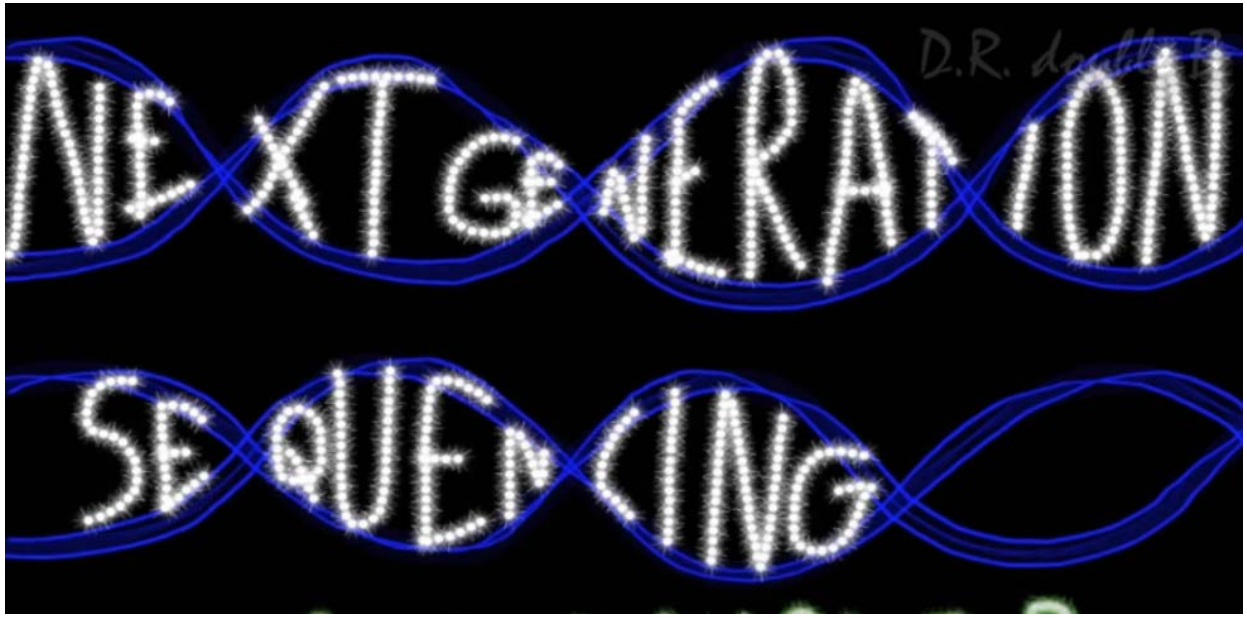
CS123A Bioinformatics Module 5 – Week 13 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ



Agenda

- Introduction to Next Generation Sequencing (NGS)
  - A Bit Of NGS History

NGS: A Working Definition

- **Next Generation Sequencing is the set of technologies and concepts that help determine the nucleotide order (i.e., sequence) of nucleic acids. NGS technologies have resulted in sequencing costs to drop precipitously over the past 4 years (with no end in sight)**

- **NGS is a "technological singularity" happening before us**
  - It is disruptive and hard to predict
  - It has already changed the world we live in
  - It will continue to change our lives, often in unpredictable ways
  - It has fundamentally changed both approaches to and the feasibility of even attempting certain scientific and technological problems

- **NGS will require a corresponding bioinformatics singularity to reach its full potential. This may already be underway…**

- What will YOU be doing in 5 years? It may well be related to NGS!

The Promise of NGS

- Personalized medicine
  - Cancer treatment
  - Discovery of causes of rare diseases (& maybe curing them!)
  - Prediction of future health issues, with recommendations for ameliorating them
  - Personalized pharmacogenomics: which drug will work best to treat a given disease in a given individual
- Characterizing entire microbial ecologies
- Molecular archaeology and anthropology
  - Discovering how humans became human & phylogeny relationships to other living organisms..
  - Tracing human population movements
- Plant & animal breeding
- Basic research (e.g. learning what "junk" DNA *really does)*
- ….

NGS History

First stabs….

- The first nucleic acid sequencing began in the mid-1960's using 2-dimensional chromatography

- The initial protocols were innovative, but inefficient by today's standards

- For example, in 1973 Gilbert and Maxam published the sequence of the 24 bp lac operator using a protocol that required 300-700 grams of bacteria, multiple purification steps, conversion of the selected DNA fragments into RNA and digestion of these sequences with different RNases (Proc Natl Acad Sci USA 1973;70:3581)

- Several properties of DNA made the initial attempts at DNA sequencing difficult:
  (1) The chemical properties of different DNA molecules were so similar that separating them appeared difficult
  (2) Compared to amino acids in proteins, DNA was much longer
  (3) No base-specific DNases were known and previous protein sequencing methods had depended upon proteases that cut adjacent to specific amino acids

Maxam-Gilbert DNA Sequencing

Maxam-Gibert DNA Sequencing

- Developed by Allan Maxim and Wally Gilbert in 1976-1977.

- Gilbert was awarded the 1980 Nobel Prize in Chemistry, shared with Frederick Sanger and Paul Berg. Gilbert and Sanger were recognized for their pioneering work in devising methods for determining the sequence of nucleotides in a nucleic acid.

- Maxam, A.; Gilbert, W. (1977). "A new method for sequencing DNA". *Proceedings of the National Academy of Sciences of the United States of America* **74** (2): 560–564.

Maxam-Gibert DNA Sequencing

- Maxam–Gilbert sequencing rapidly became popular, since purified DNA could be used directly

- DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases.

Maxam-Gibert DNA Sequencing

- The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-32P 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, A>G, C, C+T). For example, the purines (A,G) are depurinated using formic acid, the guanines (and to some extent the adenines, 5x slower) are methylated by dimethyl sulfate, and the pyrimidines (C,T) are methylated using hydrazine.

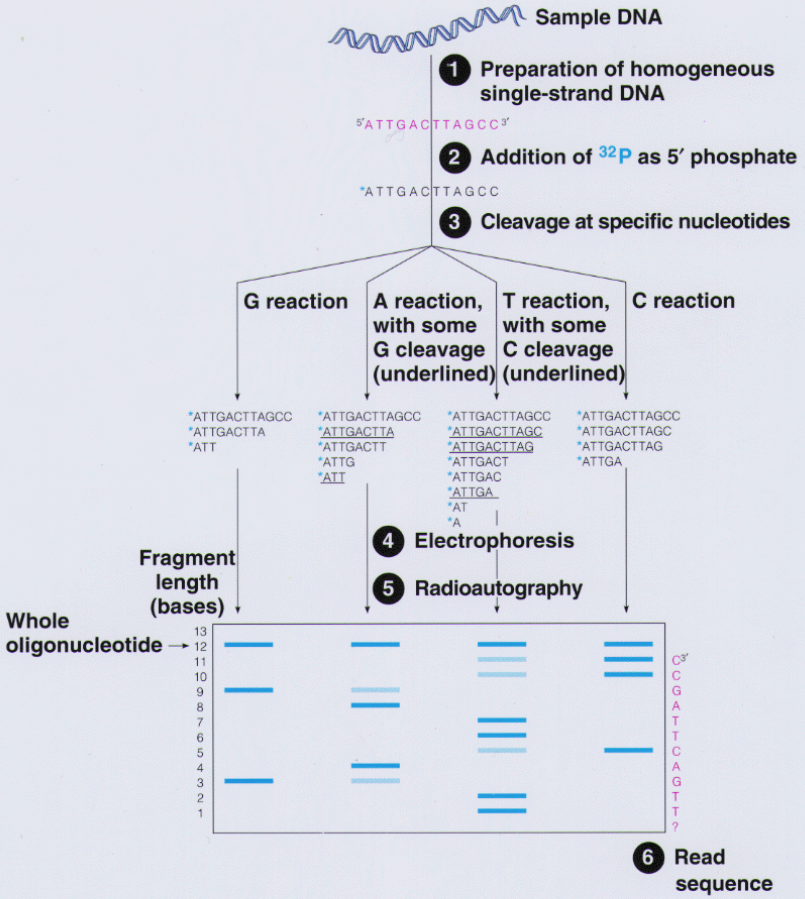
- The addition of 2M sodium chloride to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction.

- The modified DNAs are then cleaved by hot piperidine at the position of the modified base.

Maxam-Gibert DNA Sequencing

- 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.

Figure 4A.4 Sequencing an oligonucleotide by the Maxam-Gilbert method

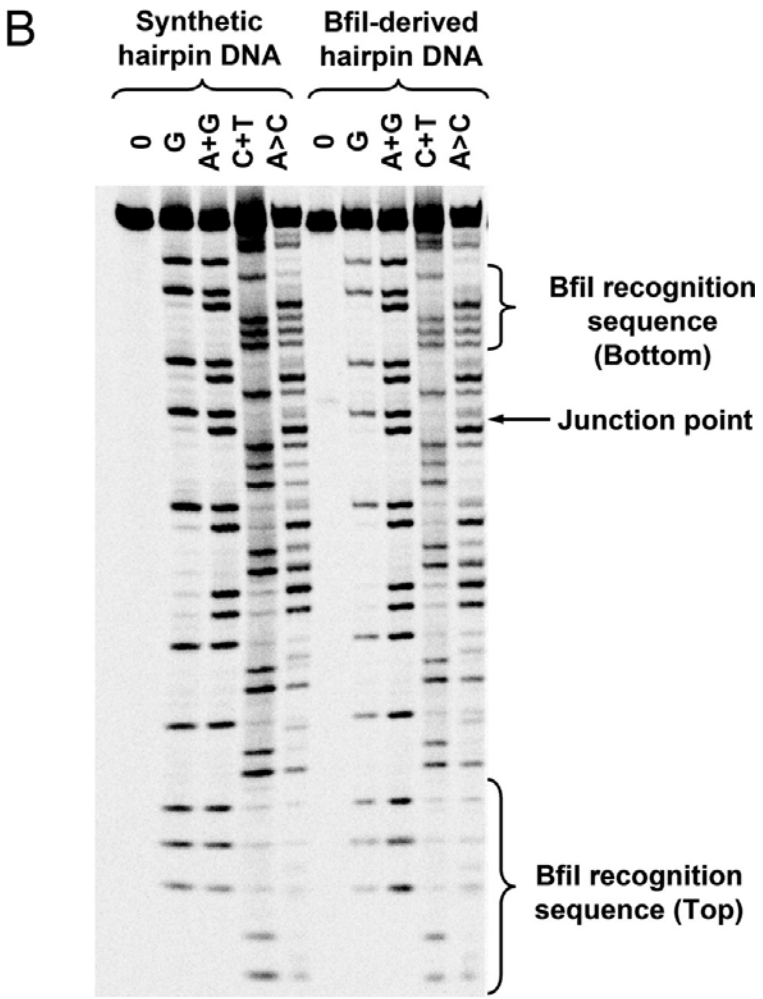


From Mathews and van Holde: *Biochemistry 2/e.* © The Benjamin/Cummings Publishing Co., Inc.

Some Observations

- Had to be done in a fume hood – very smelly, chemicals moderately dangerous.
- Gel interpretation required some mental gymnastics – lanes were G, A+G, C, C+T Therefore,
- **G** = bands in G AND A+G,
- **A**= band ONLY in A+G,
- **C** = bands in C AND C+T,
- **T** = band ONLY in C+T
- Sort of as above…some artifacts…
- Once you get used to it, you can do it relatively rapidly.
- Hard to keep your place on the autoradiogram and write down sequence – this was before PCs!! If there was any computer access at all, it was probably a terminal connected to a minicomputer, and it wasn't at your desk. One or a few terminals per floor.

Maxam–Gilbert sequencing.

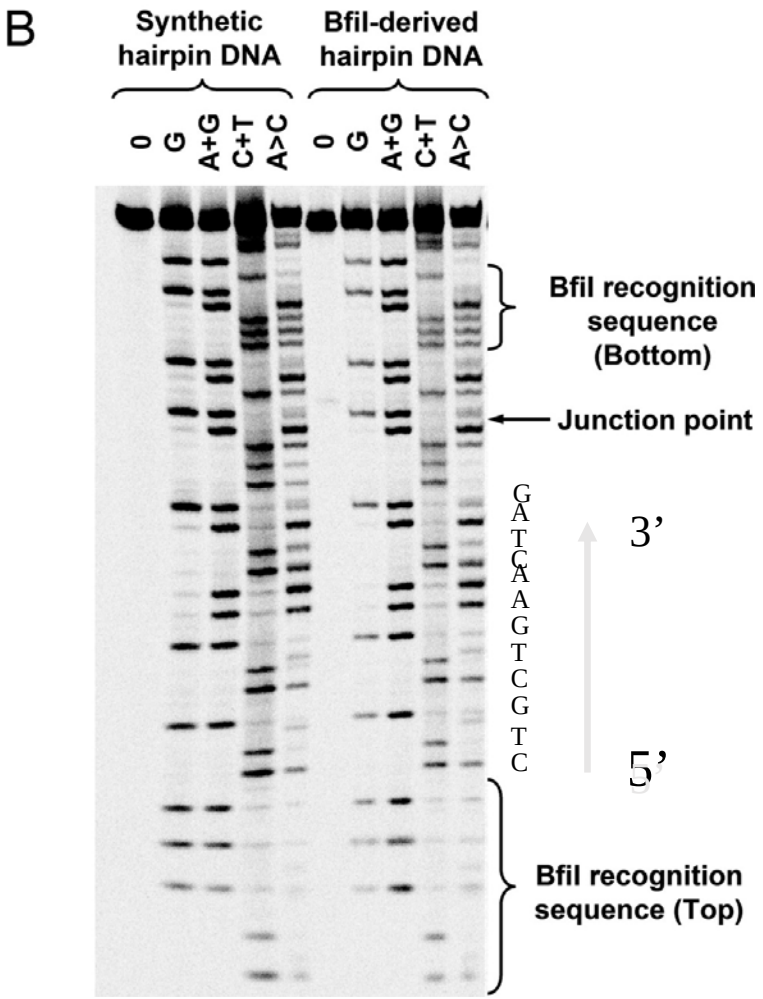


Can you read the sequence?

Sasnauskas G et al. PNAS 2007;104:2115-2120

PNAS

Maxam–Gilbert sequencing.



Can you read the sequence?

Sasnauskas G et al. PNAS 2007;104:2115-2120

PNAS

Additional Notes

- Also sometimes known as "chemical sequencing", this method led to the Methylation Interference Assay used to map DNA-binding sites for DNA-binding proteins.

- 100-250 bp/read

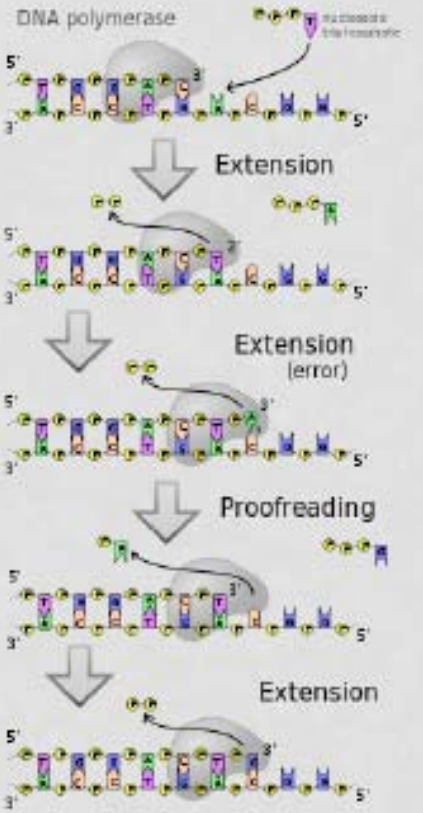
- Still used for short DNA, where you cannot synthesize a primer.

Sanger Sequencing

Quick DNA Review

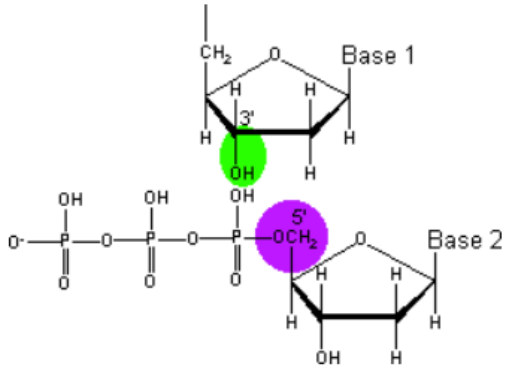
Natural Synthesis of DNA

- 5' to 3'
- By extension of an existing DNA or RNA strand (primer)
- Using a complementary (3' to 5') strand as template
- By addition of dNTP's (releasing pyrophosphate, PP_i)
- IF polymerase has 3'→5' exonuclease activity, it can "proofread" (correct errors)

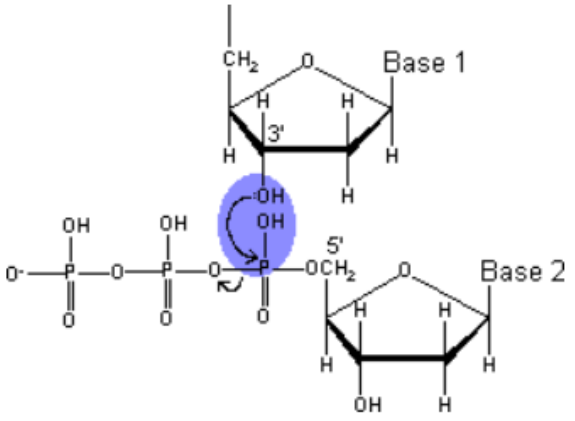


DNA Synthesis Process

- The 5’ group of a nucleotide triphosphate is held close to the 3’ hydroxyl group of a nucleotide chain.

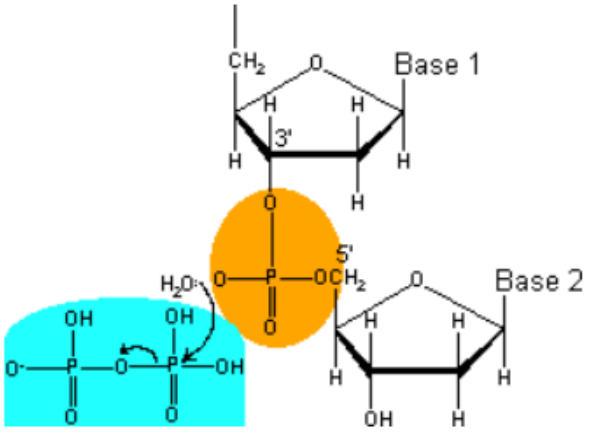


- The 3’hydroxyl group forms a bond to the phosphorous atom of the free nucleotide closest to the 5’ oxygen atom. Meanwhile the bond between the first phosphorus atom and the oxygen atom linking it to the next phosphate group breaks.

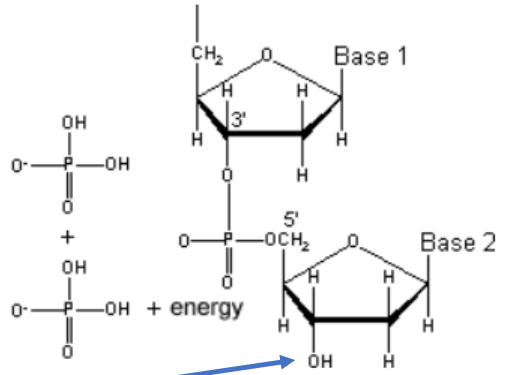


DNA Synthesis Process *(cont.)*

- A new phosphodiester bond now joins the two nucleotides. A pyrophosphate group has been liberated.



- The pyrophosphate group is hydrolyzed (split by the addition of water), releasing a lot of energy and driving the reaction forward to completion.



Remember this hydroxyl group on the 3' C. If OH is missing it is called A ddN (di-deoxy-nucleotide).

Chain-termination (Sanger) methods

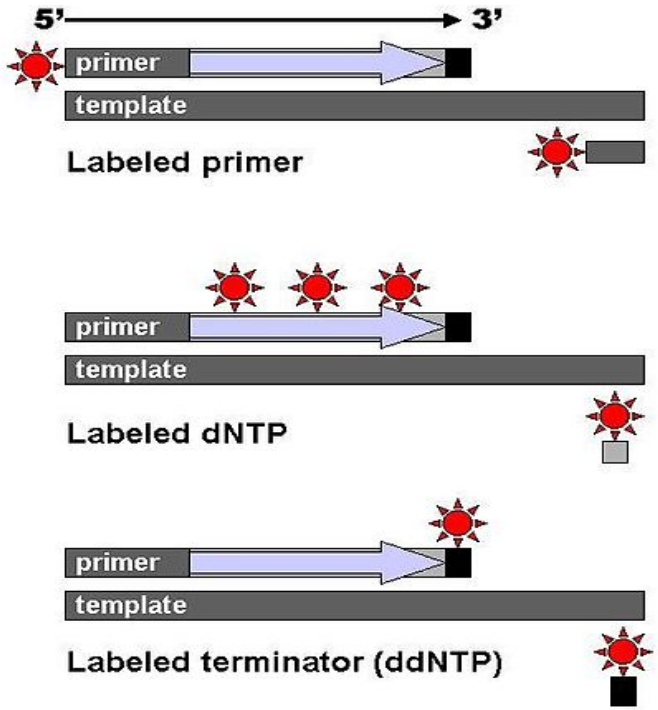
- Because the chain-terminator method (or Sanger method) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than Maxam/Gilbert, it rapidly became the method of choice.

- Sanger F, Nicklen S, Coulson AR (December 1977). "DNA sequencing with chain-terminating inhibitors". *Proc. Natl. Acad. Sci. U.S.A.* **74** (12): 5463–7 (assigned reading)

Chain-termination (Sanger) methods

- 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 were also radioactively 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, and the DNA sequence can be directly read off the X-ray film.

Chain Termination Method

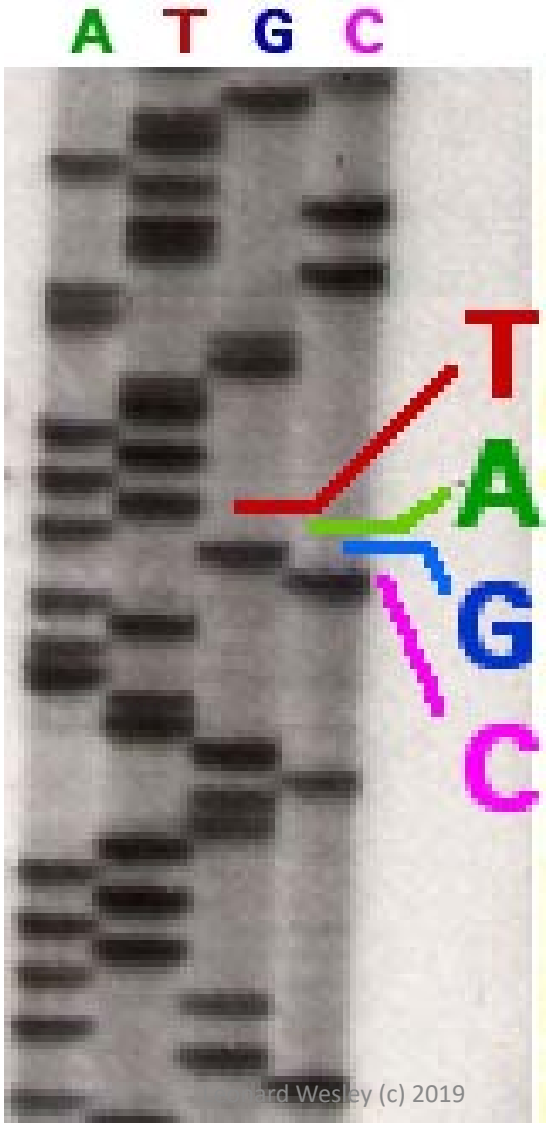


Chain-termination (Sanger) methods

- In the image on the next slide, 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.


A T G C
T
A
G
C

Chain-termination (Sanger) methods

- Chain-termination methods 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; difficulty with runs of the same nucleotide; and DNA secondary structures affecting the fidelity of the sequence.

Sanger Sequencing

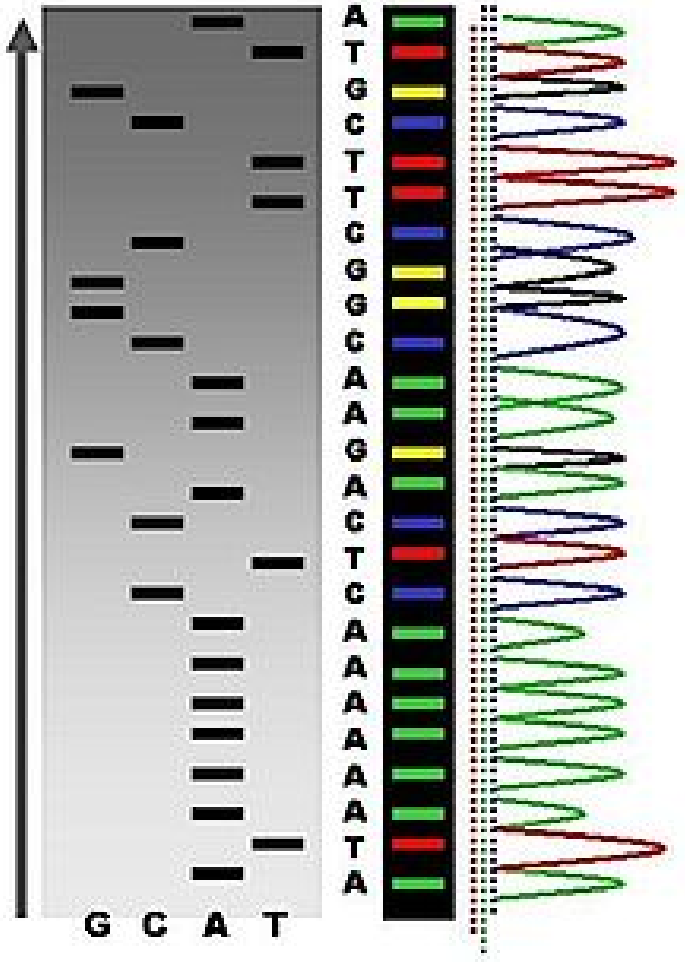
- For example, the chain-termination-based "Sequenase" kit from USB Biochemicals contains most of the reagents needed for sequencing, pre-aliquoted and ready to use.

- 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.

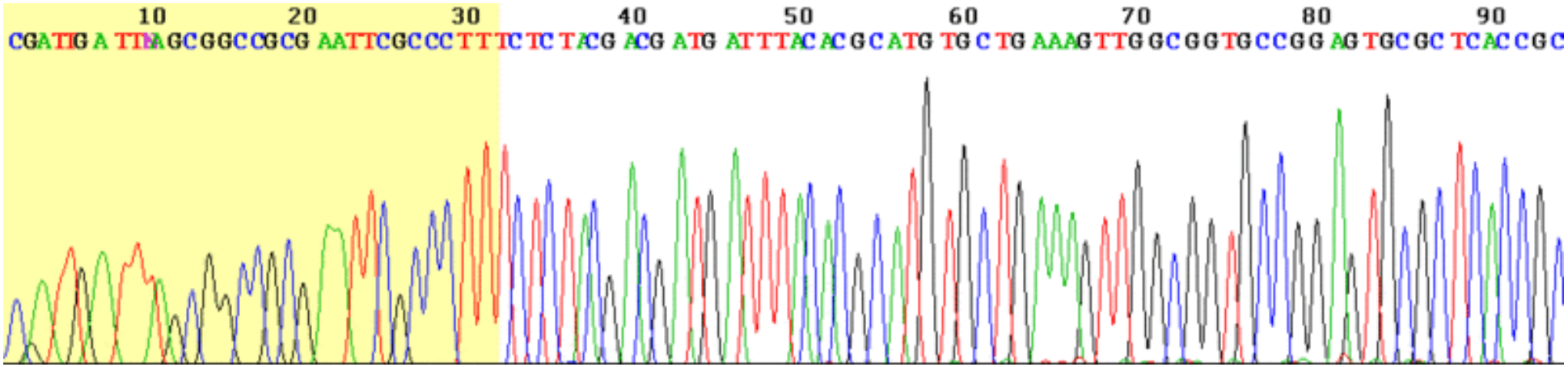
Automated Sequencing – the Beginning

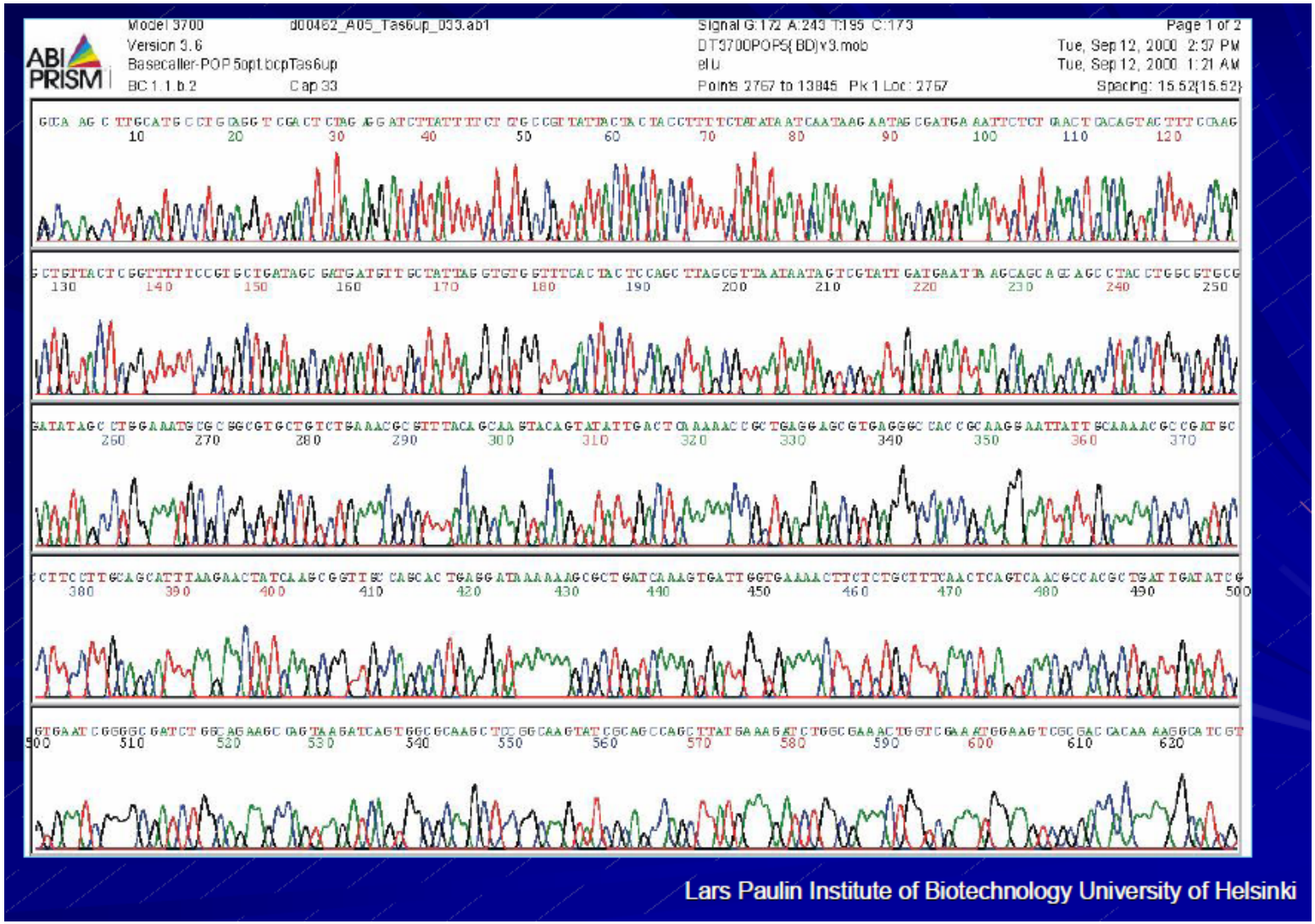
Dye-terminator sequencing

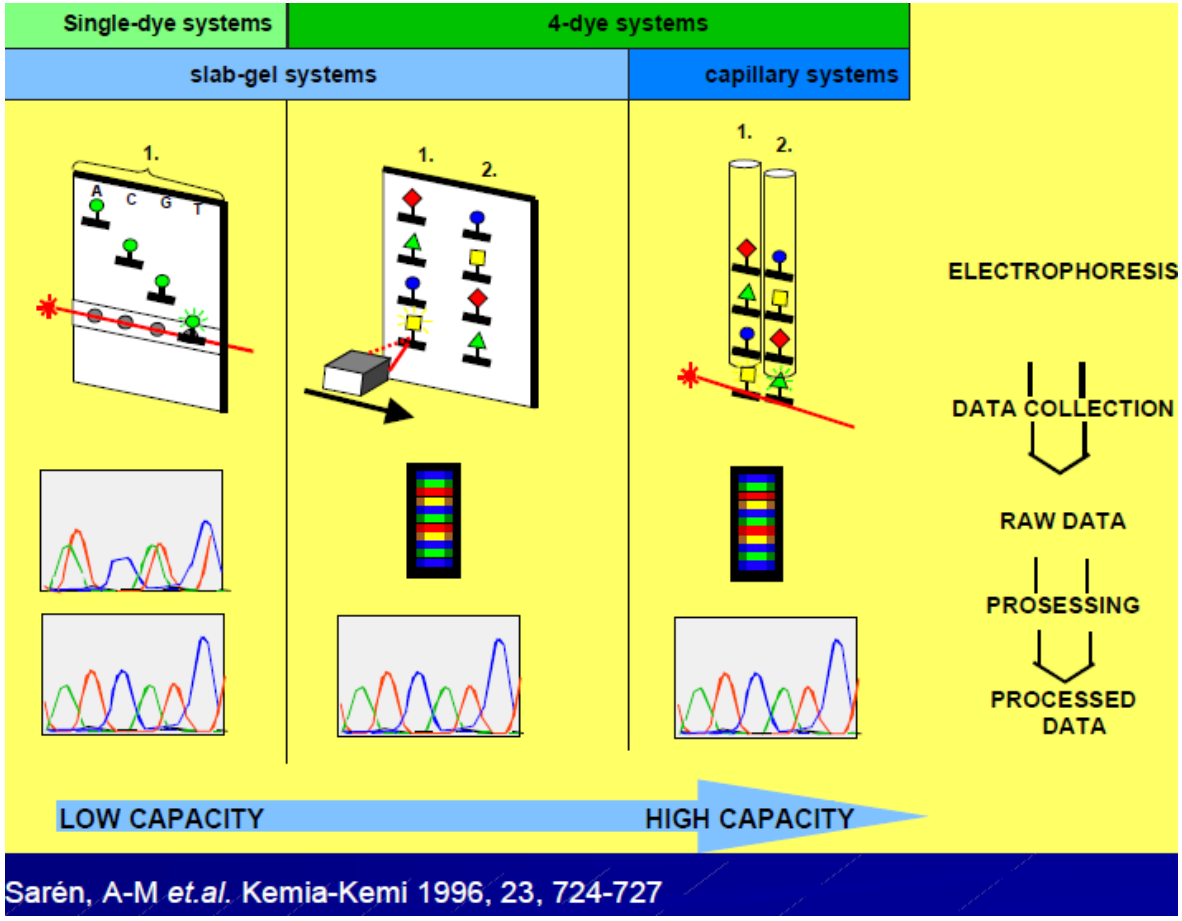
- Led to automation of DNA sequencing
- Initially used a primer labeled at the 5’ end with a fluorescent dye.
- Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation.
- Later development by L Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.


A
T
G
C
T
T
C
G
G
C
A
A
G
A
C
T
C
A
A
A
A
A
A
T
A
G C A T

Example Dye Sequencing Result




ABI PRISM
Model 3700
Version 3.6
Basecaller-POP5opt.bcpTas6up
BC 1.1.b2
d00462_A05_Tas6up_033.ab1
Cap 33
Signal G:172 A:243 T:195 C:173
DT3700POP5(BD)v3.mob
elu
Points 2767 to 13845 Pk 1 Loc: 2767
Page 1 of 2
Tue, Sep 12, 2000 2:37 PM
Tue, Sep 12, 2000 1:21 AM
Spacing: 15.52[15.62]
Lars Paulin Institute of Biotechnology University of Helsinki


Single-dye systems
4-dye systems
slab-gel systems
capillary systems
1.
A C G T
1.
2.
1.
2.
ELECTROPHORESIS
DATA COLLECTION
RAW DATA
PROSESSING
PROCESSED DATA
LOW CAPACITY
HIGH CAPACITY
Sarén, A-M et.al. Kemia-Kemi 1996, 23, 724-727

Automated Sequencing

- The first step toward this goal was achieved in 1985, when Leroy Hood at CalTech attached fluorescent dyes to the primer used in the sequencing reactions; each different color dye (blue, green, yellow, and red) was matched with a different terminator base.

- He and Michael Hunkapiller from Applied Biosystems, Inc. (ABI) built an instrument, dubbed the ABI Model 370, to read the sequence of the dyelabeled fragments. It was equipped with an argon ion laser for exciting the dyes, a flat gel laid between two glass plates (referred to as a "slab" gel) capable of sixteen-lane electrophoresis, and a Hewlett-Packard Vectra computer boasting 640 MB of memory for data analysis.

Automated Sequenicng

- Using fluorescent dyes, all four sequencing reactions could now be loaded into a single gel lane. As the fragments electrophoresed, the beam of the laser focused at the bottom of the gel made the dye-labeled fragments glow as they passed.

- The color of each dye-labeled fragment was then interpreted by the computer as a specific base (A if green, C if blue, G if yellow, and T if red). Over 350 bases could be read per lane. With this new automated approach, a technician could read more sequence in a day than could be read manually in a n entire week!

Limitations and challenges

- 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".

- 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.

Automated Sequencing Timeline

- 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 with different wavelengths of fluorescence and emission.
- 1986 -First commercialized by ABI - First Automated DNA Sequencer ABI 370 1988– Pharmacia ALF
- 1990 – ABI 373
- 1995– ABI 377, Up to 96 lanes, a four- to fivefold increase in throughput compared with that of the 373 system.
- 1996 – First Capillary DNA Sequencer ABI 310
- 1998– First 96 Capillaryinstruments: MegaBace, ABI 3700
- 2002– ABI 3730, 48 or 96 Capillary

ABI 373

On eBay for $299!!

ABI 373 XL

Increased resolution and read length with longer slab gel

Evolution

- Shortly after ABI placed its automated DNA sequencer on the market, the Dupont company introduced its own model, the Genesis 2000. Dupont had also developed a new method of labeling sequencing fragments: attaching the fluorescent dyes to the terminator bases. With this innovation, four separate sequencing reactions were no longer required; the entire sequencing reaction could be accomplished in a single tube. However, Dupont failed to effectively compete in the market and sold the rights to the dye terminator chemistry to ABI.
- ABI continued to refine its automated sequencer. More powerful computers, increased gel capacity (to 96 lanes), improvement of the optical systems, enhancement of the chemistry, and the introduction of more sensitive fluorescent dyes increased the reading capacity of the instrument to over 550 bases per lane.
- The ABI PRISM Model 377 Automated Sequencer, introduced in 1995, incorporated these changes and could read, at maximum capacity, over 19,000 bases in a day. Even at this rate, however, the sequencing of entire genomes, as that of humans (3 billion bases in length), was still not practical.

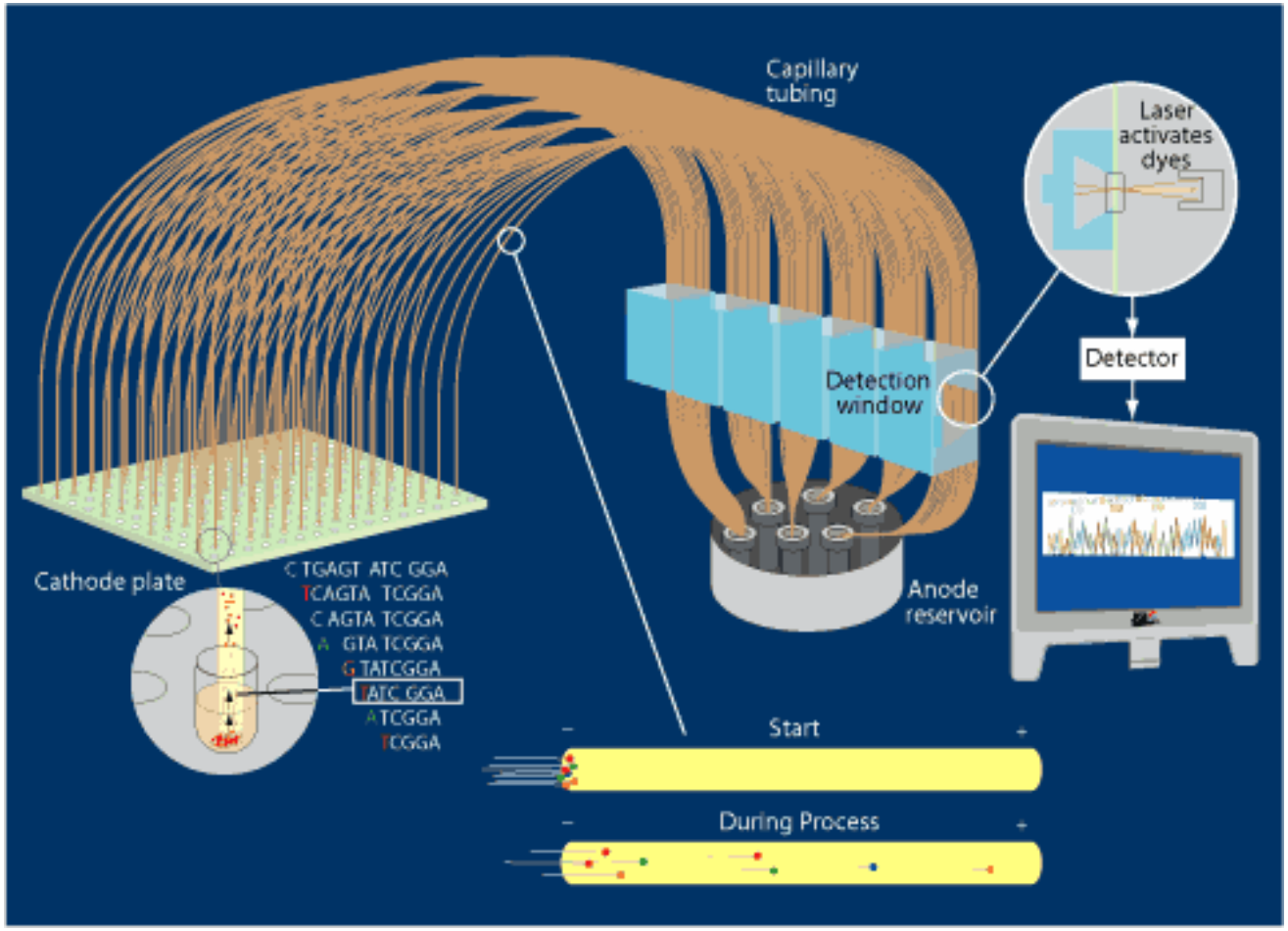
Automated Sequencing – From Genes to Genomes

410.666 - History of DNA Sequencing Part 2

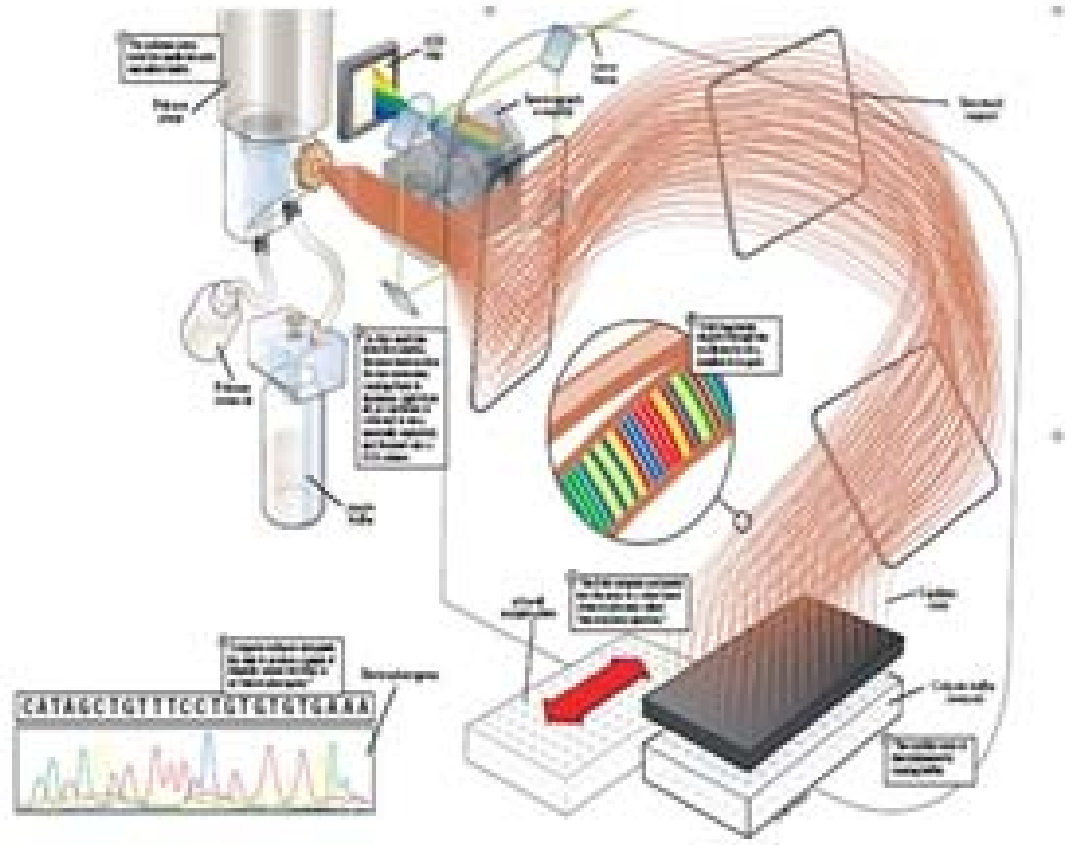
Capillary Sequencers

- Working with the Model 377 Automated Sequencer, a laboratory technician had to pour the slab gels and mount them on the instrument. This process alone was time-consuming and cumbersome. In addition, the technician had to add each sequencing reaction into each individual lane of the gel prior to the run. The MegaBase, developed by Molecular Dynamics, and the ABI Model 3700 Automated Sequencer, developed by ABI, addressed these limitations by using multiple capillaries, thin, hollow glass tubes filled with a gel polymer.

- The ABI PRISM Model 3700 Automated Sequencer, developed with the Hitachi Corporation and having a price tag of $300,000, uses ninety-six capillaries, each not much wider than a strand of human hair. The capillaries are automatically cleaned and filled with fresh gel polymer between each electrophoresis run. The instrument is also equipped with a robot arm that automatically loads the sequencing reactions into the capillaries, greatly decreasing the amount of human labor required for its operation. The Model 3700 Automated Sequencer can read over 400,000 bases in a day, a greater than twenty-fold increase over the maximum capacity of the Model 377. Beginning in September 1999 and using 300 of these instruments, the Celera Corporation had sequenced the entire human genome five times over within four months.


Capillary tubing
Laser activates dyes
Detector
Detection window
Cathode plate
C TGAGT ATC GGA
TCAGTA TCGGA
C AGTA TCGGA
A GTA TCGGA
G TATCGGA
TATC GGA
ATCGGA
TCGGA
Anode reservoir
Start
During Process

MAYO CLINIC
Mayo Medical Laboratories



Next Generation Sequencing

A Little History

- Between late 2007 and the present, DNA sequencing has experienced a “technological singularity”
  - In Sep-2007, the $1,000 human genome sequence was expected to arrive around Sep-2023
  - Instead, it is arriving NOW! Cost now $300 - $1K.
- Numerous business plans, research approaches and sequencing alternatives that made sense in 2007 have rapidly been relegated to the trash bin of history
- The technology is moving so rapidly that many of the “cutting edge” technologies mentioned in the course text (2015/2016) have already been discarded or replaced…

NEST CLASS: Additional NGS Technologies