DNA Sequencing

The Basics

Trey Ideker BE183

...different from music sequencing

image from www.synthesizers.de

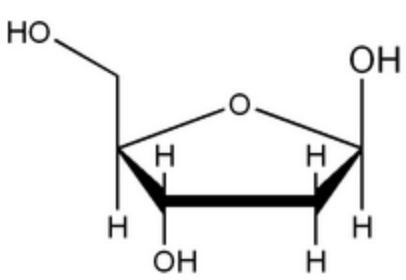
A bit of history: DNA sequencing in 1977

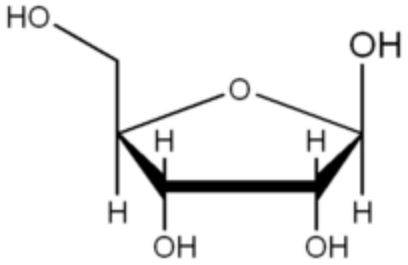
- Maxam/Gilbert (Chemical)
- Enzymatic (Sanger)
- First complete genome sequence of phage ΦX174 by Sanger
- The standard method today is the Sanger “dideoxy method”
- Nobel prize in 1980

Sanger method (Chain-termination)

- Take a strand of DNA.
- Add primer and let DNA polymerase copy the strand.
- But, throw in some molecules (ddNTPs) that will wreck the process after some random number of bases.

Ribose and reactivity

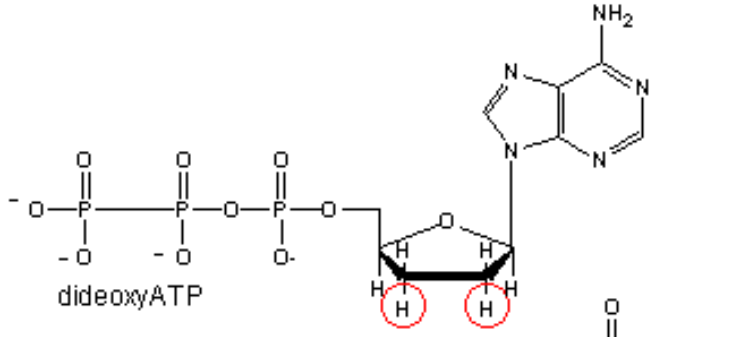


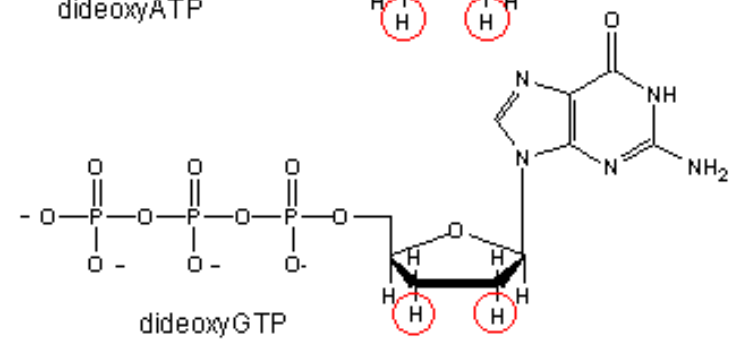


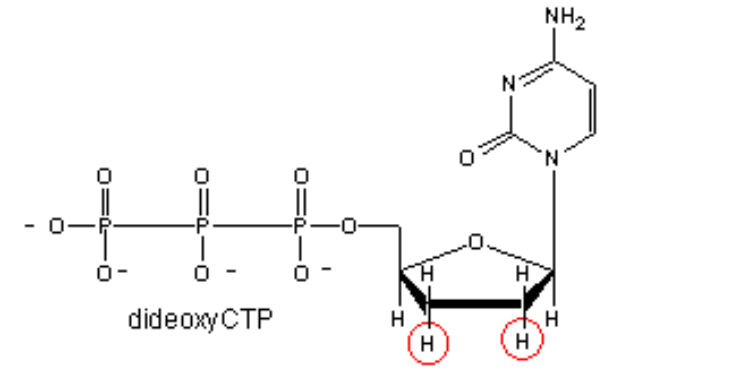
Recall: What makes DNA different from RNA is that DNA lacks the hydroxyl group on the 2' carbon of ribose. (And, thymidine has a methyl group absent in uridine)

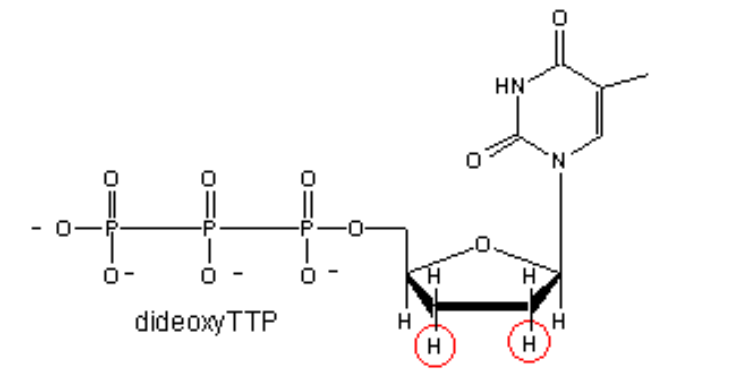
Deoxyribose (left) is less reactive than ribose (right).

This difference makes DNA more stable than RNA. RNA strands can spontaneously break by an intramolecular reaction where the 2' OH temporarily bonds to the phosphorous atom. (This stability may explain why cells evolved to use RNA **and** DNA)









Dideoxyribonucleic acids (or ddNTPs) also lack a hydroxyl group on the 3' carbon.

Replication proceeds 5'->3':
Polymerase can add ddNTPs to a strand, but after that the strand cannot be extended – replication is **halted**!

The effect of ddNTPs

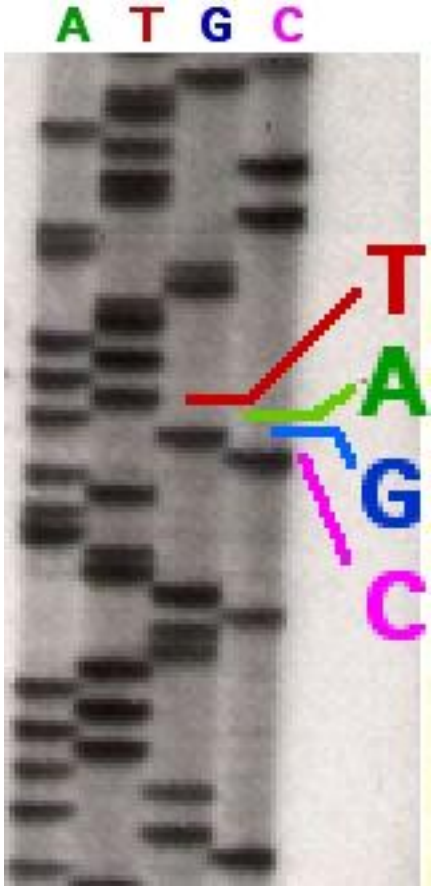
- When ddNTPs are mixed in with normal nucleotides, end products of various lengths are formed. (“Russian Roulette” replication)
- For instance, replicating TAGGATAGA with ddGTP present produces products:

  TA**G**,TAG**G**, TAGGATA**G**

Chain termination gives us sequence information

- Recall: We can measure DNA lengths with single base pair resolution using PAGE!
- In our ddGTP example, the presence of a product of length 3 means that the 3rd base is a G.
- Original method: Make radioactive dNTPs. Run four sequencing reactions, one for each ddNTP. Run products on a gel, visualize with autoradiography, and read off the sequence.

Sanger Sequencing



Sequence excerpt: TACGAGATATATGGCGTTAATACGATATATTGGAACTTC
Read from bottom of gel to top (why?)

Fluorescence improves Efficiency

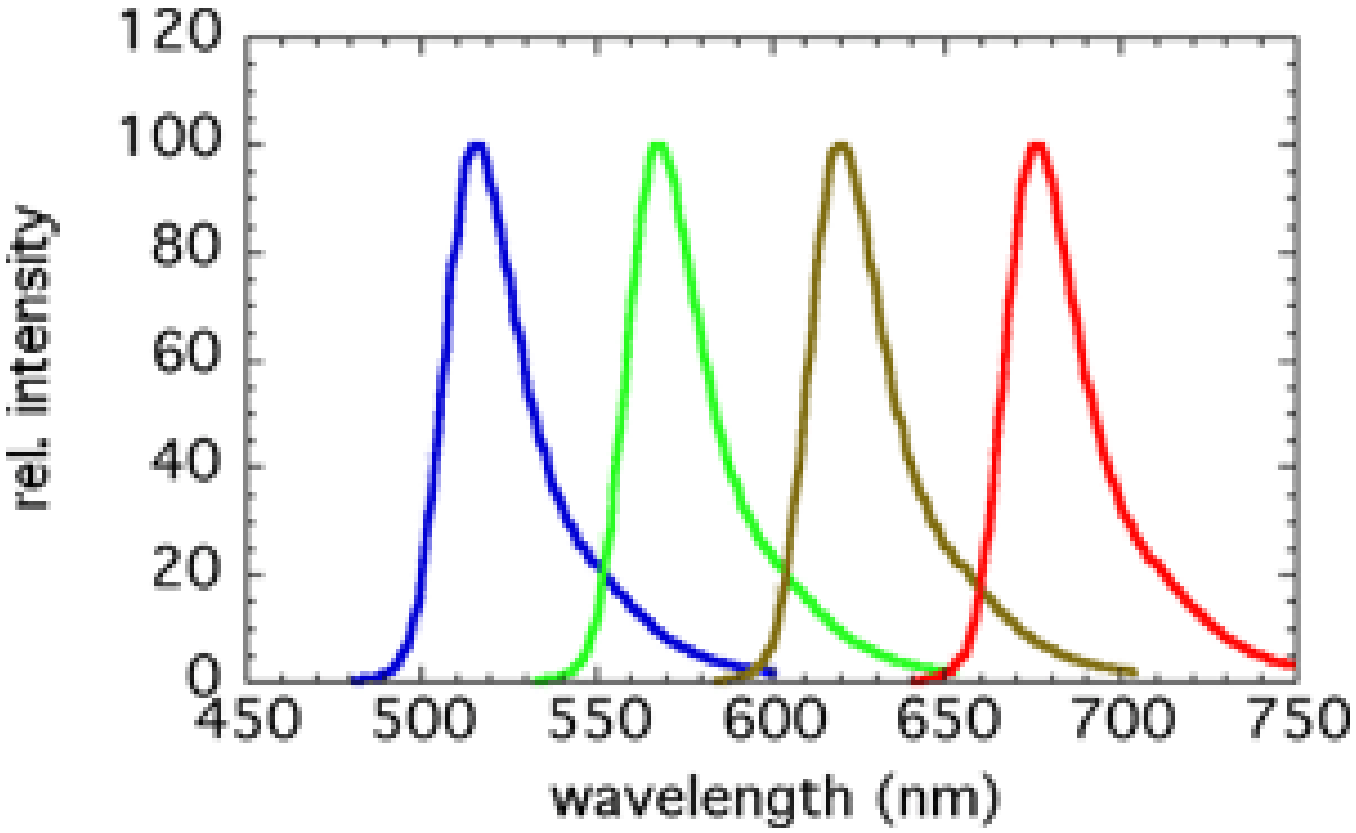
- Recent sequencing efforts use fluorescence instead of radioactivity to detect ddNTPs
- Fluorophores are attached to ddATP, ddTTP, ddGTP, ddCTP
- Safer than radioactive labeling, faster (no waiting to expose film), and cheap.
- If different fluorophores are used, sequencing can now be carried out in **one** sample instead of **four** (ABI machines work this way).

Sequencing Dyes

What are the desired spectral characteristics of fluorescent dyes used for sequencing?

Common excitation wavelength

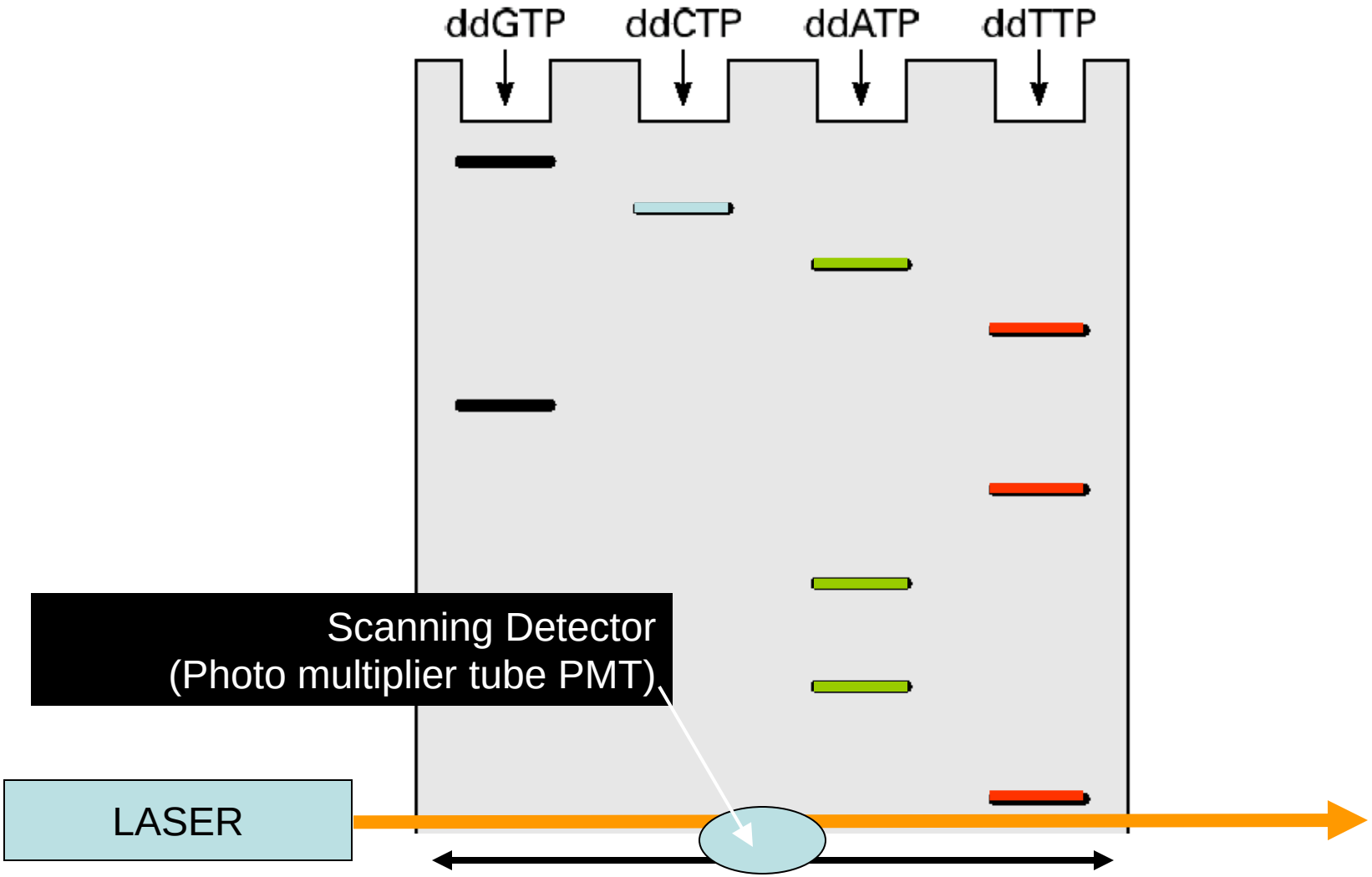
Distinct emission wavelenths



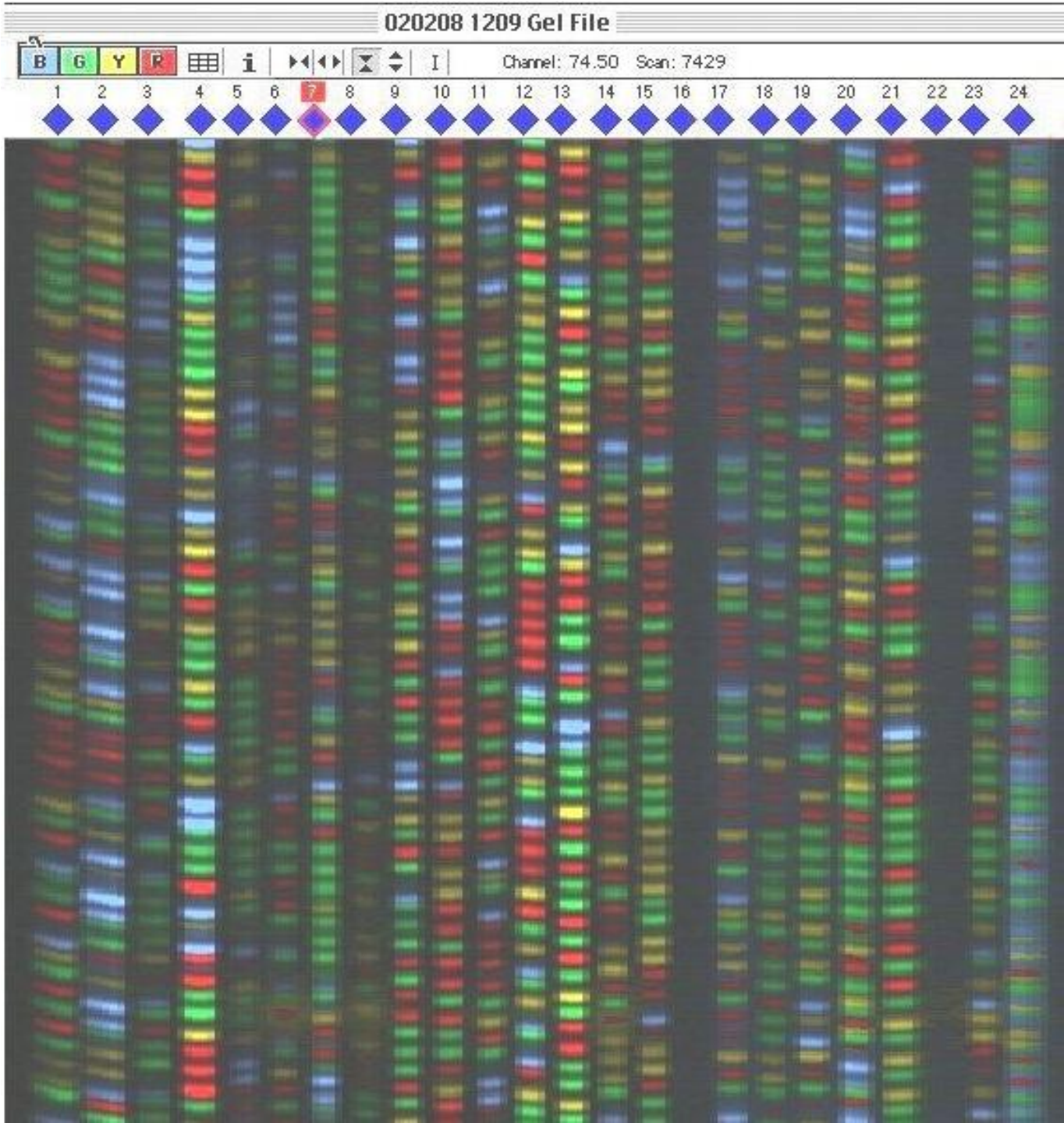
Automated sequencing

- Detection of DNA band is automated
- Uses enzymatic (Sanger) method
- Uses fluorescent labels instead of radioactivity
- Label is excited by a laser at bottom of gel
- The detector scans horizontally across the base of the gel so as to scan several sequencing lanes
- Labels are either on the primer or attached to the ddNTP terminators

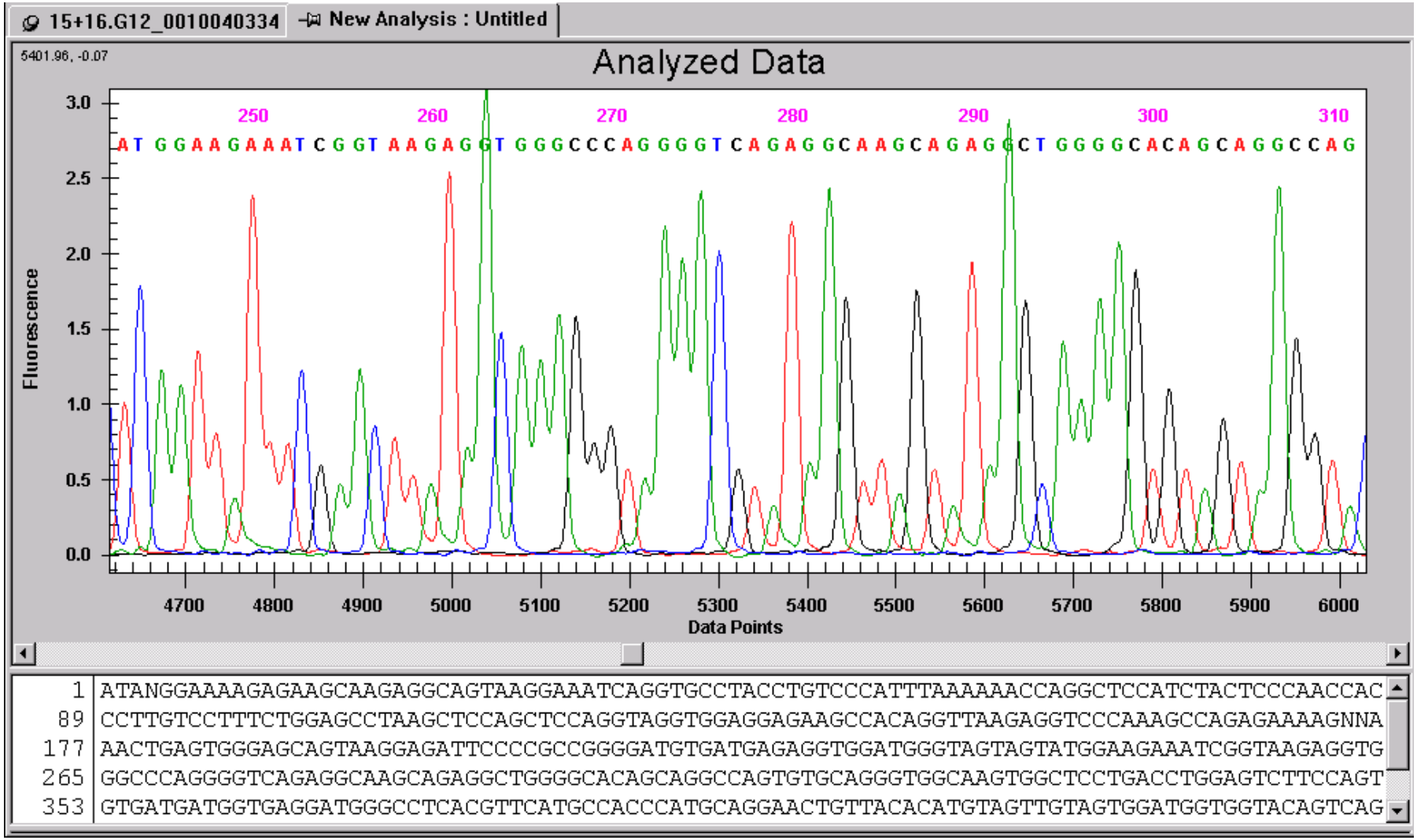
Automating the sequencing process



Fluorcnt. gel scanned image

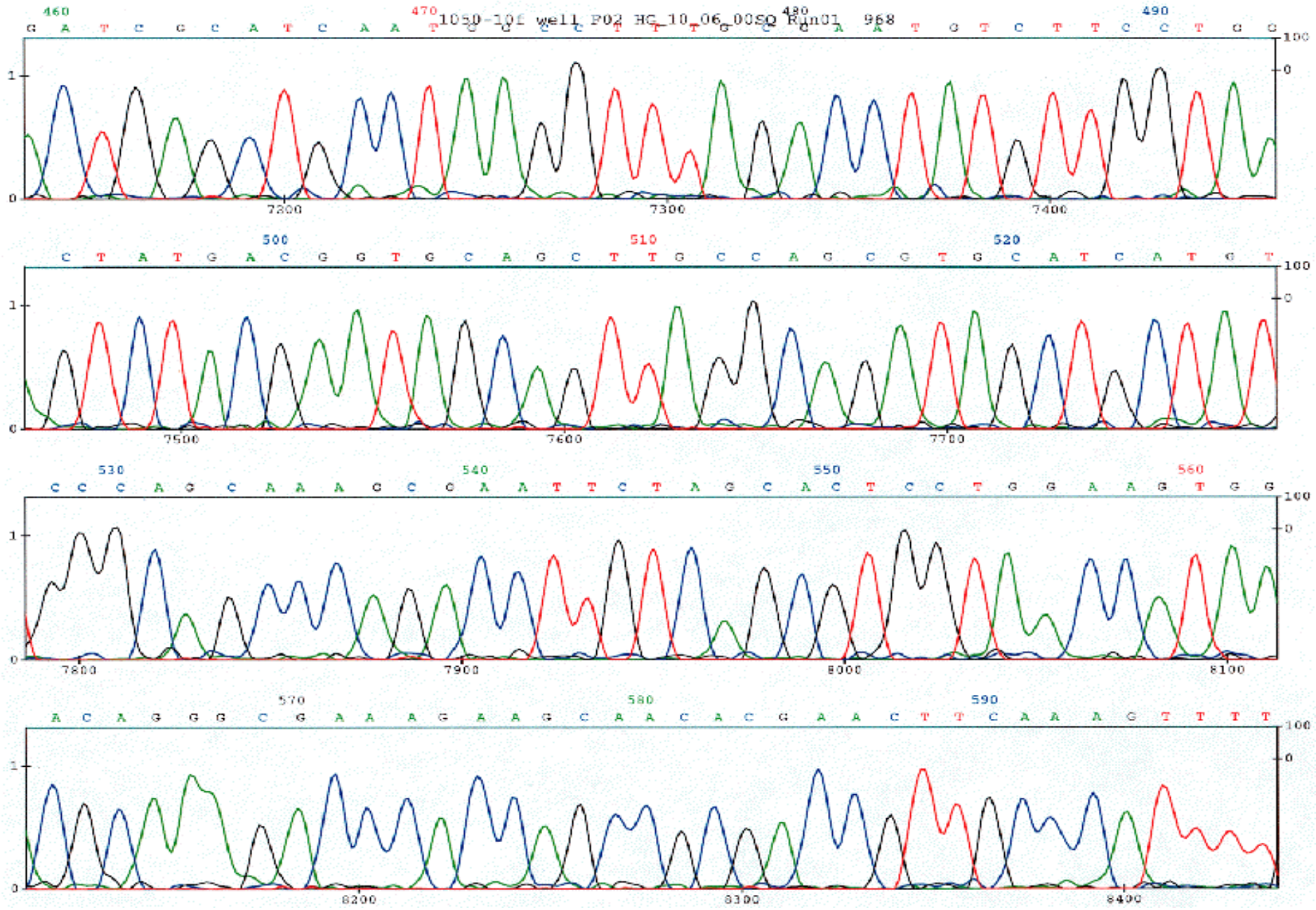


Chromatogram data



Fluorescence trace from DNA sequencing (from Beckman Coulter website)

Trace of scanned image



Base Calling

- Each peak corresponds to one length of terminated DNA chain.
- Peaks vary in height, may include noise, and may migrate at the “wrong” time due to secondary structure effects
- Software like PHRED automates base calling.

Phred overview

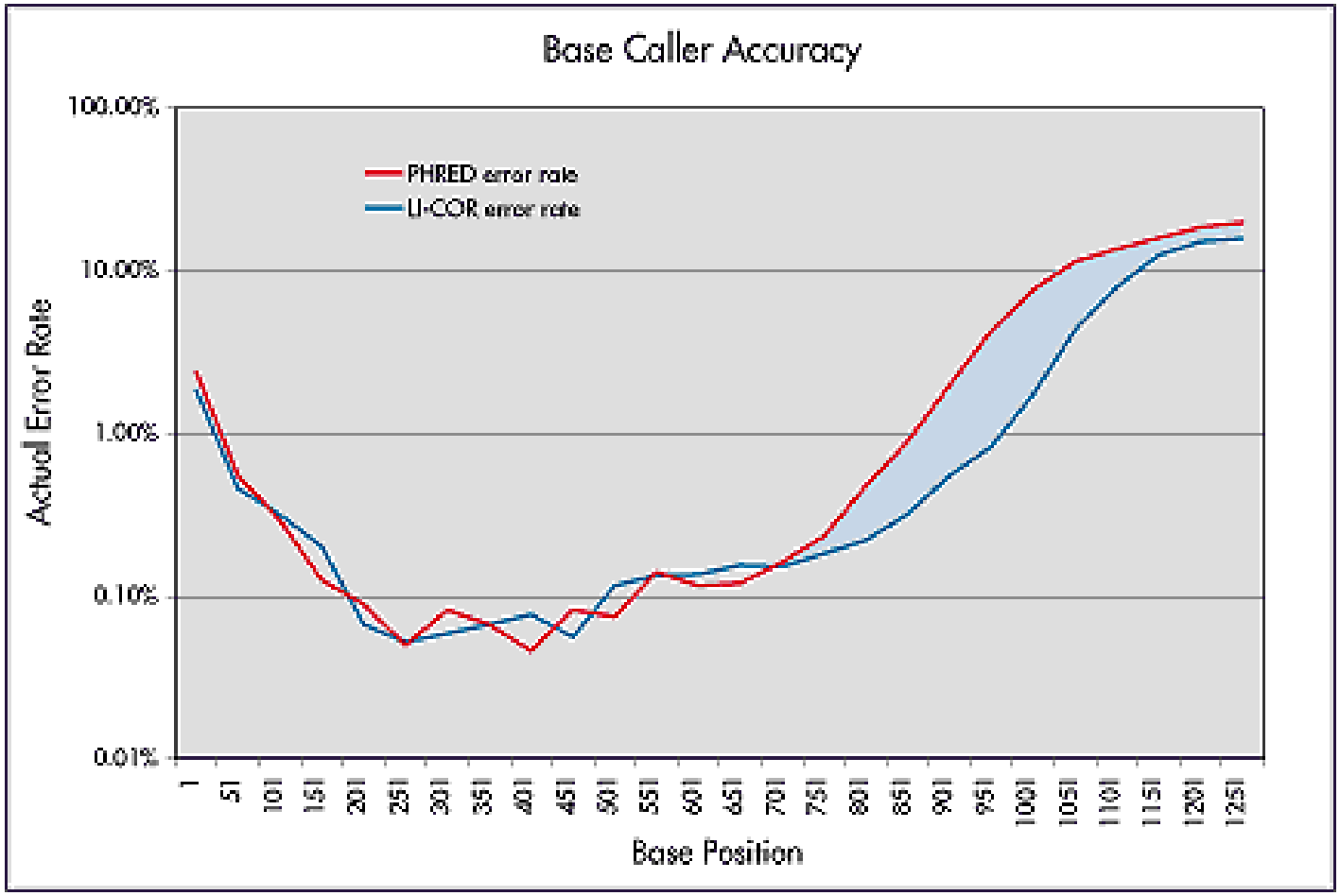
- Phred first determines peak spacing by signal processing (Fourier transform)
- Secondly, a list of high-quality observed peaks is made.
- Thirdly, observed peaks are matched to the predicted locations.
- Lastly, strong but unassigned peaks are inserted into the read.

Phred confidence score

$$q = -10 \log_{10}(P_e)$$

- Computed for each base read.
- Here, P_e is the error probability, and q is the confidence score.
- q is 0 for total uncertainty, 10 for error odds 10%, 20 for error probability 1%.

Error rates



In practice, the first ~50 bases and all bases after ~800 are discarded from further analysis

ABI 371 and 3700 sequencers

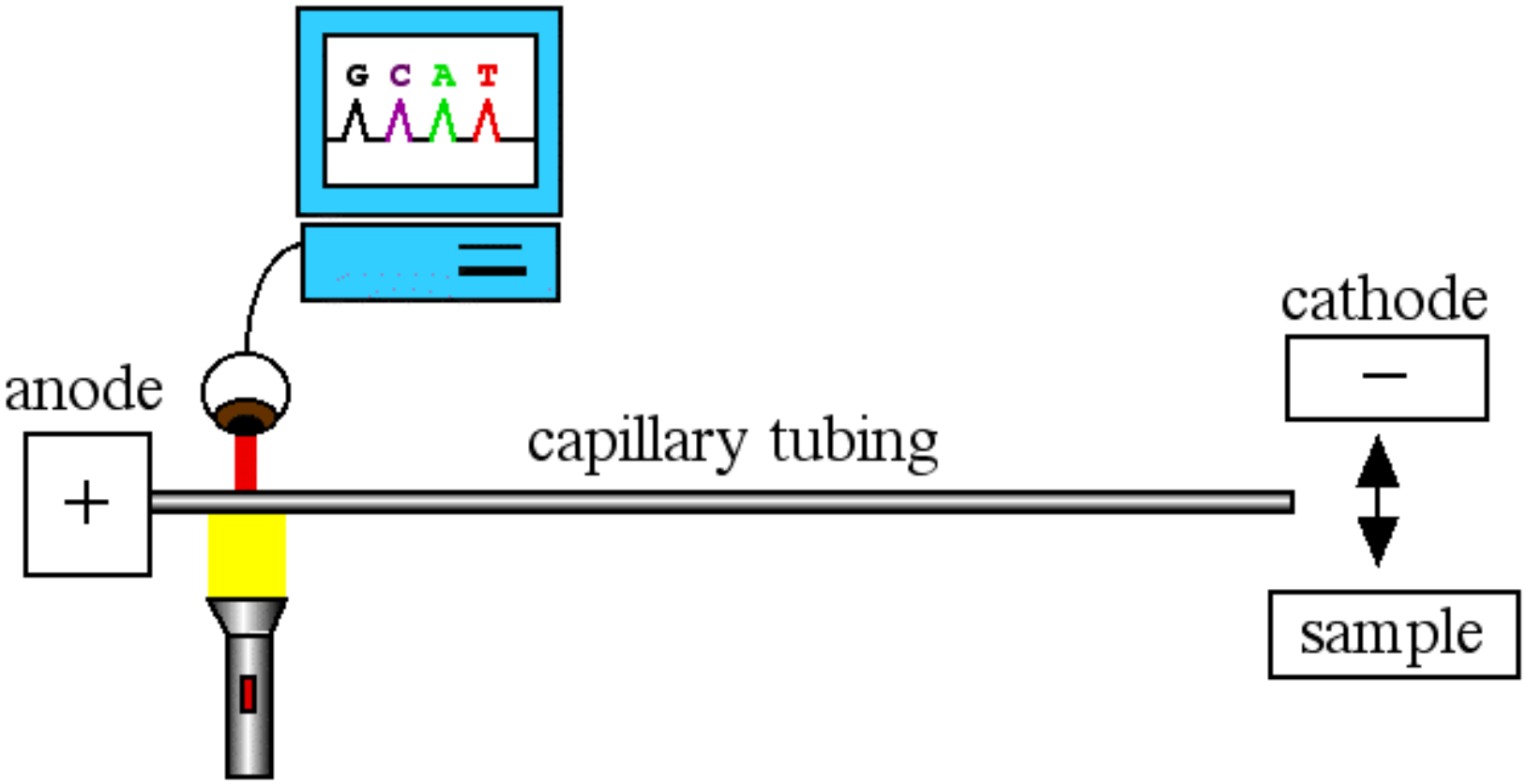
Problems with gel-based sequencing

- Preparation of gels is labor intensive
- DNA sequencing speed is related to the strength of the applied electric field
- High voltage causes Joule heating of the gel and utter meltdown
- CAPILLARY SEQUENCING partially solves these problems

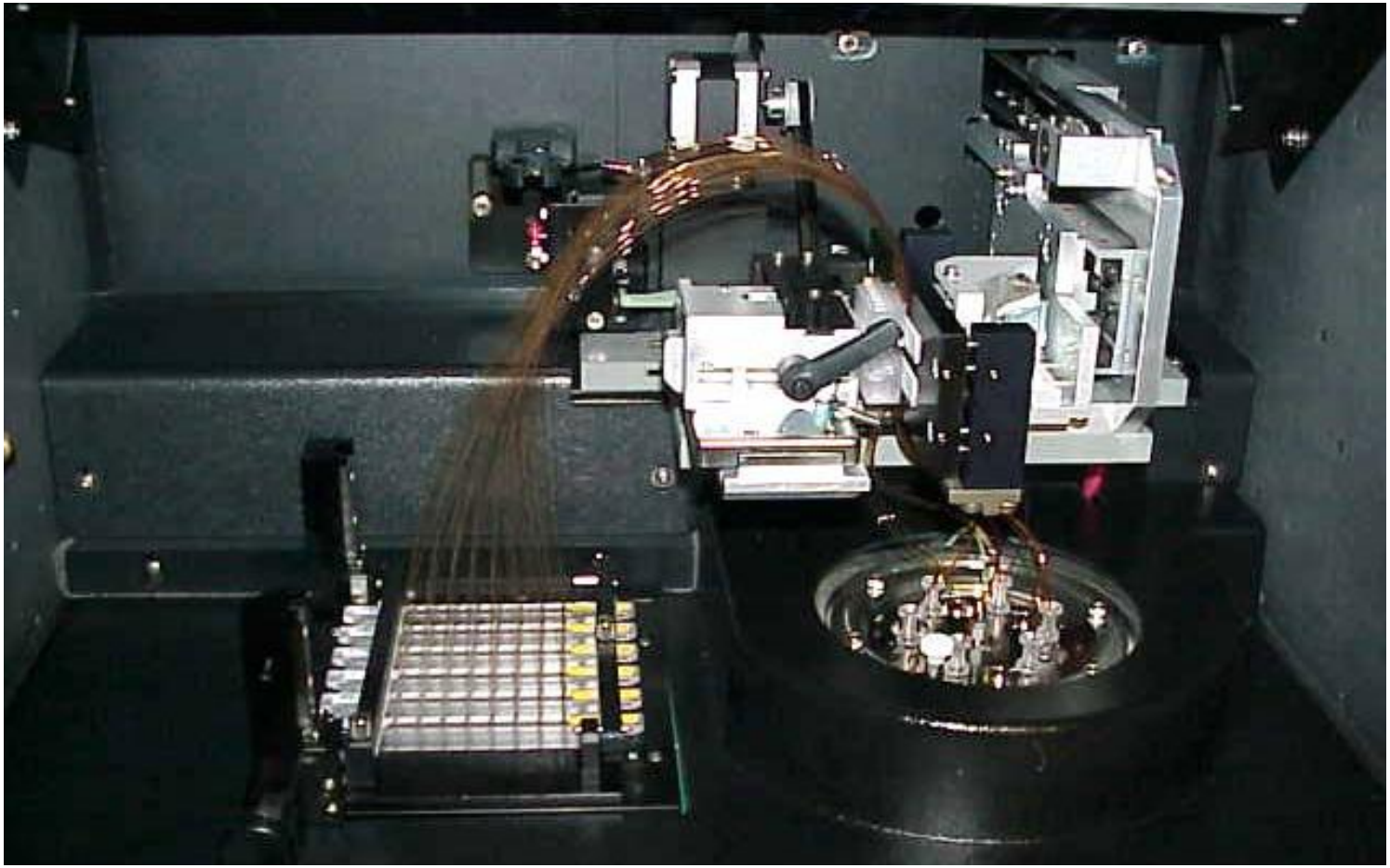
Capillaries improve Speed

- Rather than a slab of gel, DNA products run through a thin (50um) **capillary**.
- A laser excites the capillary, which fluoresces in different colors as longer and longer fragments elute through.
- Current sequencing machines have many capillaries, for several sequencing reactions in parallel.

Sequencing by capillary electrophoresis



Megabase capillary sequencer



The brown wire-like loops are the 96 individual capillary tubes. They are made of glass and coated with brown plastic. The samples are loaded from below this level on the left side and the DNA is electrophoresed towards the right. The matrix is injected into the tubes after each run and it travels from right to left when filling the tubes.