bead capping** The purpose of bead capping is to attach a “capping” oligonucleotide to the 3’ end of both unextended forward ePCR primers and the RDV segment of template DNA. The cap that being use is an amino group that prevents fluorescent probes from ligating to these ends and at the same time, helping the subsequent coupling of template DNA to the aminosilanated flow cell coverslip.

• **Coverslip arraying** First, the coverslips are washed and aminosilane-treated, enabling

the subsequent covalent coupling of template DNA on it and eliminating any fluorescent contamination. The amplified, enriched beads are mixed with acrylamide and poured into a shallow mold formed by a Teflon-masked microscope slide. Immediately, place the aminosilane-treated coverslip on top of the acrylamide gel and allow to polymerize for 45 minutes. Next, invert the slide/coverslip stack and remove the microscope slide from gel. The silane treated coverslips will bind covalently to the gel while the Teflon on the surface of microscope slide will enable the better removal of slide from the acrylamide gel. The coverslips then bonded to the flow cell body and any unattached beads will be removed.

DNA sequencing
The biochemistry of Polony sequencing mainly rely on the discriminatory capacities of polymerases and ligases. First, a series of anchor primers are flowed through the cells and hybridize to the synthetic oligonucleotide sequences at the immediate 3' or 5' end of the 17-18bp proximal or distal genomic DNA tags. Next, an enzymatic ligation reaction of the anchor primer to a population of degenerate nonamers that are labeled with fluorescent dyes is performed.
Differentially labeled nonamers:

```
5' Cy5 - NNNNNNNNT
5' Cy3 - NNNNNNNNA
5' TexasRed - NNNNNNNNC
5' 6FAM - NNNNNNNNG
```

The fluorophore-tagged nonamers selectively ligate onto the anchor primer, providing a fluorescent signal that indicates whether there is an A, C, G, or T at the query position on the genomic DNA tag. After four colour imaging, the anchor primer/nonamer complexes are stripped off and a new cycle is begun by replacing the anchor primer. A new mixture of the fluorescently tagged nonamers is introduced, for which the query position is shifted one base further into the genomic DNA tag.

```
5' Cy5 - NNNNNNNTN
5' Cy3 - NNNNNNNAN
5' TexasRed - NNNNNNNCN
5' 6FAM - NNNNNNNGN
```

Seven bases from the 5' to 3' direction and six bases from the 3' end could be queried in this fashion. The ultimate result is a read length of 26 bases per run (13 bases from each of the paired tags) with a 4 to 5 bases gap in the middle of each tag.

Analysis and software

The polony sequencing generates millions of 26 reads per run and this information needed to be normalized and converted to sequence. These can be done by the software that has been developed by Church Lab. All of the software is free and could be downloaded from the website.

Strength and Weaknesses

Polony sequencing allows for a high throughput and high consensus accuracies of DNA sequencing based on a commonly available, inexpensive instrument. Also, it is a very flexible technique that enables variable application including BAC (bacterial artificial chromosome) and bacterial genome resequencing, as well as SAGE (serial analysis of gene expression) tag and barcode sequencing. Furthermore, the polony sequencing technique is emphasized as an open system that shares everything including the software that have been developed, protocol and reagents.

However, although the raw data acquisition could be achieved as high as 786 gigabits but only 1 bit of information out of 10,000 bits collected is useful. Another challenge of this technique is the uniformity of the relative amplification of individual targets. The non-uniform amplification could lower the efficiency of sequencing and posted as the biggest obstacle in this technique.

Cost

The sequencing instrument used in this technique could be set up by the commonly available fluorescence microscope and a computer controlled flowcell. According to the calculation in the year of 2005, the set up of a complete set of instrument will cost around US$ 130,000. However, the cost could be further lower to US$ 100,000 in the near future. A biotech company, Dover, is in collaboration with the Church Laboratory of Harvard Medical School to produce an automated sequencing machine, Polonator G.007, based on polony sequencing technique. The current selling price of this machine is around US$ 170,000. According to the calculation in year 2005, every kilobase of generated raw sequence was estimated to $0.11, while omitting the paired-end tag library construction cost, the cost of every kilobase of sequence could drops to $0.08.

History

The polony sequencing is a “distant relative” of the classical polony technology which mainly developed by Dr. Rob Mitra. Together with a MD PhD student, Jay Shendure, Dr. Rob Mitra worked out ways to sequence in situ polonies using single-base extension which can achieved 5-6 bases reads. However, the existing polony sequencing technology was mainly developed by Jay Shendure and Greg Porreca. They have changed almost everything that was there in order to make this multiplex sequencing technology work.

Also, the highly parallel sequencing-by-ligation method of polony sequencing has contributed in forming the basis for ABI Solid Sequencing and others.

Chapter- 3

DNA Sequencing Theory

DNA sequencing theory is the broad body of work that attempts to lay analytical foundations for DNA sequencing. The practical aspects revolve around designing and optimizing sequencing projects, predicting project performance, troubleshooting experimental results, characterizing factors such as sequence bias and the effects of software processing algorithms, and comparing various sequencing methods to one another. In this sense, it could be considered a branch of systems engineering or operations research. The permanent archive of work is primarily mathematical, although numerical calculations are often conducted for particular problems too. DNA sequencing theory addresses *physical processes* related to sequencing DNA and should not be confused with theories of analyzing resultant DNA sequences, e.g. sequence alignment. Publications sometimes do not make a careful distinction, but the latter are primarily concerned with algorithmic issues.

Sequencing as a covering problem

All mainstream methods of DNA sequencing rely on reading small fragments of DNA and subsequently reconstructing these data to infer the original DNA target, either via assembly or alignment to a reference. The abstraction common to these methods is that of a mathematical covering problem. For example, one can imagine a line segment representing the target and a subsequent process where smaller segments are "dropped" onto random locations of the target. The target is considered "sequenced" when adequate coverage accumulates, for example when no gaps remain.

The abstract properties of covering have been studied by mathematicians for over a century. However, direct application of these results has not generally been possible. Closed-form mathematical solutions, especially for probability distributions, often cannot be readily evaluated. That is, they involve inordinately large amounts of computer time for parameters characteristic of DNA sequencing. Stevens' configuration is one such example. Results obtained from the perspective of pure mathematics also do not account for factors that are actually important in sequencing, for instance detectable overlap in sequencing fragments, double-stranding, edge-effects, and target multiplicity. Consequently, development of sequencing theory has proceeded more according to the

philosophy of applied mathematics. In particular, it has been problem-focused and makes expedient use of approximations, simulations, etc.

Early uses derived From elementary probability theory

The earliest result was actually borrowed directly from elementary probability theory. If we model the above process and take *L* and *G* as the fragment length and target length, respectively, then the probability of "covering" any given location on the target *with one particular fragment* is *L* / *G*. Note that this presumes $L \ll G$, which is valid for many, though not all sequencing scenarios. Utilizing concepts from the binomial distribution, it can then be shown that the probability that the location is covered by at least one of *N* fragments is

$$P = 1 - \left[1 - \frac{L}{G}\right]^N.$$

This equation was first used to characterize plasmid libraries, but is often more useful in a modified form. For most projects $N \gg 1$, so that, to a good degree of approximation

$$\left[1 - \frac{L}{G}\right]^N \sim \exp(-NL/G),$$

where $R = NL / G$ is called the *redundancy*. Note the significance of redundancy as representing the average number of times a position is covered with fragments. Note also that in considering the covering process over all positions in the target, this probability is identical to the expected value of the random variable *C*, which represents the fraction of the target coverage. The final result,

$$E\langle C\rangle = 1 - e^{-R},$$

remains in widespread use as a "back of the envelope" estimator and predicts that coverage for all projects evolves along a universal curve that is a function only of the redundancy.

Lander-Waterman theory

In 1988, Eric Lander and Michael Waterman published an important paper examining the covering problem from the standpoint of gaps. Although they focused on the so-called mapping problem, the abstraction to sequencing is much the same. They furnished a number of useful results that were adopted as the standard theory from the earliest days of "large-scale" genome sequencing. Their model was also used in designing the Human Genome Project and continues to play an important role in DNA sequencing.

Ultimately, the main goal of a sequencing project is to close all gaps, so the "gap perspective" was a logical basis of developing a sequencing model. One of the more

frequently used results from this model is the expected number of contigs, given the number of fragments sequenced. If one neglects the amount of sequence that is essentially "wasted" by having to detect overlaps, their theory yields

$$E\langle contigs\rangle = Ne^{-R}.$$

In 1995, Roach proposed a model that appeared to be essentially different and asserted that Lander-Waterman theory gave contradictory results for large values of R. Wendl and Waterston later showed, based on Stevens' method, that subtle differences in interpretation explained the anomalies and that both models were indeed essentially identical and consistent.

The basic ideas of Lander-Waterman theory led to a number of additional results for particular variations in mapping techniques. However, technological advancements have rendered mapping and its more esoteric theories largely obsolete.

Recent advancements

The physical processes and protocols of DNA sequencing have continued to evolve, largely driven by advancements in bio-chemical methods, hardware, and automation techniques. There is now a wide range of problems that DNA sequencing has made in-roads into, including metagenomics and medical (cancer) sequencing. There are important factors in these scenarios that classical theory does not account for. Recent work has begun to focus on resolving the effects of some of these issues. The level of mathematics becomes commensurately more sophisticated.

Multiplicity

Biologists have developed methods to filter highly-repetitive, essentially un-sequenceable regions of genomes. These procedures are important for organisms whose genomes consist mostly of such DNA, for example corn. They yield multitudes of small islands of sequenceable DNA products. Wendl and Barbazuk proposed an extension to Lander-Waterman Theory to account for "gaps" in the target due to filtering and the so-called "edge-effect". The latter is a position-specific sampling bias, for example the terminal base position has only a $1 / G$ chance of being covered, as opposed to L / G for interior positions. For $R < 1$, classical Lander-Waterman Theory still gives good predictions, but dynamics change for higher redundancies.

Paired-end sequencing

Modern sequencing methods usually sequence both ends of a larger fragment, which provides linking information for *de novo* assembly and improved probabilities for alignment to reference sequence. Researchers generally believe that longer lengths of data (read lengths) enhance performance for very large DNA targets, an idea consistent with predictions from distribution models. However, Wendl showed that smaller fragments provide better coverage on small, linear targets because they reduce the edge

effect in linear molecules. These findings have implications for sequencing the products of DNA filtering procedures. Read-pairing and fragment size evidently have negligible influence for large, whole-genome class targets.

Diploid sequencing

Sequencing is emerging as an important tool in medicine, for example in cancer research. Here, the ability to detect heterozygous mutations is important and this can only be done if the sequence of the diploid genome is obtained. In the pioneering efforts to sequence individuals, Levy *et al.* and Wheeler *et al.*, who sequenced Craig Venter and Jim Watson, respectively, outlined models for covering both alleles in a genome. Wendl and Wilson followed with a more general theory that allowed for an arbitrary number of coverings of each allele and arbitrary ploidy. These results point to the general conclusion that the amount of data needed for such projects is significantly higher than for traditional haploid projects.

Limitations

DNA sequencing theories often invoke the assumption that certain random variables in a model are independently and identically distributed. For example, in Lander-Waterman Theory, a sequenced fragment is presumed to have the same probability of covering each region of a genome and all fragments are assumed to be independent of one another. In actuality, sequencing projects are subject to various types of bias, including differences of how well regions can be cloned, sequencing anomalies, biases in the target sequence (which is *not* random), and software-dependent errors and biases. In general, theory will agree well with observation up to the point that enough data have been generated to expose latent biases. The kinds of biases related to the underlying target sequence are particularly difficult to model, since the sequence itself may not be known *a priori*. This presents a type of "chicken and egg" closure problem.

Academic status

Sequencing theory is based on elements of mathematics, biology, and systems engineering, so it is highly interdisciplinary. Although many universities now have programs in computational biology, there does not yet seem to be a strong focus at the graduate level on this topic. Academic contributions have mainly been limited to a small number of PhD dissertations.

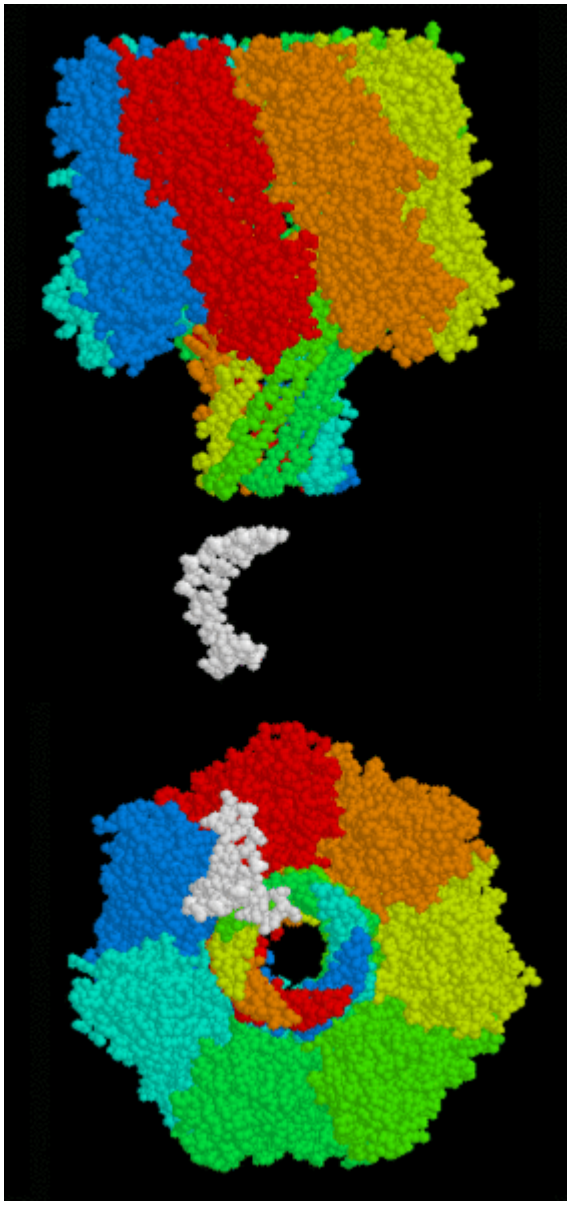
Chapter- 4

Nanopore Sequencing and Sequencing by Ligation

Nanopore sequencing

Nanopore sequencing is a method under development since 1995 for determining the order in which nucleotides occur on a strand of DNA.

A nanopore is simply a small hole, of the order of 1 nanometer in internal diameter. Certain transmembrane cellular proteins act as nanopores, and nanopores have also been made by etching a somewhat larger hole (several tens of nanometers) in a piece of silicon, and then gradually filling it in using ion-beam sculpting methods which results in a much smaller diameter hole: the nanopore.



alpha-hemolysin pore (made up of 7 identical subunits in 7 colors) and 12-mer single-stranded DNA (in white) on the same scale to illustrate DNA effects on conductance when moving through a nanopore. Below is an orthogonal view of the same molecules.

The theory behind nanopore sequencing has to do with what occurs when the nanopore is immersed in a conducting fluid and a potential (voltage) is applied across it: under these conditions a slight electric current due to conduction of ions through the nanopore can be observed, and the amount of current is very sensitive to the size and shape of the nanopore. If single nucleotides (bases) or strands of DNA pass through the nanopore, this can create a characteristic change in the magnitude of the current through the nanopore.

DNA could be passed through the nanopore for various reasons. For example, electrophoresis might attract the DNA towards the nanopore, and it might eventually pass

through it. Or, enzymes attached to the nanopore might guide DNA towards the nanopore. The scale of the nanopore means that the DNA may be forced through the hole as a long string, one base at a time, rather like thread through the eye of a needle. As it does so, each nucleotide on the DNA molecule may obstruct the nanopore to a different, characteristic degree. The amount of current which can pass through the nanopore at any given moment therefore varies depending on whether the nanopore is blocked by an A, a C, a G or a T. The change in the current through the nanopore as the DNA molecule passes through the nanopore represents a direct reading of the DNA sequence. Alternatively, a nanopore might be used to identify individual DNA bases as they pass through the nanopore in the correct order - this approach has been shown by Oxford Nanopore Technologies and Professor Hagan Bayley.

The potential is that a single molecule of DNA can be sequenced directly using a nanopore, without the need for an intervening PCR amplification step or a chemical labelling step or the need for optical instrumentation to identify the chemical label. As of July 2010, information available to the public indicates that nanopore sequencing is still in the development stage, with some laboratory-based data to back up the different components of the sequencing method, but not yet commercially available, parallelized, routineized, nor cost-effective enough yet to compete with out "next generation sequencing" methods. Nanopore-based DNA analysis techniques are being industrially developed by Oxford Nanopore Technologies (developing direct exonuclease sequencing and strand sequencing using protein nanopores, and solid-state sequencing through internal R&D and collaborations with academic institutions), NabSys (using a library of DNA probes and using nanopores to detect where these probes have hybridized to single stranded DNA) and NobleGen(using nanopores in combination with fluorescent labels). IBM has noted research projects on computer simulations of translocation of a DNA strand through a solid-state nanopore, but not projects on identifying the DNA bases on that strand.

One challenge for the 'strand sequencing' method is in refining the method to improve its resolution to be able to detect single bases. In the early papers methods, a nucleotide needed to be repeated in a sequence about 100 times successively in order to produce a measurable characteristic change. This low resolution is due to the fact that the DNA strand moves rapidly at the rate of 1 to 5μs per base through the nanopore. This makes recording difficult and prone to background noise, failing in obtaining single-nucleotide resolution. The problem is being tackled by either improving the recording technology or by controlling the speed of DNA strand by various protein engineering strategies. More recently effects of single bases due to secondary structure or released mononucleotides have been shown.. Professor Hagan Bayley, founder of Oxford Nanopore, recently proposed that creating two recognition sites within an alpha hemolysin pore may confer advantages in base recognition.

One challenge for the 'exonuclease approach', where a processive enzyme feeds individual bases, in the correct order, into the nanopore, is to integrate the exonuclease and the nanopore detection systems. In particular, the problem is that when an exonuclease hydrolyzes the phosphodieseter bonds between nucleotides in DNA, the

subsequentially released nucleotide is not necessarily guaranteed to directly move in to, say, a nearby alpha-hemolysin nanopore. One idea is to attach the exonuclease to the nanopore, perhaps through biotinylation to the beta barrel hemolsyin. The central pore of the protein may be lined with charged residues arranged so that the positive and negative charges appear on opposite sides of the pore. However, this mechanism is primarily discriminatory and does not constitute a mechanism to guide nucleotides down some particular path.

Commercialization

Agilent Laboratories was the first to license and develop nanopores but does not have any current disclosed research in the area.

The company Oxford Nanopore Technologies in 2008 licensed technology from Harvard, UCSC and other universities and is developing protein and solid state nanopore technology with the aim of sequencing DNA and identifying biomarkers, drugs of abuse and a range of other molecules.

Sequenom licensed nanopore technology from Harvard in 2007 using an approach that combines nanopores and fluorescent labels. This technology was subsequently licensed to Noblegen.

NABsys was spun out of Brown University and is researching nanopores as a method of identifying areas of single stranded DNA that have been hybridized with specific DNA probes.

Sequencing by ligation

Sequencing by ligation is a DNA sequencing method that uses the enzyme DNA ligase to identify the nucleotide present at a given position in a DNA sequence. Unlike most currently popular DNA sequencing methods, this method does not use a DNA polymerase to create a second strand. Instead, the mismatch sensitivity of a DNA ligase enzyme is used to determine the underlying sequence of the target DNA molecule.

Process

DNA ligase is an enzyme that joins together ends of DNA molecules. Although commonly represented as joining two pairs of ends at once, as in the ligation of restriction enzyme fragments, ligase can also join the ends on only one of the two strands (for example, when the other strand is already continuous or lacks a terminal phosphate necessary for ligation). DNA ligase is sensitive to the structure of DNA and has very low efficiency when there are mismatches between the bases of the two strands.

Sequencing by ligation relies upon the sensitivity of DNA ligase for base-pairing mismatches. The target molecule to be sequenced is a single strand of unknown DNA

sequence, flanked on at least one end by a known sequence. A short "anchor" strand is brought in to bind the known sequence.

A mixed pool of probe oligonucleotides is then brought in (eight or nine bases long), labeled (typically with fluorescent dyes) according to the position that will be sequenced. These molecules hybridize to the target DNA sequence, next to the anchor sequence, and DNA ligase preferentially joins the molecule to the anchor when its bases match the unknown DNA sequence. Based on the fluorescence produced by the molecule, one can infer the identity of the nucleotide at this position in the unknown sequence.

The oligonucleotide probes may also be constructed with cleavable linkages which can be cleaved after identifying the label. This will both remove the label and regenerate a 5' phosphate on the end of the ligated probe, preparing the system for another round of ligation. This cycle can be repeated several times to read longer sequences. This sequences every Nth base, where N is the length of the probe left behind after cleavage. To sequence the skipped positions, the anchor and ligated oligonucleotides may be stripped off the target DNA sequence, and another round of sequencing by ligation started with an anchor one or more bases shorter.

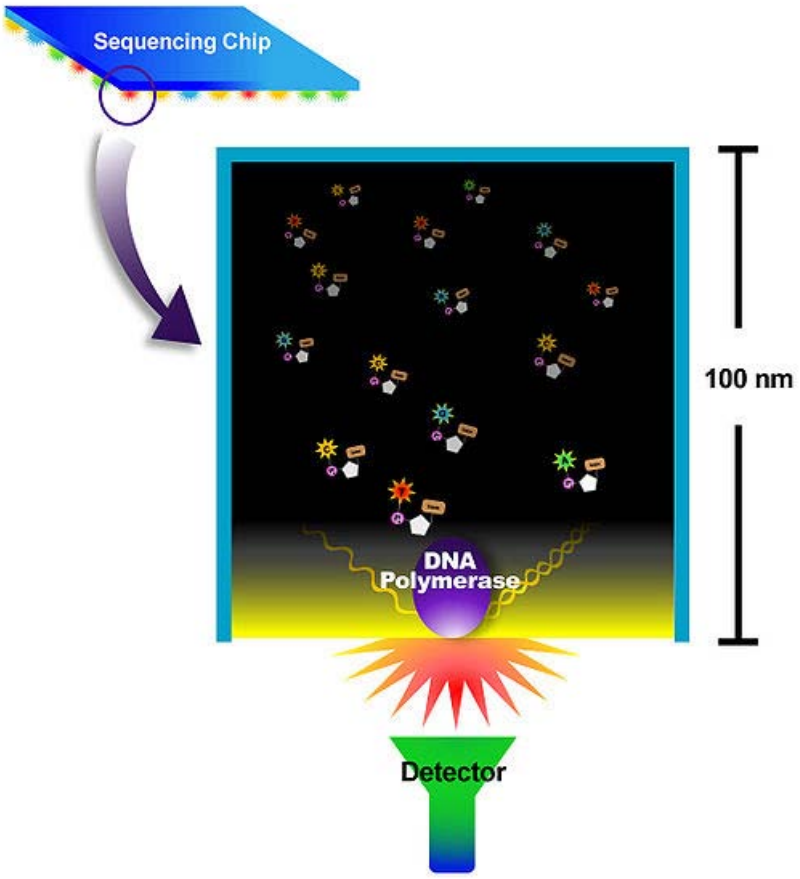
A simpler, albeit more limited, technique is to do repeated rounds of a single ligation where the label corresponds to different position in the probe, followed by stripping the anchor and ligated probe.

Sequencing by ligation can proceed in either direction (either 5'-3' or 3'-5') depending on which end of the probe oligonucleotides are blocked by the label. The 3'-5' direction is more efficient for doing multiple cycles of ligation. Note that this is the opposite direction to polymerase based sequencing methods.

Chapter- 5

Single Molecule Real Time Sequencing and Pyrosequencing

Single molecule real time sequencing



Overview of Single Molecule Real Time Sequencing

Single molecule real time sequencing (also known as SMRT) is a parallelized single molecule DNA sequencing by synthesis technology developed by Pacific Biosciences. Single molecule real time sequencing utilizes the zero-mode waveguide (ZMW), developed in the laboratories of Harold G. Craighead and Watt W. Webb at Cornell University. A single DNA polymerase enzyme is affixed at the bottom of a ZMW with a single molecule of DNA as a template. The ZMW is a structure that creates an illuminated observation volume that is small enough to observe only a single nucleotide of DNA (also known as a base) being incorporated by DNA polymerase. Each of the four DNA bases is attached to one of four different fluorescent dyes. When a nucleotide is incorporated by the DNA polymerase, the fluorescent tag is cleaved off and diffuses out of the observation area of the ZMW where its fluorescence is no longer observable. A detector detects the fluorescent signal of the nucleotide incorporation, and the base call is made according to the corresponding fluorescence of the dye.

Technology

The DNA sequencing is done on a chip that contains many ZMWs. Inside each ZMW, a single active DNA polymerase with a single molecule of single stranded DNA template is immobilized to the bottom through which light can penetrate and create a visualization chamber that allows monitoring of the activity of the DNA polymerase at a single molecule level. The signal from a phospho-linked nucleotide incorporated by the DNA polymerase is detected as the DNA synthesis proceeds which results in the DNA sequencing in real time.

Phospholinked nucleotide

For each of the nucleotide bases, there are four corresponding fluorescent dye molecules that enable the detector to identify the base being incorporated by the DNA polymerase as it performs the DNA synthesis. The fluorescent dye molecule is attached to the phosphate chain of the nucleotide. When the nucleotide is incorporated by the DNA polymerase, the fluorescent dye is cleaved off with the phosphate chain as a part of a natural DNA synthesis process during which a phosphodiester bond is created to elongate the DNA chain. The cleaved fluorescent dye molecule then diffuses out of the detection volume so that the fluorescent signal is no longer detected.

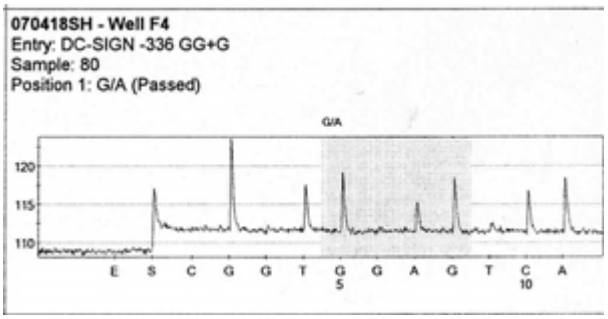
Zero-mode waveguide

The zero-mode waveguide (ZMW) is a nanophotonic confinement structure that consists of a circular hole in an aluminum cladding film deposited on a clear silica substrate. The ZMW holes are ~70 nm in diameter and ~100 nm in depth. Due to the behavior of light when it travels through a small aperture, the optical field decays exponentially inside the chamber. The observation volume within an illuminated ZMW is ~20 zeptoliters (20 X 10^{-21} liters). Within this volume, the activity of DNA polymerase incorporating a single nucleotide can be readily detected.

Sequencing performance

Pacific Biosciences expects to commercialize SMRT sequencing in 2010 or 2011. The prototype of the SMRT chip contains ~3000 ZMW holes that allow parallelized DNA sequencing. Each of the ZMW holes produces approximately 1,500 bp (base pair) read lengths at a speed of 10 bp per second.

Pyrosequencing



Example of a pyrogram showing the nucleotide sequence in a specific section of DNA. The tops represent light emission and nucleotide binding.

Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA) based on the "sequencing by synthesis" principle. It differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides. The technique was developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm in 1996.

Procedure

"Sequencing by synthesis" involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The Pyrosequencing method is based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemiluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step. The template DNA is immobile, and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template.

ssDNA template is hybridized to a sequencing primer and incubated with the enzymes DNA polymerase, ATP sulfurylase, luciferase and apyrase, and with the substrates adenosine 5´ phosphosulfate (APS) and luciferin.

1. The addition of one of the four deoxynucleotide triphosphates (dNTPs)(in the case of dATP we add dATPαS which is not a substrate for a luciferase) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi) stoichiometrically.
2. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5′ phosphosulfate. This ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a program.
3. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide.

Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA. As of 2007, pyrosequencing is most commonly used for resequencing or sequencing of genomes for which the sequence of a close relative is already available.

The templates for pyrosequencing can be made both by solid phase template preparation (streptavidin-coated magnetic beads) and enzymatic template preparation (apyrase+exonuclease).

Commercialization

The company **Pyrosequencing AB** in Uppsala, Sweden commercialized machinery and reagents for sequencing short stretches of DNA using the pyrosequencing technique. **Pyrosequencing AB** was renamed to **Biotage** in 2003 which was acquired by Qiagen in 2008. Pyrosequencing technology was further licensed to 454 Life Sciences. 454 developed an array-based pyrosequencing technology which has emerged as a platform for large-scale DNA sequencing. Most notable are the applications for genome sequencing and metagenomics. *GS FLX*, the latest pyrosequencing platform by 454 Life Sciences (now owned by Roche Diagnostics), can generate 400 million nucleotide data in a 10 hour run with a single machine. Each run would cost about 5,000-7,000 USD, with multiple fold coverage required for accuracy this pushes de novo sequencing of mammalian genomes into the million dollar range.

Chapter- 6

Nucleotide

Nucleotides are molecules that, when joined together, make up the structural units of RNA and DNA. In addition, nucleotides play central roles in metabolism. In that capacity, they serve as sources of chemical energy (adenosine triphosphate and guanosine triphosphate), participate in cellular signaling (cyclic guanosine monophosphate and cyclic adenosine monophosphate), and are incorporated into important cofactors of enzymatic reactions (coenzyme A, flavin adenine dinucleotide, flavin mononucleotide, and nicotinamide adenine dinucleotide phosphate).

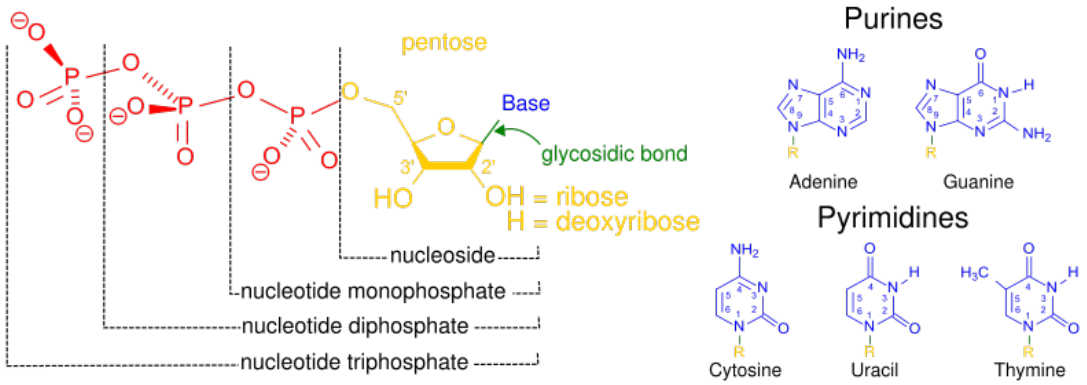


Figure 1: Structural elements of the most common nucleotides

Nucleotide structure

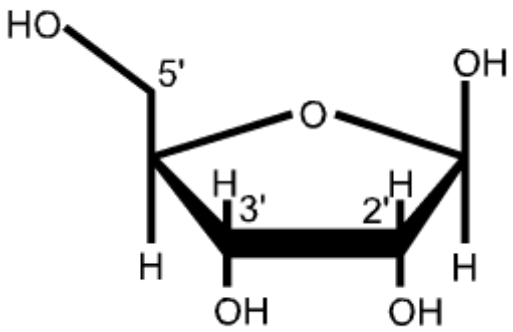


Figure 2: Ribose structure indicating numbering of carbon atoms

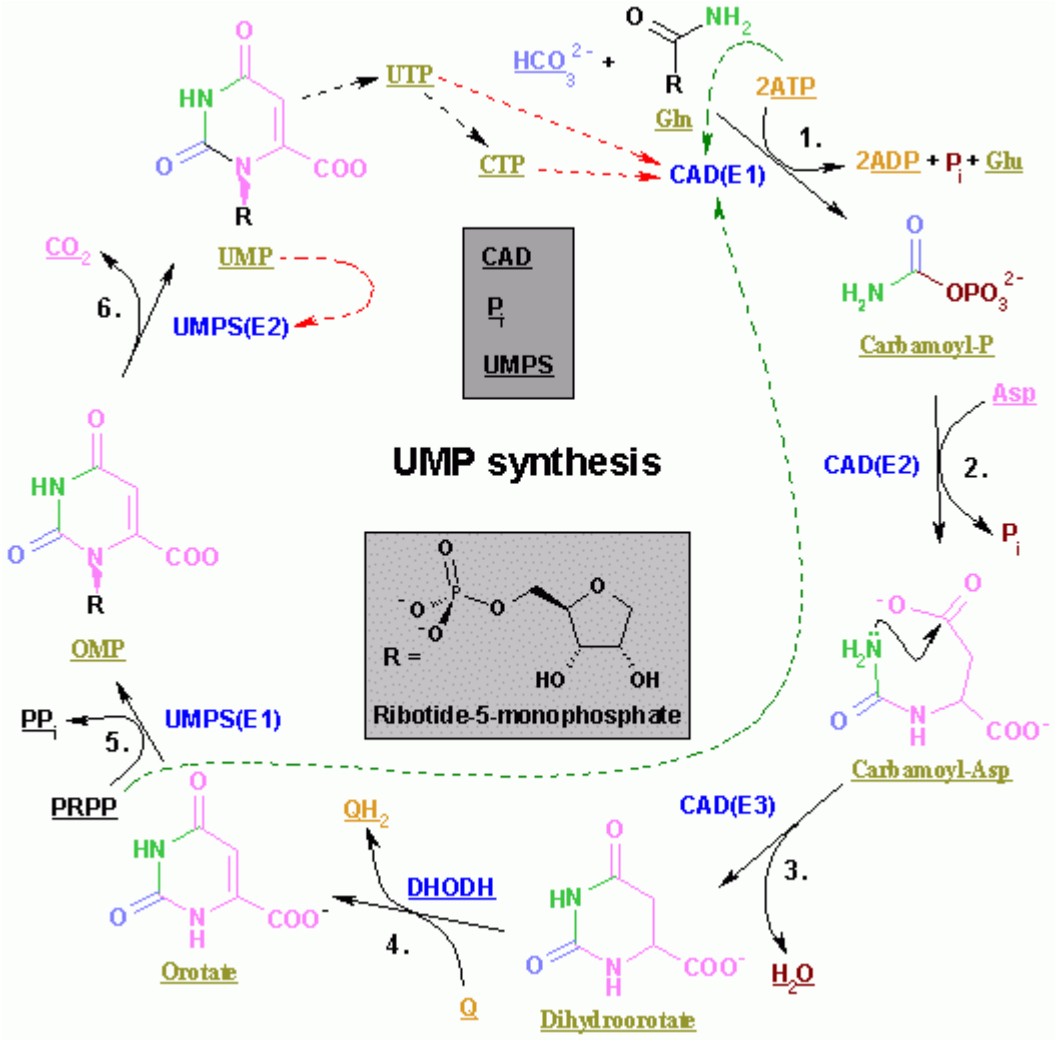
A nucleotide is composed of a nucleobase (nitrogenous base), a five-carbon sugar (either ribose or 2'-deoxyribose), and one to three phosphate groups. Together, the nucleobase and sugar comprise a nucleoside. The phosphate groups form bonds with either the 2, 3, or 5-carbon of the sugar, with the 5-carbon site most common. Cyclic nucleotides form when the phosphate group is bound to two of the sugar's hydroxyl groups. Ribonucleotides are nucleotides where the sugar is ribose, and deoxyribonucleotides contain the sugar deoxyribose. Nucleotides can contain either a purine or a pyrimidine base.

Nucleic acids are polymeric macromolecules made from nucleotide monomers. In DNA, the purine bases are adenine and guanine, while the pyrimidines are thymine and cytosine. RNA uses uracil in place of thymine. Adenine always pairs with thymine by 2 hydrogen bonds, while guanine pairs with cytosine through 3 hydrogen bonds, each due to their unique structures.

Synthesis

Nucleotides can be synthesized by a variety of means both in vitro and in vivo. In vivo, nucleotides can be synthesised de novo or recycled through salvage pathways. Nucleotides undergo breakdown such that useful parts can be reused in synthesis reactions to create new nucleotides. In vitro, protecting groups may be used during laboratory production of nucleotides. A purified nucleoside is protected to create a phosphoramidite, which can then be used to obtain analogues not found in nature and/or to synthesize an oligonucleotide.

Pyrimidine ribonucleotides

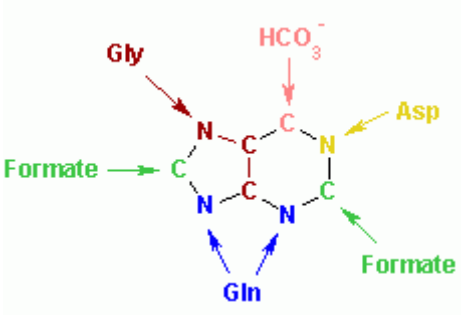


The synthesis of UMP.
The color scheme is as follows: **enzymes, coenzymes, substrate names, inorganic molecules**

Pyrimidine nucleotide synthesis starts with the formation of carbamoyl phosphate from glutamine and CO_2. The cyclisation reaction between carbamoyl phosphate reacts with aspartate, yielding orotate in subsequent steps. Orotate reacts with 5-phosphoribosyl α-diphosphate (PRPP), yielding orotidine monophosphate (OMP), which is decarboxylated to form uridine monophosphate (UMP). It is from UMP that other pyrimidine nucleotides are derived. UMP is phosphorylated to uridine triphosphate (UTP) via two sequential reactions with ATP. Cytidine monophosphate (CMP) is derived from conversion of UTP to cytidine triphosphate (CTP) with subsequent loss of two phosphates.

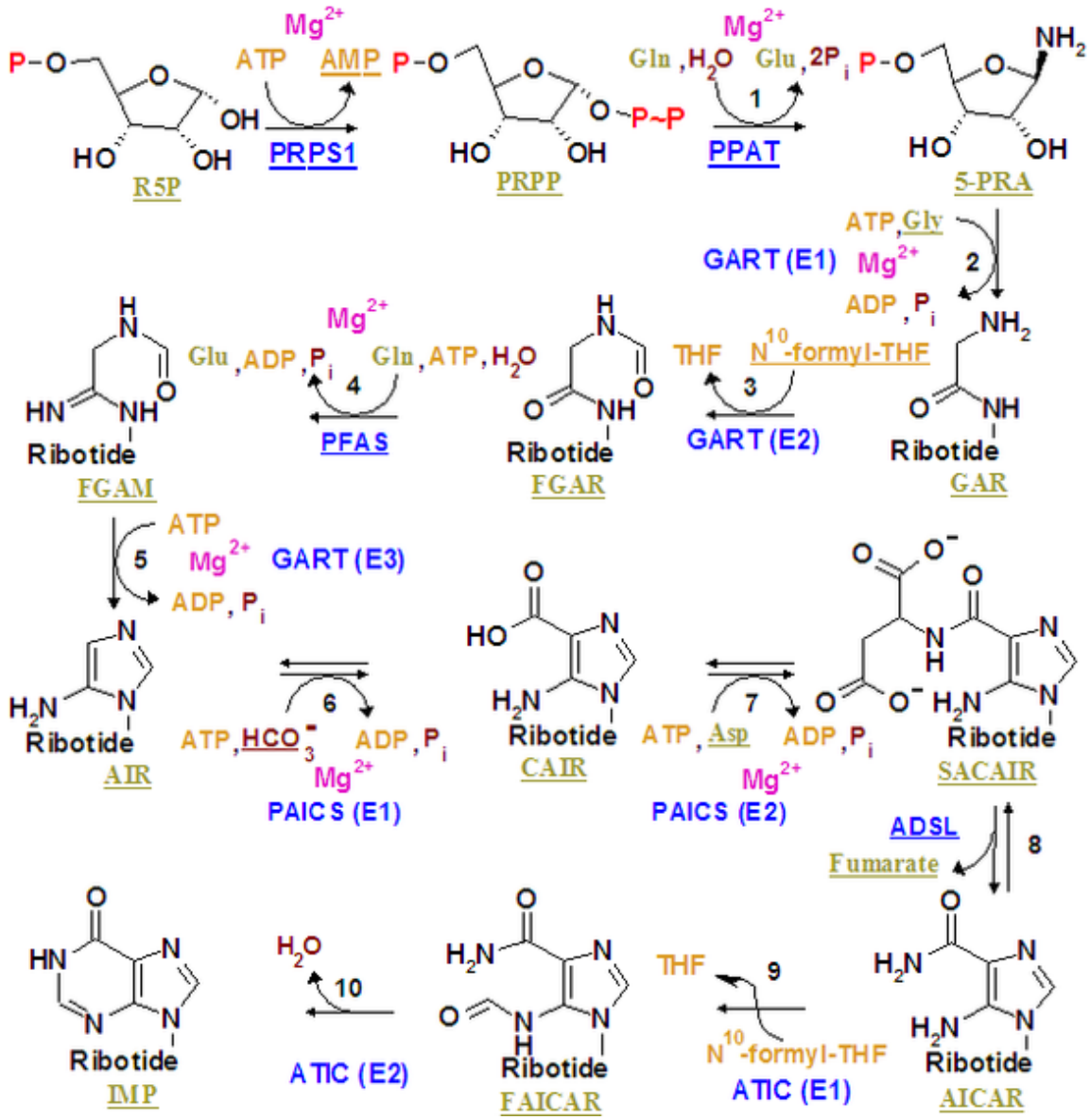
Purine ribonucleotides

The atoms which are used to build the purine nucleotides come from a variety of sources:



The biosynthetic origins of purine ring atoms

N1 arises from the amine group of Asp
C2 and C8 originate from formate
N3 and N9 are contributed by the amide group of Gln
C4, C5 and N7 are derived from Gly
C6 comes from HCO_3^- (CO_2)



The synthesis of IMP. The color scheme is as follows: **enzymes, coenzymes, substrate names, metal ions, inorganic molecules**

The de novo synthesis of purine nucleotides by which these precursors are incorporated into the purine ring proceeds by a 10-step pathway to the branch-point intermediate IMP, the nucleotide of the base hypoxanthine. AMP and GMP are subsequently synthesized from this intermediate via separate, two-step pathways. Thus, purine moieties are initially formed as part of the ribonucleotides rather than as free bases.

Six enzymes take part in IMP synthesis. Three of them are multifunctional:

- GART (reactions 2, 3, and 5)
- PAICS (reactions 6, and 7)
- ATIC (reactions 9, and 10)

Reaction 1. The pathway starts with the formation of PRPP. PRPS1 is the enzyme that activates R5P, which is formed primarily by the pentose phosphate pathway, to PRPP by reacting it with ATP. The reaction is unusual in that a pyrophosphoryl group is directly transferred from ATP to C1 of R5P and that the product has the **α** configuration about C1. This reaction is also shared with the pathways for the synthesis of the pyrimidine nucleotides, Trp, and His. As a result of being on (a) such (a) major metabolic crossroad and the use of energy, this reaction is highly regulated.

Reaction 2. In the first reaction unique to purine nucleotide biosynthesis, PPAT catalyzes the displacement of PRPP's pyrophosphate group (PP_i) by Gln's amide nitrogen. The reaction occurs with the inversion of configuration about ribose C1, thereby forming **β**-5-phosphorybosylamine (5-PRA) and establishing the anomeric form of the future nucleotide. This reaction, which is driven to completion by the subsequent hydrolysis of the released PP_i, is the pathway's flux-generating step and is therefore regulated, too.

Length unit

Nucleotide (abbreviated nt) is a common length unit for single-stranded RNA, similar to how base pair is a length unit for double-stranded DNA.

Abbreviation codes for degenerate bases

The IUPAC has designated the symbols for nucleotides. Apart from the five (A, G, C, T/U) bases, often degenerate bases are used especially for designing PCR primers. These nucleotide codes are listed here.

IUPAC nucleotide code	**Base**
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G

Y	C or T (U)
S	G or C
W	A or T (U)
K	G or T (U)
M	A or C
B	C or G or T (U)
D	A or G or T (U)
H	A or C or T (U)
V	A or C or G
N	any base
. or -	gap

Chapter- 7

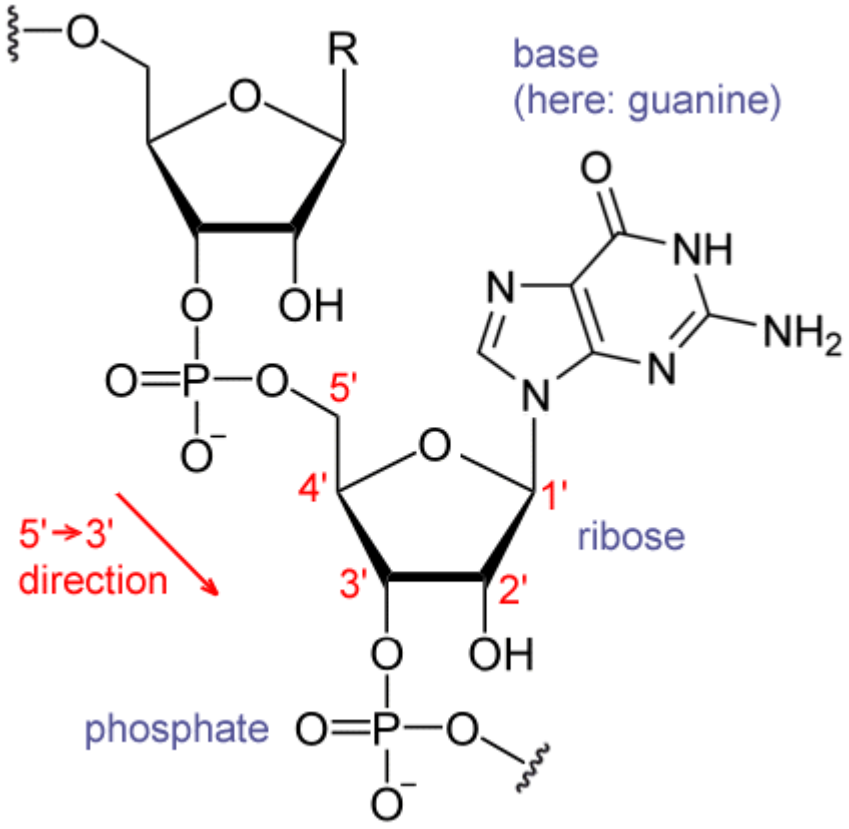
Nucleic Acid Sequence

The **sequence** or **primary structure of a nucleic acid** is the exact specification of its atomic composition and the chemical bonds connecting those atoms. As nucleic acids, e.g. DNA and RNA, are unbranched polymers, this is equivalent to specifying exact sequence of nucleotides that comprise the whole molecule. This sequence is written as a succession of letters representing a real or hypothetical DNA molecule or strand. By convention, the primary structure of a DNA or RNA molecule is reported from the 5' end to the 3' end.

The sequence has capacity to carry information. When used in reference to biological DNA, which carries the information which directs the functions of living beings, the term **genetic sequence** is often used. Sequences can be read from the biological raw material through DNA sequencing methods.

Primary structure is sometimes mistakenly termed *primary sequence*, but there is no such term, as well as no parallel concept of secondary or tertiary sequence.

Nucleotides



Chemical structure of RNA

Nucleic acids consist of a chain of linked units called nucleotides. Each nucleotide consists of three subunits: a phosphate group and a sugar (ribose in the case of RNA, deoxyribose in DNA) make up the backbone of the nucleic acid strand, and attached to the sugar is one of a set of nucleobases. The nucleobases are important in base pairing of strands to form higher-level secondary and tertiary structure such as the famed double helix.

The possible letters are *A*, *C*, *G*, and *T*, representing the four nucleotide bases of a DNA strand — adenine, cytosine, guanine, thymine — covalently linked to a phosphodiester backbone. In the typical case, the sequences are printed abutting one another without gaps, as in the sequence AAAGTCTGAC, read left to right in the 5' to 3' direction. With regards to transcription, a sequence is on the coding strand if it has the same order as the transcribed RNA.

One sequence can be complementary to another sequence, meaning that they have the base on each position is the complementary (i.e. A to T, C to G) and in the reverse order. For example, the complementary sequence to TTAC is GTAA. If one strand of the double-stranded DNA is considered the sense strand, then the other strand, considered the antisense strand, will have the complementary sequence to the sense strand.

Notation

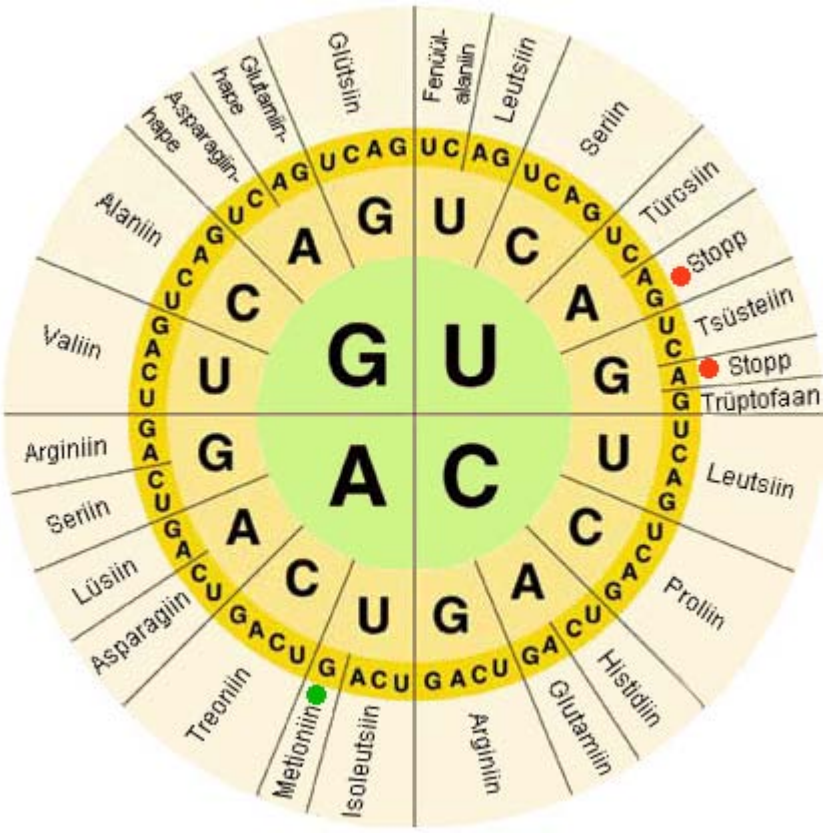
While A, T, C, and G represent a particular nucleotide at a position, there are also letters that represent ambiguity. Of all the molecules sampled, there is more than one kind of nucleotide at that position. The rules of the International Union of Pure and Applied Chemistry (IUPAC) are as follows:

- **A** = adenine
- **C** = cytosine
- **G** = guanine
- **T** = thymine
- **R** = G A (purine)
- **Y** = T C (pyrimidine)
- **K** = G T (keto)
- **M** = A C (amino)
- **S** = G C (strong bonds)
- **W** = A T (weak bonds)
- **B** = G T C (all but A)
- **D** = G A T (all but C)
- **H** = A C T (all but G)
- **V** = G C A (all but T)
- **N** = A G C T (any)

These symbols are also valid for RNA, except with U (uracil) replacing T (thymine).

Apart from adenine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), DNA and RNA also contain bases that have been modified after the nucleic acid chain has been formed. In DNA, the most common modified base is 5-methylcytidine (m5C). In RNA, there are many modified bases, including pseudouridine (Ψ), dihydrouridine (D), inosine (I), ribothymidine (rT) and 7-methylguanosine (m7G). Hypoxanthine and xanthine are two of the many bases created through mutagen presence, both of them through deamination (replacement of the amine-group with a carbonyl-group). Hypoxanthine is produced from adenine, xanthine from guanine. Similarly, deamination of cytosine results in uracil.

Biological significance

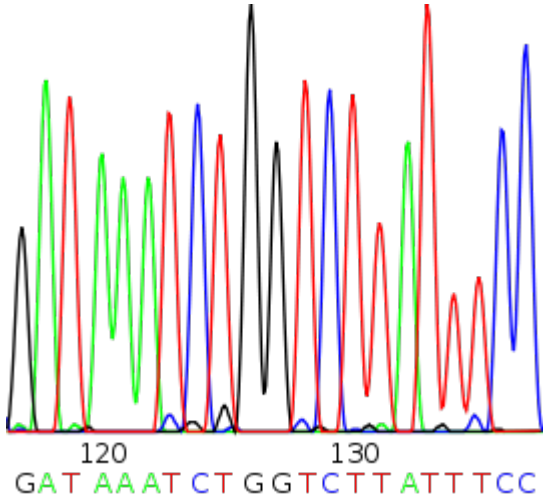


A depiction of the genetic code, by which the information contained in nucleic acids are translated into amino acid sequences in proteins.

In biological systems, nucleic acids contain information which is used by a living cell to construct specific proteins. The sequence of nucleobases on a nucleic acid strand is translated by cell machinery into a sequence of amino acids making up a protein strand. Each group of three bases, called a codon, corresponds to a single amino acid, and there is a specific genetic code by which each possible combination of three bases corresponds to a specific amino acid.

The central dogma of molecular biology outlines the mechanism by which proteins are constructed using information contained in nucleic acids. DNA is transcribed into mRNA molecules, which travels to the ribosome where the mRNA is used as a template for the construction of the protein strand. Since nucleic acids can bind to molecules with complementary sequences, there is a distinction between "sense" sequences which code for proteins, and the complementary "antisense" sequence which is by itself nonfunctional, but can bind to the sense strand.

Sequence determination



Electropherogram printout from automated sequencer for determining part of a DNA sequence

DNA sequencing is the process of determining the nucleotide sequence of a given DNA fragment. The sequence of DNA encodes the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the key nature of DNA to living things, knowledge of DNA sequence may come in useful in practically any biological research. For example, in medicine it can be used to identify, diagnose and potentially develop treatments for genetic diseases. Similarly, research into pathogens may lead to treatments for contagious diseases. Biotechnology is a burgeoning discipline, with the potential for many useful products and services.

RNA is not sequenced directly. Instead, it is copied to a DNA by reverse transcriptase, and this DNA is then sequenced.

Current sequencing methods rely on the discriminatory ability of DNA polymerases, and can therefore only distinguish four bases. An inosine (created from adenosine during RNA editing) will be read as a G, and 5-methyl-cytosine (created from cytosine by DNA methylation) will be read as a C. It is also currently difficult to sequence small amounts of DNA, as the signal will be too weak to measure. This is overcome by PCR amplification.

Digital format

Genetic sequence in digital format

Once a nucleic acid sequence has been obtained from an organism, it is stored *in silico* in digital format. Digital genetic sequences may be stored in sequence databases, be analyzed, be digitally altered and/or be used as templates for creating new actual DNA using artificial gene synthesis.

Sequence analysis

Digital genetic sequences may be analyzed using the tools of bioinformatics to attempt to determine its function.

Genetic testing

The DNA in an organism's genome can be analyzed to diagnose vulnerabilities to inherited diseases, and can also be used to determine a child's paternity (genetic father) or a person's ancestry. Normally, every person carries two copies of every gene, one inherited from their mother, one inherited from their father. The human genome is believed to contain around 20,000 - 25,000 genes. In addition to studying chromosomes to the level of individual genes, genetic testing in a broader sense includes biochemical tests for the possible presence of genetic diseases, or mutant forms of genes associated with increased risk of developing genetic disorders.

Genetic testing identifies changes in chromosomes, genes, or proteins. Most of the time, testing is used to find changes that are associated with inherited disorders. The results of a genetic test can confirm or rule out a suspected genetic condition or help determine a person's chance of developing or passing on a genetic disorder. Several hundred genetic tests are currently in use, and more are being developed.

Sequence alignment

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages. The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role.

Computational phylogenetics makes extensive use of sequence alignments in the construction and interpretation of phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The degree to which sequences in a query set differ is qualitatively related to the sequences' evolutionary distance from one another. Roughly speaking, high sequence identity suggests that the sequences in question have a comparatively young most recent common ancestor, while low identity suggests that the divergence is more ancient. This approximation, which reflects the "molecular clock" hypothesis that a roughly constant rate of evolutionary change can be used to extrapolate the elapsed time since two genes first diverged (that is, the coalescence time), assumes that the effects of mutation and selection are constant across sequence lineages. Therefore it does not account for possible difference among organisms or species in the rates of DNA repair or the possible functional conservation of specific regions in a sequence. (In the case of nucleotide sequences, the molecular clock hypothesis in its most basic form also discounts the difference in acceptance rates between silent mutations that do not alter the meaning of a given codon and other mutations that result in a different amino acid being incorporated into the protein.) More statistically accurate methods allow the evolutionary rate on each branch of the phylogenetic tree to vary, thus producing better estimates of coalescence times for genes.

Sequence motifs

Frequently the primary structure encodes motifs that are of functional importance. Some examples of sequence motifs are: the C/D and H/ACA boxes of snoRNAs, Sm binding site found in spliceosomal RNAs such as U1, U2, U4, U5, U6, U12 and U3, the Shine-Dalgarno sequence, the Kozak consensus sequence and the RNA polymerase III terminator.

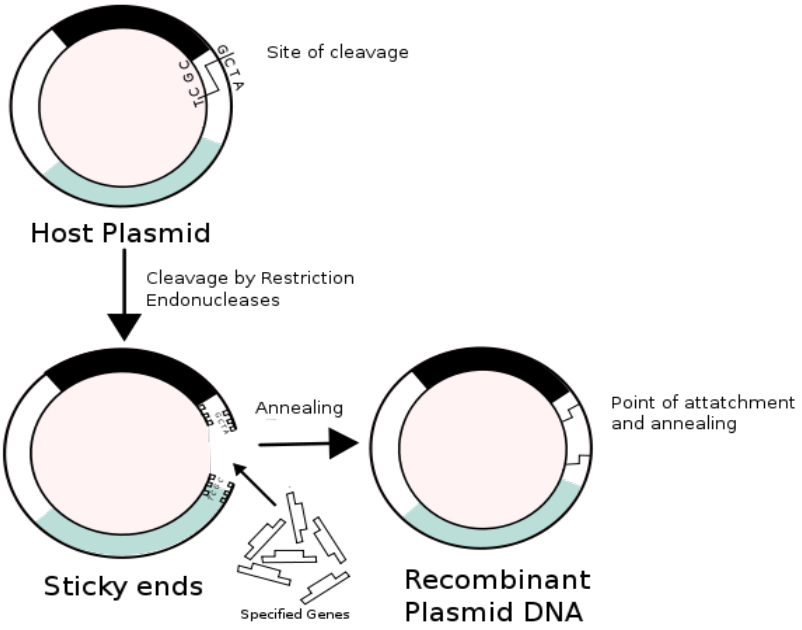
Chapter- 8

Recombinant DNA

Recombinant DNA (rDNA) is a form of artificial DNA that is created by combining two or more sequences that would not normally occur together through the process of gene splicing. In terms of genetic modification, it is created through the introduction of relevant DNA into an existing organismal DNA, such as the plasmids of bacteria, to code for or alter different traits for a specific purpose, such as antibiotic resistance. It differs from genetic recombination in that it does not occur through natural processes within the cell, but is engineered. A **recombinant protein** is a protein that is derived from recombinant DNA.

Methods

Cloning and relation to plasmids



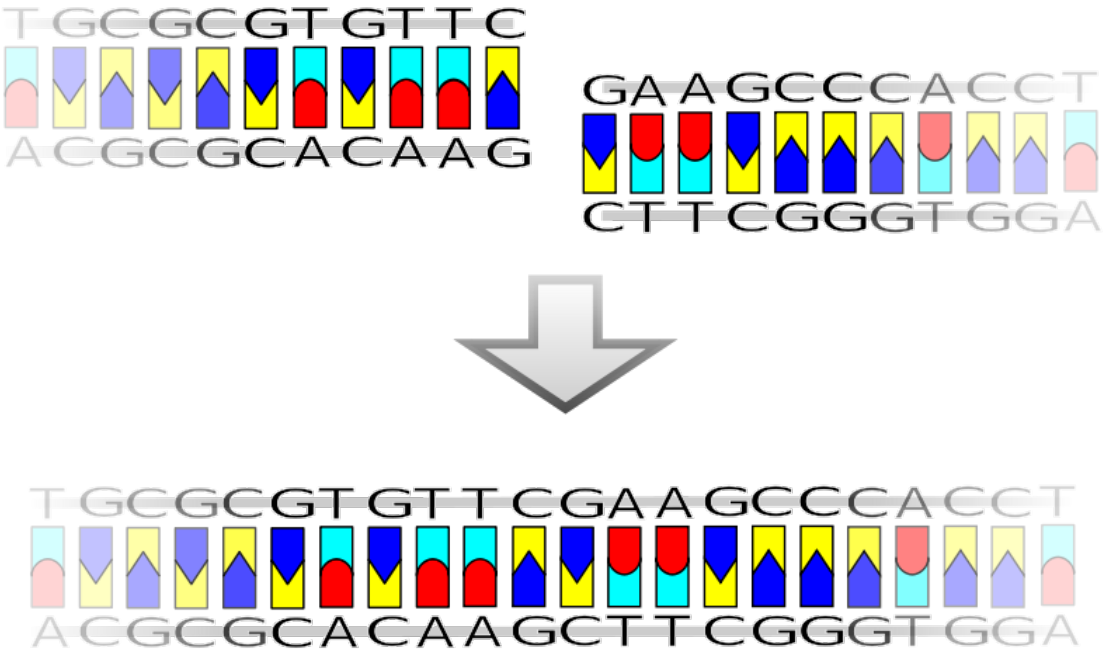
A simple example of how a desired gene is inserted into a plasmid. In this example, the gene specified in the white color becomes useless as the new gene is added.

The use of cloning is interrelated with recombinant DNA in classical biology, as the term "clone" refers to a cell or organism derived from a parental organism, with modern biology referring to the term as a collection of cells derived from the same cell that remain identical. In the classical instance, the use of recombinant DNA provides the initial cell from which the host organism is then expected to recapitulate when it undergoes further cell division, with bacteria remaining a prime example due to the use of viral vectors in medicine that contain recombinant DNA inserted into a structure known as a plasmid.

Plasmids are extrachromosomal self-replicating circular forms of DNA present in most bacteria, such as *Escherichia coli* (E. Coli), containing genes related to catabolism and metabolic activity, and allowing the carrier bacterium to survive and reproduce in conditions present within other species and environments. These genes represent characteristics of resistance to bacteriophages and antibiotics and some heavy metals, but can also be fairly easily removed or separated from the plasmid by restriction endonucleases, which regularly produce "sticky ends" and allow the attachment of a selected segment of DNA, which codes for more "reparative" substances, such as peptide hormone medications including insulin, growth hormone, and oxytocin. In the introduction of useful genes into the plasmid, the bacteria are then used as a viral vector, which are encouraged to reproduce so as to recapitulate the altered DNA within other cells it infects, and increase the amount of cells with the recombinant DNA present within them.

The use of plasmids is also key within gene therapy, where their related viruses are used as **cloning vectors** or carriers, which are means of transporting and passing on genes in recombinant DNA through viral reproduction throughout an organism. Plasmids contain three common features—a **replicator**, **selectable marker** and a **cloning site**. The replicator or "ori" refers to the origin of replication with regard to location and bacteria where replication begins. The marker refers to a particular gene that usually contains resistance to an antibiotic, but may also refer to a gene that is attached alongside the desired one, such as that which confers luminescence to allow identification of successfully recombined DNA. The cloning site is a sequence of nucleotides representing one or more positions where cleavage by restriction endonucleases occurs. Most eukaryotes do not maintain canonical plasmids; yeast is a notable exception. In addition, the Ti plasmid of the bacterium *Agrobacterium tumefaciens* can be used to integrate foreign DNA into the genomes of many plants. Other methods of introducing or creating recombinant DNA in eukaryotes include homologous recombination and transfection with modified viruses.

Chimeric plasmids



An example of chimeric plasmid formation from two "blunt ends" via the enzyme, T4 Ligase.

When recombinant DNA is then further altered or changed to host additional strands of DNA, the molecule formed is referred to as "chimeric" DNA molecule, with reference to the mythological chimera, which consisted as a composite of several animals. The presence of chimeric plasmid molecules is somewhat regular in occurrence, as, throughout the lifetime of an organism, the propagation by vectors ensures the presence of hundreds of thousands of organismal and bacterial cells that all contain copies of the original chimeric DNA.

In the production of chimeric(from chimera) plasmids, the processes involved can be somewhat uncertain, as the intended outcome of the addition of foreign DNA may not always be achieved and may result in the formation of unusable plasmids. Initially, the plasmid structure is linearised to allow the addition by bonding of complementary foreign DNA strands to single-stranded "overhangs" or "sticky ends" present at the ends of the DNA molecule from staggered, or "S-shaped" cleavages produced by restriction endonucleases.

A common vector used for the donation of plasmids originally was the bacterium Escherichia coli and, later, the EcoRI derivative, which was used for its versatility with addition of new DNA by "relaxed" replication when inhibited by chloramphenicol and spectinomycin, later being replaced by the pBR322 plasmid. In the case of EcoRI, the plasmid can anneal with the presence of foreign DNA via the route of sticky-end ligation, or with "blunt ends" via blunt-end ligation, in the presence of the phage T_4 ligase, which

forms covalent links between 3-carbon OH and 5-carbon PO_4 groups present on blunt ends. Both sticky-end, or overhang ligation and blunt-end ligation can occur between foreign DNA segments, and cleaved ends of the original plasmid depending upon the restriction endonuclease used for cleavage.

Applications

There are multitudinous proteins that are created from recombinant DNA and used as medications. Some can alternatively be produced from animal extracts or harvested from humans, such as human growth hormone (rHGH), human insulin, follicle-stimulating hormone (FSH) and factor VIII. Other proteins, when used as medication, only has recombinant DNA as a source, such as with erythropoietin.

History

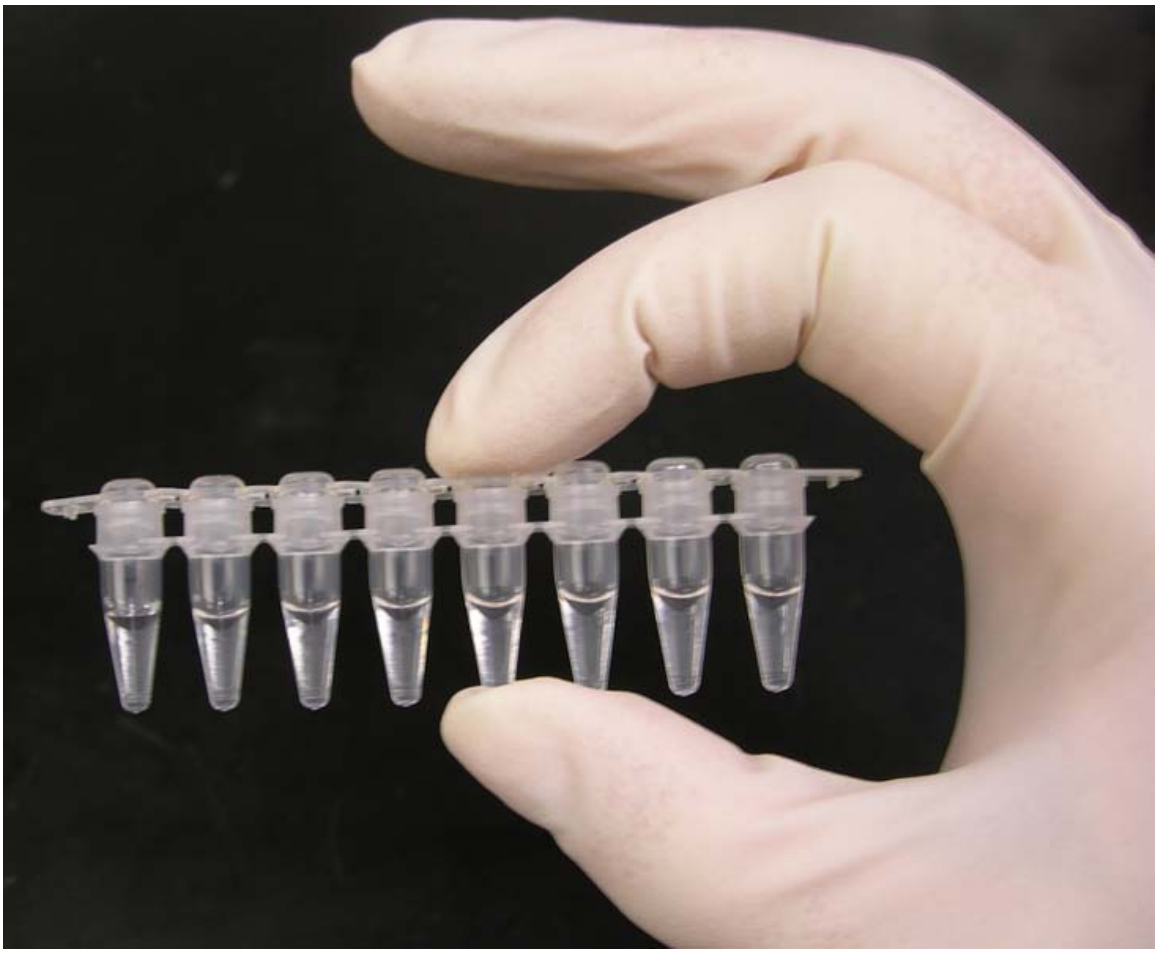
The recombinant DNA technique was first proposed by Peter Lobban, a graduate student, with A. Dale Kaiser at the Stanford University Department of Biochemistry. The technique was then realized by Lobban and Kaiser; Jackson, Symons and Berg; and Stanley Norman Cohen, Chang, Herbert Boyer and Helling, in 1972–74. They published their findings in papers including the 1972 paper "*Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli*", the 1973 paper "*Enzymatic end-to-end joining of DNA molecules*" and the 1973 paper *"Construction of Biologically Functional Bacterial Plasmids* in vitro*"*, all of which described techniques to isolate and amplify genes or DNA segments and insert them into another cell with precision, creating a transgenic bacterium.

Exploitation of recombinant DNA technology was facilitated by the discovery, isolation and application of restriction endonucleases by Werner Arber, Daniel Nathans, and Hamilton Smith, for which they received the 1978 Nobel Prize in Medicine. Cohen and Boyer applied for a patent on the Process for producing biologically functional molecular chimeras which could not exist in nature in 1974. The patent was granted in 1980.

A breakthrough in the application of recombinant DNA technology occurred in 1977 when Herbert Boyer produced biosynthetic "human" insulin in the lab. The specific gene sequence, or polynucleotide, that codes for insulin production in humans was introduced to a sample colony of the *E. coli* bacteria. It was the first medicine made via recombinant DNA technology to be approved by the FDA and commercially available under the brand name Humulin. The vast majority of insulin currently used worldwide is now biosynthetic recombinant "human" insulin or its analogs.

Chapter- 9

Polymerase Chain Reaction



A strip of eight PCR tubes, each containing a 100 µl reaction mixture

The **polymerase chain reaction** (**PCR**) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Developed in 1983 by Kary Mullis, PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR.

The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase (after which the method is named) are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array of genetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase, an enzyme originally isolated from the bacterium *Thermus aquaticus*. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary first to physically separate the two strands in a DNA double helix at a high temperature in a process called DNA melting. At a lower temperature, each strand is then used as the template in DNA synthesis by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions.

PCR principles and procedure

Figure 1a: A thermal cycler for PCR

Figure 1b: An older model three-temperature thermal cycler for PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

A basic PCR set up requires several components and reagents. These components include:

- *DNA template* that contains the DNA region (target) to be amplified.
- Two *primers* that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target.
- *Taq polymerase* or another DNA polymerase with a temperature optimum at around 70 °C.
- *Deoxynucleotide triphosphates* (dNTPs), the building blocks from which the DNA polymerases synthesizes a new DNA strand.
- *Buffer solution,* providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.

- *Divalent cations,* magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be utilized for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis
- *Monovalent cation* potassium ions.

The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below). Many modern thermal cyclers make use of the Peltier effect which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

Procedure

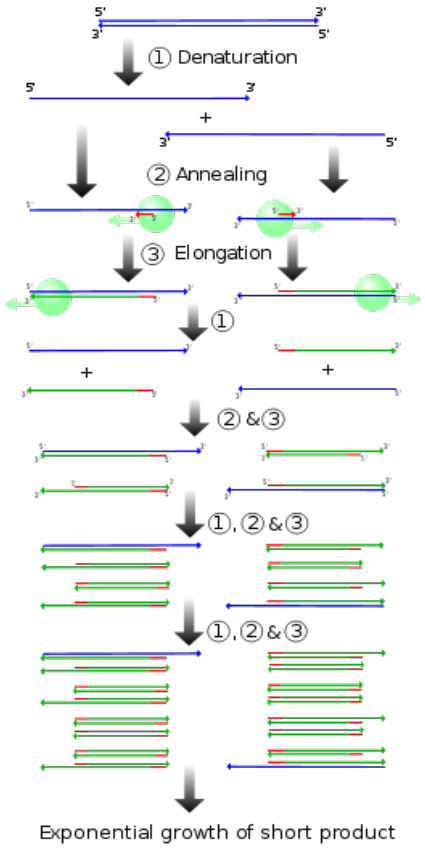


Figure 2: Schematic drawing of the PCR cycle. **(1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C**. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses.

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three (Fig. 2). The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.

- *Initialization step*: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR.

- *Denaturation step*: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

- *Annealing step*: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3-5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

- *Extension/elongation step*: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.

- *Final elongation*: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.

- *Final hold*: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

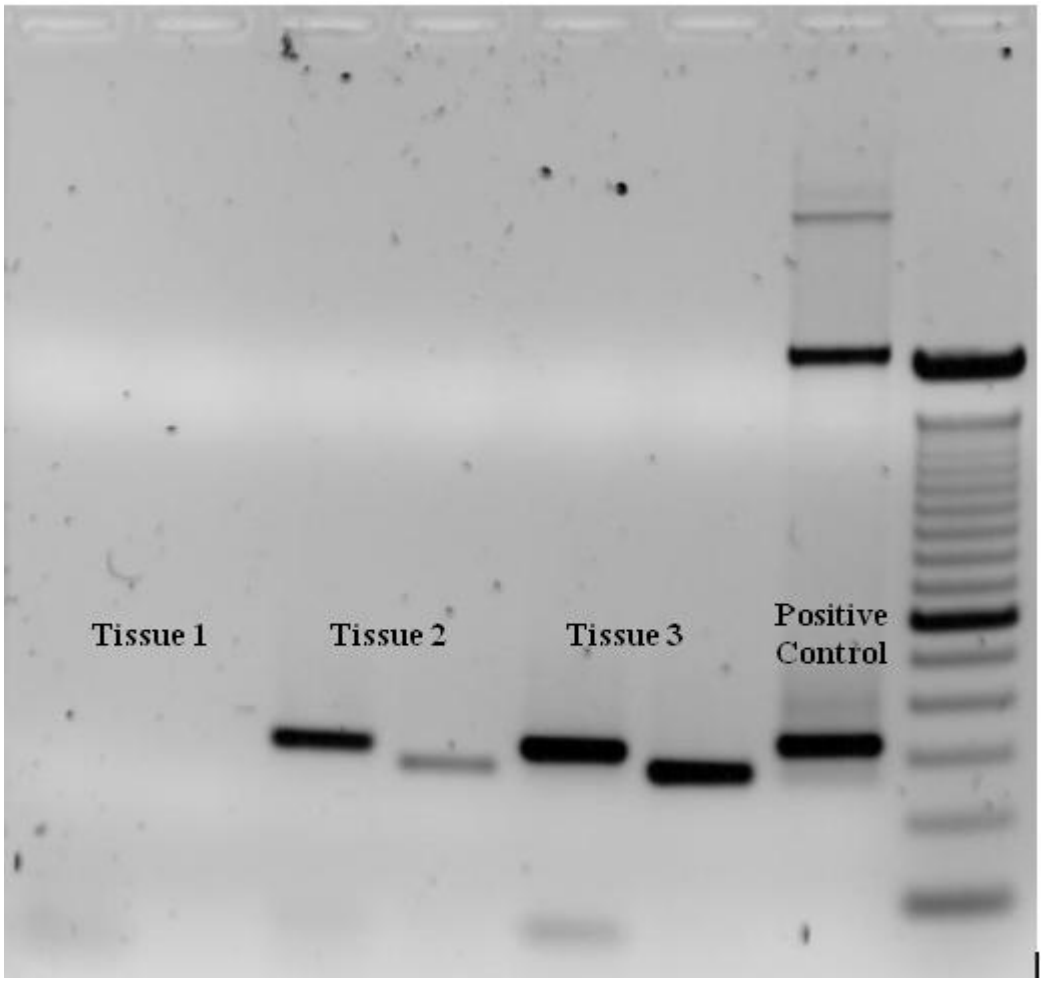


Figure 3: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products (see Fig. 3).

PCR stages

The PCR process can be divided into three stages:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA need to be present.

Levelling off stage: The reaction slows as the DNA polymerase loses activity and as consumption of reagents such as dNTPs and primers causes them to become limiting.

Plateau: No more product accumulates due to exhaustion of reagents and enzyme.

PCR optimization

In practice, PCR can fail for various reasons, in part due to its sensitivity to contamination causing amplification of spurious DNA products. Because of this, a number of techniques and procedures have been developed for optimizing PCR conditions. Contamination with extraneous DNA is addressed with lab protocols and procedures that separate pre-PCR mixtures from potential DNA contaminants. This usually involves spatial separation of PCR-setup areas from areas for analysis or purification of PCR products, use of disposable plasticware, and thoroughly cleaning the work surface between reaction setups. Primer-design techniques are important in improving PCR product yield and in avoiding the formation of spurious products, and the usage of alternate buffer components or polymerase enzymes can help with amplification of long or otherwise problematic regions of DNA. Addition of reagents, such as formamide, in buffer systems may increase the specificity and yield of PCR.

Application of PCR

Selective DNA isolation

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many methods, such as generating hybridization probes for Southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include DNA sequencing to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a plasmid or the genetic material of another organism. Bacterial colonies (E. coli) can be rapidly screened by PCR for correct DNA vector constructs. PCR may also be used for genetic fingerprinting; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between

siblings, and are used in paternity testing (Fig. 4). This technique may also be used to determine evolutionary relationships among organisms.

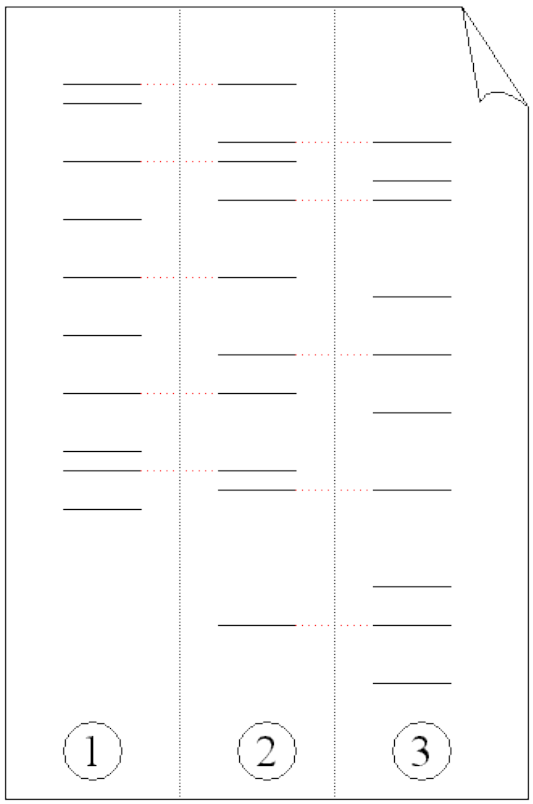


Figure 4: Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ancient DNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old mammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian tsar.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene expression. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

PCR permits early diagnosis of malignant diseases such as leukemia and lymphomas, which is currently the highest developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity which is at least 10,000 fold higher than other methods.

PCR also permits identification of non-cultivatable or slow-growing microorganisms such as mycobacteria, anaerobic bacteria, or viruses from tissue culture assays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and the PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques (see below).

Variations on the basic PCR technique

- *Allele-specific PCR*: a diagnostic or cloning technique which is based on single-nucleotide polymorphisms (SNPs) (single-base differences in DNA). It requires prior knowledge of a DNA sequence, including differences between alleles, and uses primers whose 3' ends encompass the SNP. PCR amplification under stringent conditions is much less efficient in the presence of a mismatch between template and primer, so successful amplification with an SNP-specific primer signals presence of the specific SNP in a sequence.

- *Assembly PCR* or *Polymerase Cycling Assembly (PCA)*: artificial synthesis of long DNA sequences by performing PCR on a pool of long oligonucleotides with short overlapping segments. The oligonucleotides alternate between sense and antisense directions, and the overlapping segments determine the order of the PCR fragments, thereby selectively producing the final long DNA product.

- *Asymmetric PCR*: preferentially amplifies one DNA strand in a double-stranded DNA template. It is used in sequencing and hybridization probing where amplification of only one of the two complementary strands is required. PCR is carried out as usual, but with a great excess of the primer for the strand targeted for amplification. Because of the slow (arithmetic) amplification later in the reaction after the limiting primer has been used up, extra cycles of PCR are required. A recent modification on this process, known as *L*inear-*A*fter-*T*he-*E*xponential-PCR (LATE-PCR), uses a limiting primer with a higher melting

temperature (Tm) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.

- *Helicase-dependent amplification*: similar to traditional PCR, but uses a constant temperature rather than cycling through denaturation and annealing/extension cycles. DNA helicase, an enzyme that unwinds DNA, is used in place of thermal denaturation.

- *Hot-start PCR*: a technique that reduces non-specific amplification during the initial set up stages of the PCR. It may be performed manually by heating the reaction components to the melting temperature (e.g., 95°C) before adding the polymerase. Specialized enzyme systems have been developed that inhibit the polymerase's activity at ambient temperature, either by the binding of an antibody or by the presence of covalently bound inhibitors that only dissociate after a high-temperature activation step. Hot-start/cold-finish PCR is achieved with new hybrid polymerases that are inactive at ambient temperature and are instantly activated at elongation temperature.

- *Intersequence-specific PCR* (ISSR): a PCR method for DNA fingerprinting that amplifies regions between simple sequence repeats to produce a unique fingerprint of amplified fragment lengths.

- *Inverse PCR*: is commonly used to identify the flanking sequences around genomic inserts. It involves a series of DNA digestions and self ligation, resulting in known sequences at either end of the unknown sequence.

- *Ligation-mediated PCR*: uses small DNA linkers ligated to the DNA of interest and multiple primers annealing to the DNA linkers; it has been used for DNA sequencing, genome walking, and DNA footprinting.

- *Methylation-specific PCR* (MSP): developed by Stephen Baylin and Jim Herman at the Johns Hopkins School of Medicine, and is used to detect methylation of CpG islands in genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil, which is recognized by PCR primers as thymine. Two PCRs are then carried out on the modified DNA, using primer sets identical except at any CpG islands within the primer sequences. At these points, one primer set recognizes DNA with cytosines to amplify methylated DNA, and one set recognizes DNA with uracil or thymine to amplify unmethylated DNA. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation.

- *Miniprimer PCR*: uses a thermostable polymerase (S-Tbr) that can extend from short primers ("smalligos") as short as 9 or 10 nucleotides. This method permits PCR targeting to smaller primer binding regions, and is used to amplify conserved DNA sequences, such as the 16S (or eukaryotic 18S) rRNA gene.

- *Multiplex Ligation-dependent Probe Amplification* (*MLPA*): permits multiple targets to be amplified with only a single primer pair, thus avoiding the resolution limitations of multiplex PCR (see below).

- *Multiplex-PCR*: consists of multiple primer sets within a single PCR mixture to produce amplicons of varying sizes that are specific to different DNA sequences. By targeting multiple genes at once, additional information may be gained from a single test run that otherwise would require several times the reagents and more time to perform. Annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction, and amplicon sizes, i.e., their base pair length, should be different enough to form distinct bands when visualized by gel electrophoresis.

- *Nested PCR*: increases the specificity of DNA amplification, by reducing background due to non-specific amplification of DNA. Two sets of primers are used in two successive PCRs. In the first reaction, one pair of primers is used to generate DNA products, which besides the intended target, may still consist of non-specifically amplified DNA fragments. The product(s) are then used in a second PCR with a set of primers whose binding sites are completely or partially different from and located 3' of each of the primers used in the first reaction. Nested PCR is often more successful in specifically amplifying long DNA fragments than conventional PCR, but it requires more detailed knowledge of the target sequences.

- *Overlap-extension PCR*: a genetic engineering technique allowing the construction of a DNA sequence with an alteration inserted beyond the limit of the longest practical primer length.

- *Quantitative PCR (Q-PCR)*: used to measure the quantity of a PCR product (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA or RNA. Q-PCR is commonly used to determine whether a DNA sequence is present in a sample and the number of its copies in the sample. *Quantitative real-time PCR* has a very high degree of precision. QRT-PCR methods use fluorescent dyes, such as Sybr Green, EvaGreen or fluorophore-containing DNA probes, such as TaqMan, to measure the amount of amplified product in real time. It is also sometimes abbreviated to RT-PCR (*R*eal *T*ime PCR) or RQ-PCR. QRT-PCR or RTQ-PCR are more appropriate contractions, since RT-PCR commonly refers to reverse transcription PCR (see below), often used in conjunction with Q-PCR.

- *R*everse *T*ranscription PCR (RT-PCR): for amplifying DNA from RNA. Reverse transcriptase reverse transcribes RNA into cDNA, which is then amplified by PCR. RT-PCR is widely used in expression profiling, to determine the expression of a gene or to identify the sequence of an RNA transcript, including transcription start and termination sites. If the genomic DNA sequence of a gene is known, RT-PCR can be used to map the location of exons and introns in the gene. The 5' end

of a gene (corresponding to the transcription start site) is typically identified by RACE-PCR (*Rapid Amplification of cDNA Ends*).

- *Solid Phase PCR*: encompasses multiple meanings, including Polony Amplification (where PCR colonies are derived in a gel matrix, for example), Bridge PCR (primers are covalently linked to a solid-support surface), conventional Solid Phase PCR (where Asymmetric PCR is applied in the presence of solid support bearing primer with sequence matching one of the aqueous primers) and Enhanced Solid Phase PCR (where conventional Solid Phase PCR can be improved by employing high Tm and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *Thermal asymmetric interlaced PCR (TAIL-PCR)*: for isolation of an unknown sequence flanking a known sequence. Within the known sequence, TAIL-PCR uses a nested pair of primers with differing annealing temperatures; a degenerate primer is used to amplify in the other direction from the unknown sequence.

- *Touchdown PCR* (*Step-down PCR*): a variant of PCR that aims to reduce nonspecific background by gradually lowering the annealing temperature as PCR cycling progresses. The annealing temperature at the initial cycles is usually a few degrees (3-5°C) above the T_m of the primers used, while at the later cycles, it is a few degrees (3-5°C) below the primer T_m. The higher temperatures give greater specificity for primer binding, and the lower temperatures permit more efficient amplification from the specific products formed during the initial cycles.

- *PAN-AC*: uses isothermal conditions for amplification, and may be used in living cells.

- *Universal Fast Walking*: for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer - which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

History

A 1971 paper in the Journal of Molecular Biology by Kleppe and co-workers first described a method using an enzymatic assay to replicate a short DNA template with primers *in vitro*. However, this early manifestation of the basic PCR principle did not receive much attention, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis.

At the core of the PCR method is the use of a suitable DNA polymerase able to withstand the high temperatures of >90 °C (194 °F) required for separation of the two DNA strands in the DNA double helix after each replication cycle. The DNA polymerases initially

employed for in vitro experiments presaging PCR were unable to withstand these high temperatures. So the early procedures for DNA replication were very inefficient, time consuming, and required large amounts of DNA polymerase and continual handling throughout the process.

The discovery in 1976 of Taq polymerase — a DNA polymerase purified from the thermophilic bacterium, *Thermus aquaticus*, which naturally lives in hot (50 to 80 °C (122 to 176 °F)) environments such as hot springs — paved the way for dramatic improvements of the PCR method. The DNA polymerase isolated from *T. aquaticus* is stable at high temperatures remaining active even after DNA denaturation, thus obviating the need to add new DNA polymerase after each cycle. This allowed an automated thermocycler-based process for DNA amplification.

When Mullis developed the PCR in 1983, he was working in Emeryville, California for Cetus Corporation, one of the first biotechnology companies. There, he was responsible for synthesizing short chains of DNA. Mullis has written that he conceived of PCR while cruising along the Pacific Coast Highway one night in his car. He was playing in his mind with a new way of analyzing changes (mutations) in DNA when he realized that he had instead invented a method of amplifying any DNA region through repeated cycles of duplication driven by DNA polymerase. In *Scientific American*, Mullis summarized the procedure: "Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon. The reaction is easy to execute. It requires no more than a test tube, a few simple reagents, and a source of heat." He was awarded the Nobel Prize in Chemistry in 1993 for his invention, seven years after he and his colleagues at Cetus first put his proposal to practice. However, some controversies have remained about the intellectual and practical contributions of other scientists to Mullis' work, and whether he had been the sole inventor of the PCR principle (see below).

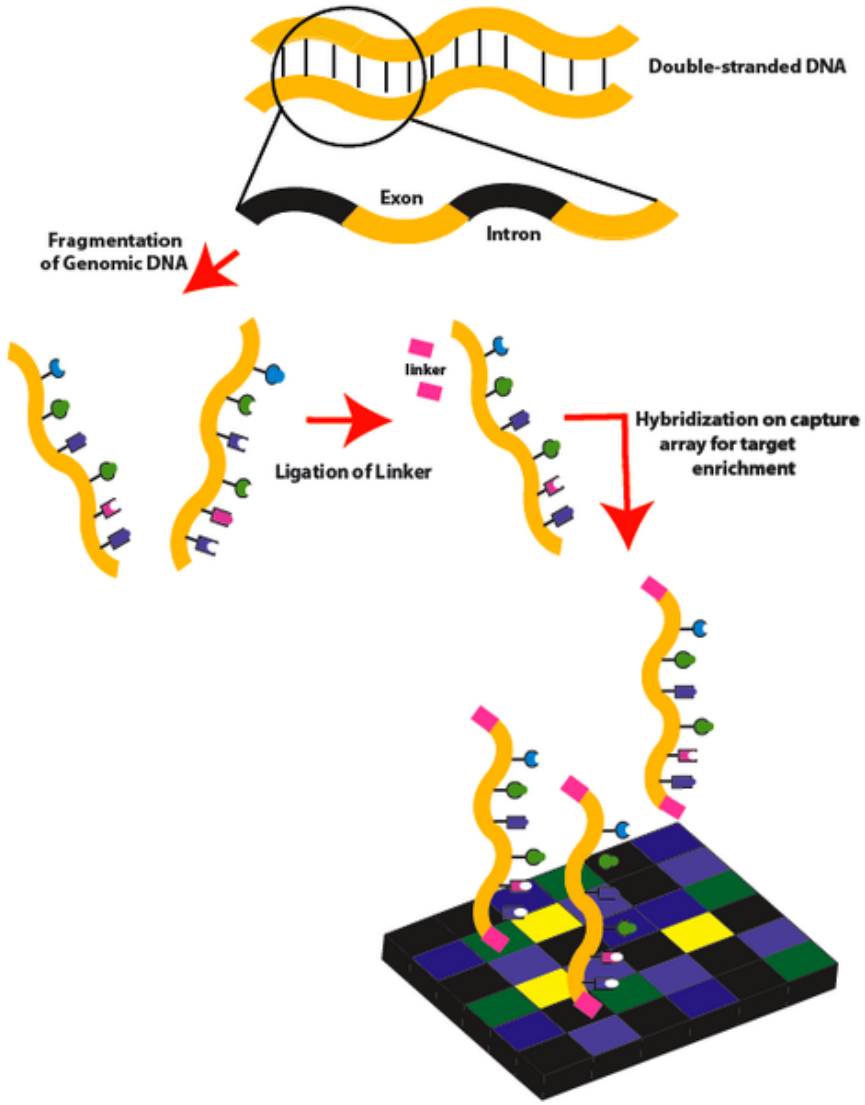
Patent wars

The PCR technique was patented by Kary Mullis and assigned to Cetus Corporation, where Mullis worked when he invented the technique in 1983. The *Taq* polymerase enzyme was also covered by patents. There have been several high-profile lawsuits related to the technique, including an unsuccessful lawsuit brought by DuPont. The pharmaceutical company Hoffmann-La Roche purchased the rights to the patents in 1992 and currently holds those that are still protected.

A related patent battle over the Taq polymerase enzyme is still ongoing in several jurisdictions around the world between Roche and Promega. The legal arguments have extended beyond the lives of the original PCR and Taq polymerase patents, which expired on March 28, 2005.

Chapter- 10

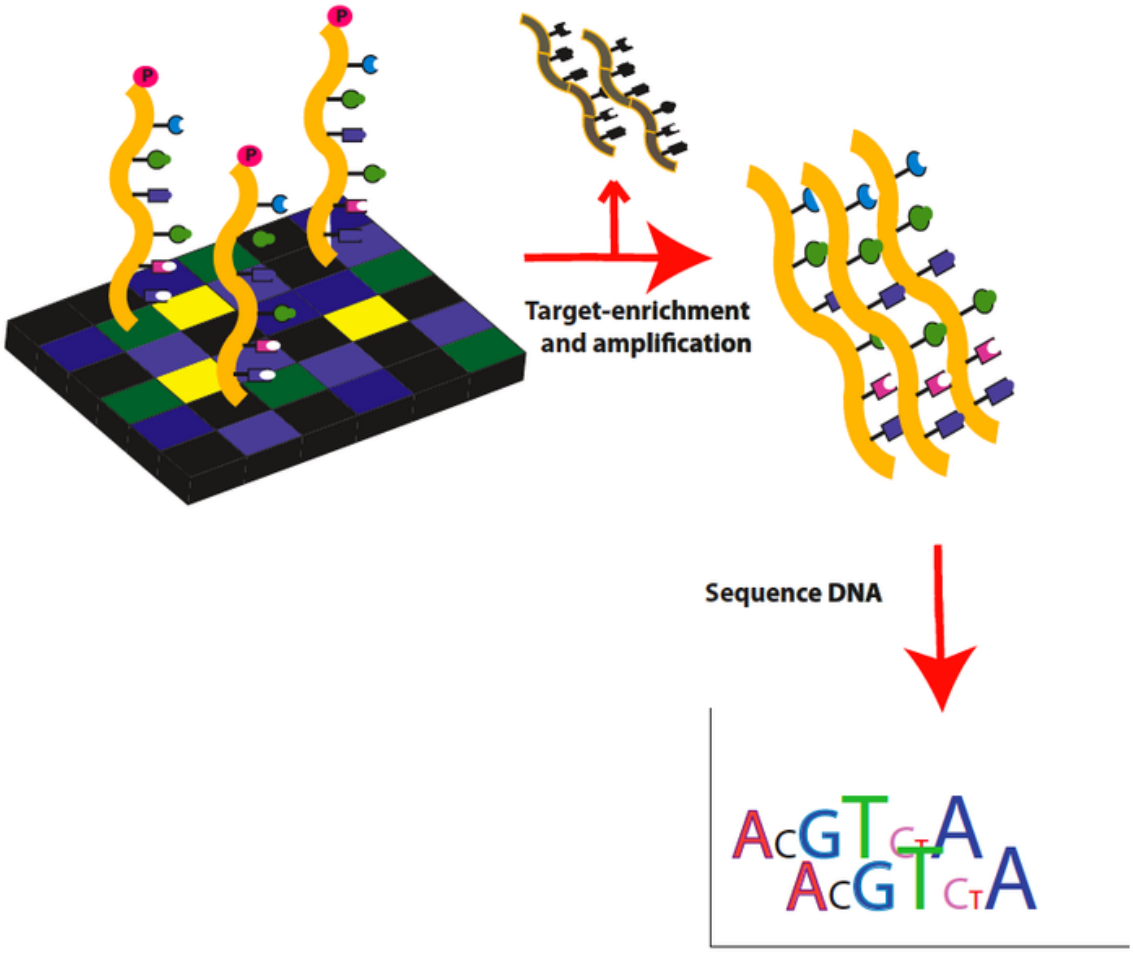
Exome Sequencing



Exome Sequencing Workflow: Part 1.

Exome sequencing (also known as **targeted exome capture**) is an efficient strategy to selectively sequence the coding regions of the human genome to identify novel genes associated with rare and common disorders. Routine whole genome sequencing of large

numbers of individuals is still not feasible partly due to the high cost associated with the technique. At present, it is necessary to use an alternative approach, in which certain regions of the genome, such as the “exome”, are targeted, enriched and sequenced, which requires ~5% as much sequencing as a whole genome. The “exome” represents all the exons in the human genome (i.e., the transcribed region of the genome). Exons are short, functionally important sequences of DNA which represent the regions in genes that are translated into protein and untranslated region flanking them (UTR). UTRs are usually not included in exome studies. In total there are about 180,000 exons found in the human genome. These protein coding regions constitute about 1% of the human genome which translates to about 30 megabases (Mb) in length. It is estimated that the protein coding regions of the human genome constitute about 85% of the disease-causing mutations.



Exome Sequencing Workflow: Part 2.

The robust approach to sequencing the complete coding region (exome) has the potential to be clinically relevant in genetic diagnosis due to current understanding of functional consequences in sequence variation. The goal of this approach is to identify the functional variation that is responsible for both mendelian and common diseases such as

Miller syndrome and Alzheimer's disease without the high costs associated with whole-genome sequencing while maintaining high coverage in sequence depth.

As an efficient strategy

Exome sequencing is an efficient strategy to identify these rare causal variants of mendelian disorders over whole genome sequencing due to few factors:

1. Positional cloning strategies have reduced power to successfully identify causal rare variants
2. The majority of genetic variants that underlie mendelian disorders disrupt protein-coding sequences
3. A large number of rare nonsynonymous substitutions are predicted to be deleterious
4. Splice sites also represent sequences in which there is high functional variation

The exome represents an enriched portion of the genome that can be used to search for variants with large effect sizes.

Mendelian disorders

Rare diseases affect less than 200,000 individuals in the United States and are of interest because the identification of the genetic basis can provide knowledge about biological pathways and therapeutic targets. It is suspected that there are more than 7,000 rare mendelian diseases which affect millions of people in the US. The majority of mendelian diseases studied to date are known to be caused by rare mutations that affect protein function. The majority of mutations that are known to cause mendelian disorders are located in protein-coding regions while non-coding regions on the other hand are likely to have weak or neutral effects.

To date, less than half of all rare monogenic disorders have been discovered. The identification of genetic variants for rare disorders is limited by a number of factors. These include sample size of affected individuals, reduced penetrance, locus heterogeneity, and alleles that impair reproductive fitness. These factors make it difficult to map these traits by linkage analysis and they reduce the power of traditional positional cloning strategies to identify these variants. For both dominant and recessive traits finding an excess of independent mutations in the same locus will provide evidence that a disease gene has been identified. Exome sequencing is a powerful technique to identify genes in rare mendelian disorders because it requires only a small number of unrelated cases to identify a causal gene.

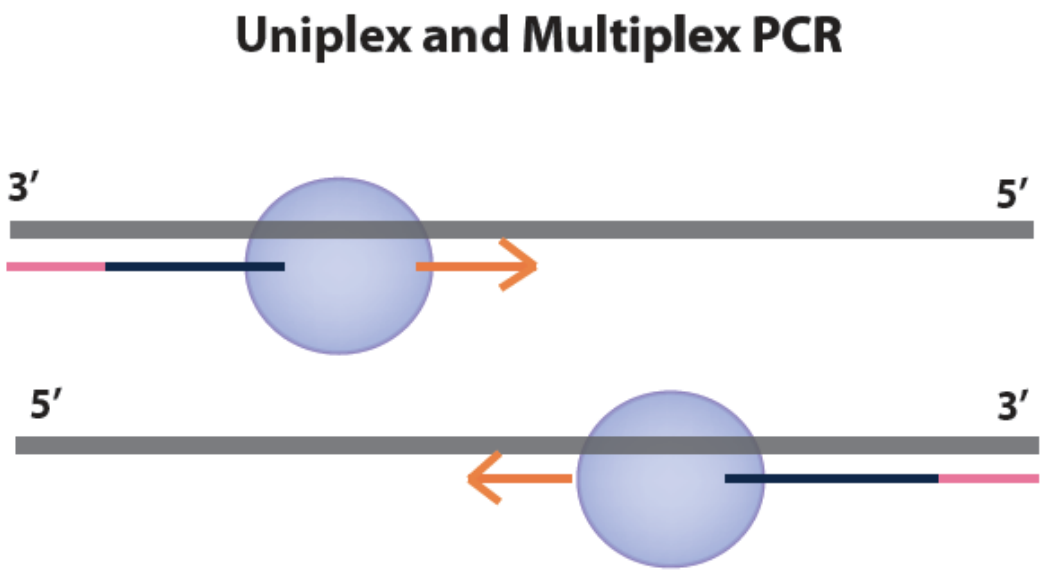
Technological platforms

The technical platforms used to carry out exome sequencing are DNA microarrays and magnetic bead based systems for the enrichment of the exome DNA and next-generation sequencing technologies.

Target-enrichment strategies

Target-enrichment methods allow to selectively capture genomic regions of interest from a DNA sample prior to sequencing. Several target-enrichment strategies have been developed.

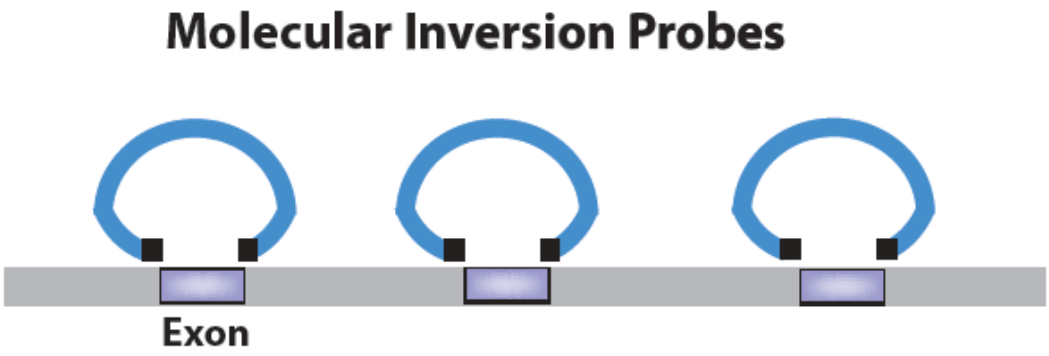
PCR



Uniplex and Multiplex PCR

PCR is one of the most widely used enrichment strategies for over 20 years. This approach is known to be useful in classical Sanger sequencing because a uniplex PCR used to generate a single DNA sequence is comparable in read length to a typical amplicon. Multiplex PCR reactions which require several primers are challenging although strategies to get around this have been developed. A limitation to this method is the size of the genomic target due to workload and quantity of DNA required. The PCR based approach is highly effective, yet it is not feasible to target genomic regions that are several megabases in size due to quantity of DNA required and cost.

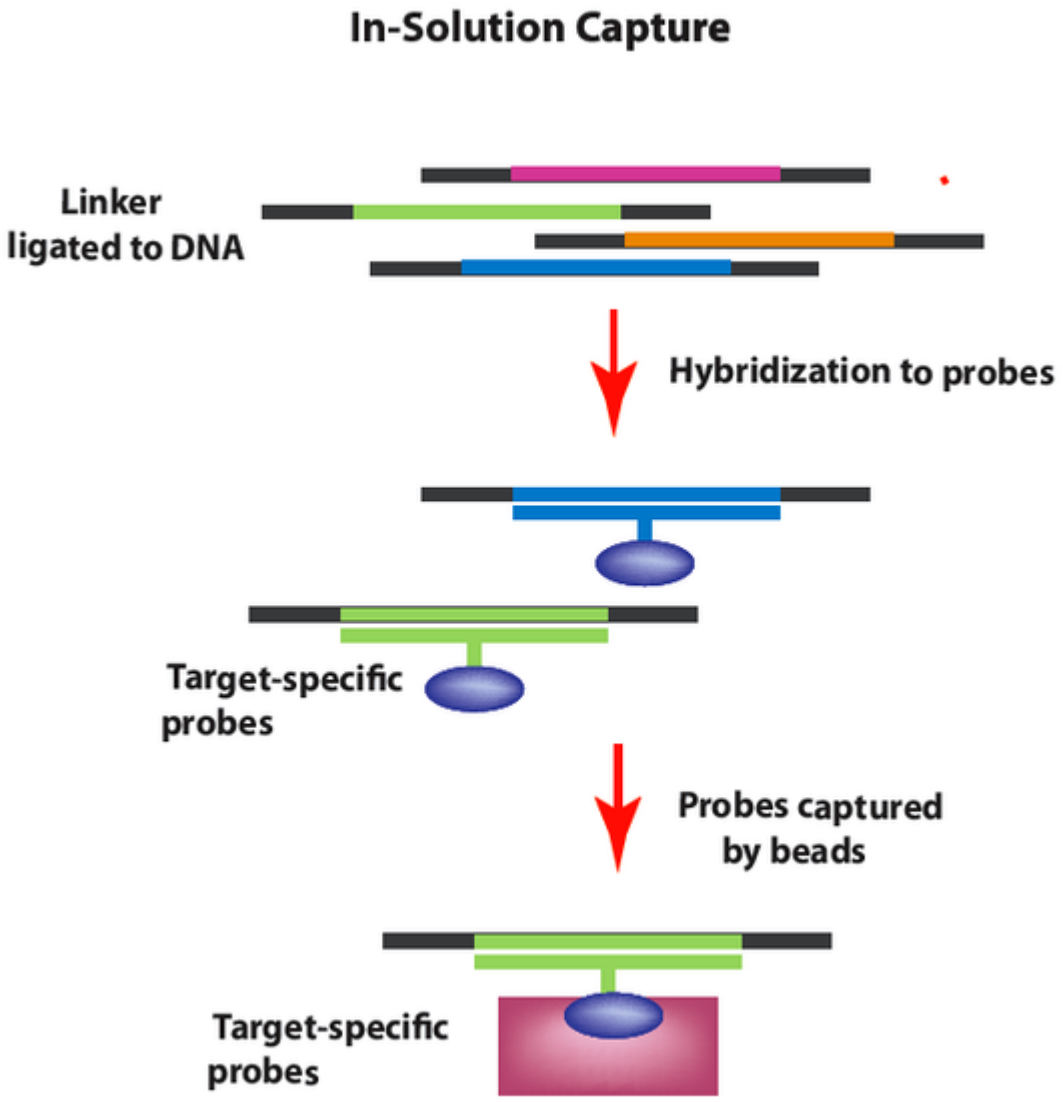
Molecular Inversion Probes (MIP)



Molecular Inversion Probes

This is an enzymatic technique that targets the amplification of genomic regions by multiplexing based on target circularization. Accurate genotypes can be achieved from massively parallel sequencing using this method. This method is suggested to be useful for small numbers of targets in a large number of samples. Major disadvantage of this method for target enrichment is the capture uniformity as well as the cost associated with covering large target sets.

Hybrid capture



In-Solution Capture

This technique involves hybridizing shotgun libraries of genomic DNA to target-specific sequences on a microarray. Roche NimbleGen was first to take this technology and adapt it for next-generation sequencing. They developed the Sequence Capture Human Exome 2.1M Array to capture ~180,000 coding exons. This method is both time-saving and cost-effective compared to PCR based methods. The Agilent Capture Array and the comparative genomic hybridization array also other methods that can be used for hybrid capture of target sequences. Limitations in this technique include the need for expensive hardware as well as a relatively large amount of DNA.

In-solution capture

To capture genomic regions of interest using in-solution capture, a pool of custom oligonucleotides (probes) is synthesized and hybridized in solution to a fragmented genomic DNA sample. The probes (labeled with beads) selectively hybridize to the genomic regions of interest after which the beads (now including the DNA fragments of interest) can be pulled down and washed to clear excess material. The beads are then removed and the genomic fragments can be sequenced allowing for selective DNA sequencing of genomic regions (e.g. exons) of interest. Several companies (e.g. FlexGen) offer custom pools of oligonucleotides or instruments to synthesize these oligopools in-house.

This method was developed to improve on the hybridization capture target-enrichment method. In solution capture as opposed to hybrid capture, there is an excess of probes to target regions of interest over the amount of template required. The optimal target size is about 3.5 Mb in length and yields excellent sequence coverage of the target regions. The preferred method is dependent on several factors including; size (bp) of region of interest, demands for reads on target, equipment in house, etc.

Sequencing

There are several sequencing platforms available including the classical Sanger sequencing. Other platforms include the Roche 454 sequencer, the Illumina Genome Analyzer II and the Applied Biosystems SOLiD, which have both been used for exome sequencing.

Significance

A study published in September 2009 discussed a proof of concept experiment to determine if it was possible to identify causal genetic variants using exome sequencing. They sequenced four individuals with Freeman-Sheldon syndrome (FSS) (OMIM 193700), a rare autosomal dominant disorder known to be caused by a mutation in the gene MYH3. Eight HapMap individuals were also sequenced to remove common variants in order to identify the causal gene for FSS. After exclusion of common variants, the authors were able to identify MYH3, which confirms that exome sequencing can be used to identify causal variants of rare disorders. This is the first reported study that used exome sequencing as an approach to identify an unknown causal gene for a rare mendelian disorder.

Subsequently, another group reported successful clinical diagnosis of a suspected Bartter syndrome patient of Turkish origin. Bartter syndrome is a renal salt-wasting disease. Exome sequencing revealed an unexpected well-conserved recessive mutation in a gene called SLC26A3 which is associated with congenital chloride diarrhea (CLD). This molecular diagnosis of CLD was confirmed by the referring clinician. This example provided proof of concept of the use of whole-exome sequencing as a clinical tool in evaluation of patients with undiagnosed genetic illnesses. This report is regarded as the

first application of next generation sequencing technology for molecular diagnosis of a patient.

A second report was conducted on exome sequencing of individuals with a mendelian disorder known as Miller syndrome (MIM#263750), a rare disorder of autosomal recessive inheritance. Two siblings and two unrelated individuals with Miller syndrome were studied. They looked at variants that have the potential to be pathogenic such as non-synonymous mutations, splice acceptor and donor sites and short coding insertions or deletions. Since Miller syndrome is a rare disorder, it is expected that the causal variant has not been previously identified. Previous exome sequencing studies of common single nucleotide polymorphisms (SNPs) in public SNP databases were used to further exclude candidate genes. After exclusion of these genes, the authors found mutations in DHODH that were shared among individuals with Miller syndrome. Each individual with Miller syndrome was a compound heterozygote for the DHODH mutations which was inherited as each parent of an affected individual was found to be a carrier.

This is the first time exome sequencing has been shown to identify a novel gene responsible for a rare mendelian disease. This exciting finding demonstrates is that exome sequencing has the potential to locate causative genes in complex diseases, which previously has not been possible due to limitations in traditional methods. Targeted capture and massively parallel sequencing represents a cost-effective, reproducible and robust strategy with high sensitivity and specificity to detect variants causing protein-coding changes in individual human genomes.

Comparison with genotyping

There are multiple technologies available to undertake methods to identify causal genetic variants associated with disease. Each technology has its own technical, financial and throughput limitations. Microarrays for example, require hybridization probes of known sequence and are therefore limited by probe design and thus prevent the identification of genetic changes that can be detected. Massively parallel sequencing technologies used for exome sequencing on the other hand makes it now possible to identify the cause of many unknown diseases by screening thousands of loci at once. This technology addresses the present limitations of hybridization genotyping arrays and classical sequencing.

Although, exome sequencing is an expensive method relative to other technologies (e.g., hybridization-based technologies) currently available, it is an efficient strategy to identify the genetic bases that underlie rare mendelian disorders. This approach has become increasingly practical with the falling cost and increased throughput of whole genome sequencing. Even by only sequencing the exomes of individuals, a large quantity of data and sequence information is generated which requires a significant amount of data analysis. Challenges associated with the analysis of this data include changes in programs used to align and assemble sequence reads. Various sequence technologies also have different error rates and generate various read-lengths which can pose challenges in comparing results from different sequencing platforms.

Limitations

Exome sequencing is able to only identify those variants found in the coding region of genes which affect protein function. It is not able to identify the structural and non-coding variants associated the disease which can be found using other methods such as whole genome sequencing. There remains 99% of the human genome that is not covered using exome sequencing. Whole genome sequencing will eventually become a standard approach and allow us to gain a deeper understanding of genetic variation found in populations. Presently, this technique is not practical due to the high costs and time associated with sequencing large numbers of genomes. Exome sequencing allows sequencing of portions of the genome over at least 20 times as many samples compared to whole genome sequencing. For translation of identified rare variants into the clinic, sample size and the ability to interpret the results to provide a clinical diagnosis indicates that with the current knowledge in genetics, exome sequencing may be the most valuable.

The statistical analysis of the large quantity of data generated from sequencing approaches is a challenge. False positive and false negative findings are associated with genomic resequencing approaches and it is a critical issue. A few strategies have been developed to improve the quality of exome data such as:

- Comparing the genetic variants identified between sequencing and array-based genotyping
- Comparing the coding SNPs to a whole genome sequenced individual with the disorder
- Comparing the coding SNPs with Sanger sequencing of HapMap individuals

Rare recessive disorders would not have single nucleotide polymorphisms (SNPs) in public databases such dbSNP. More common recessive phenotypes may have disease-causing variants reported in dbSNP. For example, the most common cystic fibrosis variant has an allele frequency of about 3% in most populations. Screening out such variants might erroneously exclude such genes from consideration. Genes for recessive disorders are usually easier to identify than dominant disorders because the genes are less likely to have more than one rare nonsynonymous variant. The system screen common genetic variants relies on dbSNP which may not have accurate information about the variation of alleles. Using lists of common variation from a study exome or genome-wide sequenced individual would be more reliable. A challenge in this approach is that as the number of exomes sequenced increases, dbSNP will also increase in the number of uncommon variants. It will be necessary to develop thresholds to define the common variants that are unlikely to be associated with a disease phenotype.

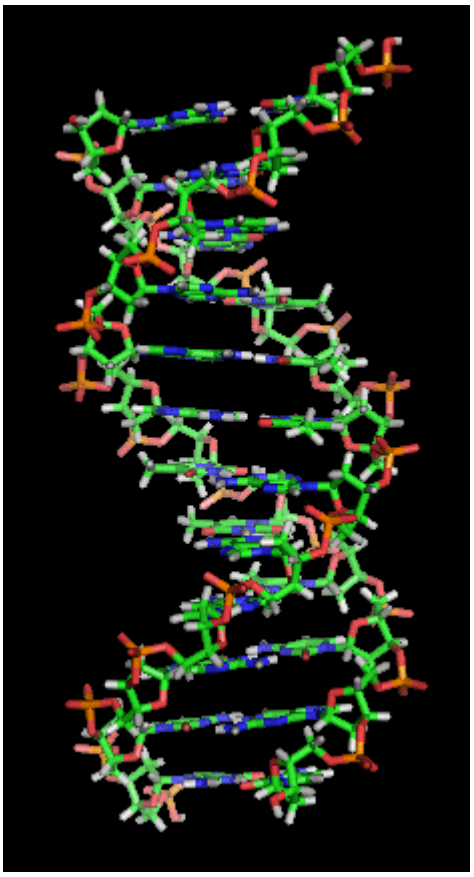
Genetic heterogeneity and population ethnicity are also major limitations as it may increase the number false positive and false negative findings which will make the identification of candidate genes more difficult. Of course it is possible to reduce the stringency of the thresholds in the presence of heterogeneity and ethnicity, however it will reduce the power to detect variants as well.

Ethical implications

New technologies in genomics has changed the way researchers approach both basic and translational research. With approaches such as exome sequencing it is possible to significantly enhance the data generated from individual genomes which has put forth a series of questions on how to deal with the vast amount of information. Should the individuals in these studies be allowed to have access to their sequencing information? Is it possible to interpret theses results for these individuals and are the identified genetic variants clinically relevant? This data can lead to unexpected findings and complicate clinical utility and patient benefit.This area of genomics still remains a challenge and researchers are looking into how to address these questions.

Chapter- 11

Nucleic Acid Double Helix



Two complementary regions of nucleic acid molecules will bind and form a double helical structure held together by base pairs.

In molecular biology, the term **double helix** refers to the structure formed by double-stranded molecules of nucleic acids such as DNA and RNA. The double helical structure of a nucleic acid complex arises as a consequence of its secondary structure, and is a fundamental component in determining its tertiary structure.

The DNA double helix is a spiral polymer of nucleic acids, held together by nucleotides which base pair together. In B-DNA, the most common double helical structure, the double helix is right-handed with about 10–10.5 nucleotides per turn. The double helix

structure of DNA contains a **major groove** and **minor groove**, the major groove being wider than the minor groove. Given the difference in widths of the major groove and minor groove, many proteins which bind to DNA do so through the wider major groove.

History

The double-helix model of DNA structure was first published in the journal Nature by James D. Watson and Francis Crick in 1953, based upon the crucial X-ray diffraction image of DNA (labeled as "Photo 51") from Rosalind Franklin in 1952, followed by her more clarified DNA image with Raymond Gosling, Maurice Wilkins, Alexander Stokes, and Herbert Wilson, as well as base-pairing chemical and biochemical information by Erwin Chargaff. The previous model was triple-stranded DNA.

Crick, Wilkins, and Watson each received one third of the 1962 Nobel Prize in Physiology or Medicine for their contributions to the discovery. (Franklin, whose breakthrough X-ray diffraction data was used to formulate the DNA structure, died in 1958, and thus was ineligible to be nominated for a Nobel Prize.)

Nucleic acid hybridization

Hybridization is the process of complementary base pairs binding to form a double helix. Melting is the process by which the interactions between the strands of the double helix are broken, separating the two nucleic acid strands. These bonds are weak, easily separated by gentle heating, enzymes, or physical force. Melting occurs preferentially at certain points in the nucleic acid. **T** and **A** rich sequences are more easily melted than **C** and **G** rich regions. Particular base steps are also susceptible to DNA melting, particularly **T A** and **T G** base steps. These mechanical features are reflected by the use of sequences such as **TATAA** at the start of many genes to assist RNA polymerase in melting the DNA for transcription.

Strand separation by gentle heating, as used in PCR, is simple providing the molecules have fewer than about 10,000 base pairs (10 kilobase pairs, or 10 kbp). The intertwining of the DNA strands makes long segments difficult to separate. The cell avoids this problem by allowing its DNA-melting enzymes (helicases) to work concurrently with topoisomerases, which can chemically cleave the phosphate backbone of one of the strands so that it can swivel around the other. Helicases unwind the strands to facilitate the advance of sequence-reading enzymes such as DNA polymerase.

Base pair geometry

The geometry of a base, or base pair step can be characterized by 6 coordinates: Shift, slide, rise, tilt, roll, and twist. These values precisely define the location and orientation in space of every base or base pair in a nucleic acid molecule relative to its predecessor along the axis of the helix. Together, they characterize the helical structure of the molecule. In regions of DNA or RNA where the "normal" structure is disrupted, the change in these values can be used to describe such disruption.

For each base pair, considered relative to its predecessor, there are the following base pair geometries to consider:

- **Shear**
- **Stretch**
- **Stagger**
- **Buckle**
- **Propeller twist**: rotation of one base with respect to the other in the same base pair.
- **Opening**
- **Shift**: displacement along an axis in the base-pair plane perpendicular to the first, directed from the minor to the major groove.
- **Slide**: displacement along an axis in the plane of the base pair directed from one strand to the other.
- **Rise**: displacement along the helix axis.
- **Tilt**: rotation around this axis.
- **vRoll**: rotation around this axis.
- **Twist**: rotation around the helix axis.
- **vx-displacement**
- **y-displacement**
- **inclination**
- **tip**
- **pitch**: the number of base pairs per complete turn of the helix.

Rise and twist determine the handedness and pitch of the helix. The other coordinates, by contrast, can be zero. Slide and shift are typically small in B-DNA, but are substantial in A- and Z-DNA. Roll and tilt make successive base pairs less parallel, and are typically small. A diagram of these coordinates can be found in 3DNA website.

Note that "tilt" has often been used differently in the scientific literature, referring to the deviation of the first, inter-strand base-pair axis from perpendicularity to the helix axis. This corresponds to slide between a succession of base pairs, and in helix-based coordinates is properly termed "inclination".

DNA helix geometries

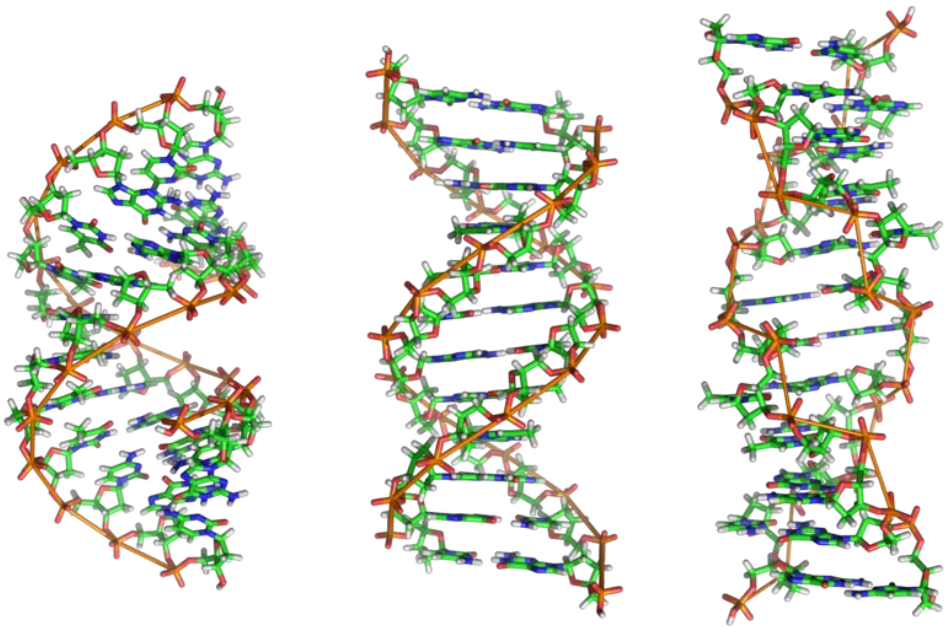
At least three DNA conformations are believed to be found in nature, A-DNA, B-DNA, and Z-DNA. The "B" form described by James D. Watson and Francis Crick is believed to predominate in cells. It is 23.7 Å wide and extends 34 Å per 10 bp of sequence. The double helix makes one complete turn about its axis every 10.4-10.5 base pairs in solution. This frequency of twist (known as the helical *pitch*) depends largely on stacking forces that each base exerts on its neighbours in the chain.

Other conformations are possible; A-DNA, B-DNA, C-DNA, E-DNA, L-DNA (the enantiomeric form of D-DNA), P-DNA, S-DNA, Z-DNA, etc. have been described so far. In fact, only the letters F, Q, U, V, and Y are now available to describe any new DNA

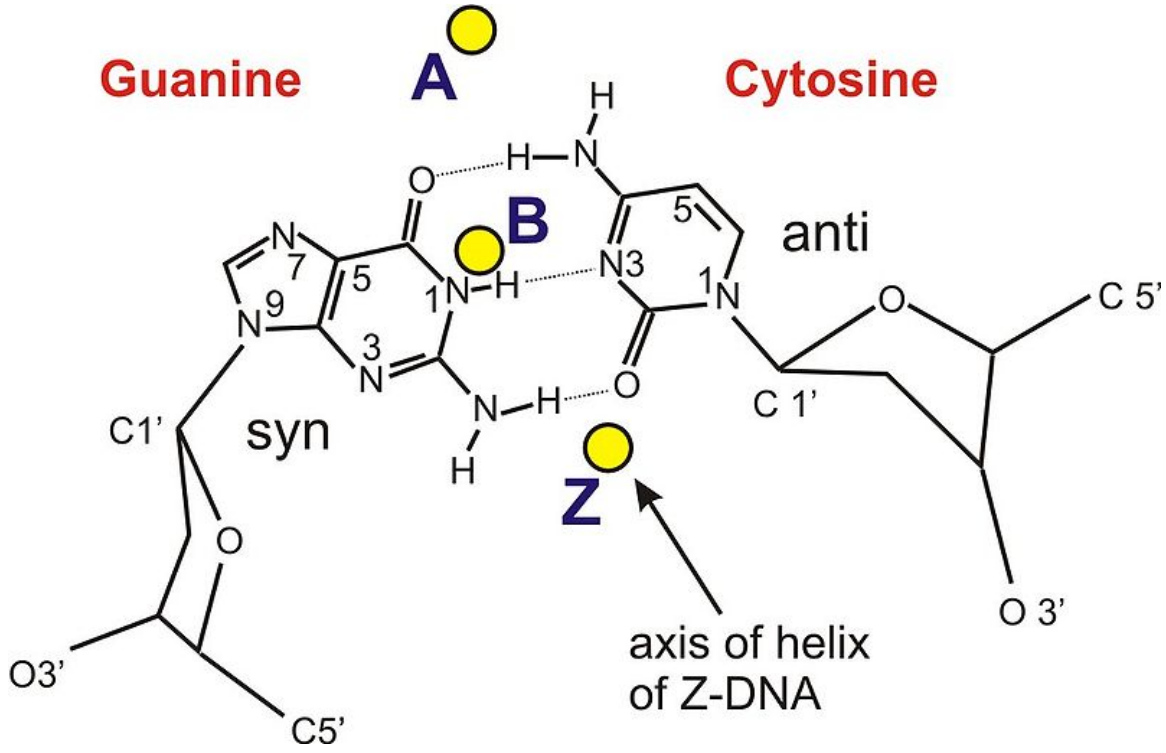
structure that may appear in the future. However, most of these forms have been created synthetically and have not been observed in naturally occurring biological systems. Also note the triple-stranded DNA possibility.

A- and Z-DNA

A-DNA and Z-DNA differ significantly in their geometry and dimensions to B-DNA, although still form helical structures. The A form appears likely to occur only in dehydrated samples of DNA, such as those used in crystallographic experiments, and possibly in hybrid pairings of DNA and RNA strands. Segments of DNA that cells have methylated for regulatory purposes may adopt the Z geometry, in which the strands turn about the helical axis the opposite way to A-DNA and B-DNA. There is also evidence of protein-DNA complexes forming Z-DNA structures.



The structures of A-, B-, and Z-DNA



The helix axis of A-, B-, and Z-DNA

Structural features of the three major forms of DNA

Geometry attribute	A-DNA	B-DNA	Z-DNA
Helix sense	right-handed	right-handed	left-handed
Repeating unit	1 bp	1 bp	2 bp
Rotation/bp	32.7°	35.9°	60°/2
bp/turn	11	10.5	12
Inclination of bp to axis	+19°	−1.2°	−9°
Rise/bp along axis	2.3 Å (0.23 nm)	3.32 Å (0.332 nm)	3.8 Å (0.38 nm)
Pitch/turn of helix	28.2 Å (2.82 nm)	33.2 Å (3.32 nm)	45.6 Å (4.56 nm)
Mean propeller twist	+18°	+16°	0°
Glycosyl angle	anti	anti	C: anti, G: syn
Sugar pucker	C3'-endo	C2'-endo	C: C2'-endo, G: C2'-exo
Diameter	23 Å (2.3 nm)	20 Å (2.0 nm)	18 Å (1.8 nm)

Supercoiled DNA

The B form of the DNA helix twists 360° per 10.4-10.5 bp in the absence of torsional strain. But many molecular biological processes can induce torsional strain. A DNA

segment with excess or insufficient helical twisting is referred to, respectively, as positively or negatively "supercoiled". DNA *in vivo* is typically negatively supercoiled, which facilitates the unwinding (melting) of the double-helix required for RNA transcription.

Non-helical forms

Other non-double helical forms of DNA have been described, for example side-by-side (SBS) and triple helical configurations. Single stranded DNA may exist *in statu nascendi* or as thermally induced despiralized DNA.

DNA bending

DNA is a relatively rigid polymer, typically modelled as a worm-like chain. It has three significant degrees of freedom; bending, twisting and compression, each of which cause particular limitations on what is possible with DNA within a cell. Twisting/torsional stiffness is important for the circularisation of DNA and the orientation of DNA bound proteins relative to each other and bending/axial stiffness is important for DNA wrapping and circularisation and protein interactions. Compression/extension is relatively unimportant in the absence of high tension.

Example sequences and their persistence lengths (B DNA)

Sequence	Persistence Length /base pairs
Random	154±10
$(\mathbf{CA})_{repeat}$	133±10
$(\mathbf{CAG})_{repeat}$	124±10
$(\mathbf{TATA})_{repeat}$	137±10

Persistence length/Axial stiffness

DNA in solution does not take a rigid structure but is continually changing conformation due to thermal vibration and collisions with water molecules, which makes classical measures of rigidity impossible. Hence, the bending stiffness of DNA is measured by the persistence length, defined as:

> "The length of DNA over which the time-averaged orientation of the polymer becomes uncorrelated by a factor of *e*".

This value may be directly measured using an atomic force microscope to directly image DNA molecules of various lengths. In an aqueous solution, the average persistence length is 46-50 nm or 140-150 base pairs (the diameter of DNA is 2 nm), although can vary significantly. This makes DNA a moderately stiff molecule.

The persistence length of a section of DNA is somewhat dependent on its sequence, and this can cause significant variation. The variation is largely due to base stacking energies and the residues which extend into the minor and major grooves.

Models for DNA bending

Stacking stability of base steps (B DNA)

Step	Stacking ΔG /kcal mol^{-1}
T A	-0.19
T G or **C A**	-0.55
C G	-0.91
A G or **C T**	-1.06
A A or **T T**	-1.11
A T	-1.34
G A or **T C**	-1.43
C C or **G G**	-1.44
A C or **G T**	-1.81
G C	-2.17

The entropic flexibility of DNA is remarkably consistent with standard polymer physics models such as the *Kratky-Porod* worm-like chain model. Consistent with the worm-like chain model is the observation that bending DNA is also described by Hooke's law at very small (sub-piconewton) forces. However for DNA segments less than the persistence length, the bending force is approximately constant and behaviour deviates from the worm-like chain predictions.

This effect results in unusual ease in circularising small DNA molecules and a higher probability of finding highly bent sections of DNA.

Bending preference

DNA molecules often have a preferred direction to bend, ie. anisotropic bending. This is, again, due to the properties of the bases which make up the DNA sequence - a random sequence will have no preferred bend direction, i.e. isotropic bending.

Preferred DNA bend direction is determined by the stability of stacking each base on top of the next. If unstable base stacking steps are always found on one side of the DNA helix then the DNA will preferentially bend away from that direction. As bend angle increases then steric hindrances and ability to roll the residues relative to each other also play a role, especially in the minor groove. **A** and **T** residues will be preferentially be found in the minor grooves on the inside of bends. This effect is particularly seen in DNA-protein binding where tight DNA bending is induced, such as in nucleosome particles.

DNA molecules with exceptional bending preference can become intrinsically bent. This was first observed in trypanosomatid kinetoplast DNA. Typical sequences which cause this contain stretches of 4-6 **T** and **A** residues separated by **G** and **C** rich sections which keep the A and T residues in phase with the minor groove on one side of the molecule. For example:

```
 |                 |                 |                 |
|                 |
G A T T C C C A A A A A T G T C A A A A A A T A G G C A A A A A A T G C
C A A A A A A T C C C A A A C
```

The intrinsically bent structure is induced by the 'propeller twist' of base pairs relative to each other allowing unusual bifurcated Hydrogen-bon ds between base steps. At higher temperatures this structure, and so the intrinsic bend, is lost.

All DNA which bends anisotropically has, on average, a longer persistence length and greater axial stiffness. This increased rigidity is required to prevent random bending which would make the molecule act isotropically.

DNA circularization

DNA circularization depends on both the axial (bending) stiffness and torsional (rotational) stiffness of the molecule. For a DNA molecule to successfully circularize it must be long enough to easily bend into the full circle and must have the correct number of bases so the ends are in the correct rotation to allow bonding to occur. The optimum length for circularization of DNA is around 400 base pairs (136 nm), with an integral number of turns of the DNA helix, i.e. multiples of 10.4 base pairs. Having a non integral number of turns presents a significant energy barrier for circularization, for example a 10.4 x 30 = 312 base pair molecule will circularize hundreds of times faster than 10.4 x 30.5 ≈ 317 base pair molecule.

DNA stretching

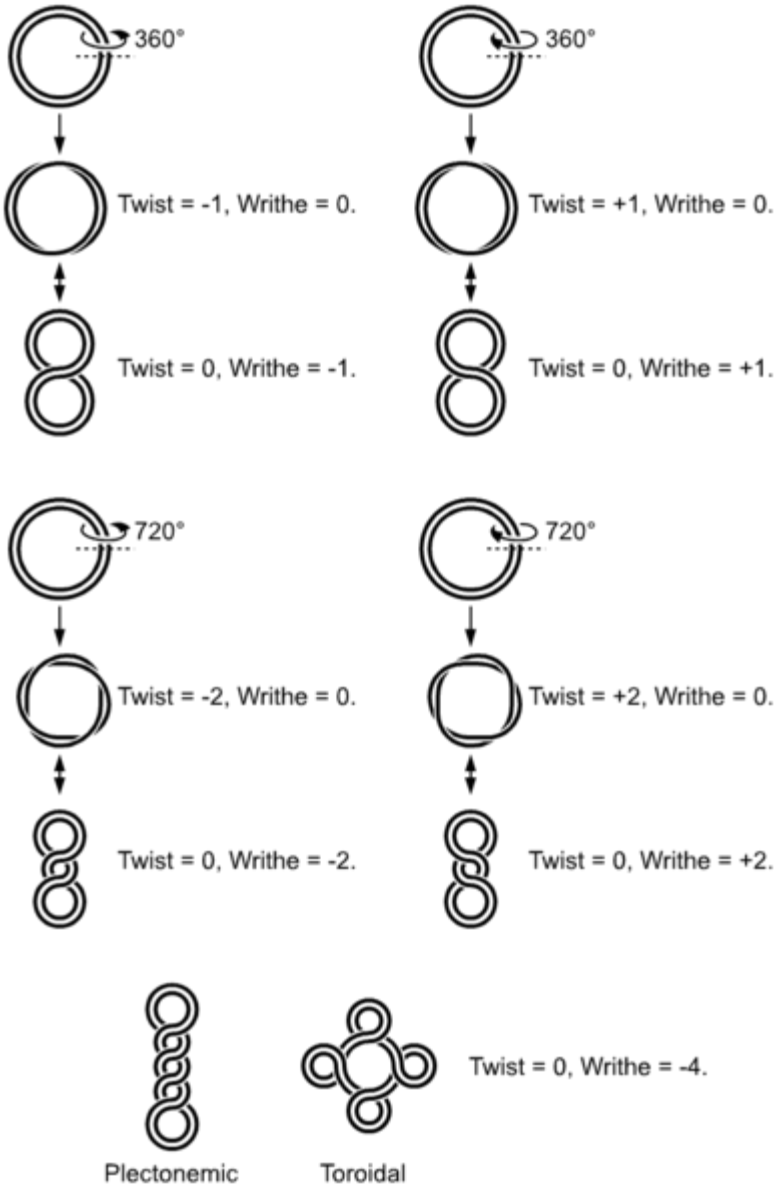
Longer stretches of DNA are entropically elastic under tension. When DNA is in solution, it undergoes continuous structural variations due to the energy available in the thermal bath of the solvent. This is due to the thermal vibration of the molecule combined with continual collisions with water molecules. For entropic reasons, more compact relaxed states are thermally accessible than stretched out states, and so DNA molecules are almost universally found in a tangled relaxed layouts. For this reason, a single molecule of DNA will stretch under a force, straightening it out. Using optical tweezers, the entropic stretching behavior of DNA has been studied and analyzed from a polymer physics perspective, and it has been found that DNA behaves largely like the *Kratky-Porod* worm-like chain model under physiologically accessible energy scales.

Under sufficient tension and positive torque, DNA is thought to undergo a phase transition with the bases splaying outwards and the phosphates moving to the middle.

This proposed structure for overstretched DNA has been called "P-form DNA," in honor of Linus Pauling who originally presented it as a possible structure of DNA

The mechanical properties DNA under compression have not been characterized due to experimental difficulties in preventing the polymer from bending under the compressive force.

DNA topology



Supercoiled structure of circular DNA molecules with low writhe. Note that the helical nature of the DNA duplex is omitted for clarity.

Within the cell most DNA is topologically restricted. DNA is typically found in closed loops (such as plasmids in prokaryotes) which are topologically closed, or as very long molecules whose diffusion coefficients produce effectively topologically closed domains. Linear sections of DNA are also commonly bound to proteins or physical structures (such as membranes) to form closed topological loops.

Francis Crick was one of the first to propose the importance of linking numbers when considering DNA supercoils. In a paper published in 1976, Crick outlined the problem as follows:

In considering supercoils formed by closed double-stranded molecules of DNA certain mathematical concepts, such as the linking number and the twist, are needed. The meaning of these for a closed ribbon is explained and also that of the writhing number of a closed curve. Some simple examples are given, some of which may be relevant to the structure of chromatin.

Analysis of DNA topology uses three values:

L = linking number - the number of times one DNA strand wraps around the other. It is an integer for a closed loop and constant for a closed topological domain.
T = twist - total number of turns in the double stranded DNA helix. This will normally tend to approach the number of turns that a topologically open double stranded DNA helix makes free in solution: number of bases/10.5, assuming there are no intercalating agents (e.g., chloroquine) or other elements modifying the stiffness of the DNA.
W = writhe - number of turns of the double stranded DNA helix around the superhelical axis
$L = T + W$ and $\Delta L = \Delta T + \Delta W$

Any change of T in a closed topological domain must be balanced by a change in W, and vice versa. This results in higher order structure of DNA. A circular DNA molecule with a writhe of 0 will be circular. If the twist of this molecule is subsequently increased or decreased by supercoiling then the writhe will be appropriately altered, making the molecule undergo plectonemic or toroidal superhelical coiling.

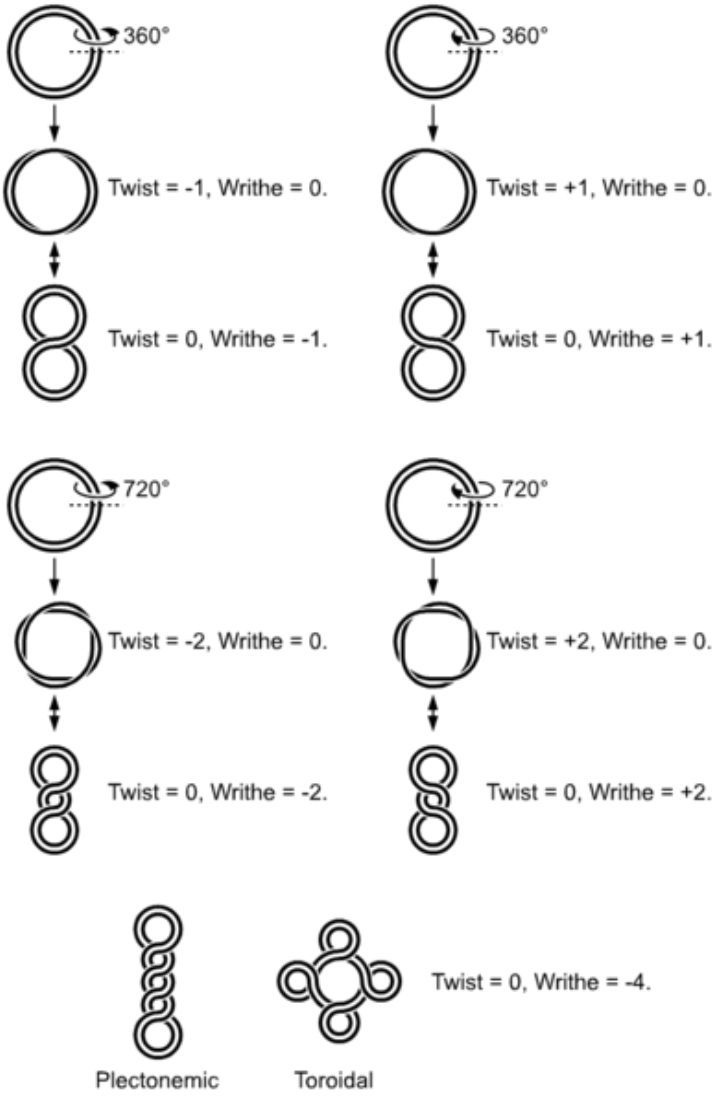
When the ends of a piece of double stranded helical DNA are joined so that it forms a circle the strands are topologically knotted. This means the single strands cannot be separated any process that does not involve breaking a strand (such as heating). The task of un-knotting topologically linked strands of DNA falls to enzymes known as topoisomerases. These enzymes are dedicated to un-knotting circular DNA by cleaving one or both strands so that another double or single stranded segment can pass through. This un-knotting is required for the replication of circular DNA and various types of recombination in linear DNA which have similar topological constraints.

The linking number paradox

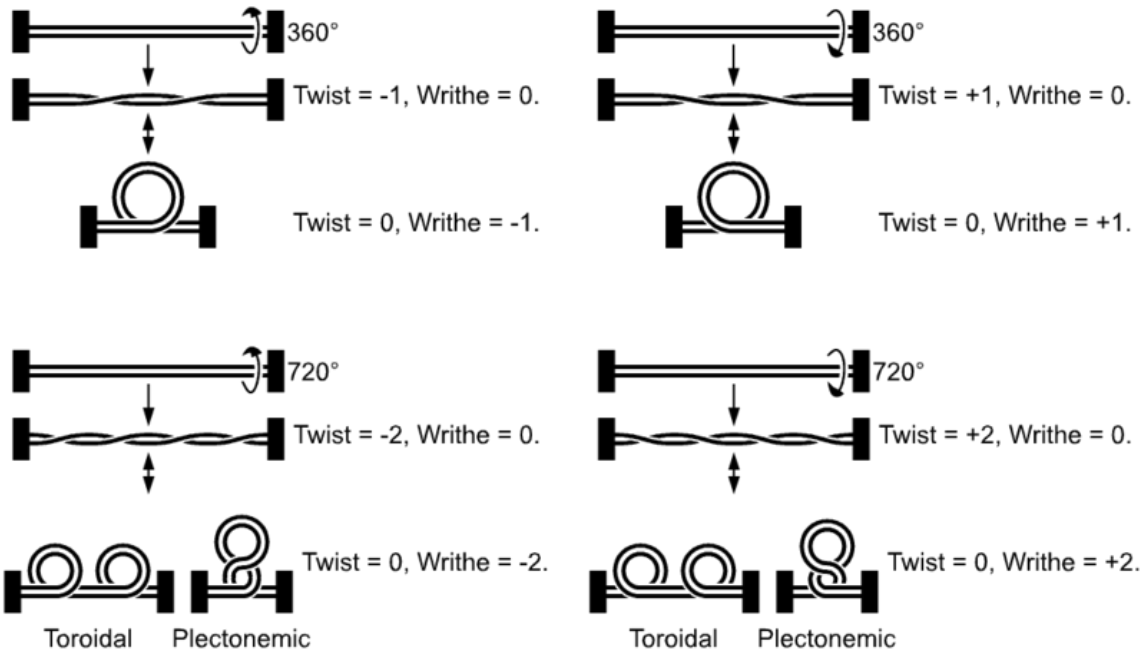
For many years, the origin of residual supercoiling in eukaryotic genomes remained unclear. This topological puzzle was referred to by some as the "linking number paradox". However, when experimentally determined structures of the nucleosome displayed an overtwisted left-handed wrap of DNA around the histone octamer, this "paradox" was solved.

Chapter- 12

DNA Supercoil



Supercoiled structure of circular DNA molecules with low writhe. Note that the helical nature of the DNA duplex is omitted for clarity.



Supercoiled structure of linear DNA molecules with constrained ends. Note that the helical nature of the DNA duplex is omitted for clarity.

In a "relaxed" double-helical segment of **B-DNA**, the two strands twist around the helical axis once every 10.4-10.5 base pairs of sequence. Adding or subtracting twists, as some enzymes can do, imposes strain. If a DNA segment under twist strain were to be closed into a circle by joining its two ends and then it is allowed to move freely, the circular DNA would contort into a new shape, such as a simple figure-eight. Such a contortion is a **supercoil**.

The simple figure eight is the simplest supercoil, and is the shape a circular DNA assumes to accommodate one too many or one too few helical twists. The two lobes of the figure eight will appear rotated either clockwise or counterclockwise with respect to one another, depending on whether the helix is over or underwound. For each additional helical twist being accommodated, the lobes will show one more rotation about their axis.

The noun form "supercoil" is rarely used in the context of DNA topology. Instead, global contortions of a circular DNA, such as the rotation of the figure-eight lobes above, are referred to as *writhe*. The above example illustrates that twist and writhe are interconvertible. "**Supercoiling**" is an abstract mathematical property representing the sum of twist and writhe. The twist is the number of helical turns in the DNA and the writhe is the number of times the double helix crosses over on itself (these are the supercoils).

Extra helical twists are positive and lead to positive supercoiling, while subtractive twisting causes negative supercoiling. Many topoisomerase enzymes sense supercoiling

and either generate or dissipate it as they change DNA topology. DNA of most organisms is negatively supercoiled.

In part because chromosomes may be very large, segments in the middle may act as if their ends are anchored. As a result, they may be unable to distribute excess twist to the rest of the chromosome or to absorb twist to recover from underwinding--the segments may become *supercoiled*, in other words. In response to supercoiling, they will assume an amount of writhe, just as if their ends were joined.

Supercoiled DNA forms two structures; a plectoneme or a toroid, or a combination of both. A negatively supercoiled DNA molecule will produce either a one-start left-handed helix, the toroid, or a two-start right-handed helix with terminal loops, the plectoneme. Plectonemes are typically more common in nature, and this is the shape most bacterial plasmids will take. For larger molecules it is common for hybrid structures to form - a loop on a toroid can extend into a plectoneme. If all the loops on a toroid extend then it becomes a branch point in the plectonemic structure.

Occurrence of DNA supercoiling

DNA supercoiling is important for DNA packaging within all cells. Because the length of DNA can be thousands of times that of a cell, packaging this genetic material into the cell or nucleus (in eukaryotes) is a difficult feat. Supercoiling of DNA reduces the space and allows for a lot more DNA to be packaged. In prokaryotes, plectonemic supercoils are predominant, because of the circular chromosome and relatively small amount of genetic material. In eukaryotes, DNA supercoiling exists on many levels of both plectonemic and solenoidal supercoils, with the solenoidal supercoiling proving most effective in compacting the DNA. Solenoidal supercoiling is achieved with histones to form a 10nm fiber. This fiber is further coiled into a 30nm fiber, and further coiled upon itself numerous times more.

DNA packaging is greatly increased during nuclear division events such as mitosis or meiosis, where DNA must be compacted and segregated to daughter cells. Condensins and cohesins are *Structural Maintenance of Chromosome* proteins that aid in the condensation of sister chromatids and the linkage of the centromere in sister chromatids. These SMC proteins induce positive supercoils.

Supercoiling is also required for DNA/RNA synthesis. Because DNA must be unwound for DNA/RNA polymerase action, supercoils will result. The region ahead of the polymerase complex will be unwound; this stress is compensated with positive supercoils ahead of the complex. Behind the complex, DNA is rewound and there will be **compensatory** negative supercoils. It is important to note that topoisomerases such as DNA gyrase (Type II Topoisomerase) play a role in relieving some of the stress during DNA/RNA synthesis.

Modeling using mathematics

DNA supercoiling can be described numerically by changes in the 'linking number' Lk. The linking number is the most descriptive property of supercoiled DNA. Lk_o, the number of turns in the relaxed (B type) DNA plasmid/molecule, is determined by dividing the total base pairs of the molecule by the relaxed bp/turn which, depending on reference is 10.4-10.5.

$$Lk_o = bp / 10.4$$

Lk is merely the number of crosses a single strand makes across the other in a planar projection. The topology of the DNA is described by the equation below in which the linking number is equivalent to the sum of TW, which is the number of twists or turns of the double helix, and Wr which is the number of coils or 'writhes'. If there is a closed DNA molecule, the sum of TW and Wr, or the linking number, does not change. However, there may be complementary changes in TW and Wr without changing their sum.

$$Lk = Tw + Wr$$

The change in the linking number, ΔLk, is the actual number of turns in the plasmid/molecule, Lk, minus the number of turns in the relaxed plasmid/molecule Lk_o.

$$\Delta Lk = Lk - Lk_o$$

If the DNA is negatively supercoiled $\Delta Lk < 0$. The negative supercoiling implies that the DNA is underwound.

A standard expression independent of the molecule size is the "specific linking difference" or "superhelical density" denoted σ. σ represents the number of turns added or removed relative to the total number of turns in the relaxed molecule/plasmid, indicating the level of supercoiling.

$$\sigma = \Delta Lk / Lk_o$$

The Gibbs free energy associated with the coiling is given by the equation below

$$\Delta G / N = 10RT\sigma^2$$

Examples

Since the linking number L of supercoiled DNA is the number of times the two strands are intertwined (and both strands remain covalently intact), L cannot change. The reference state (or parameter) L_0 of a circular DNA duplex is its relaxed state. In this state, its writhe $W = 0$. Since $L = T + W$, in a relaxed state $T = L$. Thus, if we have a 400

bp relaxed circular DNA duplex, $L \sim 40$ (assuming ~10 bp per turn in B-DNA). Then $T \sim 40$.

- Positively supercoiling:

  T = 0, W = 0, then L = 0
  T = +3, W = 0, then L = +3
  T = +2, W = +1, then L = +3

- Negatively supercoiling:

  T = 0, W = 0, then L = 0
  T = -3, W = 0, then L = -3
  T = -2, W = -1, then L = -3

Negative supercoils favor local unwinding of the DNA, allowing processes such as transcription, DNA replication, and recombination. Negative supercoiling is also thought to favour the transition between B-DNA and Z-DNA, and moderate the interactions of DNA binding proteins involved in gene regulation.