

Analytical Technologies in Biotechnology
Prof. Dr. Ashwani K. Sharma
Department of Biotechnology
Indian Institute of Technology, Roorkee

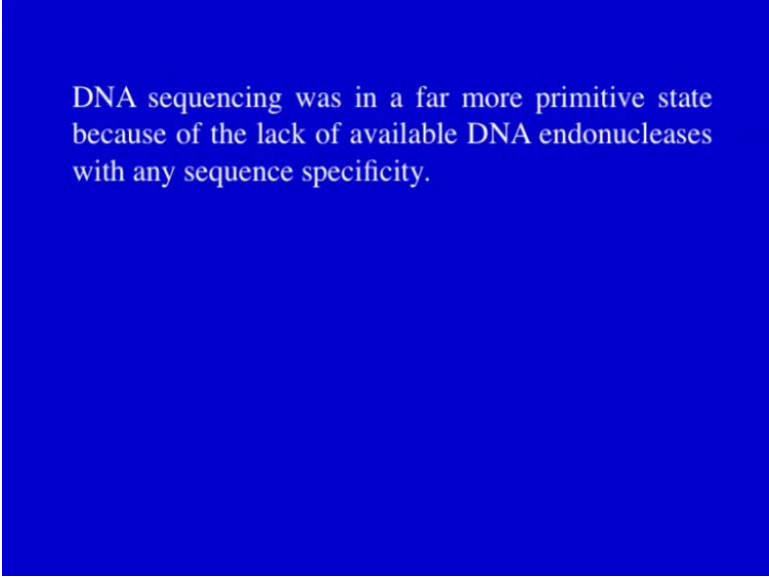
Module - 7
PCR, DNA Sequencing and ELISA
Lecture - 2
DNA sequencing methods

In this lecture, we are going to discuss about various DNA Sequencing Methods, that is the strategy to sequence nucleic acids DNA or RNA. Now, the specific degradation and fractionation of the polynucleotide of interest to fragment small enough to be fully sequenced that is the first requirement of DNA sequencing, that you could degrade very big fragments into a smaller fragments and then sequence them; another important point in basic strategy is the sequencing of the individual fragments.

First thing is to fragment the DNA fragment, DNA sample; that is larger DNA, then sequencing of that material and then ordering of the various sequences of the fragments that is obtained to arrange or to yield the polynucleotide sequence, actually through overlap region set. So, these are the basic three points about designing or you know planning a particular nucleic acid sequencing. Now, little bit in history the first nucleic acid sequence, it was obtained in almost like 65 by Robert Holley, and it was 76 nucleotide molecule of yeast aliening transfer RNA.

Some 12 years after Frederick Sanger had determined the amino acid sequence of insulin, like there was sequencing of numerous other species of TRNA, and say 5 S ribosomal RNA from several organisms was done. So, the art of RNA sequencing reached like in it is Zenith in 1976 with the sequencing by Walter fires of the entire 3569 nucleotide genome of the bacteriophage MS2.

(Refer Slide Time: 02:46)



DNA sequencing was in a far more primitive state because of the lack of available DNA endonucleases with any sequence specificity.

Now, in contrast to RNA sequencing, DNA sequencing was in a far more primitive state, because of the lack of available DNA endonucleases with any sequence specificity. Now, the first complete genome sequence, which was determined was of the gram negative bacterium, haemophilus influenza, which was reported in around 95 by Craig Venter. But, 2010 the genome sequences of over 1000 prokaryotes had been reported, and many more is being determined now, as well as sequences of over one twenty eukaryotes and many more in progress were reported like the those of humans also. And many other vertebrates insects worms, plants and fungi, a dramatic progress has been made in a nucleic acid sequencing methods after 75.

And there were three advances made this possible actually, one is discovery of restriction endonucleases to enable the cleavage of DNA at specific sequences, then the development of molecular cloning techniques to permit, the acquisition of almost any identifiable DNA segment in the amounts required for sequencing, then the development of DNA sequencing techniques, various methods which have been developed, which we are going to discuss in here.

(Refer Slide Time: 04:23)

Chemical Cleavage method

Now, first method which was developed was chemical cleavage method, and it was developed by Maxam and Gilbert, that is Allan Maxam and Walter Gilbert and it involved following steps, which is like, first thing was to radioactively label one end of the DNA. And that is the one, first step in the chemical cleavage method to radio label one and usually it is used to be the 5 prime end and it was labeled with the ^{32}P . Now, if DNA already has 5 prime phosphate group, then it has to be removed by the alkaline phosphates treatment and then 5 prime terminus is labeled in a reaction with γ ^{32}P , ^{80}P catalyzed by polynucleotide kinase.

So, through this particular reaction the ^{32}P or the radio label P is incorporated, then second step was to cleave the DNA in a base specific manner. Now, basic strategy for this chemical cleavage method is to, specifically cleave the end labeled DNA at only one type of nucleotide. So, under conditions that each molecule is broken at an average of one randomly located susceptible bond, so in this only the cleavage occurs at each base or a particular base according to the method, according to the strategy involved.

Now, what will happen, this will produce a set of radioactive fragments whose members extend from the ^{32}P labeled and that is which is labeled from one end, to one of the positions occupied by the chosen base, where the cleavage has occurred. Now, always labeling can also be done on 3 prime and so both strategies could be adopted.

(Refer Slide Time: 06:26)

- For example, if the DNA to be sequenced is

$^{32}\text{P-TGTAGGAGCT}$

Cleavage on the 5' side of the G residues would produce the following set of 5'-labeled fragments:

$^{32}\text{P-TGTAGGA}$
 $^{32}\text{P-TGTAG}$
 $^{32}\text{P-TGTA}$
 $^{32}\text{P-T}$

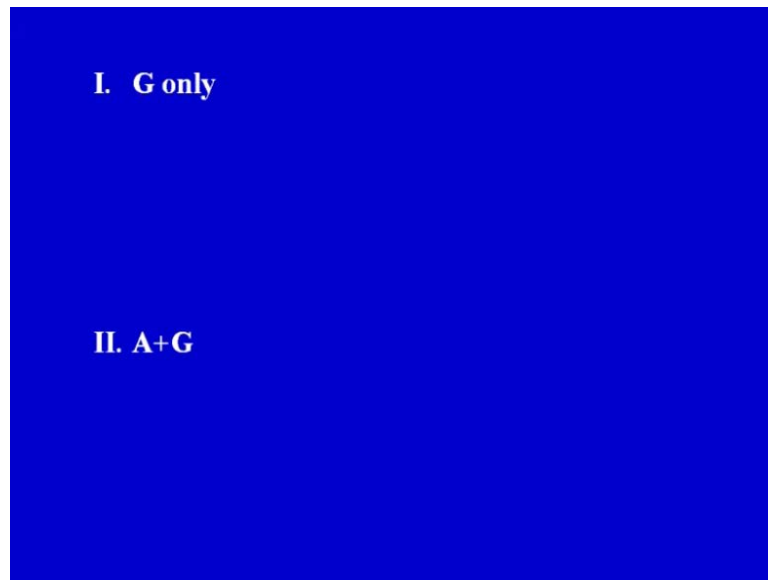
Now, for example, if the DNA has to be sequenced, if you can see on your screen, if this is the DNA which needs to be sequenced, then it is labeled at 5 prime end ^{32}P always remember left hand side is 5 prime and 3 prime is the right hand side, unless it is specified. So, what will happen cleavage on the 5 prime side of the, say for example, we can take one example G residue, then you will get following fragments, because like one cleavage will be occur after T and before G that is at five prime side of G. And you will get this particular, like different fragments you are going to get here, which is smallest one will be T.

Because, if it is occurring at this one if it occurs at this G then you will get another bigger fragment, another G we will get this and finally, when it is here then you will get this fragment. That is what you can see on the gel, when you are running the acrylamide gel you can see only those which are labeled fragments, the other fragments which have gone will not be able to be seen here. Now, polyacrylamide gel electrophoresis will separate these fragments according to size, and remember as we have discussed earlier this can even differentiate between a single nucleotide.

So, the position of the G residue in the DNA identified from the relative positions on the gel of their corresponding ^{32}P labeled fragments as revealed by autoradiography. Now, in order for this method to work, the gel must be sufficiently resolving power have sufficient resolving power to separate unambiguously fragments, that differ in length by

only one nucleotide, and as I said it is possible in polyacrylamide gel. So, the DNA to be sequenced may be cleaved at a specific basis by subjecting it in a separate aliquots that is 4 different treatments are done, and then those fragments 4 reactions are run on the gel, so here if you go for each base in ideally, it should be A T G C.

(Refer Slide Time: 08:47)



But, practically what is done is first, there is a reaction for cleaving at G only, so what is done is DNA is treated with dimethyl sulphate, and which methylates G residue at N 7. And rendering the glycosidic bond of the methylated residue, susceptible to hydrolysis and subsequent treatment by piperidine cleaves the polynucleotide chain, before the depurinated residue. Then second reaction will be for A plus G that is either A or G, so the treatment with TMSF or dimethyl sulfate will preferentially methylate A residue at only N 3 rather than N 7.

And so above treatment cleave the DNA at A residue at only about 1, 5th rate it does at G residue, but if this whole reaction is treated with acid, then both A and G are released remember, without acid only G is released frequently, but in acidic treatment both A and G will be released at comparable rate. So, that is the another reaction and then piperidine treatment causes strength cleavage before both A and G, and A residues are identified by comparing the position of G and A plus G cleavage.

(Refer Slide Time: 10:09)

III. C+T

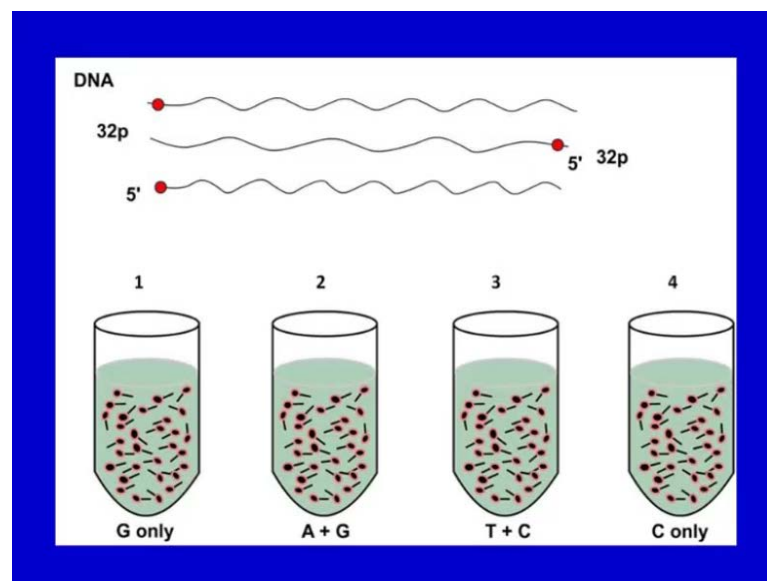
The third reaction could be C plus T, now the reaction of DNA with hydrogen followed by piperidine treatment cleaves DNA before both, it is C and T residues. And then IV reaction allocate will be for only cleavage at C, now here DNA is treated with hydrogen in salt actually 1.5 molar NaCl, and where cleavage will occur only at C residue. So, this way you can compare C and C plus T and then, find out which residue it is, now in all 4 reaction the conditions are adjusted, so that the strengths are cleaved at an average of one randomly located position each.

(Refer Slide Time: 10:59)

3. Cleavage fragments are separated according to size gel.

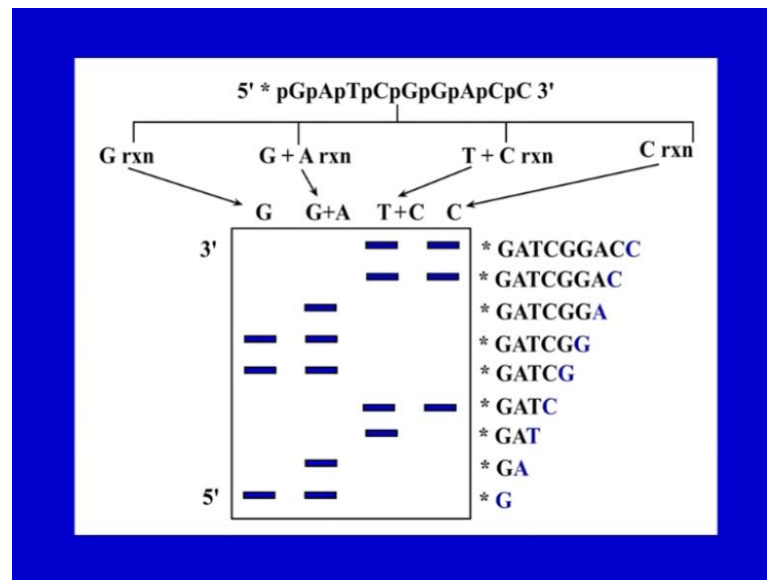
Now, cleavage fragments which are obtained, they will be separated on the basis of size, so the 4 differently fragmented sample of DNA that is a plus G, C and C plus T will be simultaneously electrophoreses. And then it is a sequencing gel as we have discussed earlier, it is done at very high temperature and 8 molar urea to avoid any base pairing, and ensure that DNA fragments are separated only according to their size, and this could be read directly of an autoradiogram of the sequencing gel.

(Refer Slide Time: 11:39)



So, if you can see on your screen, this is basic strategy here is that first thing is you convert it to single stranded DNA, mostly here single stranded DNA's done in Maxam and Gilbert method. Then it is labeled at 5 prime end and 4 tubes that contains 4 reaction, that is to cleave at base G or A out G T and C and base C will be performed as we have discussed and then finally, it will be run on the electrophoresis.

(Refer Slide Time: 12:07)



So, if you can see these are like, the lanes for each of the reactions and if you can see here this particular one, this was the sequence and as we have gone through, you can see the sequence could be build by starting from the bottom. Now, remember when the cleavage will occur at 5 prime side of this G here, earlier we have taken the sequence which was T, which was different little bit it was TGT, but just to understand that.

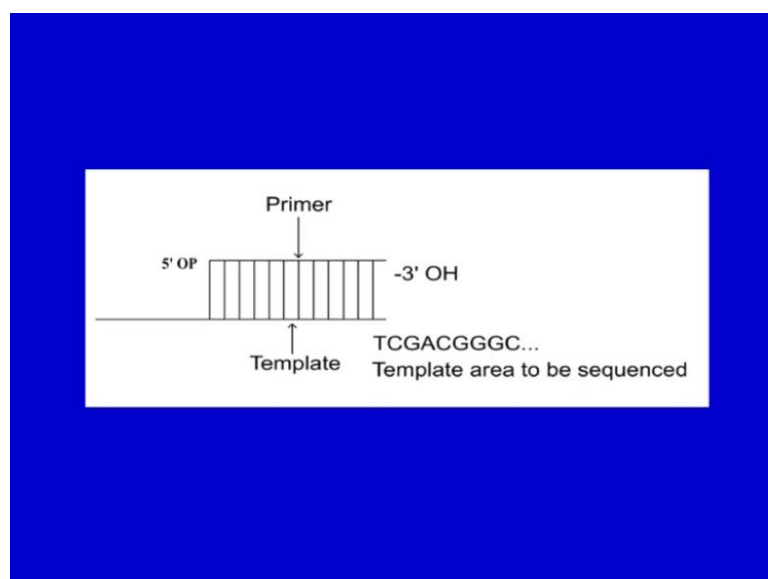
If it is like here, then they will be only labeled P actually and no base, so that will sometimes is difficult to seen here, so that way the whole sequence can be generated except for first base actually. Now, this was Maxam and Gilbert method, and it was preferred method at that time for sequencing and but, it is no longer used actually.

(Refer Slide Time: 13:15)

The Sanger Method

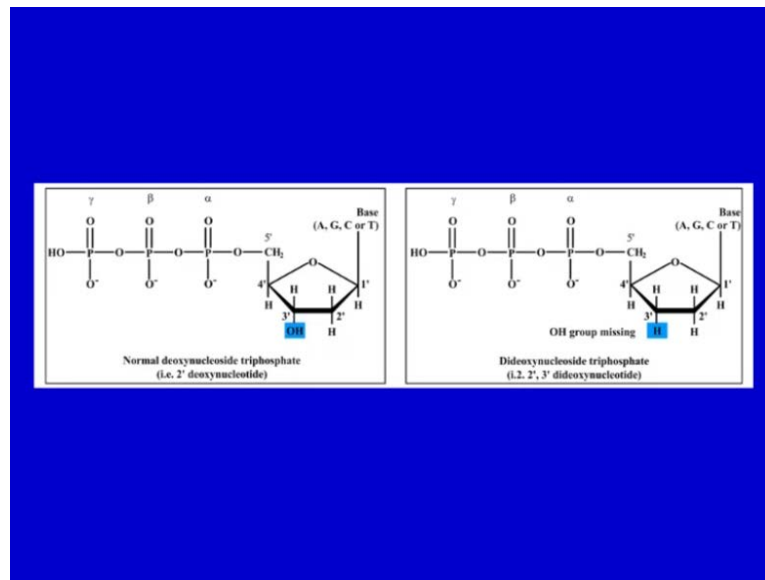
After that the Sanger method was very popular, and it is called chain termination method or dideoxy chain termination method, what it does is utilizes the Klenow fragment of DNA polymerase to synthesis complementary copies of the single stranded DNA, means sequenced, and it lacks five prime to 3 prime exonuclease activity. now DNA to be sequenced is incubated with the Klenow fragment, suitable primer like here PCR is required like as we've discussed in earlier lecture, and 4 DNTP's. Now, here what happens is that either at least 1 DNTP or the primer is labeled here, so that this could be read actually; so either primer could be labeled or DNTP's labeled.

(Refer Slide Time: 14:13)



So, what is done here like, if this is primer and this is template as it extend across, then a new strand will be synthesized on this primer. So, in this case, what is done is in this method a small amount of 2 prime, 3 prime dideoxynucleotides triphosphate is added to the reaction mixture.

(Refer Slide Time: 14:39)



So, what will happen, so these are like if you see on your screen, this is normal DNTP and this is where you have 2 H here at both 2 prime and 3 prime side, so the extension will not occur, the next reaction will not occur actually in chain extension. So, when the dideoxy analog is incorporated in the growing polynucleotide, in place of the corresponding normal nucleotide chain growth is terminated, because of the absence of 3 prime OH group. So, by using only a small amount of dideoxy nucleotide, a series of truncated chains are generated and each of which is terminated by a dideoxy analog, at one of the positions occupied by the corresponding base.


Now, each of the 4 D DNTP's is reacted in a separate vessel, now the 4 reaction mixtures are simultaneously electrophoresed, like it was done in Maxam and Gilbert method in parallel lines on a sequencing gel. And this is a long thin gel, and it contains like earlier 7 molar urea and run at 70 degree Celsius, to avoid the hydrogen bonding and ensure the DNA fragments separate only according to their size. The sequence of the DNA that is complementary to the template DNA, can then be directly read of an autoradiogram of the sequencing gel from bottom to top, like it was done in Maxam and Gilbert method. If

you see the smallest one at was at the bottom and then, the largest fragment which is like you go in that particular sequence.

However a single gel is incapable of resolving much more than 300 to 400 consecutive fragments, the limitation is you can be generated like, you can run more than two sets of gels, one run for a longer time and higher voltage than the other. To obtain the sequences up to say 800 base pair DNA fragment, that is the different strategies like one can fragment and many fragments of 300 to 400 base pairs could be run or it could be other way that you can have longer one.

The absolute requirement is that DNA to be sequenced is in a single stranded form, as it is like it could be run on the gel, and to facilitate the isolation of single strands DNA to be sequenced may be cloned into one of the clustered cloning sites in the region of M 13 M P series of vectors. Feature of these vectors is that cloning into the same region can be mediated by anyone of a large selection of restriction enzymes, but still permits the use of single sequencing primer.

(Refer Slide Time: 17:43)

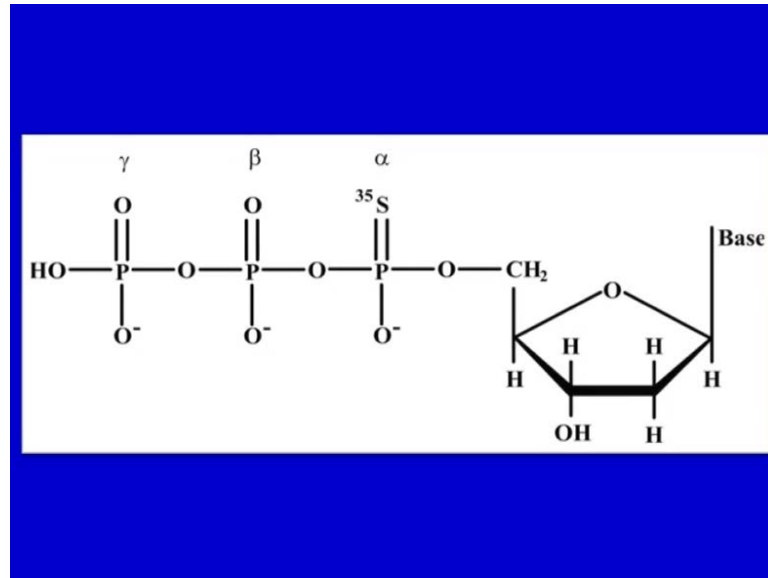


Modifications of chain-termination sequencing method

They could be many modifications of chain termination sequencing method, one the sharpness of the auto radiographic image can be improved, by replacing the ^{32}P radio labeled with the much lower energy like say ^{33}P or ^{35}S . In the case of ^{35}S , this is achieved by including an alpha ^{35}S deoxynucleoside triphosphate in the sequencing reaction, and this modified nucleotide is accepted by DNA polymerase, so there is no

problem about that. And it is incorporated into the growing DNA chain, now known isotopic detection methods have also been developed with chemiluminescent, chromogenic or fluorogenic reporter systems.

(Refer Slide Time: 18:44)



Now, although the sensitivity of these methods will not be as great as with radio labels, but they are adequate for many different purpose. You can see here this ^{35}S incorporated in your at alpha position, there other technical improvements to Sanger's original method have been made by replacing say Klenow fragment of E coli DNA polymerase one. Natural or modified forms of the phase T 7 DNA polymerase have found favored, has the DNA polymerase of the thermophilic bacterium like, tech polymerase and other polymerases have also come in picture.

The T 7 DNA polymerase is more processive than Klenow polymerase that is, it is capable of polymerizing a longer run of nucleotide, before releasing them from the template. It is incorporation of dideoxynucleotides is less affected by local nucleotide sequences, and so the sequencing ladder comprises a series of bands with more even intensities.

Taq DNA polymerase can be used in chain termination reactions, carried out at high temperatures and this minimizes, chain termination artifacts caused by secondary structures in the DNA. Then Tabor and Richardson in 95 shown that, replacing a single phenylalanine residue of Taq DNA polymerase with the tyrosine residue results in a

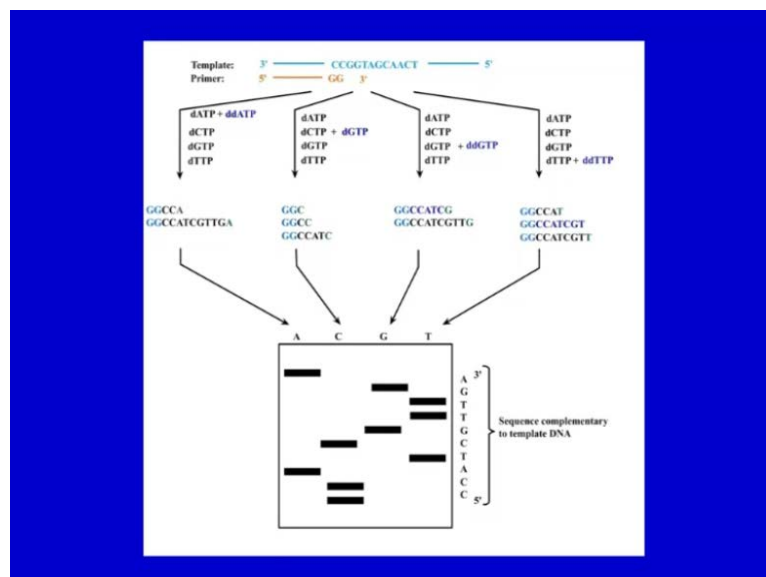
thermostable sequencing enzyme; that is no longer discriminates between dideoxy and deoxy nucleotides.

Combination of chain terminator sequencing and M 13 vectors to produce single stranded DNA is very powerful, very good quality sequencing is obtained with this technique, especially when the improvements given by 35 S labeled precursors and T 7 DNA polymerases are exploited. So, there were further modifications have been done, which allow sequencing of double stranded DNA, double stranded input DNA is denatured by alkali and neutralized. And then one strand is annealed with a specific primer, for the actual chain terminator sequencing reaction.

This approach has gained in popularity as the convenience of having universal primer, has grown less important with the wide spread availability of oligonucleotide synthesizers. So, with this development Sanger sequencing has been liberated from its attachment to M 13 cloning system for example, polymerase chain reaction amplified DNA sequencing can be sequenced directly also. One variant of the double stranded approach of an employed in automated sequencing is cyclic sequencing, this involves a linear amplification of sequencing reaction, using say 25 cycles of denaturation annealing of a specific primer to one extend only.

And extension in presence of Taq DNA polymerase plus labeled dideoxynucleotids, now alternatively labeled primers can be used with unlabeled dideoxynucleotids.

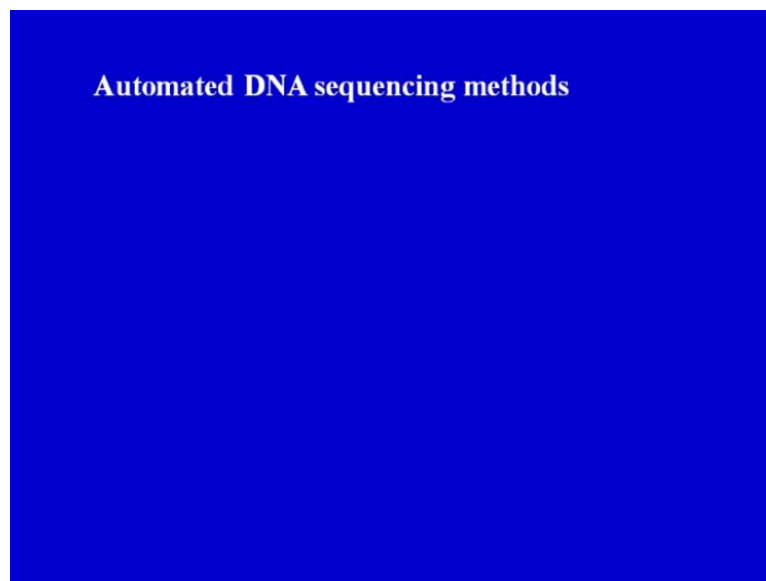
(Refer Slide Time: 22:03)



So, here this schematic shows, this particular method here, so what you have is a templates and a primer which is on one side, so this will be extended by the polymerase. And what you get is according to the particular dideoxy nucleotide will get these fragments here, because they will be like, it will be interrupted at different forms different bases actually. So, if you have ddATP say, then it will be like a will be these fragments will be generated as complementary to this template fragment likewise for G for C and for G and for TDD, dideoxy T.

So, that way and if you run on these fragments, then you will get different sized fragments which are terminated fragments, and as you read it from bottom to top like you can make out this particular sequence, which is CCGG, CCATC from 5 prime to 3 prime. So, this is like as you have seen here, this is GG, so it will come CC here then A, it is A, then T, then CCGG and likewise you will get the sequence. So, reading from top to bottom excluding the primer you will be able to read the sequence, from the sequencing gel on autoradiograph.

(Refer Slide Time: 23:56)



Now, sequencing methods, DNA sequencing methods have been automated and lot of manual intervention is lessened. So, in order to sequence very large tracts of DNA such as entire chromosome, the Sanger method has been greatly accelerated through automation. This required that the above described radio labeling techniques which are not readily automated be replaced by say fluorescence labeling techniques. So, it has

benefit of eliminating the health hazards and storage problems of using radio labeled nucleotides.

In the most widely used such technique each of the 4 ddNTP's, used to terminate chain extension is covalently linked to a different fluorescence dye that is with different colors. And the chain extension reactions are carried out in a single vessel, containing all 4 of these labeled ddNTP, so it is not separately done, it is not 4 reactions, but since you can see the colors that is 4 different dyes, it is done in the same vessel. The resulting fragment mixture is subjected to sequencing gel electrophoresis in a single lane, now as each fragment exits the gel its terminal bases identified according to its characteristic fluorescence spectrum, by a laser activated fluorescence detection system.

The fluorescence detectors used in these devices, which have error rates of say 1 percent are computer controlled and hence data acquisition is automated. And in most of the advanced systems, such systems the sequencing gel is contained in an array of say up to 96 capillary tubes, rather than in a slab shaped apparatus sample preparation. And loading are performed by robotic systems, and electrophoresis and data analysis are fully automated; for high sensitivity DNA detection in 4 color sequencing and high accuracy based calling.

One would ideally like the certain criteria to be met, one is each of the 4 dyes to be exhibit strong absorption at a common laser wavelength, so that is important that all the 4 dyes are detected at same level, to have an emission maximum at a distinctly different wavelengths. That they emit a distinctly different wavelengths and to introduce the same relative mobility shift of the DNA sequencing fragments. Now, recently dyes with these properties have been identified and successfully applied to automated system.

(Refer Slide Time: 27:00)

Advantages of automated DNA sequencing

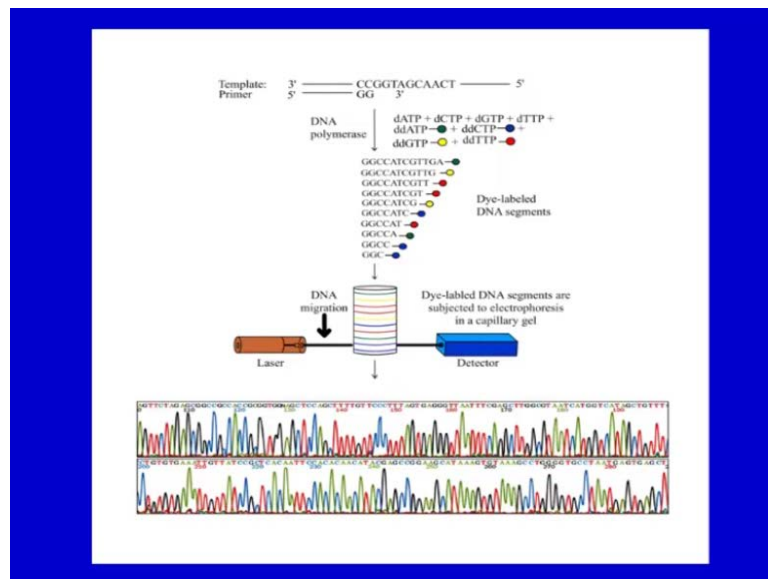
There are a lot of advantages of automated DNA sequencing method, one is that manual sequencing can generate excellent data, but even the best sequences in laboratories due to like say poor autoradiographs are frequently encountered. And that will make sequence read out difficult or impossible, usually the problem is related to the need to run different termination reactions, in different tracks of the gels, even skilled DNA sequencers ignore bad sequencing tracks, but many laboratories do not.

Now, this leads to poor quality sequencing data, and the use of single gel track for all 4 dideoxy reactions means that, this problem is less accurate in automated, so one is that many times you might get poor data due to certain practical problems. So, in automated sequencing this will be little less, never the less it is desirable to sequence a species of DNA several times. One should do it to confirm the sequence, and many times the two strands are also sequenced to confirm that the obtained sequence is right, and this eliminates errors quite a lot.

It should be noted that long runs of the same nucleotide or a high GC content can cause compression of the bands on a gel. If there is a same nucleotide, long stretch of that or high GC content, then it could be a problem necessitating manual reading of the data. Even with an automated system this could be a problem, note also that multiple tandem shot repeats, which are common in the DNA of a higher eukaryotes can reduce the fidelity of the DNA copy, particularly with Taq DNA polymerase.

Second advantage of automated DNA sequencing is that, the output from them is in machine readable form, so this eliminates the errors that arise when DNA sequences are read and transcribed manually. In the new generation of the slab gel, like in new generation of sequence slab gel is replaced with like I said 48 or 96 capillaries filled with gel matrix, the equipment has been designed for use with robotics. Therefore, minimizing hands on time and increasing the throughput, with a 96 capillary sequencer you can sequence very long fragments per day, many many nucleotides per day.

(Refer Slide Time: 29:49)



So, on your screen if you can see, here when you take 4 florescent dyes attached, different color dyes are test to each base or each dideoxy analogs, then what you get you get fragments here, as they are of different sizes, as they PCR goes on, and these fragments you can see when they run on the same gel and they are in the same like extensions are done in the same vessel. So, what you get is as they run, you can see that colors could be monitored through the laser here and data could be recorded, and on based on the color of the dye the base could be identify.

And you will get a chromatogram in this form, where you will get different colors as they move out of the base of the gel and they could be a proper and very accurate sequence could be obtained.

(Refer Slide Time: 30:56)

Genome Sequencing

Now, genome sequencing is one another challenge, so like we have seen in Sanger's method fragments could be done, and certain DNA fragments could be sequenced without adding any problem that you can fragment them, and you can sequenced those fragments and overlap them on a smaller scale. But, when you come at a genome label there is a major technical challenge in sequencing a genome, and it is like when you are sequencing a genome, there are the DNA sequencing it is like assembling the tens of thousands to tens of millions of sequenced segments here, depending on the size of the genome. It is not like 1 DNA or a few DNA fragment, it is a 1000 of DNA fragments has to be arranged and two contiguous blocks, are called consigs.

And assigning them to the proper position in the genome, it is its much more complex task, much more challenging task then sequencing of longer DNA fragment. Now, one way that contigs might be ordered is through chromosome, walking however to do so far eukaryotic genome would be prohibitively time consuming and expensive. So, there are certain techniques which have been developed for genome sequencing, and as you heard about human genome and other eukaryotic and other genomes have come up, they have been done the two strategies has been employed, one was map based genome sequencing.

So, this is efficient technique of genome sequencing and it was developed in the late 80's it is a role, so what is the low resolution, physical maps of each chromosome are

prepared by identifying sheared landmarks on overlapping say 250 kilo base insert that are cloned in each artificial chromosomes. Now, these landmarks often take the form of 200 to 300 base pair segments, which are known as sequence tagged sites or STS, and whose exact sequence occurs nowhere else in the genome hence two clones that contain the same STS must overlap.

Now, the STS containing inserts are, then randomly fragmented into say 40 KB segments that are sub-cloned into cosmic vectors, so that a high resolution map can be constructed by identifying their landmarks overlaps. Now, the cosmid inserts are then randomly fragmented in to overlapping 5 to 10 kilo base or 1 KB segments, for insertion into plasmid or say M 30 vectors. Now, these inserts which could be like say 800, M 13 clones per cosmid, are then sequenced like 400 based pairs per clone.

And the resulting so called reads are assembled computationally into contigs to yield the sequence of the parent cosmid insert, with the redundancy of say 400 base pair per clone into 800 clones per cosmid, it could be you can get redundancy of 8, which is you have 40,000 base pairs per cosmid, so it could be like you can get redundancy of around 8. Finally, the cosmid inserts are assemble through cosmid walking, using their landmark overlaps and these landmarks ideally spaced at intervals of say 100 kilo base or less. And you will sequence of the east artificial chromosome inserts, which are then assembled using their STS to yield the chromosomes sequences.

Now, genome of most of the complex eukaryotes contain numerous tracts of repetitive sequences, and such repetitive sequences greatly excurvate the difficulty of finding properly spaced STS. To partially circumvent the later difficulty, STS like sequences of CDNA's which are known as experienced expressed sequence tags are used in place of STS. Now, since the MRNA's from which CDNA's are obtained which are reverse transcribed and they encode proteins, they are unlikely to contain repetitive sequences.

So, this is one method which is map based sequencing, and as you have seen it is like first test to clone it in certain fragments in artificial, each artificial chromosome, then cosmid, then certain vectors and finally, there is those sequences has to be assembled. There is another approach that is called whole genome shotgun assembly strategy, the whole genome shotgun assembly strategy is formulated by Craig Venter Hamilton, Smith and Leroy hood.

(Refer Slide Time: 36:15)



**The Whole Genome Shotgun Assembly
Strategy (WGS)**

And this strategy in this strategy, a genome is randomly fragmented into a large number of clone fragments, first it is fragmented and then, large number of clone fragments are then sequenced. And then genome is assembled by identifying the overlaps between pairs of fragments, so here it is a random shotgun method adopts a random fragmentation, cloning and then sequencing and then, identifying the overlaps to arrange the sequences in a proper order.

(Refer Slide Time: 37:06)



For bacterial genomes:

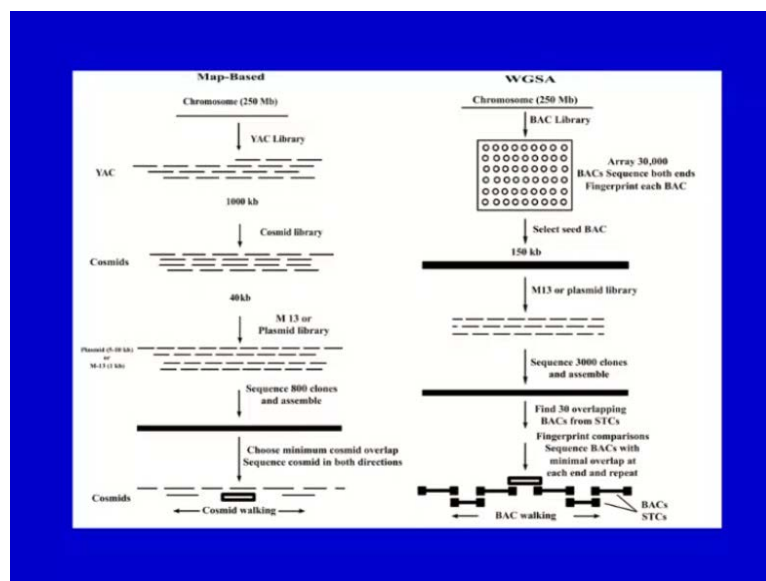
For eukaryotic genomes

For bacterial genomes, the bacterial genomes this shotgun approach is carried out straight forwardly by sequencing 10's of 1000's of fragments and assembling them, then in a task known as finishing. The gaps between contigs are filled, and by several techniques including synthesizing PCR primers complimentary to the ends of the contigs, and using them to isolate the missing segments. For eukaryotic genomes there are much great to sizes are there, and require strategy to be carried out in stages, the shotgun strategy which is carried out in stages.

So, here a bacterial artificial chromosome library of say 150 kilo base inserts is generated, the insert in such of these like artificial chromosome clones is identified by sequencing 500 base pair in from each of the end to yield segments; known as sequence tact connectors, or BSE ends. Now, one BSE insert is then fragmented and shotgun cloned into plasmid or M 13 vectors, the fragments are sequenced and assembled into contigs. Now, sequence of this sheet BSE is then compared with the database of STC's to identify the 30 overlapping BAC, say up to 30 overlapping fragments.

Here 30 overlapping BAC clones, the two with minimal overlap at either end are then selected sequenced, and the operation repeated until the entire chromosome is sequenced. This is called you can say BAC walking, bacterial artificial chromosome walking, which for the human genome required 27 million sequencing seeds.

(Refer Slide Time: 39:12)



So, these are two strategies here compared, in say map based strategies like we have discussed, there is a chromosome which is 250 million base pairs, this is like first artificial eased chromosome, eased artificial chromosome library is made. And then finally, cosmid libraries made and then, they are fragmented and cloned into M 13 or plasmid library, which are then chosen to be sequenced. And then cosmid walking is performed and finally, the sequences are obtained and they are properly arranged.

In another method where that is shotgun approach, were chromosome is taken and bacterial artificial chromosome library is prepared, then after that M 13 or plasmid library is prepared. Then lot of these clones are sequenced and assembled, here it is randomly fragmented and then this overlapping, this is like called BAC walking and then this particular whole sequences obtained.

(Refer Slide Time: 40:26)



**Advantages of WGS strategy over
map-based strategy**

There are certain advantages of shotgun strategy over the map based strategy, the shotgun strategy is readily automated through robotics, and hence is faster and less expensive, then the map based strategy. Most known genome sequences have been determined using the shotgun approach many in a matter of say few months, and it is advent reduce the time to sequence the human genome also by several years. Never the less is appears that for eukaryotic genomes, most of the residual errors in WGS or you can say shotgun based genomes sequencing, can be eliminated by finishing it through the use of some of the techniques of the map based strategy.


((Refer Time: 41:20)) Now, if we just little bit discuss about human genome, human genome has been sequenced which rough draft was reported in 2001 by 2 independent groups. Publically funded international human genome sequence consortium, which is like 6 countries were involved, and were led by Francis, Collins, Eric, Lander and John Sulston and which used the map based strategy. And then a privately funded group from Celera genomics led by Craig Venter, which used the shotgun approach. So, the first one determined genome sequence was our conglomerate from numerous individuals, where is that from Celera genomics was derived from five individuals, but mainly from Craig Venter.

Now, these draft sequences like ten percent of the gene rich chromosomal regions, known as euchromatin and much of the largely unexpressed chromosomal regions known as constitutive heterochromatin. Both draft assemblies had sequencing errors rates of say 1 percent, and contained around 100000 gaps, so that 160000 gaps around that. So, that the order and the orientation of many contigs within the local regions had not been established, since 2004 the public funded consortium reported the finished sequence, it cover almost 99 percent of the euchromatin genome. And with a rare error rate of 0.001 percent and had only 281 gaps.

In 2007, Venter reported the finish sequence of his own deployed genome, and these stunning like, so it is a very like great achievement, and these stunning achievements the culmination of over a decade of intense efforts by 100 of scientists, has really revolutionized the way both biochemistry and medicine will be viewed, and they will be practiced. Few of the major observations here were that about 40 percent of the human genome consist of repeating sequence of various lengths, and only 28 percent of the genome is transcribed to RNA.

Then only 1.2 percent of the genome that is 4 percent of the transcribed RNA and coats protein, and it is said that human genome appears to contain around say 23,000 odd gens or protein encoding gens, which are open reading frames; earlier it was thought that more than 50,000 open reading frames are present. Only a small fraction of human protein families is unique to vertebrates, and most occur in other life forms also; two randomly selected human genomes differ on an average by only one nucleotide per 1000. And that is the two people, if we compare they will have 99.9 percent genetical identity, or they will be identical, so this is like only 0.1 percent make.

(Refer Slide Time: 44:56)



<http://www.ncbi.nlm.nih.gov/projects/mapview/>

Genomes of eukaryotes including that of homo sapiens can be explored at this given website, and so this was really a great achievements and many genomes are being sequenced and reported. Now, these were methods, which were advanced like map based or shotgun approach where advance DNA sequencing method. But, now there are next generation methods of DNA sequencing have come in, which is like second generation you can say third generation methods have come in, we will discuss some of them here.

(Refer Slide Time: 45:34)



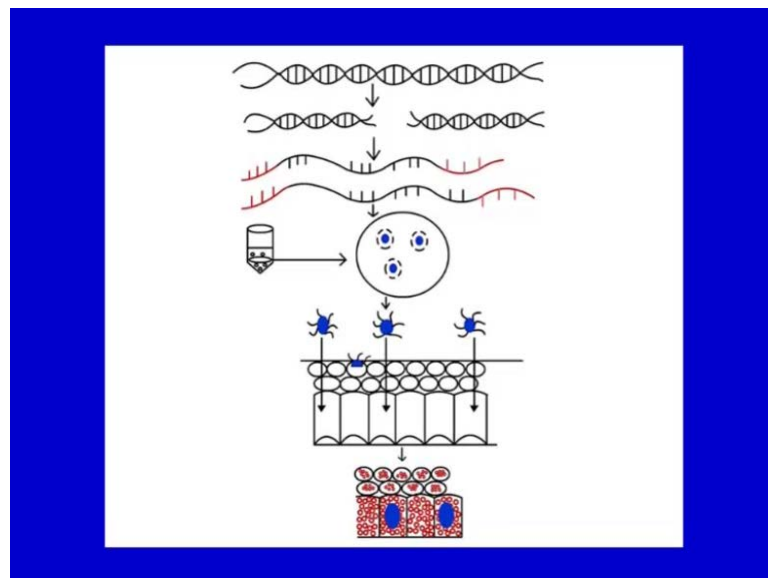
Next Generation DNA Sequencing Technologies

1. The 454 Sequencing System

One is the 454 sequencing system, the Watson genome was sequenced using a system developed by 454 life sciences that employs the following methodology. The methodology implied is here that genomic DNA is randomly sheared to small say 300 to 500 base pair fragments and ligated to adaptors which in turn are specifically bound by say 30 micrometer diameter DNA captured beads under dilution conditions such that, at most 1 DNA fragment is bound to each bead. Now, the beads are suspended in a PCR mixture containing DNTP's primers complimentary to the adaptors and Taq DNA polymerase.

The suspension is emulsified with oil such that, each aqueous droplet contains only one bead and that is each bead is contained in it is own micro reactor, thus preventing the introduction of competing or contaminating sequences. The PCR is carried out by thermocycling, until 10 million identical DNA fragments are bound to each DNA capture bead. The emulsion is broken by the addition of isopropanol, the DNA is denatured and the resulting single stranded DNA bearing beads are deposited into 5 peculator wells, on a fiber optic slide with one bead per well, and the slide containing 1.6 million wells, so this is how it is done.

(Refer Slide Time: 47:13)



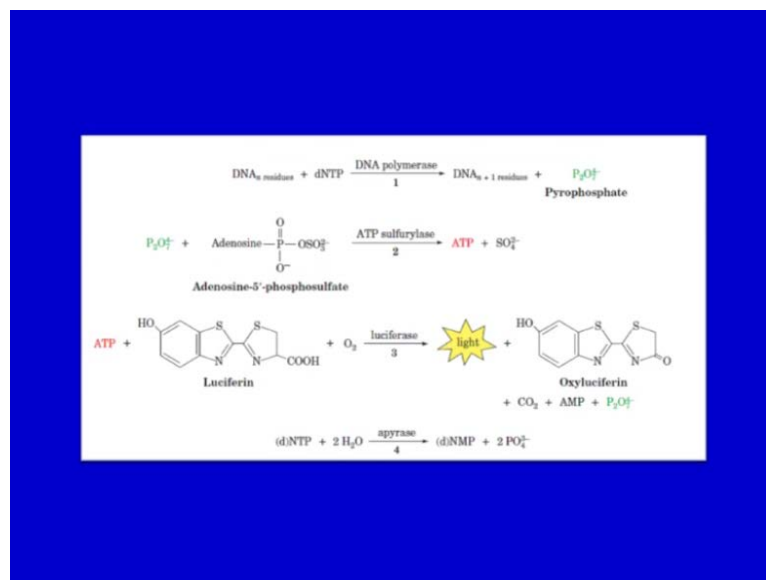
These are like fragments, there is a bead here on which each one fragment is processed and these are wells where these beads are put in, the DNA on each of these beads is sequenced using a series of coupled enzymatic reactions, and these are collectively

known as pyrosequencing. So, in pyrosequencing a solution containing only one of the 4 dNTP'S is flowed over the bead containing slide, now if that dNTP is complimentary to the first unpaired based on a template strand, DNA polymerase catalyzes addition to the primer strand and releases pyrophosphate ion.

In a reaction catalyzed by the enzyme ATP sulfurylase, the pyrophosphate ion reacts with adenosine 5 prime phosphosulfate to yield ATP. Now, in a reaction catalyze then by F enzyme luciferase, the ATP reacts with luciferin and O₂ to yield oxy luciferin and a flash of visible light the well from, which the light flash emanates together with it is intensity is recorded by an imaging system. And thus identifying those wells in which the forgoing nucleotide was added to the primer strand, the intensity of the light is propositional to the number of nucleotides reacted. So, that when two or more consecutive nucleotides of the same type are added to the primer strained, there number is determined.

In preparation for the next reaction cycle, any unreacted dNTP'S and ATP'S are hydrolyzed to mononucleotides, and phosphate ion by wash containing the enzyme apyrase.

(Refer Slide Time: 49:05)



So, these are the way these are different reactions as we have discussed, through various like first DNA polymerase, then ATP sulfurylase. And then finally, luciferase where light is emitted and then finally, washing steps are performed in this particular method. Now, this series of reactions is automatically iterated by sequentially using all 4 dNTP'S, and

then repeating the entire process; in this way the sequence of say around 400 DNA fragments can be simultaneously determined to a length of 400 nucleotides, each with an accuracy of 99 percent in 1 say 4 hour run.

(Refer Slide Time: 49:56)

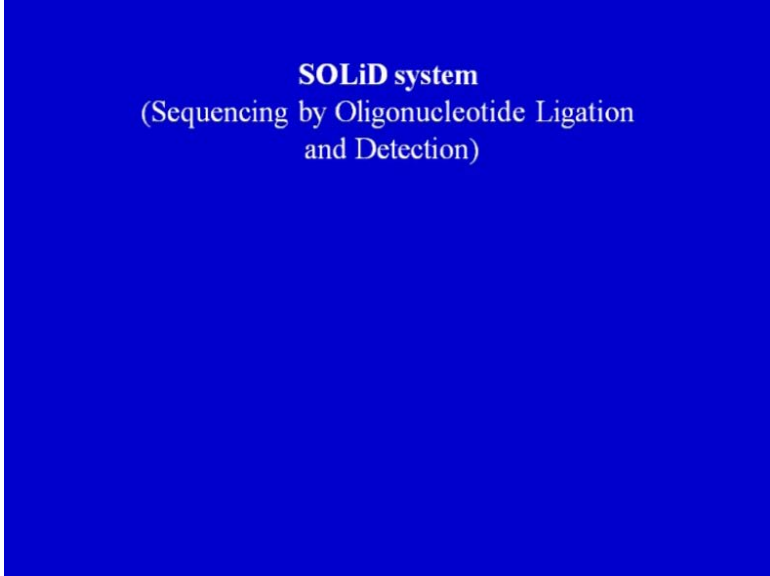
2. Other DNA Sequencing Technologies

The First Next Generation Instruments Genome Analyzer (GA)

So, hence the 454 system is over 300 fold faster than the Sanger method of sequencing, then there other DNA sequencing platforms, are there like first next generation instrument genome analyzer was developed by solex inc, which was later purchased by illumine. It utilizes the strategy for massively parallel DNA sequencing, were by individual DNA fragments with appropriate linkers are first amplified, on a slide matrix similar to microscopic slide using a solid phase PCR process.

The slide with those so called clusters of amplified fragments is then placed on to the genome analyzer, and the 4 bases are added one at a time to the entire slide in a repetitive cycle, then genome analyzer utilizes fluorescently labeled terminators that allow detection of florescent, by a sensitive camera of single base incorporation events into growing DNA strand. The terminator molecules are designed such that, two sequential bases cannot be added in the same reagent additional cycle.

(Refer Slide Time: 51:08)



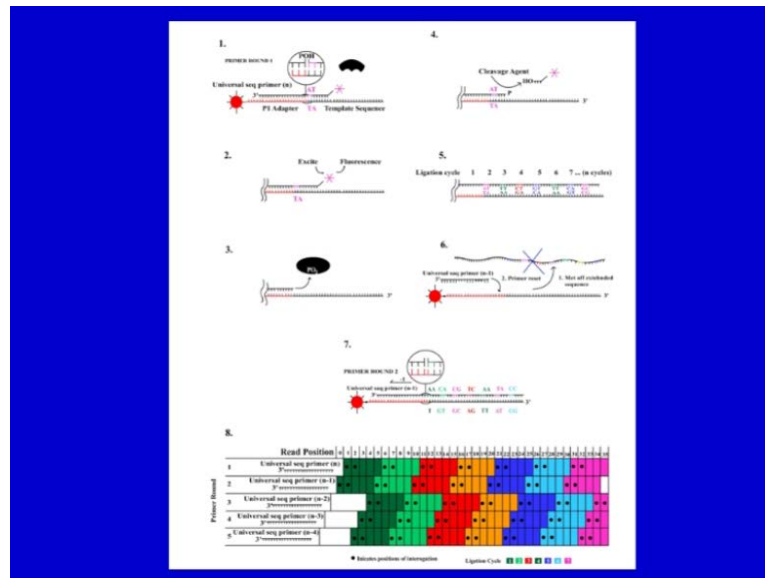
SOLiD system
(Sequencing by Oligonucleotide Ligation
and Detection)

There is another system called sequencing by oligonucleotide ligation and detection, this is from applied biosystem; here this can simultaneously sequenced say 180 million DNA fragments with read lengths up to say 50 nucleotide each for a total of 9 billion nucleotide in a single run. A library of DNA fragments is prepared here from the sample to be sequenced, individual DNA fragments have sequencing primers ligated on to them, and then are amplified using emulsion PCR on tiny beads. Beads with amplify DNA on them are then purified and covalently linked to a solid to a slide surface.

Smaller beads size enable much higher number of simultaneous DNA sequences to be analyzed, then the previous two platforms, the ABI chemistry utilizes sequencing by ligation, in this method fluorescently labeled oligonucleotide probes are ligated to the primer, only if they are perfectly matched to the upstream sequences. This ligated piece of DNA now serves as a primer, and a next labeled probe is ligated to this, if it matches the upstream sequence.

So, significantly higher specificity is obtained and higher accuracy also, then the sequencing by the synthesis approach, so this is how this reaction is carried out, these are like beads here on which it is attached and PCR is carried out.

(Refer Slide Time: 52:45)



And this is the real reaction here, where priming and ligation are two methods, so what is done if you can see on your screen, there is an adaptor and there is a primer against that. Now, this fragment which is the probe you can say, and this is like what you call fluorescence material is in here, so this will be matching and it will as per like it is matching after the primer it will sit in here, and these are various probes in here. So, one which matches here will be complementary then this here gap is filled in by the ligase, so ligation is done after ligation this is excited and fluorescence could be obtained for this.

And then this is removed here an extended strand and other things will be removed, second is the floor containing a few bases is cleaved off and then, these steps from 1 to 4 are repeated. So, what you get if you see here, these 2 bases here which could be which are known or in the probe, then there are other two bases, and other two bases likewise you will get sequence through priming and ligation. That is each time these fragments, different types containing 2 dye, you can say there are two bases probes, dye-based probes you can say.

And these are 6 to 8 base long and when they are attached through ligation and finally, they are removed here could fluorescence could be read, and then these are repeated these steps. Finally, primer reset is done where this whole thing is removed and the primer is taken which is $n - 1$ base pairs of the original primer, and again these steps are repeated and almost like 5 cycles are done, with each time one nucleotide is removed,

let like $n - 1$, $n - 2$ likewise it will be done.

This is summarized in here if you can see, this is the original primer then this $n - 1$ primer, $n - 2$ primer and if you can see in each cycle there are certain overlaps, like these 2 bases were. The first one these two bases here, and then next two bases were again these were the next two bases likewise, there is a overlap here of these sequences. And finally, you will get the full sequence as you go each time you ligation and primary reaction with one short, nucleotide shorten primers and different probes are performed.

(Refer Slide Time: 55:31)



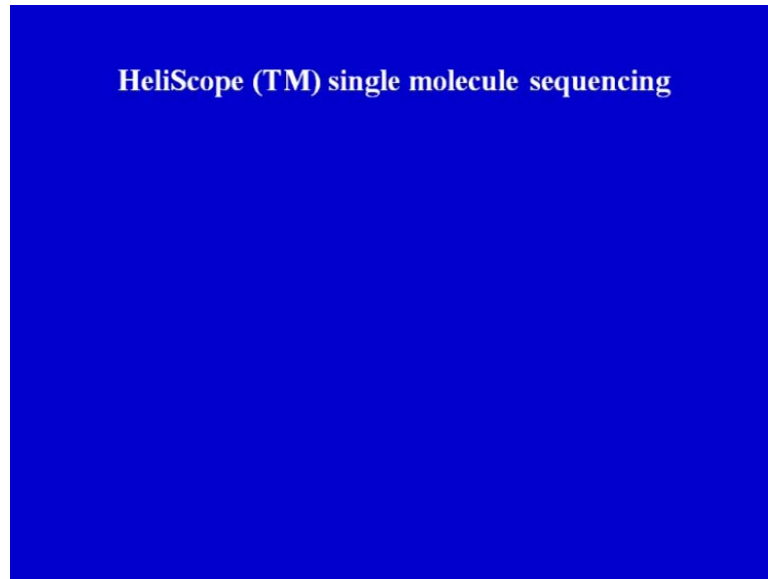
The Second Next Generation Instruments: Single Molecule Sequencers

The second next generation instruments, which is single molecule sequencers the dramatic output of DNA sequence achieved by the first next generation DNA sequences, was obtained by utilizing massively parallel DNA sequencing. And very small reaction volume significantly reduced the reagent cost, but this was previously done; however, all of these platforms depend on PCR to create sufficient mass of the DNA fragments to be analyzed. Now, sequencing of these amplified individual templates causes two problems, one is PCR is relatively expensive and time consuming.

And second is the amplification process is not perfect, and can in rare cases introduce erroneous bases; and in amplified products and these errors are perpetuated in the DNA sequences obtained from these amplified products, and will increase the error rates. So, one solution to both problems is the development of sequencing platforms that can analyze individual single molecules of native DNA, without the need for PCR

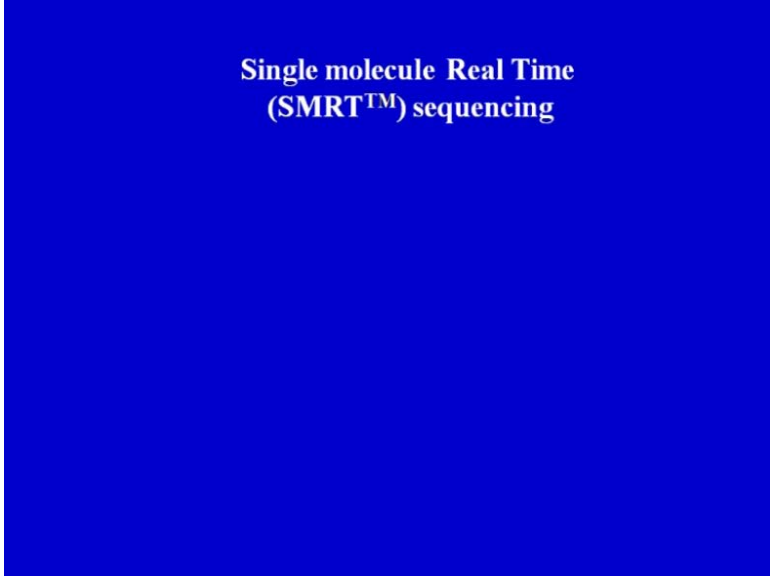
amplification, so this second next generation methods do that.

(Refer Slide Time: 56:57)



And there are two methods we are going to discuss, one is heliscope single molecule sequencing, now the first single molecule sequencing machine was developed by helicobios, this platform uses individual DNA fragments, that have been tailed with poly DA. And then annealed to a slide containing a lawn of oligo DT primers, the slide is done sequentially exposed to each of the 4 bases, and image to determine which DNA fragments have incorporated a specific nucleotide. This machine produces very short sequence reads, but on 100 of millions of templates yielding multiple Gega bases of DNA sequences per run.

(Refer Slide Time: 57:41)



**Single molecule Real Time
(SMRT™) sequencing**

There is another method called single molecule real time sequencing and it is produced by Pacific Biosciences, this sequencing platform tethers a single DNA polymerase at the bottom of an optical chamber; known as a 0 mode waveguide. This waveguide is the structure that creates an eliminated observation volume, that is small enough to observe by laser induced fluorescence. As a single nucleotide of DNA is being incorporated by DNA polymerase, the 4 nucleotides are fluorescently labeled and each as they are incorporated by the DNA they could be, by the DNA polymerase the cleavage of the fluorescent terminator by the polymerase can be detected, prior to the next base being incorporated.

This instrument can obtain sequence information at the processivity rate of DNA polymerase, which is several hundred bases per second. So, these are also called like second or third generation methods, it depends on how you call which you call first generation or you can call these as next generation platforms. So, this completes our lecture, I think one could look into details of the two methods we discussed in the end.

But, we have discussed in this lecture in detail about various methods, like map based methods or shotgun approach likewise, we have discussed Sanger's dideoxy method, and also Maxam, Gilbert very old method. All these methods and new generation platforms have really enhanced the efficiency of sequencing, and the time has been reduced quite a lot for sequencing. This is helping in sequencing many complex genomes of eukaryotes

as well as prokaryotes. And will certainly help in understanding many different questions, which are raised regarding in this particular area of biotechnology; and these sequences will certainly answer them to satisfactory level.

Thank you.