6 DNA Sequencing

This exposition is based closely on the following sources, which are all recommended reading:


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

DNA sequences play an important role in many different fields, such as:

- diagnostics,
- biotechnology,
- forensic biology and
- biological systematics.

6.0.1 Sanger Sequencing

In 1975, Sanger introduced the *plus and minus* method for DNA sequencing (Sanger and Coulson, 1975).

Maxam and Gilbert (1977) developed a DNA sequencing method that was similar to the Sanger and Coulson method.

Both methods were soon replaced by the *chain-terminator* method (Sanger, Nicklen, Coulson, 1977), also known as *Sanger sequencing* (after its developer Frederick Sanger) or *dideoxy sequencing*, as it was more efficient and used fewer toxic chemicals and lower amounts of radioactivity.

The key principle of the Sanger method is the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-terminator method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, radioactively or fluorescently labeled nucleotides, and modified nucleotides that terminate DNA strand elongation.

The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase.

In each of the reactions, only one of four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is added. These nucleotides lack the 3'-OH group required for the formation of a phosphodiester bond between two nucleotides and thus terminate DNA strand extension and result in DNA fragments of varying length:
The synthesized DNA fragments are separated by size using gel electrophoresis, 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 above, 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 used to read the DNA sequence. The obtainable sequence length is \( \approx 100 \) nucleotides.

### 6.0.2 Modern Sanger Sequencing

In the **dye-terminator** variant of Sanger sequencing, each of the four dideoxynucleotide chain terminators is labelled by a different fluorescent dye, permitting sequencing in a single reaction rather than four:

In 1986 Leroy Hood and Applied Biosystems (ABI) introduced the dye-terminator variant of Sanger sequencing in the first report on automation of DNA sequencing (Smith et al, 1986). They showed that sequencing data could be collected directly to a computer.

Sanger sequencing using dye-terminators became the dominant sequencing technique until the introduction of so-called **next-generation** sequencing technologies in 2005.

The obtainable sequence length for modern Sanger sequencing is \( \approx 1000 \) nucleotides.

The **ABI 370A DNA sequencer** was used in 1987 to determine the sequence of a gene by Craig Venter and colleagues at NIH (National Institutes of Health in Bethesda, MD). At NIH, Venter set up a sequencing facility with six automated sequencers and two ABI Catalyst robots.

In 1992 Venter established *The Institute for Genomic Research (TIGR)* to expand his sequencing operation to a facility with 30 ABI 373A automated sequencers and 17 ABI Catalyst 800 robots.

This was a real factory with teams dedicated to different steps in the sequencing process such as template preparation, gel pouring and sequencer operation.

Data analysis was integrated into the process so that problems in earlier steps could be detected and corrected as soon as possible.
6.1 Cellular genomes

Before 1995, only small viral and organelle DNA molecules had been completely sequenced. That year Craig Venter and co-workers reported the complete genome sequences of two bacterial species, *Haemophilus influenzae* (Fleischmann et al, 1995) and *Mycoplasma genitalium* (Frazer et al, 1995).

The *H. influenzae* sequence was the first complete instruction set of a living organism.

The methods used to obtain these sequences were as important for subsequent events as the biological insights they revealed.

Sequencing of *H. influenzae* introduced the whole genome shotgun (WGS) method for sequencing cellular genomes.

In this method, genomic DNA is fragmented randomly and cloned to produce a random library in *E. coli*.

Clones are sequenced at random and the results are assembled to produce the complete genome sequence by a computer program that compares all of the sequence reads and aligns matching sequences.

Sanger and colleagues used this general strategy to sequence the lambda phage genome (48.5 kb), published in 1982. However, no larger genome was shotgun sequenced until *H. influenzae* (1.83 Mb).

Outline of the basic WGS strategy:

Another critical factor in the application of shotgun sequencing to cellular genomes was the TIGR assembler (Sutton et al, 1995).

Previous assembly programs were not designed to handle thousands of sequence reads involved in even the smallest cellular genome projects.

However, the TIGR assembler that had been designed to assemble vast amounts of data was adequate for the job.

Many genomes were subsequently sequenced and assembled, such as:

- The bacteria *E. coli* and *Bacillus subtilis*,
- the yeast *Saccharomyces cerevisiae*,
- the nematode *C. elegans*,
- the fruit fly *Drosophila melanogaster*,

![DNA extraction](source.png)
• the plant *Arabidopsis thaliana*, and
• the human genome.

### 6.2 The Human Genome

The US Human Genome Project established goals of mapping and sequencing the human genome, as well as some other model organisms.

In 1994, a detailed genetic map of the human genome was published including 5840 mapped loci with a mean spacing of 0.7 cM (1 centiMorgan $\approx 10^6$ bp).

In 1998 the public project, now in a race with the private company Celera Genomics, adopted the new ABI Prism 3700 capillary sequencers.

In 1999 the Human Genome Project celebrated passing the billion base-pair mark, and the first complete sequence of a human chromosome was reported (chromosome 22).

The Human Genome Project was using a slow-but-sure approach of sequencing and assembling the human genome piece by piece.

The approach adopted at Celera was to sequence and assemble the whole human genome in one large computation, emphasizing bioinformatics and simplified lab protocols.

Human genome sequencing began in September 1999 and continued until June 2000, when data collection was completed and an initial assembly was achieved.

The Celera data provided approximately 5-fold coverage of the genome. An additional 3-fold coverage of unordered and unoriented BAC sequences from the public effort was included in the assembly. This work demonstrated that the WGS strategy is applicable to large genomes.

Bioinformaticians led by Gene Myers at Celera developed the first computer program capable of assembling a mammalian-sized genome, establishing the whole genome shotgun approach as the method of choice for such projects.

(An article in “Die Zeit” asks whether Gene Myers is the one that deserves a Nobel prize: “Es wird sich noch zeigen, ob der legendäre Genom-Kartierer Craig Venter den Nobelpreis verdient hat oder nicht doch sein Chefmathematiker Gene Myers.”. [http://www.zeit.de/2004/50/C-Mathematik](http://www.zeit.de/2004/50/C-Mathematik))

On 25 June 2000, President Clinton, Francis Collins (NIH) and Craig Venter publicly announced the completion of the first draft versions of the human genome sequence from both the publicly funded project and from Celera.

In February 2001 the Celera and the public draft human genome sequences were published the same week in Science and Nature.

The race was officially a tie, but it was clear that the entry of Celera had shorted the process by several years.
6.3 Next Generation Sequencing

Since 2005, a number of new methods have emerged for sequencing.

The common feature of these methods is that they are **massively parallel**, meaning that the number of sequence reads from a single experiment is vastly greater than the 96 obtained with modern capillary electrophoresis-based Sanger sequencers.

At present, this very high throughput is achieved with sacrifices in length and accuracy of the individual reads when compared to Sanger sequencing. However, assemblies of such data can be highly accurate because of the high degree of sequence coverage obtainable.

The methods are designed for projects that employ the WGS approach. They are often applied to **resequencing**, in which sequence data is aligned with a reference genome sequence in order to look for differences from that reference.

6.3.1 Roche 454 Sequencing

The first massively parallel method to become commercially available was developed by 454 Life Sciences and is based on the **pyrosequencing** technique (Margulies et al, 2005).

This system allows shotgun sequencing of whole genomes without cloning in *E. coli* or any host cell. First DNA is randomly sheared and ligated to linker sequences that permit individual molecules captured on the surface of a bead to be amplified while isolated within an emulsion droplet.

A very large collection of such beads is arrayed in the 1.6 million wells of a fiber-optic slide. As with the Sanger method, sequencing is carried out using primed synthesis by DNA polymerase. The array is presented with each of the four dNTPs, sequentially, and the amount of incorporation is monitored by luminometric detection of the pyrophosphate released (hence the name “pyrosequencing”).

A CCD imager coupled to the fiber-optic array collects the data. In sequencing across a homopolymer run, the run length is estimated from the amount of pyrophosphate released, which is proportional to the number of residues incorporated. Errors that result from misjudging the length of homopolymer runs result in single-base insertions and deletions (indels). These constitute the major source of errors in 454 data.

The 454 Genome Sequencer FLX is reportedly able to produce 100 Mb of sequence with 99.5% accuracy for individual reads averaging over 250 bases in length.

**Emulsion-based clonal amplification:**

![Emulsion-based clonal amplification](image-url)
Deposit DNA beads into a PicoTiterPlate:

Parallel sequencing of fragments:

Base-calling is performed by analyzing a series of images produced in the flow cycles:

6.3.2 Illumina (Solexa) Sequencing

The second next-generation sequencing technology to be released (in 2006) was Illumina (Solexa) Sequencing (Bennett, 2004).
A key difference between this method and the 454 is that it uses chain-terminating nucleotides. The fluorescent label on the terminating base can be removed to leave an unblocked 3’ terminus, making chain termination a reversible process. The method reads each base in a homopolymer run in a separate step and therefore does not produce as many indels within such runs as the 454.

Because the reversible dye terminator nucleotides are not incorporated efficiently, the read length of the Solexa method is less than for 454. Also more base-substitution errors are observed due to the use of modified polymerase and dye terminator nucleotides. The method sequences clusters of DNA molecules amplified from individual fragments attached randomly on the surface of a flow cell.

Due to the very high densities of clusters that can be analyzed, the latest sequencers can produce billions of (paired) reads of length $\geq 100$ in a single run.

1. Produce fragments and add adapters:

   - [Diagram]

   **Prepare genomic DNA sample**
   - Randomly fragment genomic DNA and ligate adapters to both ends of the fragments

2. Attach fragments to surface:

   - [Diagram]

   **Attach DNA to surface**
   - Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Add nucleotides for amplification:

Make fragments double stranded:

Denature double strands:
Repeat and complete amplification:

Add all four labeled nucleotides:

Measure intensities of nucleotide colors:
Add next nucleotides:

Repeat measurement:

Repeat cycles to sequence bases:
Solexa workflow:

The Illumina HiSeq2000 can produce nearly three billion reads in a single run.

The Weigel lab at the MPI for Developmental Biology here in Tübingen has a number of Illumina sequencers.

6.3.3 ABI’s SOLiD sequencing

Applied Biosystems has also developed a massively parallel sequencer, its Supported Oligonucleotide Ligation and Detection system (SOLiD), released in 2008.

The technology is based on a hybridization-ligation chemistry (Shendure et al, 2005). The sample preparation aspect of this technology including library preparation, clonal amplification of the target DNA by emulsion PCR on beads is very similar to the 454 processes in principle.

However, the size of the beads used for emPCR (1 µm versus 26 µm) and the array format (random versus ordered) are different. These differences afford the SOLiD technology the potential of generating a significantly higher density sequencing array (potentially over a few hundred fold higher), as well as more flexibility in terms of sample input format.

The sequence interrogation is done through repeated cycles of hybridization of a mixture of sequencing primers and fluorescently labeled probes, followed by ligation of the sequencing primers and the probes, then the detection of the fluorescent signals on the probes which encode the bases that are being interrogated. Although it has a short read length of about 25 – 35, it can generate $\approx 2 – 3$ Gb of sequence per run.

The CeGaT company here in Tübingen uses SOLiD sequencers for analyzing medical samples.
6.3.4 Third Generation Sequencing

Third generation sequencing methods (such as Pacific Biosciences SMRT sequencing) are able to sequence individual DNA molecules, achieving read lengths that can be in the order of kilo bases. Current technologies are promising, but have high error rates.

6.4 Genomic Medicine

The sequencing of the human genome as well as all major pathogens is beginning to have a major impact on the diagnosis, treatment and prevention of diseases. Genome sequences have provided potential targets for drug therapy as well as vaccine candidates. An era of personalized medicine, informed by information concerning the patient’s genotype, has been widely predicted. The Human Microbiome Project seeks to understand the relationship between humans, their microbiome and health, by sequencing the microbes associated with the human body.

6.5 What Next?

The amount of nucleotide sequence in the databases has increased logarithmically by nine orders of magnitude over the 40-year period from 1965 to 2005:

This amounts to an average doubling time of about 16 months. The bottleneck of sequencing projects has shifted from the sequencing technology to the capacity for analyzing the data, provided by bioinformatics.