

Applied Environmental Microbiology
Dr. Gargi Singh
Department of Civil Engineering
Indian Institute of Technology, Roorkee

Lecture – 21
Environmental Genomics I

Dear students. So, in this lecture we are going to explore environmental genomics. So, we do know now the basic tenant of microbiology is that the DNA stores the essential information for cell and then it is transcribed into RNA, messenger RNA, which then is taken to certain parts of the cell and it is used as a template for making proteins which is translation or expression of the gene. So, DNA stores information and then it undergoes the cell undergoes various steps so that a protein corresponding to the DNA can be made.

Now, if I want to understand the functional characteristics of the cell, if I want to understand the behavior and the preferences of the cell, then I need to know what it is DNA is. So, genomics is when I can I should look at the all the genes of a cell and then predict the behavior of the cell. So, think of it this way; if you know the code of the computer if you know the software code which was written by some computer algorithmist and scientists. If you know the code you can know what kind of work your computer can do, what kind of behavior to expect from your computer; same is with the cell. If it microbes we can know their genomes, we can guess we can make a better guess at what functions and expressions are to be expected from the microbes.

So, this in when we apply this technology to environmental systems, environmental microbial communities we call it environmental genomics. So, that is what we will be exploring this week and let us see, what we have here today.

So, all boils down into sequencing the DNA, so, once I know the sequence of DNA ATG whatever it is, I can understand what amino acids it would codon for; because, ATG for example, would be transcribed into an RNA where T would be replaced by U. And, then this RNA would go to where the proteins are going to be made and then the protein ATG whatever this codon codes for that particular amino acid will be attached to the amino acid chain. So, if I know the sequence of genes or if I know the sequence of DNA, I can know the sequence of amino acids and if I know the sequence of amino acids, I can do in silico

modeling and understand what its secondary structure would be, what its tertiary structure would be, what its coronary structure would be and thus what its behavior would be.

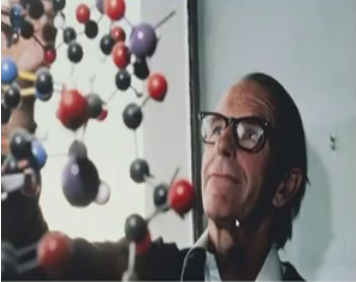
And, usually what we have done in past few decades is, that we have grown microbes in lab, we have tested them under different conditions. So, we know for example, this microbe will eat glucose, but not lactose or we know this microbe is resistant to some carbapenem antibiotic, but not resistant to something else. So, when we know these behaviors we sequence the genome and we can say, if a microbe is perfectly similar to this, if the genome of another microbe is perfectly similar to genome of the one we grew in the lab we know that there perhaps same microbe or they are very closely related to each other and thus we understand we can also predict without even growing the other microbe from the environment, we can predict its behavior.

For example, if I have a material Methicillin-resistant *Staphylococcus aureus* in my lab that I am growing MRSA and I take an environmental sample and I know that there is only one bacteria let us say and I sequence its entire genome and then I match it with the genome of the microbe that I have in my lab the MRSA in my lab and if I see that they are 98 percent similar to each other, very highly similar to each other I can guess that maybe the microbe that I have isolated from the environment is also methicillin-resistant *Staphylococcus aureus* or MRSA or it is very closely related to each other. So, just by looking at the genetic typing I can say that maybe it is a very good chance that my environmental isolate is also resistant to methicillin.

So, how do I sequence that is a major question. If you remember the DNA molecules are very small and they are very tightly coiled to each other they make a very beautiful double helix which is very tightly coiled to each other, it is called supercoiling. Now, sequencing means not only do I need to look at them nucleotide by nucleotide, but I also need to have a mechanism when I am reading nucleotide by nucleotide, a signal is created which can be read by my instrument which can be translated into meaningful data for me, but also should be done rapid enough pace. It makes sense for me; otherwise I might have to wait months and months to get holding of single bacteria. So, the first groundbreaking work was done by Frederick Sanger.

(Refer Slide Time: 05:02)

First generation DNA sequencing: The Sanger Dideoxy Method



- Use DNA synthesis method
- ddNTP to block chain extensions
- Using labelled precursors for detection

Fred Sanger
Nobel Laureate

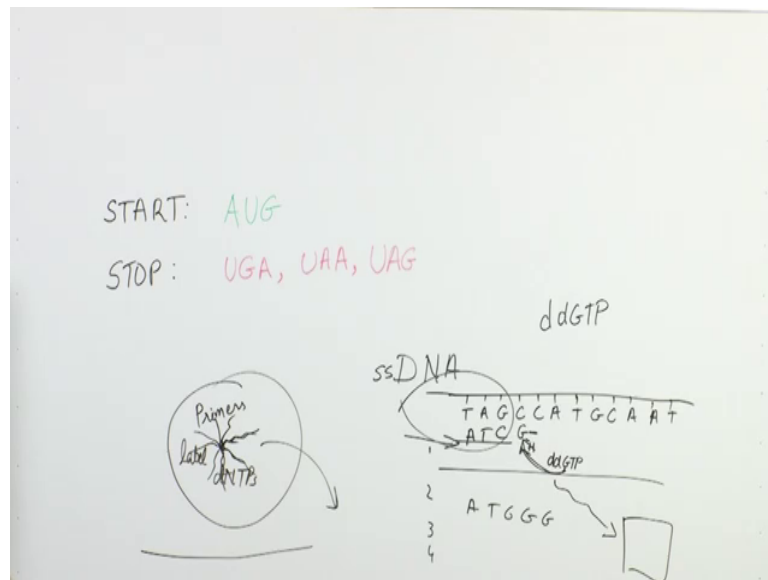
IT ROOKEE NPTEL ONLINE CERTIFICATION COURSE 2

Here is a picture of Frederick Sanger on the left panel and when he made his when he developed his Sanger dideoxy method for sequencing genes, he actually got Nobel prize for it and even though this is the first generation DNA sequencing by the way. Even though now we have second generation, third generation and fourth generation DNA sequencing techniques, the Sanger sequencing is still used from many processes because it is so helpful. One of the advantages of Sanger sequencing is that you can sequence really long fragments of DNA. The other thing is that DNA sequencing technique that Frederick Sanger invented it was based on 3 major principles and those principles are still used by many sequencing technologies right now.

So, the first principle that Sanger worked on proposed for his Sanger sequencing was using the DNA synthesis. So, if I want to read the sequence or something the intuitive approach is to break it down one by one you know break down a nucleotide this is ATP, so, A; this is GTP, so G; this is CTP, so, C. This is one approach to go, but he said, no, not by breaking down, but by synthesis. So, I have a template of like I have a DNA, I denature it now it is single stranded and now I am I have a DNA Taq polymerase which wants to make its replica, more complementary strand. So, I put DATP and I say, A attached; so, the other side should have been T, G attached; so, it should have been C, C attached, so should have been G. So, by synthesis by seeing what is being attached when I am making a complementary strand I can do DNA sequencing, that is the major philosophy of Sanger sequencing.

This next thing beautiful thing that he did was he used ddNTP dideoxy nucleic acids. Now, dideoxy nucleic acids they do not have an OH radical in the third position, but they have an H. So, when they have an H not an OH they cannot continue form a chain when something is added. So, let us say I have it would look something like this.

(Refer Slide Time: 07:12)



So, let us say this is a part of the DNA that I have and it is a single stranded DNA, part of single stranded DNA, let us say this is the direction at which I am sequencing my DNA. Now, remember the first underlying principle of Frederick Sanger sequencing technique was sequence it by DNA synthesis. So, I might have my Taq polymerase here and it wants to do it by synthesis. So, now, let us say these have already been paired up. Now, here I have C. So, when I add now here ideally it should pair up with gene. It will form a very strong hydrogen triple hydrogen bond with G. So, instead of adding ddGTP dideoxy GTP, it has one OH ridiculous. So, when ddGTP comes here and attaches here, it cannot because it has dideoxy instead of OH it has H in third position, it cannot form a further link with the G that is going to attach here. So, the reaction stops here. So, the reaction will stop here.

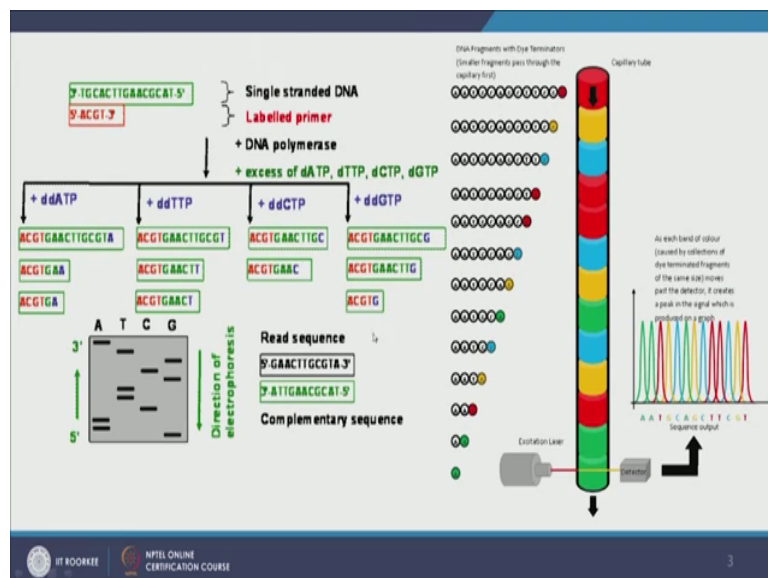
So, this is how the Sanger sequencing works and every step. So, now, the reaction is stopped here we know that G will come. So, now, we have this ATCG. So, we divided it into 4 wells and I will show you how we do that. We divide it into 4 wells, now we already have ATCG there is room for another G to attach. So, when in the well that I put ddGTP again the G will attach here again. So, now, the reaction will terminate. So, every step the reaction terminates.

So, I need to have many wells in which this reaction is carried out. Earlier this reaction was very painstakingly slow because we had there was lot of manual work to do, but now we do Sanger sequencing in automatic robotic instruments, so, this is relatively much faster.

The other thing is every time the G attaches here initially when Fred Sanger worked on it they were radioactive labeled, so, we could catch the radioactive signal and we knew the G that responded at T responded or A it responded or C it responded and we could create a sequence for this single stranded DNA, but now we do not use radioactive material anymore we use fluorescence instead. So, every time some dideoxy nucleotide comes and attaches to it a fluorescent signal is released and this is read by our detector. So, the detector then tells of ATGC and so and so forth.

So, the 3 principles of Fred Sanger sequencing, Sanger dideoxy method for sequencing are used DNA synthesis method instead of DNA breaking method use ddNTP to block chain extension. So, every time a nucleotide is added the reaction terminates right there because you are using dideoxy nucleotides and use label for precursors for detection and as I mentioned in times of Fred Sanger we used to use radioactively labeled precursors, but now we use fluorescent ones.

(Refer Slide Time: 11:00)



This is a picture showing how the sequencing happens. So, I have this sequence here and I add in we divided into 4 wells and in each well we add a particular dideoxy nucleic acids. So,

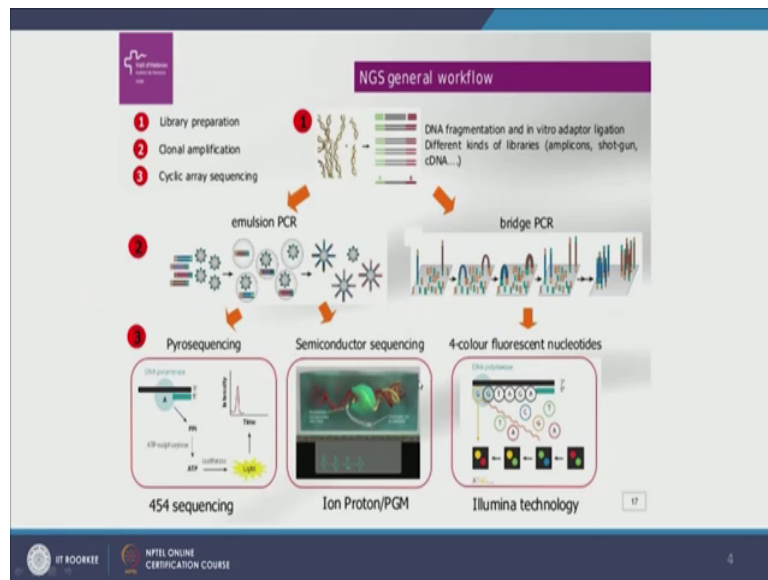
I have ddATP in one, ddTTP in one, ddCTP in one and ddGTP in one. So, if I get a signal from the first one I will call it A signal, second I will call it T signal, third C signal, fourth G signal and so on and so forth. So, remember extension happens 5 prime to 3 prime and then when A is added I get a signal here, when T is added I get a signal here and so on and so forth and then I can read my signal first A was added then T was added then T G A C G C A T and so on and so forth and I can know what my read sequences and I can know what my complementary sequences and this is how the sequencing happens.

So, now, the way it works is my dear is, that let us say this is the fragment that I want to sequence, G A C T G C T G T. So, this is what I want to sequence, this are we attach it to a primer. These primers are short oligonucleotides; oligo means small. So, they have their short length nucleotides polymer and they are in lab attached to the sequence the genes that I or the DNA element that I want a sequence and these are labeled. Earlier, they used to be reductively labeled, now they are fluorescent labeled and then the reaction completes here does not complete here I know what has happened. So, I had a ddATP it terminated here I read in A. So, A is read.

Next step again A is read, next step here C would be read, next step we have T, so, T would be read here and then T would be read here, then we have a G, so, G would be read here, then we have a C, so, C would be read here and then we have a G, so, G would be read here and then we have a read sequence ready. So, it actually looks like this we have die term. Now, what we do is instead of breaking it into 4 different wells and doing it painstakingly slow, this is a small fragment of what it looks like, but we have pictures you can Google them, I encourage you to look it up, pictures of Fred Sanger reading his electrophoresis sheets and they are very long sheets of these blots and then you have to read A D A A G G C C very long time to write it down.

But, now we have a very wonderful method we run them through a capillary tube. So, it is like capillary electrophoresis and they are sequenced with dye terminators. So, when we have A and then its distraction stopped at A, distraction stopped at T A G C and so and so forth and when they are when they pass through the capillary tube there is an excitation laser and it excites these fluorescent Taqs, the terminator Taqs and as they pass we get a signal. So, when we get signal of 2 intensity like twice the green signal we do not green 1, green 2; we get twice the intensity of A or green, so, we know that green stands for A here. So, we have two A, if you get thrice we know there 3 is.

(Refer Slide Time: 14:08)



Now, coming for this is called the first generation sequencing method, first generation DNA sequencing. Now, coming to second generations DNA sequencing; in second generation DNA sequencing the major advancement was that we could multiplex C sequences. So, here we are sequencing a single strand at one go. So, it is very slow, but in second generation we could multiplex them. So, we found ways to make micro tubes, we found ways to make micro wells in each well we could put one fragment of DNA and then we could sequence it and then the robotic arm could reach it could read from each well and know what the sequences. So, this was developed in the second generation sequencing. Now, the advantages of second generation sequencing is that it was the first time and it actually was very near to past not very long time ago, in near past the second generation sequencing techniques have highly reduced the price of sequencing per base pair.

So, now we can get earlier we had to spend lot of money to get Sanger sequencing, but now we get very rapid and very cheap sequences per base pair by second generation sequencing. There are 2 basic platforms in which we use for second generation sequencing; one is emulsion PCR, the others bridge PCR. So, let us talk about both of them, but let us start with the first step. The first step is library preparation. So, in library preparation whatever DNA that you want to sequence now, this could be whole genome. It does not have to be a part of DNA particular gene it could be whole genome and the first step is that we would we shotgun this DNA.

So, we break it into small fragments now each of these fragments are then individually sequenced. And, we usually when we prepare the library, when we break them apart we add certain primers, the adapters, markers that tell us this is the right side, this is what I want to do and when I sequence I know what I am sequencing. So, this is also called shotgun sequencing, because we basically shotgun the whole genome and then we take the right size of amplicon and we sequence that.

So, step one is library preparation this is usually done in vitro. So, we fragment the DNA. So, the DNA we had we fragmented. So, the blue one is a fragmented one, then we add adapters and in this picture we are doing 2 adapters, green and red. They both are important because they both will tell us, when I detect A, when the sequencer detects it knows these are the readings, the following readings I have to take. When it detects B it knows the reading before is what I have to read. So, this is in vitro ligation. There are many ways of doing it; one is amplicon, the other is shotgun, the other cDNA techniques of preparing library.

Now, amplicon sequencing would look like this that we take these fragments and then we clone them in different E.coli and then, we have created multiple copies, you have created amplicon, multiple copies of each of these fragments, now when each of these fragments have been amplified I can then sequence them individually. In shotgun as I mentioned we just break it into fragments, we add I will get the adapter and we are ready to sequence.

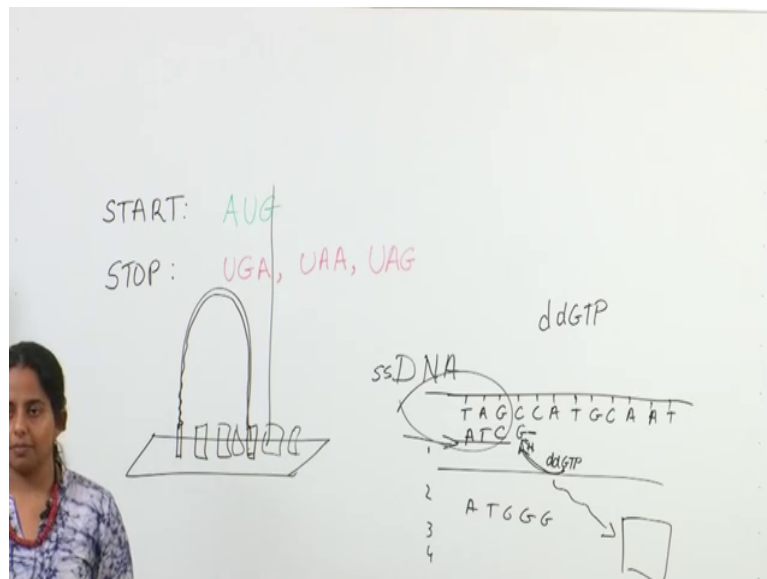
Now, the second step is clonal amplification. We can either do it in clones or we can do immersion PCR or in bridge PCR. They are usually faster than clonal amplification because clones take time to grow and you have to pick so many colonies. In emulsion PCR is like a general PCR, but instead of having a single well where the PCR reaction takes place we create micro oil emulsions. So, these are micro emotions of oil and each of these emulsion they carry all the essential ingredients for PCR. So, they are primers, they have the DNTPs, they have the gene fragment that we have onto sequence and then at the end of emulsion PCR each of this emulsion that had the starting DNA material will have multiple copies of your portion of DNA that you want to sequence.

Now, usually emulsion PCR will carry beads in them. So, the way it works is that we have a bead. So, the way it works is that in emulsion PCR, we have beads. Now, this bead will be a part of your DNA that you want to sequence will attach to this bead, it will be part of your emulsion, micro emulsion which will have the primers you want, will have all the DNTPs

you want, the labels you want and then it will undergo reaction and when the reaction have converts you will notice that you will have multiple copies of this particular DNA fragment and then in the next step you can separate these beads and the sequencer can read them and because it reads a multiple time because their multiple copies it can ensure the accuracy of what you are reading.

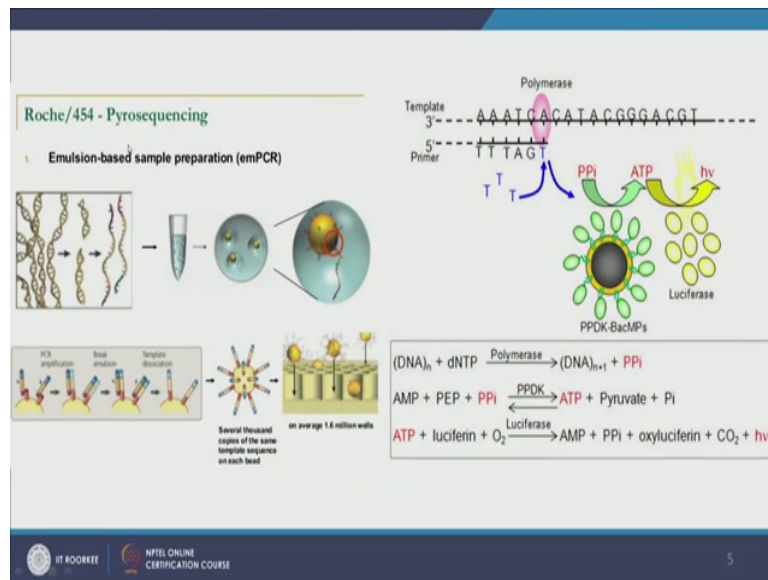
In bridge PCR, instead of doing emulsion PCR what we have is we have a bridge, a plate. Now, the adapters they attach to one end they attach to the bridge to the plate and the other end also attaches and then the DNA Taq will come and sequence them. So, it looks something like this.

(Refer Slide Time: 19:34)



So, this is the plate you have. So, one adapter attaches here and this is your DNA you want to sequence, the other adapter end attaches here. The DNA Taq will come, it will create a, this is single stranded nor this double standard. So, it has sequenced it as it synthesized to it and in the end what you will have is once it comes here this will be released and you have a long sequence chain here which and you can create multiple copies. So, you have multiple adapters here ready to attach to and DNA fragments. So, this is bridge PCR. The emulsion PCR are used by the iron proton sequencing technique by 4 by 4 pyrosequencing and by illumina sequencing and I think I will be briefly going through them to tell you what are the differences in pyrosequencing, semiconductor sequencing and fluorescent nucleotide based illumina sequencing.

(Refer Slide Time: 20:34).



So, let us look at pyrosequencing. Pyrosequencing first step is library preparation and after library has been prepared we do emulsion base sample PCR. So, first we had the genetic material, we split it into small pieces. Now, they were converted into single stranded and look they have blue and purple adapters. So, these libraries will then be distributed in emulsion. So, these are emulsion, each immersion has bead and a genetic fragment. Then, they undergo PCR amplification and when they undergo PCR amplification they will create another strand of genetic material, then they dissociate with each other and now we have 2 strands. You see how the dissociation happens in denaturing step and at the end of it we will have several thousand copies of this genetic material sticking with the bead. Now, each of these bead is then robotically put into a micro well. So, these are millions of wells. So, here we have on average 1.6 million wells. So, they are put in millions of wells where they are sequenced.

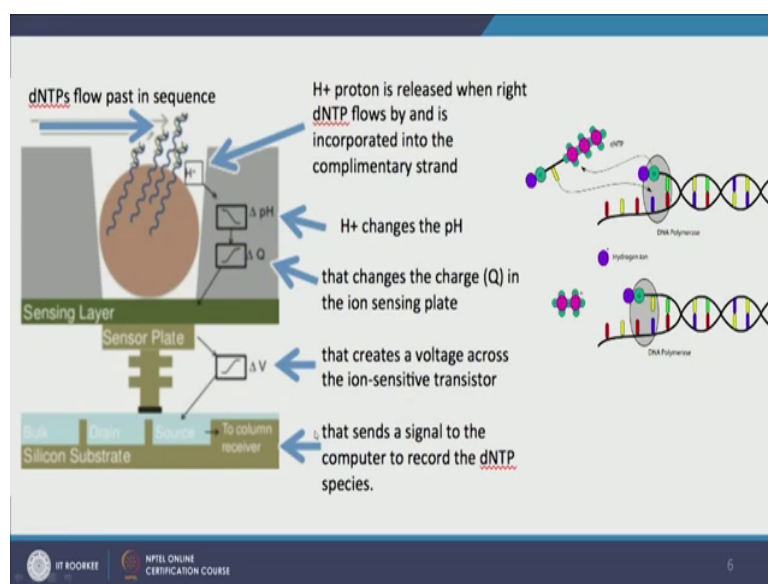
Now, the way the technology of pyrosequencing works this way. So, once they go into these wells, you have your DNA and now another step of amplification will happen. When this step of amplification happens, when the ddNTP attaches, again this is Sanger it uses the same DNA synthesis principle of Sanger sequencing, then it attach, let us say DTP is attaching, when T attaches it releases a pyrophosphate. This pyrophosphate is consumed by AMP to make ATP and when ATP is formed ATP is consumed by luciferase, which is an enzyme, when this enzyme comes in touch with ATP it consumes a tip and releases light and this light is seen. So, let us say I am another well I add G, but I do not see any light. So, I know that no

pyrophosphate it was released as G was not accepted. So, then I will add T. When I add T, I noticed light, observed light. So, I know that pyrophosphate was released hence T was accepted.

So, I know that now the next nucleotide that just has been added is T. Thus in the template I have A. So, now, this way I have sequence this pyrosequencing. Obviously, you can guess, the name is derived from pyrophosphate that is released every time a nucleotide is added. So, basically we have DNA, we are dNTP polymerase and it reduce pyrophosphate attaches with AMP and PEP forms ATP and pyruvate and it will, this ATP attaches with luciferin, I get oxidized in presence of luciferase enzyme and releases light. Now, luciferase enzyme is the same enzyme that allows fireflies and many deep sea microbes and deep sea mammals to release light. So, this is pyrosequencing for you.

Now, this is really need to summarize we prepare the library. We have emulsion PCR. Now, at the end of emulsion PCR the beginning of emulsion PCR ideally in each emulsion we have all the things we need for PCR we have a bead and there is one genetic strand attached to the bead. With multiple steps we have several thousands of the same genetic elements. Now, replica of these attached to the bead. Each bead goes to a micro valve and in this way we generate millions of sequences because there are 1.6 million valves. So, we will generate millions of sequences per run.

(Refer Slide Time: 23:54)



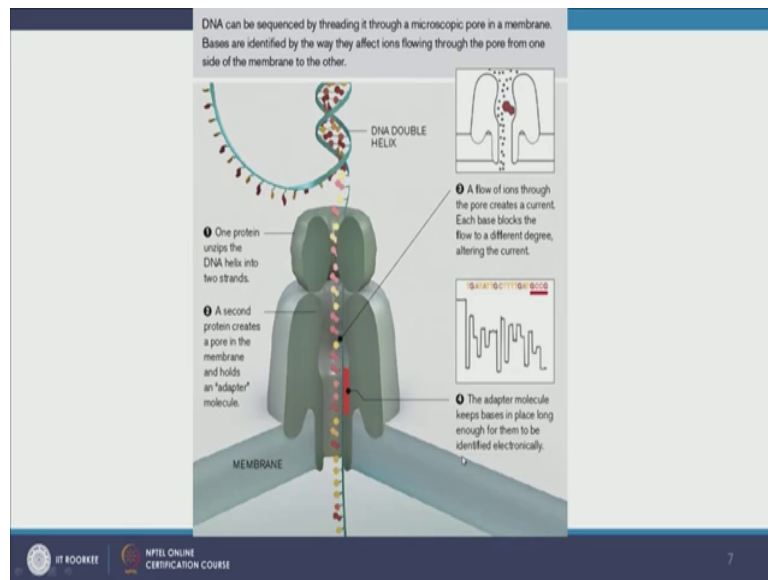
So, here you notice that the dNTPs of, now this is slightly different. Now, I think is part of the third generation sequencing because if you notice here in pyrosequencing we are relying on optical measurement, we are relying on how we can measure the light. So, optical cameras which are so tuned to such small amount of light generated usually by the way this is not so tiny amount of light because remember we have several thousand copies per bead said that when we add dNTP for example, in this case each of these would ideally release some light. So, collectively we will get a considerable signal, but to have such a fine camera and such a small optical recorder or such as for such high throughput sequencing is very expensive.

So, the third generation sequencing techniques are the ones that are past optical detectors. Now, they use pH electro they use electro chemical detectors. So, this is the dNTPs are passing in sequence as the sequencing is happening. The pH changes you know a proton is released that proton is detected. So, if I pass dNTPs the dNTPs. So, I know the T was accepted if a proton is released.

So, in fact, ion torrent sequencing which uses this third generation sequencing method is they are they dub their sensor as the world's smallest pH meter. So, proton is very easy to detect pH is very easy to detect compared to the optical signal. This is a advantage of the third generation sequencing technique and ion torrent is one of them. I am actually looking forward to working with some ion torrent sequencing in near future. So, by the time you read this lecture, you hear this lecture, I have probably already worked with ion torrent sequencing. So, when the pH is measured we know that following nucleotide is has been accepted and here it is how it works, let us read through this because it is really well written.

So, a proton is released when right dNTP flows by and is incorporated into the complementary strand. This proton released changes the pH because remember pH is minus log concentration of H^+ which changes the charge in the ion sensing plate. So, ion torrent, here the ions and same plate and it changes the charge which create creates a voltage across the ion sensitive transistor that sends a signal to computer to record the dNTP species.

(Refer Slide Time: 26:18)



Again, this is very high throughput, this is third generation sequencing. So, remember first generation sequencing we have Sanger sequencing; second generation sequencing we have we have multiplexed it, so, we have high amount, we are creating high amount of base pair data for very short, small amount of cost in very little time; third generation sequencing we have passed the optical measurements and we are looking at the electron measurements or proton measurements. So, this electrochemical sensing makes things cheaper. So, ion torrents sequencing is much more cheaper than the optical base sequencing techniques.

In fourth generation sequencing and this is the example of nanopore sequencing. It is very different. In nanopore sequencing for example, we have this protein across a membrane. So, this is a transmembrane protein and what happens is that the DNA double helix is in a way denatured, uncoiled and one of the single strand passes through the protein and as it passes through the protein factor it is excreted through the protein. When it passes, it changes the electrical signal and the signal depends on what is passing through it. So, let us read through this so that you understand this more clearly and when the signal is captured we know what has passed through.

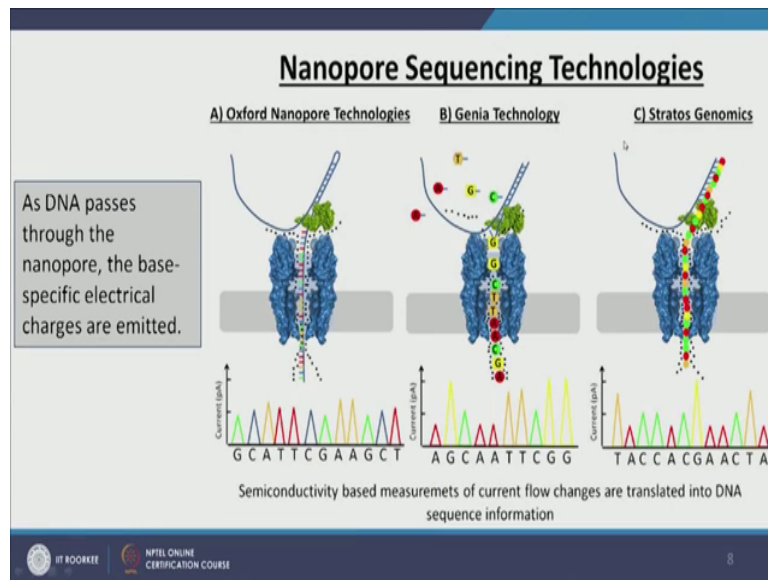
The advantage of nanopore sequencing and fourth generation sequencing is that we can generate data for longer genetic fragments. For example, ion torrent it does really good job around 200 to 300 base pair, pyrosequencing will go to 400 to 500 maybe even 600 base pair,

Sanger sequencing can go from anywhere from 500 to 2000 base pair, but the nanopore sequencing can do a much better job.

So, this is an example of by the way Oxford nanopore based sequencing. So, DNA can be sequenced by threading it through a microscopic pore in a membrane. Phases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other. So, we have a DNA double helix uncoiled passing through here. So, one protein will unzip the DNA helix into 2 strands. It will reject one strand and pull the other strand in. The second protein will create a pore in the membrane and it holds an adapter molecule. Now, the ions are also flowing through this. As the ions are flowing through this a current happens; current is basically flow of electrons or flow of ions and each base as it passes will block the flow. Now, ATGC all of them block the flow to different degrees. By looking at this impedance if blockage and flow we know what was present.

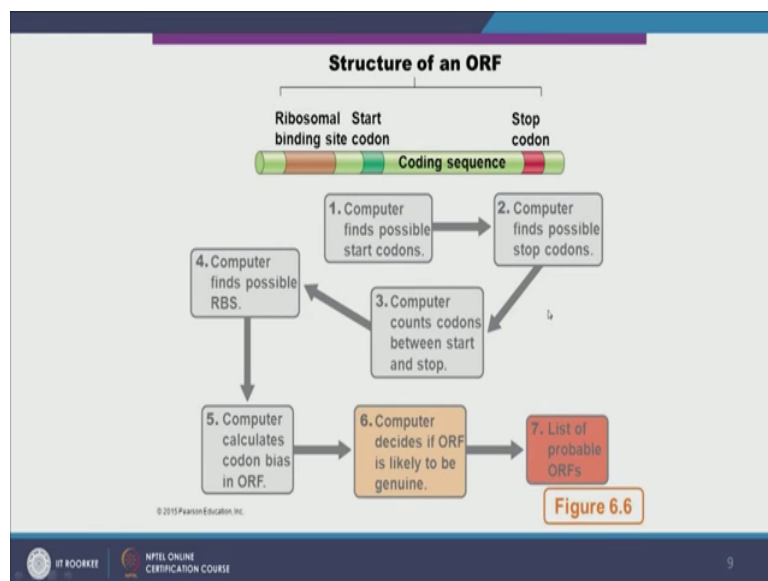
So, if you look at this spectrograph here we notice T has the highest block, G blocks this much. So, every time I notice this amount of blocking, I notice G. Every time I notice this amount of blocking I know it is T and what the adapter molecule does is, it keeps the basis it checks their speed of it is secretion or excretion. So, it makes sure they pass slowly that we get a signal. This is nanopore sequencing and last 2 years the company has done a very good show of trying to promote it and the beauty of (Refer Time: 29:21) generation sequencing is how portable the instruments are. So, ion torrent for example, it requires an air conditioned room, it requires to sit on the right bench and it needs to be maintained in the right way; however, nanopore sequencing is the size of a flashlight. So, you have a flashlight you just carry nanopore sequencing with you, put your sample here and it can be so fast, you can sequence the entire human genome in few hours.

(Refer Slide Time: 29:49)



So, this is nanopore sequencing for you and there are 3 different technologies right now in market; one is the Oxford nanopore technology whose example I was showing you here and then there is Genia technology and Stratos genomics. So, in all of them all of them use semi conductivity based measurements to know how current is flowing through the pore and then they translate that into DNA sequence, but there are slight difference in each of them.

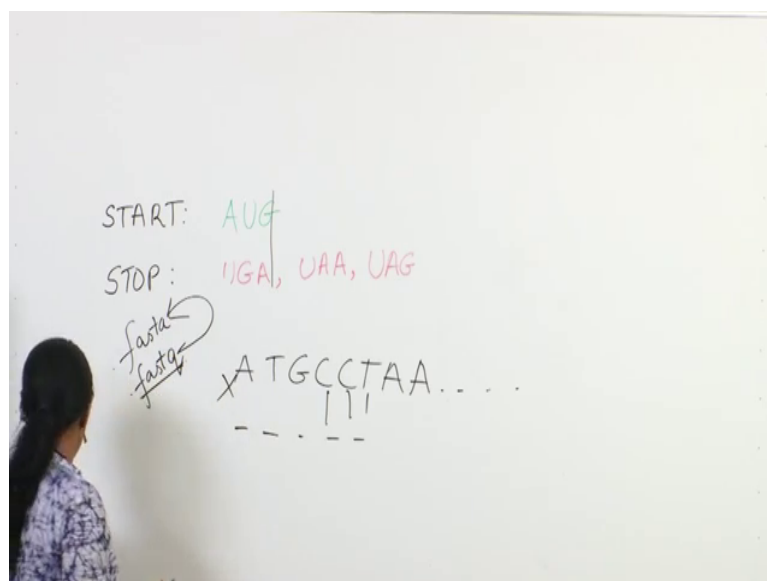
(Refer Slide Time: 30:15)



Now, once we have generated the sequence, the next challenge is how we are going to analyze the sequence. I have a long sequence ATGCGCTTA whatever, but I do not know how to make sense out of it. So, what is the difference between GCATTCGAAAGCT is it even meaningful data or is it just noise or just meaningless data? Because, we notice that in a cell in it is DNA it carries informational genetic codes and it carries non informational genetic codes and just by looking at it there is no way for us to know whether it is informational or non informational.

So, for example, remember I talked about start codon and stop codon. So, between start codon and stop codon ideally we have an ORF Open Reading Frame. So, we have genetic material, informational material and near start codon there should be ribosomal attachment site binding site. So, in the ribosome will attach there and then it will allow translation of the MRNA that will be formed by this ORF, but between multiple ORFs there are still sequences and they do not code for anything, they are nonsense in this way. So, how do we tell them? We tell them by feeding the sequence of data that we have generated here this and typically they are in form of fast queue file or fasta file. So, the sequence that will generate either from the first generation Sanger sequencing, second generation illumina sequencing, third generation ion torrent sequencing, or fourth generation nanopore sequencing would look like this and so on and so forth.

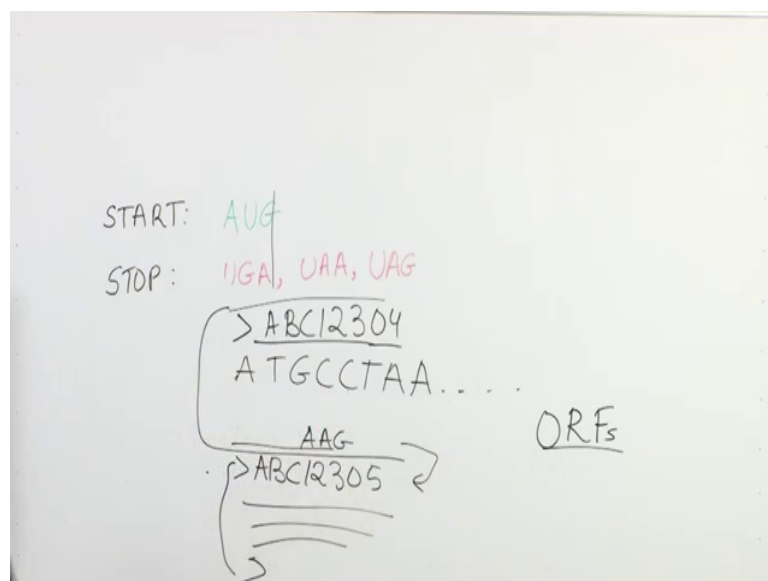
(Refer Slide Time: 31:52)



And, the file the output file that you will get will either be in the form of fasta, fasta file or dot fasta or dot fastq. The difference between fastq and fasta file is that the fastq also carries the quality information. So, not only will it know ATGCCTAA and so on so forth the information, but it will also have quality scores for each of the base pair. So, if the quality score is less than I can ignore this, saying that this is poor quality this might just be a chimera or this might be a mislead, let us ignore, let us reject the sequence.

So, in most of the sequences will give you fastq file and then you can put fastq file in your computer and you can ask your computer to eliminate the poor quality reads and then you are left with a good quality fastq file. Now, in order to analyze we the computer does not need quality scores. Once you have already eliminated poor quality score nucleotides and reads, you do not need the quality information. So, most programmers will convert this into fasta file.

(Refer Slide Time: 33:20)

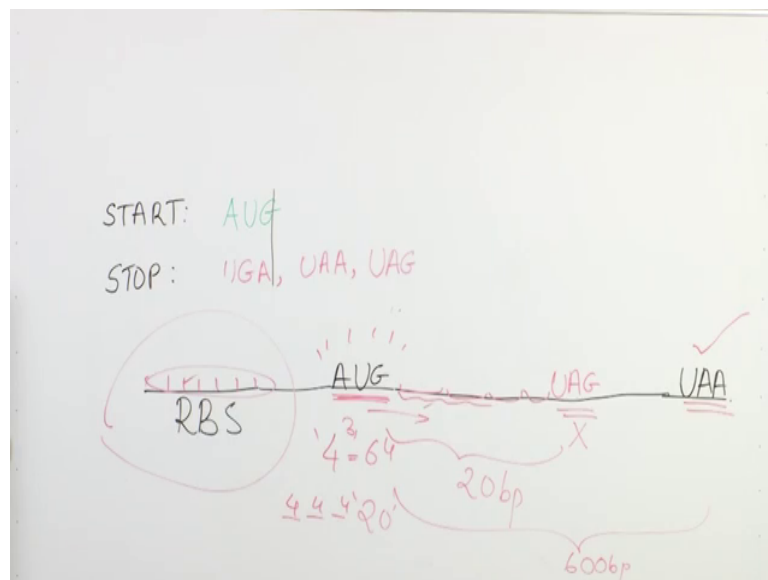


Now, let us take a look how a typical fasta file will look. A typical fasta file will have a mark like this, it will have the name of this read. So, it could be ABC12304 something like this and then in the next line you will have the sequence. It might be many lines of sequence and then when the sequence ends, for example, you will have another line and you will have and then in next line you have view more sequences. So, in the one of the home works for this week what you will have is you will get a fasta file and I will ask you to annotate it.

So, this is your fasta file and in fact another homework that I am giving you is I will give you 2-3 different sequences and I will ask you to create a fasta file. So, even if I have only one sequence, let us say Sanger sequencing, I am doing and I only have one sequence. This is still a fasta file or standalone fasta file. I can combine 2 fasta file by just writing the components of the second fasta file below and then in the next line here, the next line of where the sequence of the first faster file ended. So, I can combine 2 fasta files. Now, once the computer gets a fasta file, it needs to identify what are called as ORFs – Open Reading Frames. So, once it has identified open reading frames only then it can go ahead and try to make sense out of the sequence.

Now, how does it identify open reading frames let us take a look. Now, in order to identify open reading frame, the computer undergoes a particular algorithm and it looks like this; computer finds possible start codons, then computer finds possible stop codons, then computer counts codons between start and stop codons, then computer finds possible ribosomal binding sites. It calculates the codon bias in operational in the tentative ORF and then it decides if the ORF is genuine or not and then it creates a list of probable ORFs. So, let us take a look.

(Refer Slide Time: 35:23)



So, let us say this is one of the sequences and I have left the places where I do not want to read the sequence and blank. So, the computer the first step for computer would be it will go

through all the sequences and look for start codon. There is only one start codon that we use microbes use and that is AUG.

So, the computer will go through and it will look for all ATGs, all AUGs and when it has identified the start codons, the next step it will do is it will look if there is a stop codon after the start codon. So, it will search for stop codons, now the stop there 3 stop codons UGA, UAA and UAG. So, you know why it went for start codons first, because this will minimize the number of sequences that it needs to analyze. So, then it will look for stop codons.

So, the second step is stop codon. Once it has identified the stop codon and you know there might be multiple stop codons, let us say UAA is here UAG is here. So, there are 2 stop codons after the start codon in this particular sequence. So, the computer identified 2 stop codons after the start codon. The third step it does is it calculates how many base pairs are between start and stop codon. Let us say, this is 20 base pair, let us say this is 600 base pair which is more likely to be the actual ORF, the actual open reading frame that carries information. 20 base pair there is no gene no protein that is so small, so, very less likely that this will be an actual ORF. So, the computer will reject this as a stop codon. It will reject this that this is not a right stop codon either there is a sequencing error or it is just not a right ORF.

Now, if it looks at UAA this is 600 base pairs far away. So, there are some 200 amino acids which is a decent size for a protein, 200 amino acid protein and thus it will say this might be my right stop codon. In the next step what computer does, once it has found a start codon part is found the right stop codon it has confirmed that the base pairs between start and stop make sense. It will look for ribosomal binding site. So, it has a particular sequence and this sequence if it matches the sequence for ribosomal binding there is the ribosome can come in sit and bind ribosome also has some sequences then it will say I have identified an ORF.

The next thing, what the next thing what the computer does it does before it confirms that this is an ORF. So, see computer has a lot of data, millions of reads and it is trying to reduce the number of data to meaningful data, sensible data. So, first those who do not have start codons reject them, those who do not have start and stop both reject them, those who do not have stop at the right position reject them, those who have start and stop and at the right positions with right amount of base pairs in between, but do not have RBS reject them. Those who have all of them, let us find out the codon bias.

So, in biology we have code remember codons. So, 3 nucleotides together the code for an amino acid this is a codon. So, some there are some codons that code for a singular amino acid. So, because we have 4 different kinds of nucleotides and codons are based on 3 nucleotides. So, in first position we can have 4, second we can have 4, third we can have 4. So, basically we can have 4 to power 3 codons sorry 4 to power 3 codons. So, 64 codons, but we only have 20 some amino acids. So, you can see on average each amino acid is coded by 3 codons so, but each bacteria, each microbe has a preference. I like to use this particular codon for coding for this amino not the other 2 or not the other 3 or other 4. So, this codon bias, they have a codon preference.

So, what this computer will do is then it will find out what codons am I reading in pairs of 3 and are these the preferred codons of rest of the microbe if you compares with other sequences it has got from this micro or not. If it is very different then it will reject this because the microbial preference for codon or the codon bias in a microbe is specific to the entire microbe. Now, in some cases; now, please note this very clearly, in some cases let us it is possible that this is an actual ORF. It actually codes for sensible data and it is expressed into a protein, but it has a distinct codon bias from rest of the genome. In that case the computer can say maybe it is possible that this particular gene, this particular component of genetic element was not originally of the microbe, but the microbe incorporated it from its environment or from other microbes, other bacteria, other protozoans. So, this is horizontal gene transfer and we will talk about it soon.

(Refer Slide Time: 41:03)



So, this is how our computer decides structure of an ORF and the next step here is now I have the ORF, this is my unknown genome by the way and the next generation sequencing machine made small fragments of it, then identified the right ORFs and once it has identified the right ORFs this is what I am left with. And, then we have genome assembly software. So, these are assemblers; these assemblers will what they will do is they will align these with each other. So, all the ORFs are aligned to each other and whenever there are significant overlaps, we say yeah, look there are overlaps, so, these are overhangs. So, because they are overhangs they were initially part of the same genetic material, but then there it was split from different places and then it is sequenced. So, it will reconstruct genome. Often we cannot reconstruct the entire genome, but we can convert these sequences into scaffolds which are longer strands and the more meaningful thus and at times very rarely we can also reconstruct the whole genome, but reconstructing the whole genome is much more expensive because they are often there are gaps that are left and filling these gaps is very intensive and expensive process.

So, dear students, this is all for today. In the next class we will go ahead and we will to explore more about environmental genomics, more about the tools that are used for environmental genomics and are actually and I must mention that this is still our growing field right now. In fact, it is one of the latest things. It is called as bioinformatics making

sense of the information from biology and in the next class we will talk about them. That is all for today.

Thank you.