RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF AGRICULTURE SCIENCES AND ALLIED INDUSTRIES

(Principles of Biotechnology)

For

M.Sc. Ag (GPB)

Course Instructor

Dr Shiv Prakash Shrivastav

FASAI(Genetics and Plant Breeding)

Rama University, Kanpur

DNA Sequencing

Introduction

Prior to the mid-1970's no method existed by which DNA could be directly sequenced. Knowledge about gene and genome organization was based upon studies of prokaryotic organisms and the primary means of obtaining DNA sequence was so-called **reverse genetics** in which the amino acid sequence of the gene product of interest is back- translated into a nucleotide sequence based upon the appropriate codons. Given the degeneracy of the genetic code, this process can be tricky at best. In the mid-1970's two methods were developed for directly sequencing DNA. These were the **Maxam-Gilbert** chemical cleavage method and the **Sanger** chain-termination method.

Maxam-Gilbert

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** [1]. 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 1). 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 2). The process of **base calling** would involve interpreting the banding pattern relative to the four chemical reactions. For

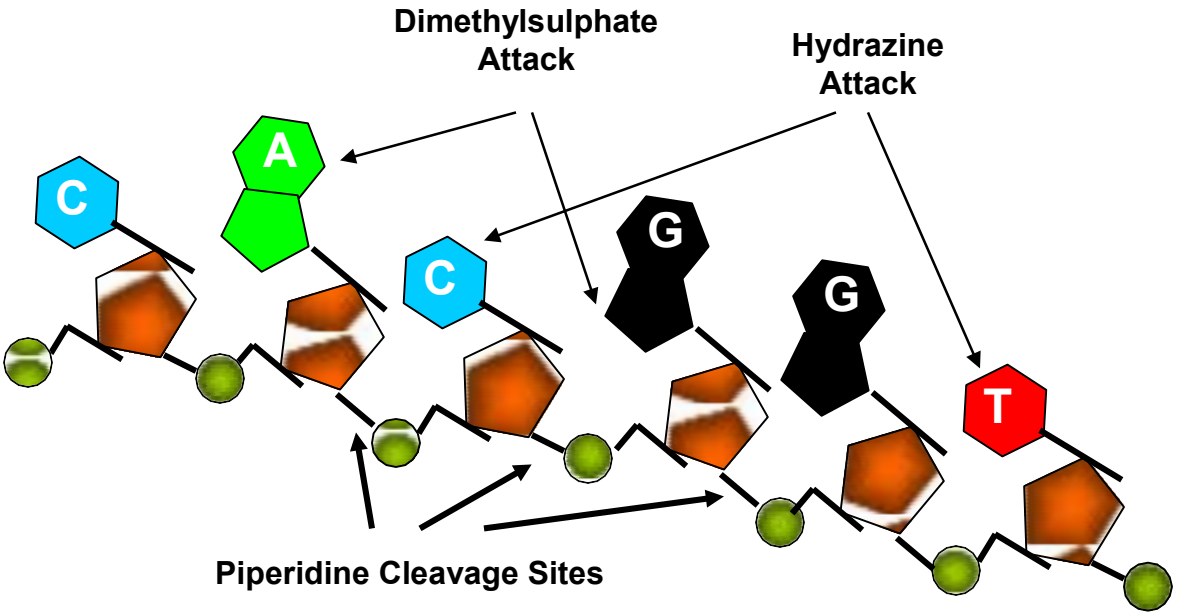

Figure 1. Chemical targets in the Maxam-Gilbert DNA sequencing strategy. Dimethylsulphate or hydrazine will attack the purine or pyrimidine rings respectively and piperidine will cleave the phosphate bond at the 3' carbon.

example, a band in the lanes corresponding to the C only and the C + T reactions would be called a C. If the band was present in the C + T reaction lane but not in the C only reaction lane it would be called a T. The same decision process would obtain for the G only and the G + A reaction lanes. Sequences would be confirmed by running replicate reactions on the same gel and comparing the autoradiographic patterns between replicates.

If all went well, that is, if the radioactive labeling process worked, if the cleavage reactions performed as expected, the gel set up properly, the electrophoresis worked, the gel was not torn or otherwise destroyed during transfer, and the X-ray film developer did not break down, you could expect to get 200-300 bases of confirmed DNA sequence every few days. The exchange for this priceless information was that you had to use rather large amounts of radioactive material, either ^{35}S or ^{32}P, you had to constantly be pouring large, paper thin acrylamide gels, and hydrazine just happens to be a neurotoxin. In spite of the obstacles, however, DNA sequences started to accumulate from a host of organisms and genes and one of the very first discoveries was that the assumption that eukaryotic gene organization was the same as prokaryotic gene organization came crashing down. Breathnach et al. [2] and Jeffries and Flavell [3] announced the discovery that the gene encoding ovalbumin in chicken and the gene

encoding β-globin in rabbit respectively contained non-coding gaps in the coding regions. These gaps were flanked by the same dinucleotides in the two genes; GT on the 5' end of the gaps and AG on the 3' end of the gaps. Soon, Breathnach and Chambon [4] reported that this GT/AG rule was adhered to in a host of coding sequence gaps and the terms **intron** and **exon** were added to the genetic lexicon to describe the coding and non-coding regions of eukaryotic genes.

Figure 2. 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.

Sanger

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 3, 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**.

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 [5]. If the primer was radiolabeled instead of the substrate, the resulting fragment populations

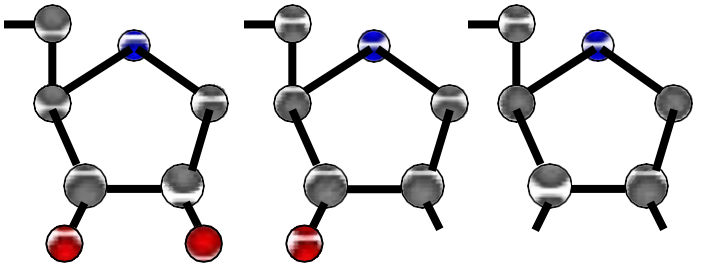

Figure 3. The structure of the three five carbon sugars ribose, deoxyribose, and dideoxyribose. The hydroxyl groups are shown in red.

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 (Figure 4). A major improvement ushered in by Sanger sequencing was the elimination of some of the dangerous chemicals, like hydrazine. The most important improvement, however, was in efficiency. All things being equal, when dealing with nucleic acids, enzymatic processes are more efficient than chemical processes. Case in point, Taq polymerase makes DNA strands off of a template at 500 bases per minute whereas chemical synthesis of a 25-mer oligonucleotide takes more than two hours. If things went well the Sanger method could deliver two to three times as much confirmed data in the same amount of time as Maxam-Gilbert sequencing.

The advent of Sanger sequencing gave a boost to DNA sequencing in general and led to an even more rapid accumulation of sequence data for various genes and organisms.
This increase in sequence data in the scientific literature also resulted in the establishment of the first DNA sequence repository by Walter Goad at Los Alamos National Laboratories in 1979. This repository has since become GenBank [6].

Automated Fluorescence Sequencing

The most dramatic advance in sequencing and the one that carried DNA sequencing into a high throughput environment was the introduction of automated sequencing using fluorescence-labeled dideoxy-terminators. In 1986, Leroy Hood and colleagues reported on a DNA sequencing method in which the radioactive labels, autoradiography, and manual base calling were all replaced by fluorescent labels, laser induced fluorescence detection, and computerized base calling [7]. In their method, the primer was labeled with one of four different fluorescent dyes and each was placed in a separate sequencing reaction with one of the four dideoxynucleotides plus all four deoxynucleotides. Once the reactions were complete, the four reactions were pooled and run together in one lane of a polyacrylamide sequencing gel. A four-color laser induced fluorescence detector scanned the gel as the reaction fragments migrated past. The fluorescence signature of each fragment was then sent to a computer where the software was trained to perform base calling. This method was commercialized in 1987 by Applied Biosystems.

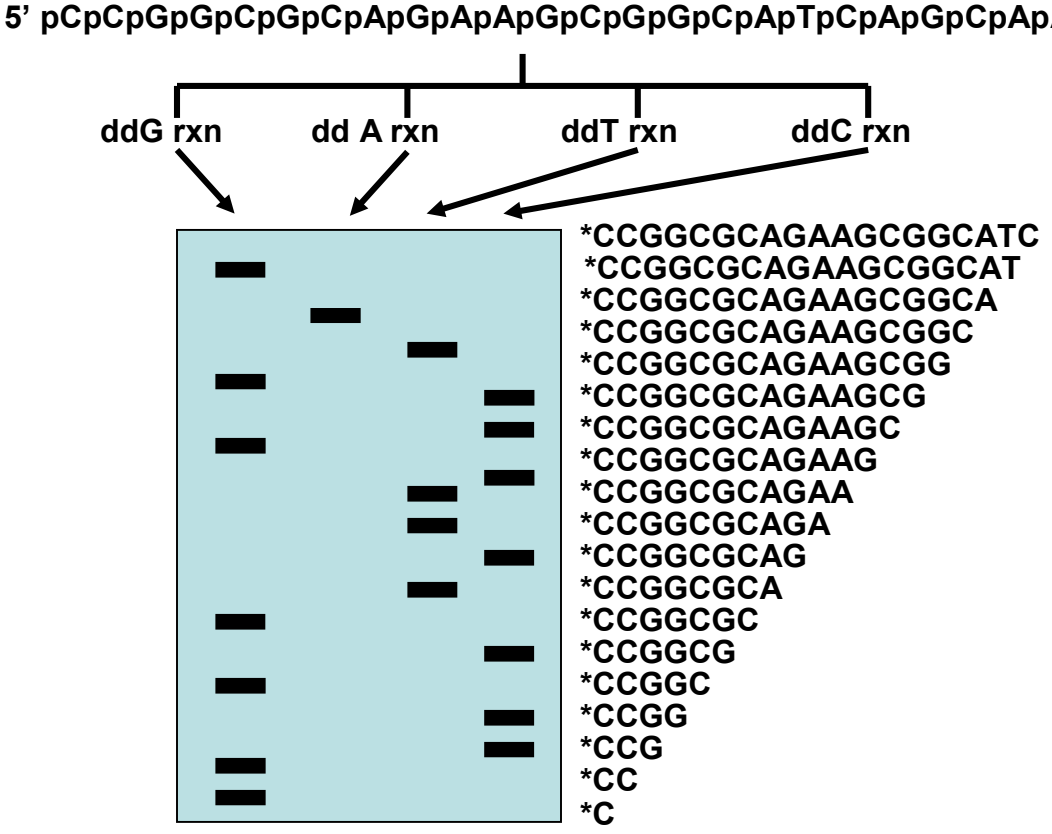

Figure 4. A Sanger sequencing scheme. Here, a sequencing primer is radiolabeled and the reaction involves generation of sequence fragments that are the complement of the template DNA. The sequence fragments are resolved on a polyacrylamide gel and the gel is autoradiographed. Base calling for the template is the complement of the gel band.

James M. Prober and colleagues at DuPont took the fluorescent sequencing method to the next level by developing “a more elegant chemistry” [8]. Instead of fluorescence- labeled primers, they labeled the terminators themselves. The first “dye set” was based upon succinylfluorescein. Slight shifts in the emission wavelengths of the dyes were achieved by changing the side groups. The dyes SF505, SF512, SF519, and SF526 were attached to dideoxy terminators ddG, ddA, ddC, and ddT respectively. The four dyes and their emission spectra are shown in figure 5. All four dye labeled terminators could be excited by an argon ion laser at 488nm and each would produce a peak emission that could be distinguished by the detector. This detection system meant that the sequencing reaction could be carried out in a single tube with all four terminators present and fragment resolution would require only one gel lane [9]. DuPont commercialized this technology themselves for a brief period and then sold the license to Applied Biosystems.

Applied Biosystems continued to refine both the terminator chemistries and the detection/ base calling systems into the 1990’s. Major refinements of the chemistry involved changing the dye labels on the terminators and improving fragment resolution. The fluorescent dyes were changed to a series of rhodamine derivatives; ddG was tagged with dichloroROX, ddA with dichloroR6G, ddC with dichloroR110, and ddT with dichloroTAMRA. Fragment resolution was improved by substituting deoxyInosine triphosphate (dITP) for dGTP and deoxyUridine triphosphate (dUTP) for dTTP. The former helped eliminate band compression on the gels and the latter helped with ddT incorporation in the sequencing reactions. Even though these improvements led to significant increases in DNA sequencing throughput, they were still acrylamide gel-based

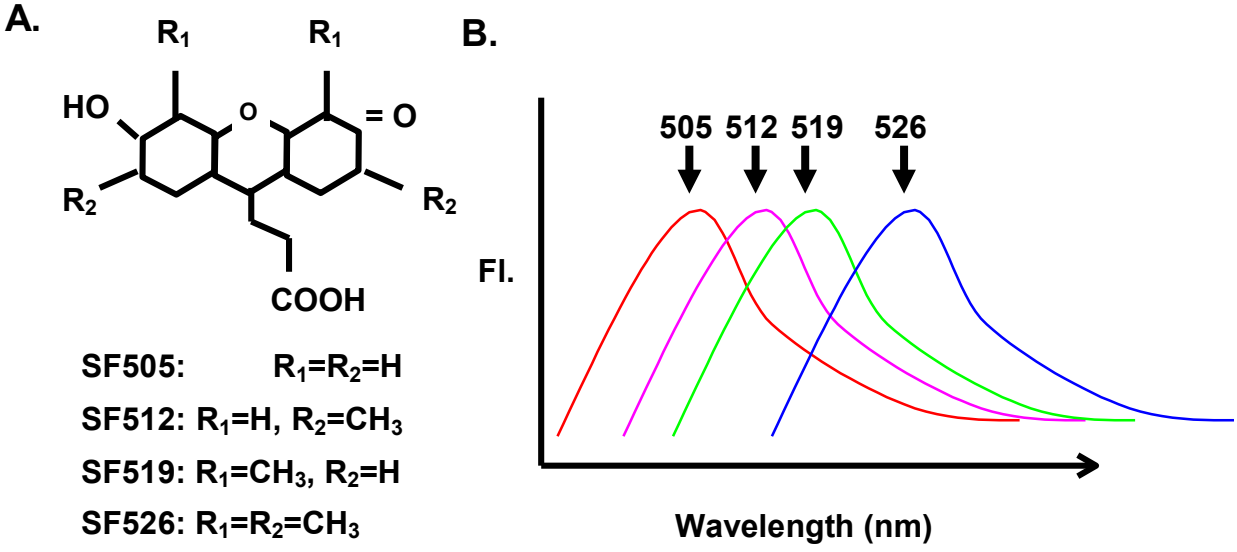

Figure 5. A. Chemical structure of the four succinylfluorescein dyes developed at DuPont. **B.** Normalized fluorescence emission spectra for each of the four dyes following excitation at 488nm. Shifts in the spectra were achieved by changing the side groups R_1 and R_2.

systems. In spite of the improvements in the reactions, detection and data interpretation, gel-based sequencing was still labor intensive and not well suited to a high throughput environment. In the early 1990's Harold Swerdlow and colleagues reported on the use of capillaries to obtain DNA sequences [10, 11]. Capillary electrophoresis was a well established technique in analytical chemistry in the late 1980's. Capillaries are small, a 50μm inner diameter, and they dissipate heat very efficiently due to their high surface area to volume ratios. This means that a capillary- based system can be run with much higher voltages thus dramatically lowering the run times. Most importantly, capillary systems can be automated, a major limitation in gel- based systems. In 1993, B.L. Karger and colleagues reported on the use of a low viscosity separation matrix that could be pumped into capillaries at relatively low pressure [12]. This matrix could replace cross-linked polyacrylamide and remove the final obstacle to the development of a truly automated DNA sequencing platform. With cross-linked polymers the capillary could not be reused. The low viscosity non-cross-linked polymer could be flushed out after a run and replaced for the next run without having to touch the capillary. Studies of thermal stability by Zhang et al. [13] established that a non- cross-linked polymer would be stable at 60^{o}C and would deliver high quality sequence data. Here, then, were all of the elements required for the development of a fully automated, high throughput DNA sequencing platform.

DNA sequencing reactions can be carried out in a single reaction tube and be prepared for loading once the reaction reagents had been filtered out. The capillary system is set up to deliver new polymer to the capillary, load the sequencing reaction into the capillary, apply a constant electrical current through the capillary, and have the resolved fragments migrate past an optical window where a laser would excite the dye terminator, a detector would collect the fluorescence emission wavelengths, and software would interpret the emission wavelengths as nucleotides (Figure 6). At the present time such systems can deliver 500−1000 bases of high quality DNA sequence in a matter of a few hours.

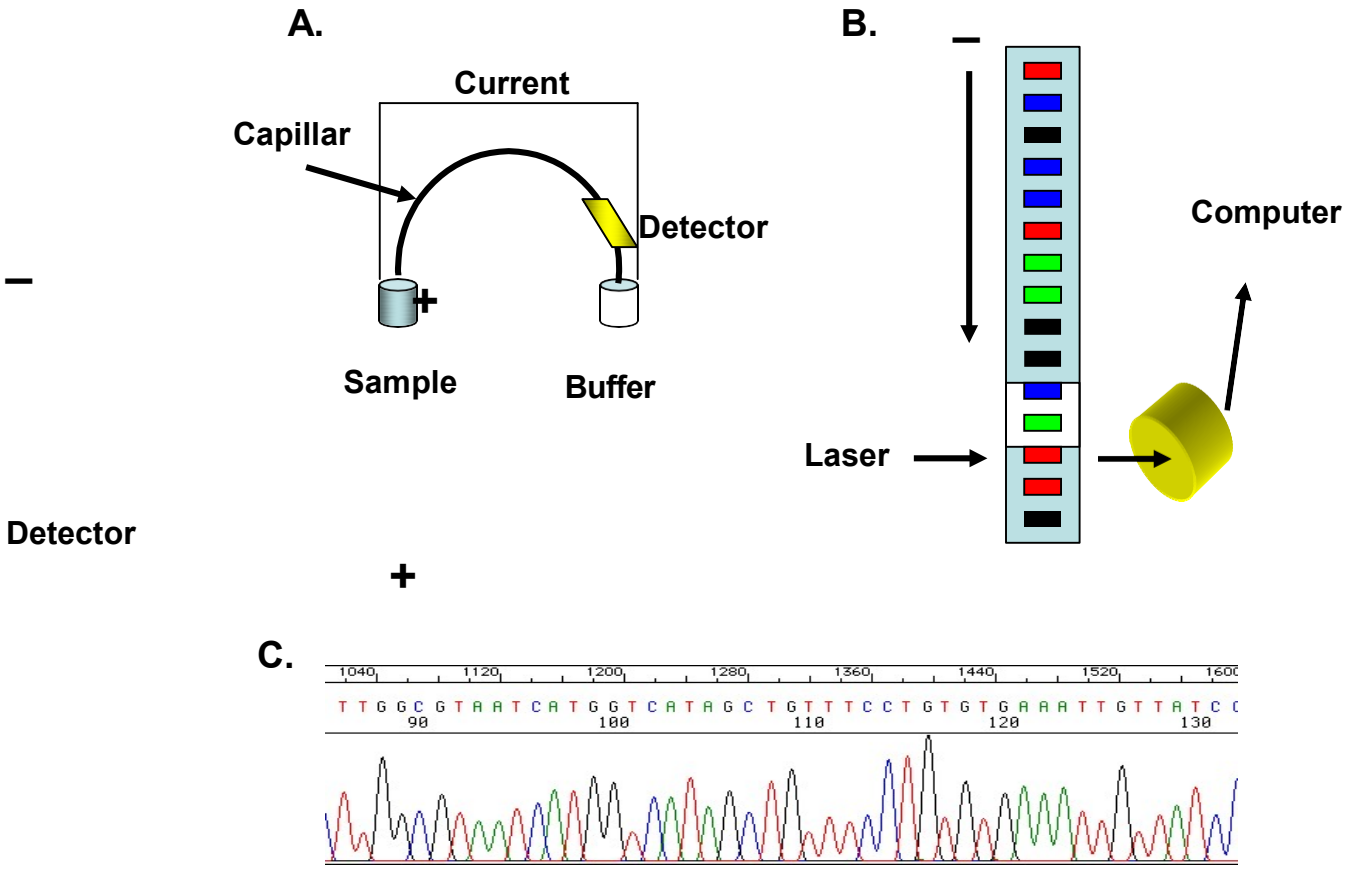

Figure 6. Schematic representation of a capillary-based DNA sequencing system. A. The basic set up of a capillary system. The sequencing reaction is placed in the sample holder, the electrophoresis buffer is held in the second holder, the capillary has been filled with polymer, and the electrophoresis current is applied through the capillary. **B.** False color representations of the resolving sequence fragments running past the optical window, being excited by the laser, the detector reading emission wavelengths and sending that information to the computer where the software has been trained for base calling. **C.** The final sequence electropherogram output.