DNA Sequencing:

Introduction:

The term *DNA sequencing* refers to methods for determining the order of the nucleotides 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*.

Fred Sanger

Gilbert

Figure1: Founders of sequencing technology (Nobel Prize in Chemistry in 1980)

A sequencing can be done by different methods :

1. Maxam – Gilbert sequencing
2. Chain-termination methods
3. Dye-terminator sequencing
4. Automation and sample preperation
5. Large scale sequencing strategies
6. New sequencing methods.

1- Chemical cleavage method (Maxam & Gilbert /chemical method, 1977)

Allan Maxam and Walter Gilbert developed a method for sequencing single-stranded DNA by taking advantage of a two-step catalytic process involving **piperidine** and two chemicals that selectively attack **purines** and **pyrimidines**. Purines will react with **dimethyl sulfate** and pyrimidines will react with **hydrazine** in such a way as to break the glycoside bond between the ribose sugar and the base displacing the base (Step 1).

Piperidine will then catalyze phosphodiester bond cleavage where the base has been displaced (Step 2). Moreover, dimethyl sulfate and piperidine alone will

selectively cleave guanine nucleotides but dimethyl sulfate and piperidine in **formic acid** will cleave both guanine and adenine nucleotides. Similarly, hydrazine and piperidine will cleave both thymine and cytosine nucleotides whereas hydrazine and piperidine in 1.5M **NaCl** will only cleave cytosine nucleotides (Figure 2).

Figure 2: Chemical targets in the Maxam-Gilbert DNA sequencing strategy. Dimethylsulphate or hydrazine will attack the purine or pyrimidine rings respectively and piperidine will cleave the phosphate bond at the 3' carbon.

The use of these selective reactions to DNA sequencing then involved creating a single-stranded DNA substrate carrying a radioactive label on the 5' end.
This labeled substrate would be subjected to four separate cleavage reactions, each of which would create a population of labeled cleavage products ending in known nucleotides.
The reactions would be loaded on high percentage polyacrylamide gels and the fragments resolved by electrophoresis. The gel would then be transferred to a light-proof X-ray film cassette, a piece of X-ray film placed over the gel, and the cassette placed in a freezer for several days. Wherever a labeled fragment stopped on the gel the radioactive tag would expose the film due to particle decay (**autoradiography**).

Since electrophoresis, whether in an acrylamide or an agarose matrix, will resolve nucleic acid fragments in the inverse order of length, that is, smaller fragments will run faster in the gel matrix than larger fragments, the dark autoradiographic bands on the film will represent the 5'→3' DNA sequence when read from bottom to top (Figure 3).
The process of **base calling** would involve interpreting the banding pattern relative to the four chemical reactions.

Figure 3: The Maxam-Gilbert manual sequencing scheme. The target DNA is radiolabeled and then split into the four chemical cleavage reactions. Each reaction is loaded onto a polyacrylamide gel and run. Finally, the gel is autoradiographed and base calling proceeds from bottom to top.

2-Sanger Method (Dideoxynucleotide chain termination)

Defination: Sanger sequencing is a DNA sequencing method in which target DNA is denatured and annealed to an oligonucleotide primer, which is then extended by DNA polymerase using a mixture of deoxynucleotide triphosphates (normal dNTPs) and chain-terminating dideoxynucleotide triphosphates (ddNTPs).

ddNTPs lack the 3' OH group to which the next dNTP of the growing DNA chain is added. Without the 3' OH, no more nucleotides can be added, and DNA polymerase falls off. The resulting newly synthesized DNA chains will be a mixture of lengths, depending on how long the chain was when a ddNTP was randomly incorporated.

At about the same time as Maxam-Gilbert DNA sequencing was being developed; Fred Sanger was developing an alternative method. Rather than using chemical cleavage reactions, Sanger opted for a method involving a third form of the ribose sugars. As shown in (Figure 4), Ribose has a hydroxyl group

on both the 2' and the 3' carbons whereas deoxyribose has only the one hydroxyl group on the 3' carbon. This is not a concern for polynucleotide synthesis in vivo since the coupling occurs through the 3' carbon in both RNA and DNA. There is a third form of ribose in which the hydroxyl group is missing from both the 2' and the 3' carbons. This is dideoxyribose. Sanger knew that, whenever a dideoxynucleotide was incorporated into a polynucleotide, the chain would irreversibly stop, or terminate. Thus, the incorporation of specific dideoxynucleotides in vitro would result in selective chain termination.

Figure 4: The structure of the three five carbon sugars ribose, deoxyribose, and dideoxyribose. The hydroxyl groups are shown in red.

Sanger proceeded to establish a protocol in which four separate reactions, each incorporating a different dideoxynucleotide along with the four deoxynucleotides, would produce a population of fragments all ending in the same dideoxynucleotide in the presence of a DNA polymerase if the ratio of the dideoxynucleotide and the corresponding deoxynucleotide was properly set. All that was needed for the reactions to be specific was an appropriate **primer** for the polymerase. If the primer was radiolabeled instead of the substrate, the resulting fragment populations would be labeled and could be resolved on polyacrylamide gels just like Maxam-Gilbert fragments.

Unlike Maxam-Gilbert fragments each lane would be base-specific. Auto-radiography was the same but base calling was easier. The one new twist was that the sequence fragments on the gel were the complement of the actual template.

Manual DNA sequencing example:

• First, anneal the primer to the DNA template (must be single stranded):

5' – GAATGTCCT T TCTCTAAG
3'- GAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

• Then split the sample into four aliquots including the following nucleotides:

"G" tube: **All four** dNTPs, one of which is radiolabeled, **plus ddGTP** (low concentration)
"A" tube: All four dNTPs, one of which is radiolabeled, **plus ddATP**
"T" tube: All four dNTPs, one of which is radiolabeled, **plus ddTTP**
"C" tube: All four dNTPs, one of which is radiolabeled, **plus ddCTP**

• When a DNA polymerase (e.g. Klenow fragment) is added to the tubes, the synthetic reaction proceeds until, by chance, a dideoxynucleotide is incorporated instead of a deoxynucleotide. This is a **"chain termination"** event, because there is a 3' H instead of a 3' OH group. Since the synthesized DNA is labeled (classically with 35S-dATP), the products can be detected and distinguished from the template.

Note that the higher the concentration of the ddNTP in the reaction, the shorter the products will be, hence, you will get sequence CLOSER to your primer. With lower concentrations of ddNTP, chain termination will be less likely, and you will get longer products (sequence further AWAY from the primer).

If, for example, we were to look only at the "G" reaction, there would be a mixture of the following products of synthesis:

5'- GAATGTCCT T TCTCTAAGTCCTAA**G**
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'- GAATGTCCT T TCTCTAAGTCCTAAGTCC TCC**G**
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG**G**
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

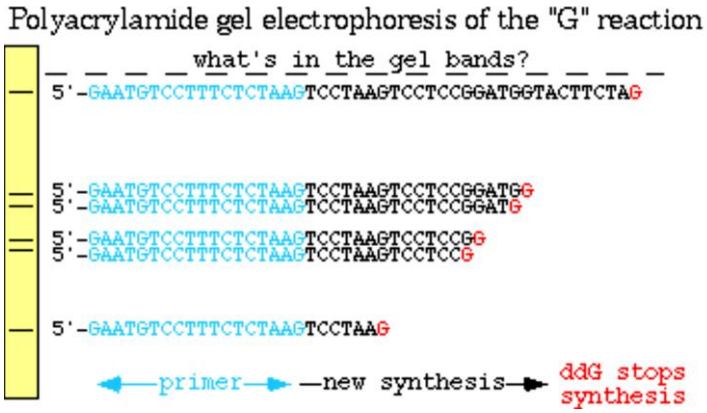
5 '-GAATGTCCT T T CTCTAAGTCCTAAGTCCTCCGGAT**G**
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG**G**
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

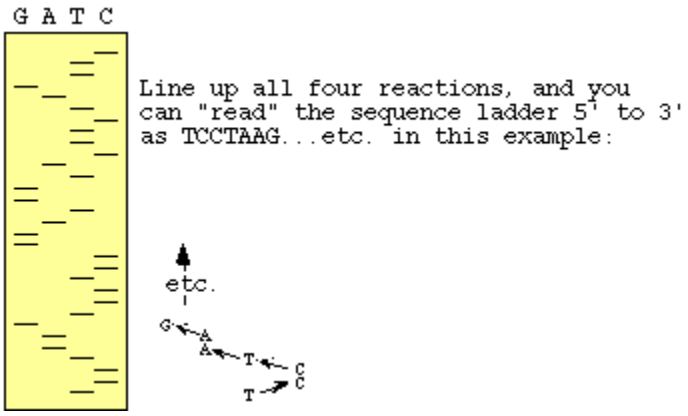
5'- GAATGTCCTTTCTCT AAGTCCTAAGTCCTCCGGATGGTACTTCTA**G**
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

(and so on, if the DNA being sequenced continues to the right)

Each newly synthesized strand at some point had a ddGTP incorporated instead of dGTP. Chain termination then occurred (no more polymerization). Because ddGTP incorporation is random, all possible lengths of DNA that *end in* G are produced.
These products are denatured into single stranded DNA molecules and run on a polyacrylamide/urea gel.
(Polyacrylamide gels, unlike agarose, allow resolution of DNA molecules that differ in size by only one nucleotide.) The gel is dried onto chromatography paper and exposed to X-ray film. Since the template strand is not radioactively labeled, it does not generate a band on the X-ray film. Only the labeled top strands generate bands, which would look like this:



As you can see from this one reaction (the "G" reaction) the chain termination events produce individual bands on a gel. The chain terminations closest to the primer generate the smallest DNA molecules (which migrate further down the gel), and chain terminations further from the primer generate larger DNA molecules (which are slower on the gel and therefore remain nearer to the top).
When similar chain termination reactions are run for each nucleotide, the four reactions can be run next to each other, and the sequence of the DNA can be read off of the "ladder" of bands, 5' to 3' sequence being read from bottom to top:



The resolution of the gel electrophoresis is very important in DNA sequencing. Molecules that are 50, 100, or 200 bases in length must be separable from molecules that are 51, 101, or 201 bases in length (respectively).

To accomplish this:

1- Polyacrylamide, not agarose, is used.

2- The gels must be quite large so that the molecules migrate further and are better resolved.

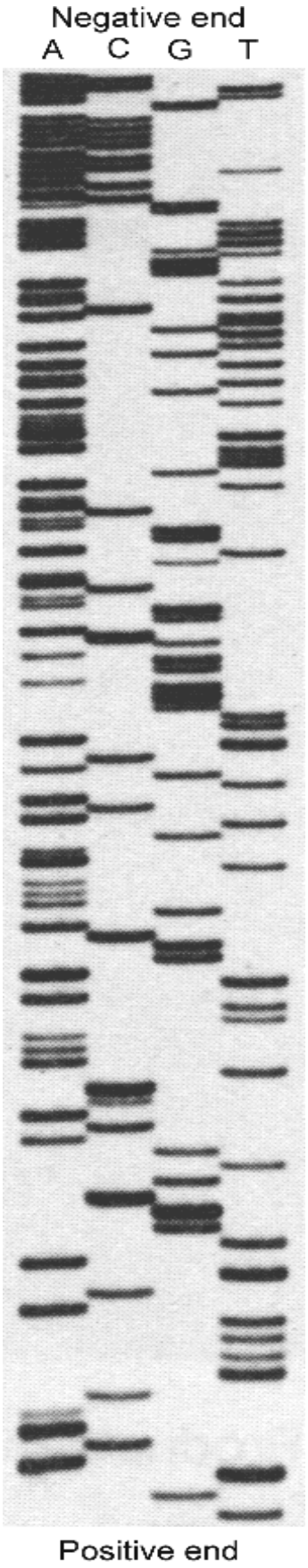
3- Samples are denatured before they are loaded, and the gels must contain a high concentration of urea (7 to 8 molar) to prevent folding of the molecules and formation of secondary structures by hydrogen bonding that would alter the mobility of the molecule.

4- The gels are run at higher temperature (about 50 C), also to prevent *H bond* formation.

Note that this example is for demonstration only: you can't really obtain usable sequence information that close to the end of the primer because few termination events will have occurred so soon. If you increased the ddNTP:dNTP ratio, you would get more sequence close to the primer, but make it more difficult to read sequence 200 to 300 nucleotides further down, because most of the synthetic products would have terminated earlier.

Sequence on gel below :

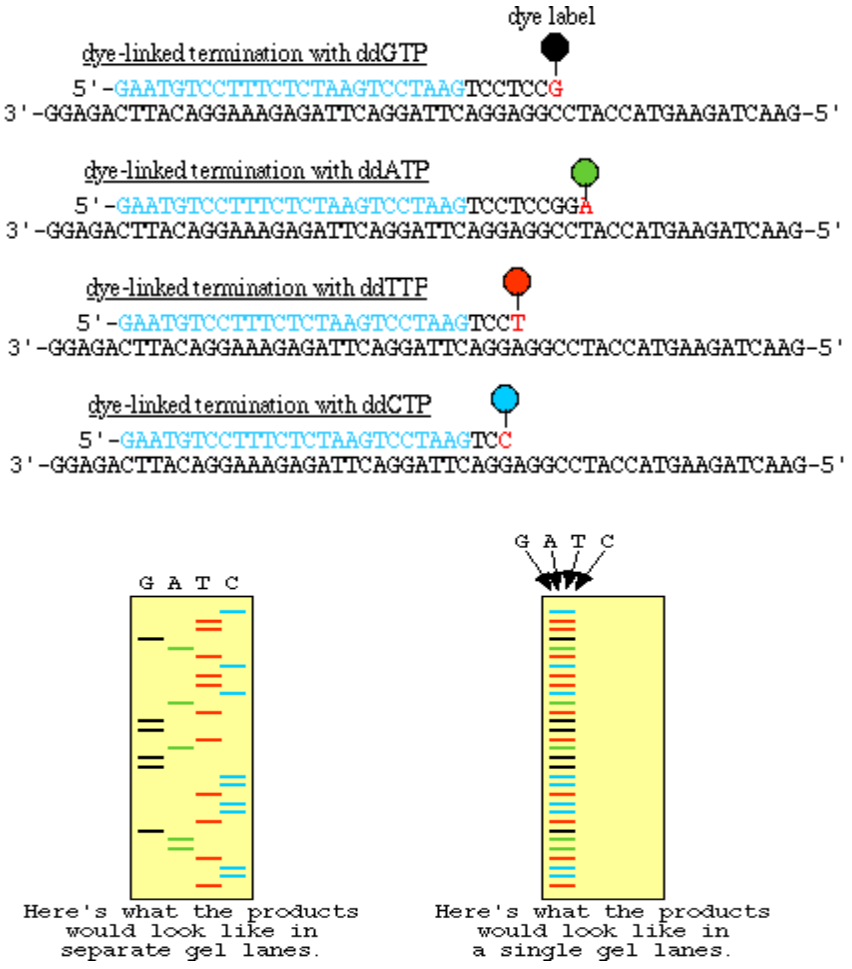
{TG}TACAACTTTTACTATGGCGTGACACCTAAATTATAGGCAGAAA…


Negative end
A C G T
Positive end

3-Dye termination sequencing

Most DNA sequencing is now automated. In the dye-terminator variant of Sanger sequencing ,the Sanger method chain termination reactions are still used, but pouring, running, & reading polyacrylamide gels has been replaced by automated methods. Instead of labeling the products of all 4 sequencing reactions the same (with a radioactive deoxynucleotide), each *dideoxynucleotide* is labeled with a *different fluorescent* marker. When excited with a laser, the 4 different kinds of products are detected and the fluorescence intensity translated into a data "**peak**".

Thus all four chain termination reactions can be performed in the same tube, and run on a single lane on a gel. A machine scans the lane with a laser. The wavelength of fluorescence from the label conjugated to the ddNTPs can be interpreted by the machine as an indication of which reaction (ddG, ddA, ddT, or ddC) a particular DNA band came from.



dye terminator sequencing is now the mainstay in automated sequencing. Its limitation 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.

The 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.

Note: The fluorescence output is stored in the form of a **chromatogram**:

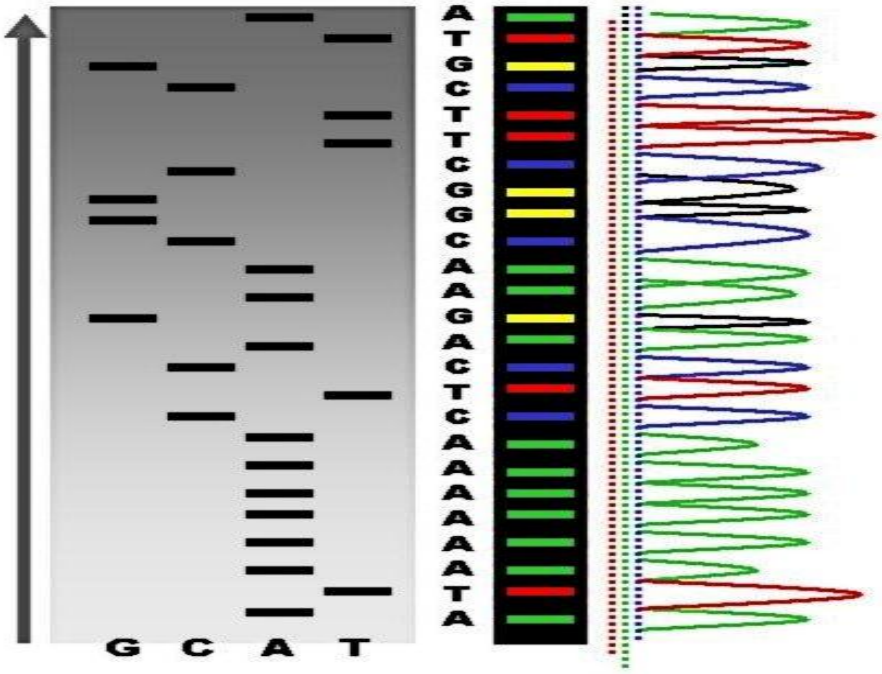


Figure 5. Sequence ladder by radioactive sequencing compared to fluorescent peaks.

In 1986 Smith et al, published the first report of automation of DNA sequencing which established the dye-terminator variant of Sanger sequencing. This initial report 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*/or *The 2nd generation of sequencing* technologies beginning in 2005(such as : Roche/454: pyrosequencing, Solexa (Illumina) &SOLID (ABI)). The obtainable sequence length is now _ 1000 nucleotides.

4- Automated DNA sequencing

Automated DNA sequencing instruments (***DNA sequencers***) can sequence upto 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size seperation, detection and recording of dye fluorescence, and data output as fluorescent peak trace

chromatograms. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programmes score the quality of each peak and remove low quality base peaks (generally located at the ends of the sequence).

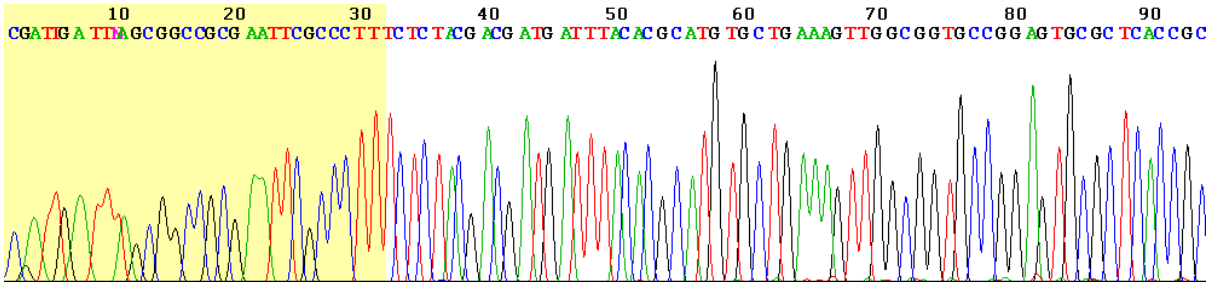


Figure 6. View of the start of an example dye-terminator read.