

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

Lecture – 22
Environmental Genomics II

Dear students, in the previous lecture on environmental genomics I taught you about 4 different generations of sequencing techniques. The first generation sequencing Sanger technique developed by Fred Sanger, Nobel laureate, the second generation sequencing techniques will be multiplexed and thus reduce the net cost of sequencing a base pair, the third generation sequencing technique where we moved from optical measurements to electrochemical measurements. The ion torrent for example, uses what is called as the world's smallest pH probe, pH meter to sequence or to get the signal when we are sequencing the genetic element and then the fourth generation sequencing technique of nanopore sequencing where the DNA sequence as it passes through a nanopore, a very small pore in a protein, in a membrane and then I went ahead and told you about ORFs. How ORFs are identified by computers and then how computers proceed forward to analyze the genes, to analyze what is present.

Now, in order to understand how computers move ahead after they have identified the ORFs and after they have assembled, we briefly talked about assembly. The next step usually is alignment. So, in environmental genomics we do not need to fill in the gaps as we need to do if you were working with pure microbiological questions or pure biotechnological questions we just need to understand what genes are present and what sequence they are present, so, we will get an idea of what sequence they are translated in and if they are triggered at the same time, if they are up regulated down regulated together and what these fancy up regulation down regulation terms are.

So, for applied environmental microbiological questions we do not need to fill in the gaps in the genome. If we get the genetic information, the sequences we want to assemble them. So, they are more meaningful. So, we get multiple genes in a single scaffold. So, that is the assembly part which I was talking about last time.

So, let us start from there. So, in assembly we use different processors, different algorithms, some of them look at the similarity in different sequences look at the overlaps and how

frequently the overlaps happen. If there is one overlap that happens only one time it might probably just be a coincidence, but if there is another overlap that happens thousands of times in a pool of genetic sequences we might assume that they were together and then they were split from different places and then they were amplified and sequenced. So, probably we can attach these 2 together we can attach the 2 genetic sequences together and we talked about it.

The other option in when it comes to assembly is de novo sequencing. The other option in sequencing is we look at algorithms that will predict what the next nucleotide should be. So, there is some really good algorithms that can actually predict the next nucleotide and we ask them to predict, what should follow g, what should follow let us say a should follow g, what should follow a, c, t. So, when they predict this we look for where in this pool of genetic sequences are we seeing same sequence that it is predicting and if it is we find it and we notice that there is also some overlap there is some statistical significance and it makes sense to combine assemble them together then they are assemble together into a long sequence.

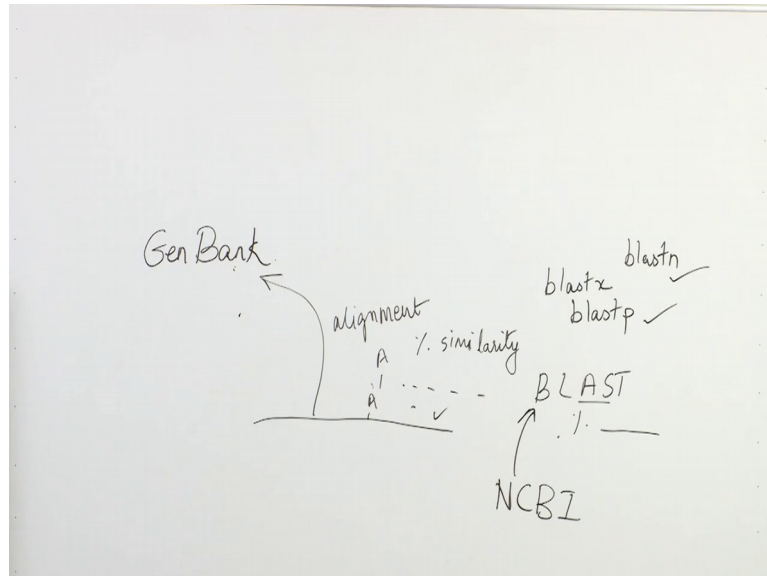
So, remember most generation most sequencing techniques will give you anywhere from 60 base pairs to 600 base pairs may be in turn in case of Sanger up to 2000 base pair, but then you need to assemble them and make really long scaffolds. So, once the assembly has been done, the next step for an environmental microbiologist is to align it with the given databases. So, what people have done very painstakingly over last few decades is that they grew microbes in the lab, they cultured them and they understood their behavior, they sequenced them and they made libraries of their sequence and they what they found out was they found each ORF in those complete sequences, they express the protein specific to the ORF and they found out the nature of the protein.

So, they know this particular ORF encodes for protein that helps in lactose metabolism. This helps in that, that helps, so, they found this help this makes efflux pumps which get rid of heavy metals in the cell; the heavy metals in the environment that find a way inside it effluxes them out. So, for each ORF or each protein they found out the sense. So, a lot of it is already known for microbes that have been well cultured in life for many years now.

So, all this information is collectively and very nicely arranged together in databases and some of them are very well curated databases. So, we are very assured of their quality. Also nowadays it is required that if you sequence something you need to submit your sequence to

GenBank. So, I will write these name down name down here because I will write these names here because they are very important.

(Refer Slide Time: 05:18)



So, you have GenBank and as the name suggests GenBank is a bank of all genetic sequences. So, if you sequence something to publish your work or to communicate it with public and with other scientist you are required to submit your sequences in GenBank and they have a very streamlined process for different kind of sequences and I one of the home works were actually required going through the GenBank looking at their submission processes for different kinds of sequences trying to understand them and answering some of the questions. So, this is GenBank for you. So, after you have.

So, now, what people can do is when you submit your data and GenBank you also submit metadata. For example, you will inform what kind of sample is it? Is it an environmental lake sample or is it something you grew in the culture. It is something you isolated from a petroleum reservoir or it is something that you was growing on in your bathroom tile. So, you give some information.

So, if I sequence something and I assemble a very long sequence and let us say it matches really well. So, this is called alignment. So, I do alignment what alignment means is that base pair by base pair. So, base pair by base pair the sequences are aligned the one that I am submitting or the one I have recently sequenced with the one that is already present in some

database or GenBank or other similar bank of genetic sequences. And, then they will calculate percent similarities and let us say. So, then we can calculate percent similarity let us say it is very similar to something that has been submitted to GenBank and was found in a groundwater aquifer which had perchlorate contamination.

So, then I know that this sequence was found in groundwater aquifer that had perchlorate sequence and maybe there is then I will get some information about what I can expect about the sequence that I just sequenced. Let us say I got this sequence from a groundwater aquifer, but I do not know this particulate contamination or not. But, I noticed that every time people submitted a similar sequence to GenBank it was informed perchlorate contaminated soils and aquifers, then I can say there is a very high probability or there is at least some good probability that this well from where I got this aquifer from where I got my sequence also has perchlorate contamination.

So, the GenBank serves a very nice purpose. It gives us a very bird's eye view idea of whereas such sequences have been found. However, GenBank is not very well curated. Its purpose is not alignment like this; its purpose is to just collect all the sequences to maintain transparency in research and to prepare a repository of earth microbiome. So, earth microbiome very briefly is a project where we are trying to sequence everything that is alive on earth, every sequence that is available on earth, whether it is a alive bacterial, eukaryote, archaea, higher order of life like ourselves. We have been fully sequenced by the way or whether it is the sub alive, sub dead viral particles. So, everything that can be sequenced they want to sequence and they want to make a bank out of it.

So, we know what the genetic code of life on earth is. There is a whole purpose of GenBank. So, this alignment we do it by data bases that are there whose purpose is to align and to give information. One of the most common one is blast. Another homework from the in this week is based on blast, where I will give you the fasta file. First you have to combine the fasta file and then you have to blast it which means you have to align it and then you align it will give you an idea of the percent similarities with different submissions. What the submissions are in many cases people are submitting from pure cultures and they know this is clostridium thermocellum. So, if it has a very good match with clostridium thermocellum, you know it is probably close to clostridium thermocellum or something very similar. So, this way you can give an identity to your sequence or you can get some idea about your sequence.

Let us say, your sequence; now, in blasting we can do in 2 ways we can do blast x, we can do blast p and you can do blast n. So, let us look at blast n and blast p because they are 2 most widely used commands in blasts. So, coming I want to take a seg way back to the blast in general. You can align your sequences using blast repository either online on their web page if you have limited number of sequences, but if you have created lot of sequences by the wonderful new generation next generation sequencing techniques then the recommendation is you download the database the blast NCBI data, by the way yeah the blast database is maintained by NCBI. So, you can download the NCBI blast database and then do the alignment locally on your server on your computer.

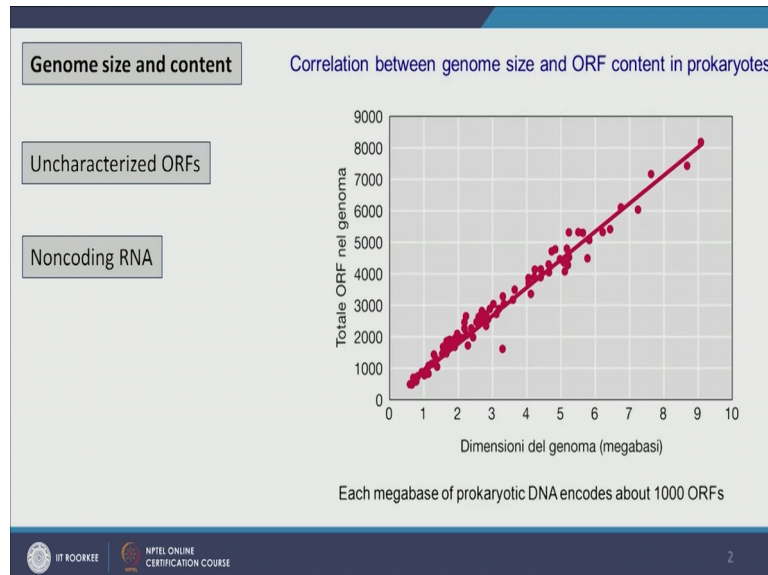
So, blast n and blast p; blast n you look for alignment in the nucleotide. So, a for a, t for t g for g, c for c. In blast p you do a protein query. So, you take your ORF you can translate it into protein in silico; obviously, and then you look for protein matches. This is blast p is very helpful when we have degeneracy. So, multiple codons coding for a single amino acid we know that codon bias. In degeneracy, what happens is that it does not matter if a codon changes here and there not, but the amino acid would be same. So, the blast n will tell you the similarity in the nucleotide it is not very much, but if you do blast p you can notice that the protein similarity is very high. So, that is blast p and blast n for you.

And, now move on to what we notice in our real life samples when you are trying to do alignment and annotation. I should mention this it alignment is looking for similarities which we do for blast and notation is when we try to give it a taxonomic identity. So, we have some really well curated databases and like green genes for example, we can use RDP – ribosomal database project, we can use green genes and silva, we can use these databases for annotating our sequences after we have aligned. So, sometimes in NCBI plus we aligned, but we cannot annotate because many of the sequences entries in NCBI blast are unknown bacterium from gulf of Mexico sand, unknown bacterium from arctic. So, we cannot annotate, we cannot give it a name, but green genes for example, silva for example, RDP they maintain a very well curated repository of known classified and well annotated microbes.

Now, when we sequence and we create multiple not multiple a large amount of sequence data, we run into an issue that, are the sequencing deep enough, which is are we capturing the entire diversity, are we capturing all kind of ORFs that are present in the sample or not. The other possibility is that I capture a lot of ORFs, but the ORFs I do not have the information

about them. So, I captured the ORF, but I do not know what kind of protein it encodes for. It is also possible that it is a nonsensical ORF.

(Refer Slide Time: 13:08)



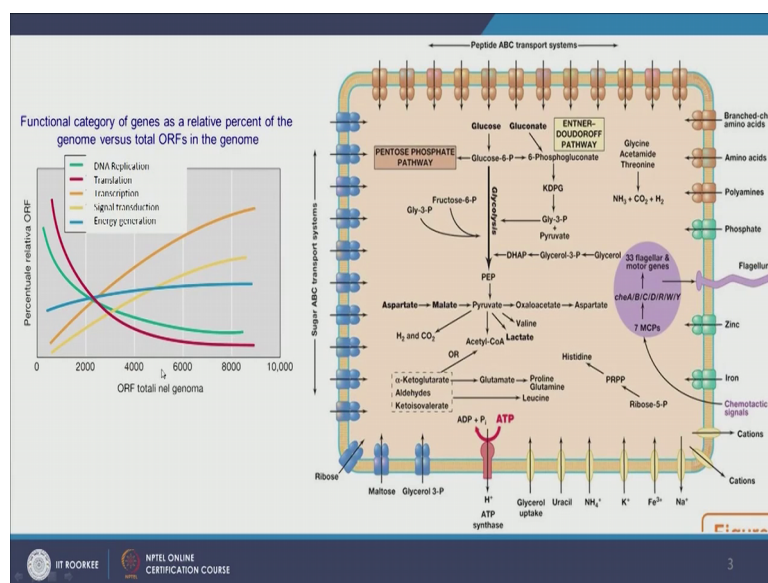
So, let us look here one of the things about number of ORFs that will capture from a sequencing is what is the size of the genome that we are sequencing. For example, if we have an organism that is 3 mega base pair is we have genome has 3 mega base pairs then we can predicts less than 300 total ORFs, but if we are sequencing something that has more than 8 mega base pair size of genome then we should expect up to 7000 ORFs. So, that as a genome size increases the number ORF increases. However, after some time we notice there is not the case. For example, here there are very few members that have very high base pair and have very high number of ORF.

Now, in this the other problem is uncharacterized ORF. The alignment and annotation work on the principle that if the genetic sequence is the same or the protein amino acid sequence is the same in case of blast p, then the function will also be same. I t works on this assumption. So, if let us say characterized genome from a bacterial genome and I characterize in ORF and I decided this ORF is a very important for SOS system – Save Our Soul signal. So, this is when the bacteria undergoes a panic mode and it turns on certain genes to increase it is survival rate.

Now, let us say in a new mammalian species the same ORF is found. So, we can assume that it has same function in a eukaryotic system, just because their nucleotide sequence or their amino acid sequence is highly similar. But, many a times speed as I mentioned earlier we do not have the information about the ORFs we are noticing especially this is a trouble now, because we are generating such high amount of genetic sequence and we are characterizing so many ORFs. We do not know, we do not have complete information about many of them. The other problem is we find an ORF, but it is a non coding RNA. A very classic example would be 16S rRNA gene. Now, 16S rRNA gene is a ribosomal nucleic acid obviously, but ribosomal RNA, but it is a non coding, in sense that it is not code for a protein that is also possible.

The other thing we have to understand is, that let us say I am doing the whole genome sequencing of a particular microbe and I am interested in genes that trigger lactose degradation, but most of the genes in the bacteria are not related to lactose degradation most of them will be housekeeping genes that are used for day to day business or that are used for making efflux pumps, now that are used for replication. So, the my genes of interest are very small proportion usually of the entire genome. Thus, another question arises do we really need to do whole genome sequencing or do can we just do amplicon sequencing and figure what we need to figure out.

(Refer Slide Time: 16:11)

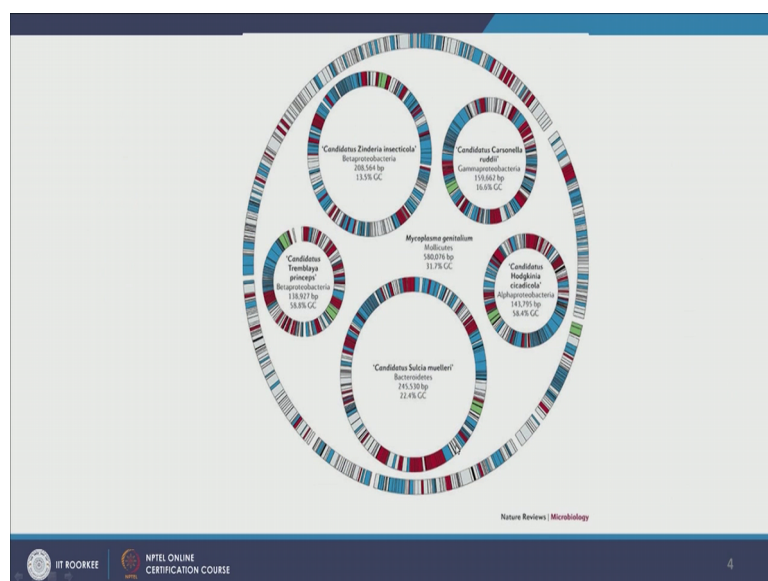


So, on the left here we have a panel that is showing as the number of ORFs increases. How number of ORFs related to DNA replication, translation, transcription, signal transduction and energy generation changes. So, for smaller ORF per genome we have most of the ORFs dedicated for translation, few for transcription and for DNA replication. Those cells say my priority is that I should be able to replicate and whatever mRNA I make I should be able to translate it very quickly. So, one single copy of mRNA should make enough proteins. I do not need to make multiple mRNA. So, ORFs for transcription are smaller, same is for signal transduction and energy in regeneration remains constant because all of them no matter how big the genome is, how many ORFs they have, they require everybody knows energy is important.

Now, here we have a very nice, very beautiful pathway, pentose phosphate pathway showing how glucose undergoes different degradation steps to give energy, to create energy carriers in the cell. Now, this is a very complex system, but this is a very simple schematic of even more complex biochemical reaction chain and this is to give you an idea that at each step there are multiple enzymes involved. Each of these enzymes are encoded by certain ORFs and thus just for this pentose phosphate pathway we have multiple ORFs that we will have.

So, in fact, one of the limitations we talked about when we talk about metagenomics the whole sample genomics in environmental engineering is there are functional genes of interest a very small proportion of the entire genome. Most of it is this housekeeping genes.

(Refer Slide Time: 17:52)



And, now the other thing is we are talking about the size of the gene. So, the size of gene varies, size of genome varies. So, in the inside here we have different kinds of bacteria gammaproteobacteria, molecules betaproteobacteria, vector IDTs and alphaproteobacteria. So, look even among the proteobacteria we notice that the size of genome changes. These 2 are better proteobacteria, but the size changes and then this the big one is the mycoplasma and look how big the genome size is. So, then we have a larger genome size will have more ORFs and many of them would probably be uncharacterized if we are lucky enough they are characterized already, if not we have to do it ourselves and the other thing is that we have very small proportion of functional genes.


Now, let us when you talk about functional genes let us move to functional genomics. As environmental engineers environmental scientists environmental students or students of environmental science we are very interested in functional genomics. Because, it really does not matter for us a lot how if all the pentose phosphate how many copies a pentose phosphate system are present how many ORFs related to pentose phosphate system are present in the microbe, but it is very important for us to know; for example, is this microbe resistant to heavy metals, is this microbe resistance to mercury methyl mercury or is it resistant to some anti microbials, antibiotics. So, that kind of information is more important. So, we go into functional genomics, we are more interested in the function.


(Refer Slide Time: 19:35)

Functional Genomics

Omics terminology	
DNA	Genome
	Metagenome
	Epigenome
	Methylome
RNA	Transcriptome
Protein	Proteome
	Translatome
	Interactome

Metabolites	Metabolome
	Glycome
Organisms	Microbiome
	Virome
	Mycobiome

 IIT ROORKEE

 NPTEL ONLINE
CERTIFICATION COURSE

5

Now, before I go into functional genomics let us look at the omics terminology. So, whether it is metagenomics or genomics we are talking about omics and all these have suffixes as omics. So, let us look at them one by one.

When we are sequencing DNA we have 4 different kinds of omics; genomics, metagenomics, epigenomics, and methylomics. So, genomics is for a particular cell let us say *Clostridium thermocellum* strain 1 2 3, something like that. When I sequence each and every order of each and every nucleotide in its entire genetic material it is called genomics. When I take an environmental microbial community and I sequence all the genes all ORFs all the nucleotides present in that environmental community it is called metagenomics. So, metagenomics would be, for example, all the bacteria are sequenced, all the archaea are sequenced. Now, epigenomics, where all the epigenomes are sequenced. Methylome genomics when I am looking at all methylated DNA.

Now, when it comes to RNA sequencing I can also do RNA sequencing and RNA sequencing is very useful, think of it this way. DNA is the potential to make a protein, but DNA does not necessarily say that protein is being made because it is possible that the gene that makes a particular protein is down regulated and not allowed to express or it is damaged and it will never express or it until it gets its signal it will not be expressed. So, just because the DNA is present does not mean that cell carries out the activity. For example, a particular cell might have the gene required for degrading beta lactamase by antibiotics, but it is not expressing them. So, it would not have resistance to beta lactamase based antibiotics through that mechanism degrading them.

So, if you look at RNA we know that which genes are being up regulated. Which genes are actively being transcribed into mRNA. So, that is why RNA sequencing is very important for us and when we sequence all RNA present in the microbe it is called transcriptomics. Transcriptomics is more informative than DNA, analyses like genomics or metagenomics, but there is a big challenge here. In transcriptomics, one of the biggest challenges that most RNA are short lived and then RNA vary in their longevity too. Some RNA will destroy faster some mRNA will live longer. So, as time proceeds you know the delay in our sequencing proceeds between sampling and actually sequencing it, we will have a very different picture maybe than what was actually present when we sampled it. So, there are some limitations with transcriptomics and there are some ways out we add some regions to prolong the life of

RNA, we store RNA at minus 80 to prevent its degradation, but if successfully done transcriptomics is very promising.

Now, let us look at protein. Protein, we look at proteomics, translomics and interactomics. Now, proteomics is I look at all the proteins in the system and I sequence all the amino acids, this is proteomics. If I do proteomics analysis for all microbes present in an environment it is called meta proteomics and this is very tricky because it is very hard to isolate protein, purify them and then do meta proteomics on them a proteomics and then translomics is when I look at all the proteins that are freshly translated. So, what is being translated right now, what activities are being done right now?

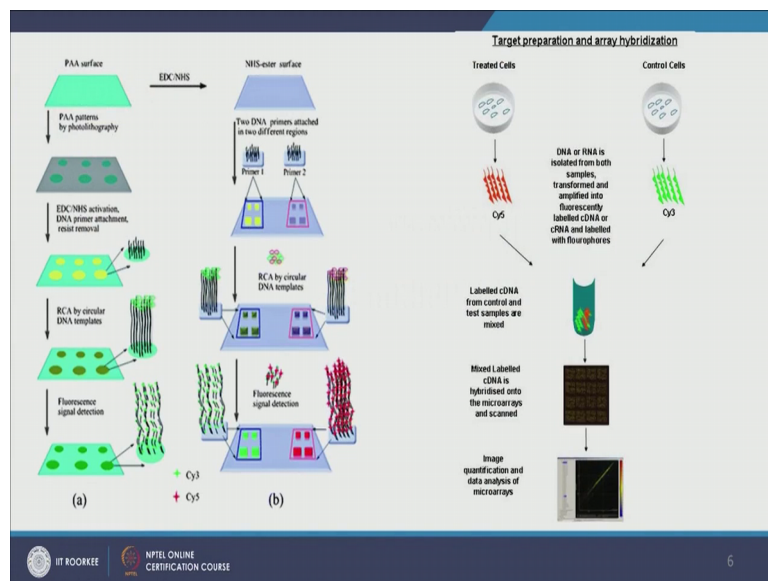
Interactomics is when I look at the interactions between the proteins RNA and DNA. So, interactomics actually includes the DNA analysis, the RNA analysis and the protein analysis and this is very important when I am interested in regulation because some of the regulation is carried by pro enzymes. So, there are some enzyme inhibitors that I do not allow other enzymes other proteins to do their job sometimes it is RNA molecules, sometimes it is other molecules that actually up regulated or down regulate proteins, genes. So, interactomics looks at the in the interaction between all 3.

Then, we have metabolites. So, metabolites would be for example, acetate, glucose. So, when I look at these metabolites, there are the breakdown products or products of anabolism or catabolism and I characterize them completely it is called metabolomics. Now, metabolomics is very important for environmental engineers and biotechnologists. Think about it this way, I want to make butyric acid out of cow dung. I definitely need to do need to do metabolomics to understand usually what are the different degradation products of cow dung, what microbes are involved in it, what are the pathways that chemical pathways they are going through and then I can try to tweak them and to make sure that they make what I want them to make. So, this is why metabolomics is very important and then we have glycomics when we are only looking at the glucose. So, we are characterizing glucose in a microbe when it is making how much it making etcetera at any given time that is glycomics.

Now, when I am looking at organisms from a higher picture I have microbiome, virome and mycobiome. So, microbiome is when I am characterizing not just when I am characterizing microbes from different domains. So, 3 domains prokaryotes, eukaryotes, and archaea. So, when I am characterizing all of them, when I am characterizing all the bacteria, all the

archaea and all the eukaryotes including higher order of life like fishes and algae, then it is called microbiome. So, for example, I often talk about drinking water microbiome in my classes. So, what I am referring to is all the organisms present in drinking water it could be protozoa, they could be virus, they could be bacteria, they could be algae, they could be anything. Virome is when I am looking at all the viral particles present in my sample and microbiome is when I am looking at all the fungal particles present in my sample.

(Refer Slide Time: 25:27)



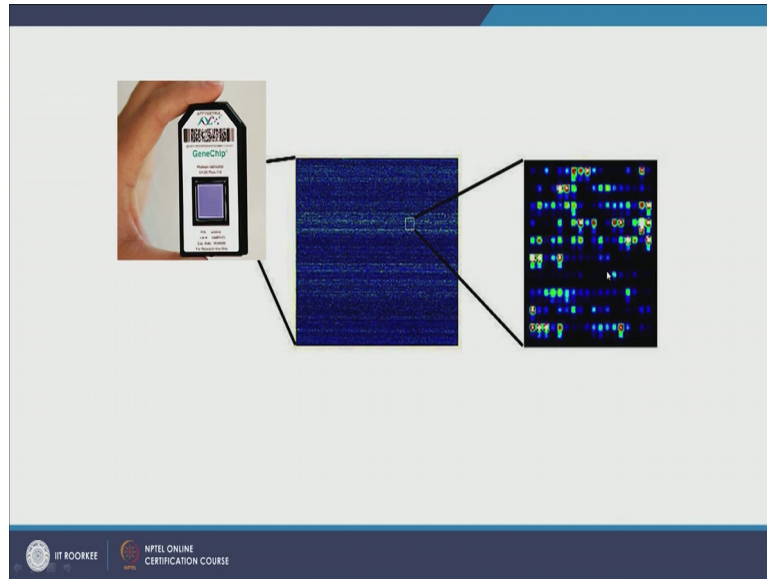
Alright dear friends, now we have a wonderful method of getting an idea of expression. So, I am interested in which gene is getting expressed or not. So, this is microarray preparation and I will very briefly go through this. So, in microarray preparation we do not need to sequence, but what we already know is we have certain genes of interest, when these genes are transcribed they make RNA and the RNA how they look like we know. So, we take a chip like this is the PAA surface, we use photolithography to etch them, activate them. We attach these small primers to them and we allow synthesizing of the RNA that are of interest and when I attach, I can in a same singular plate what I can do is I can attach multiple primers. So, I can have multiple templates. Now, when I flow my sample through it and they will the complementary strain would attach with these RNA strains and I will know which what is being transcribed, what is being expressed, what is being translated by getting a signal. And, there are many ways of doing it, how the microarray system undergo (Refer Time: 26:28), but basically it works on DNA hybridization.

So, for example, let us say this is a DNA that I am interested in and I know it is perfect sequence. So, I can have a single stranded DNA of the right sequence and this is the other environmental DNA that I have isolated and I want to know if it is this DNA of my interest. So, what I can do is, I can allow them to interact and if they are perfect complement to each other they will hybridize. So, they will form hydrogen bonds and they will form one double stranded DNA and when this is formed if my single stranded was fluorescent probe when they form bonded double stranded DNA, the fluorescent probe will be released and the fluorescence would be measured. So, this is how microarray system works.

Now, in my microarray what I can do is, I can make these genes of interest, multiple genes of interest and I can attach them ligate them to my surface and as I flow my sample whatever has the perfect complement will attach and the corresponding fluorescent signal would be released. So, then I can read my chip and I can say already here we saw signal, here we saw signal. So, in this sample these expressions were happening, these RNA molecules were present.

So, on the right panel you have treated cells, on the left and here you have control cells. So, DNA or RNA is isolated from both samples it is transformed and amplified into fluorescently labeled cDNA. So, if it is RNA it will be converted into fluorescently label cDNA, if it is DNA just fluorescently labeled DNA and then they have label force as the name suggest, then they are from the control and this is hybridization by the way in the right panel. The controlled and the treated cells are mixed with each other. When they hybridize they are hybridized on microarray plates and they are scanned wherever I get a signal I know hybridization happen. So, the match was found.

(Refer Slide Time: 28:26)

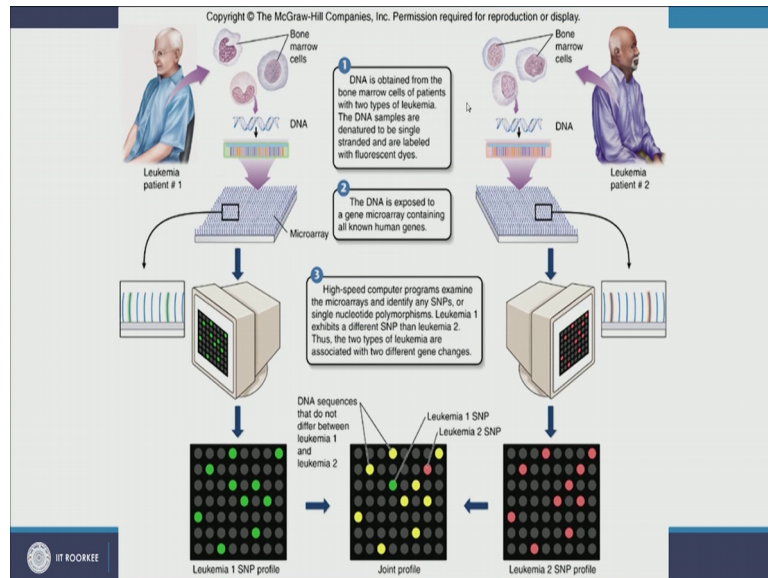


Now, this is a very beautiful technology it is a gene chip technology and I think it is very important to understand this. Where we are headed now is to a highly personalized medical system. We know certain diseases such as cancer for example, does not manifest similarly into people. The immune system of 2 different cancer patients will behave very differently, when it is trying to body is trying to feed the cancer this how cancer works and when we add cancer drugs to kill the cancer cells, the body finds ways to protect the cancer cells because it treats cancer cells as it is own very important cells. So, when this happens, the genetic sequence of the cancer cells mutates and undergoes and different changes and these changes determine the resistance of cancer cells to any anti cancer drugs.

So, in gene chip what we can do is we will have the genome of the entire human put here. So, I have the genome of entire human put here and then I can read the genome and I know what the what is going on in the human cells, what is going on in the cancer cell, what RNA is being translated, what gene is being transcribed, expressed. So, in this gene chip I will have thousands and thousands of complementary RNA molecules attached, very small and when I flow the sample over it, so this is gene chip, when I flow the samples over it, a human sample, blood sample, stool sample, fluid sample whatever and then the hybridization would happen the perfect compliment would hybridize and release a fluorescent signal like this. So, this is a blown up image of a very small portion of this. So, wherever I get a signal the computer can read it and say these RNAs are being expressed. So, if I know some filthy act

fishy activity is happening here I will know exactly what the resistance mechanism is there of this particular cancer patient.

(Refer Slide Time: 30:20)



So, let us look at this particular example from the McGraw Hill company book. This is there 2 leukemia patients, leukemia patient 1 and leukemia patient 2. I take their bone marrow cells, I take the DNA out of it and I put it on my microarray which have these thousands of complementary not thousands of oligonucleotides that might be complemented with this DNA and wherever the hybridization happens the computer checks. Similarly, here in leukemia patient 2, I can again take his samples; his fluid bone marrow cells put them on the same chip and notice what his signals look like. So, this is the SNP profile.

Now, as the leukemia is progressing their mutation is happening in the genes. So, what I am getting mutation usually start with single nuclide well there many ways of mutation. One of the ones that is presented here is single nucleotide polymorphism, where a single nucleotide changes. So, wherever there is a single nucleotide change there you have a green signal here you have a red signal. So, what you can do is you can find out the similarities. In both leukemia 1 and leukemia 2 patient these yellow ones are similar and the green and red ones are unique. So, here you can get an idea of how leukemia progresses overall. If you do it enough patience you can also get an idea of how leukemia expresses itself very differently for each of the patient.

So, my dear students, I will end it here and in the next lecture we will go ahead and we will talk more about environmental genomics that is all for today.

Thank you.