

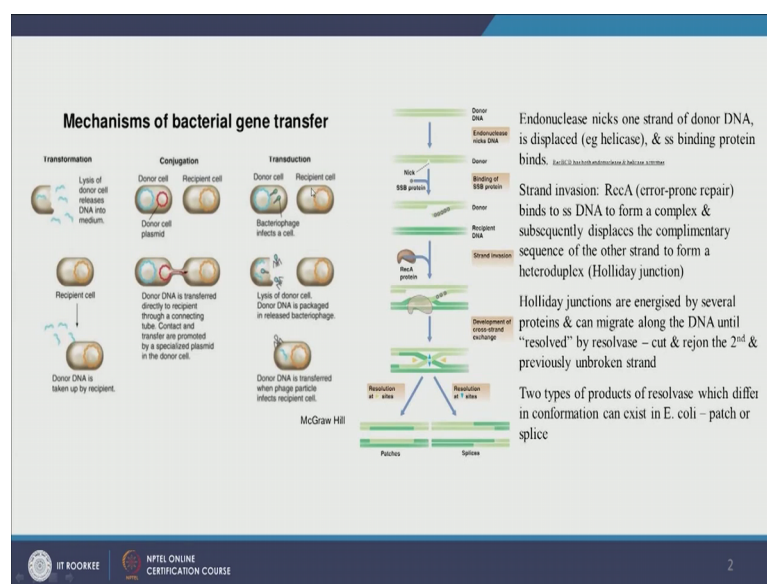
Applied Environmental Microbiology
Dr. Gargi Singh
Department of Civil Engineering
Indian Institute of Technology, Roorkee

Lecture – 25
Environmental Genomics V

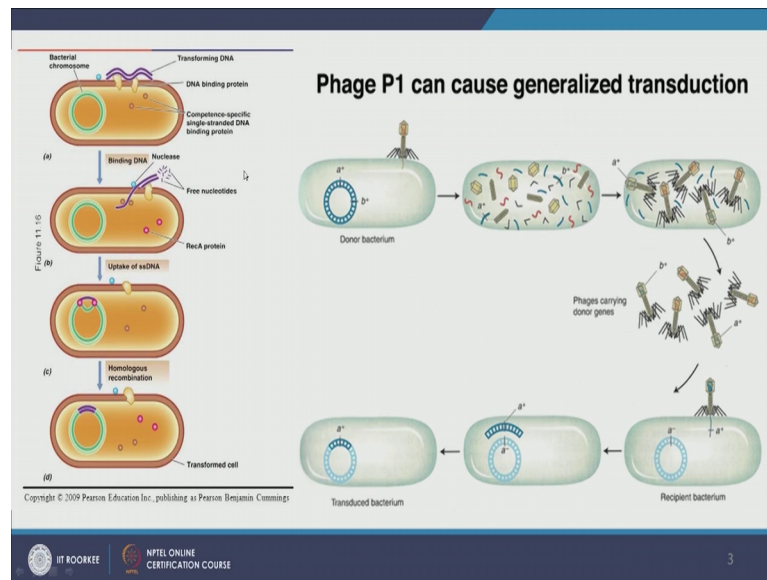
Dear student, in the previous lecture, we talked about how genetic evolution happens in the cell whether it is vertically, horizontally or via recombination. Today, I am going to revise two major ways of horizontal gene transfer just so that you are clear we are sort of talk enough about transformation for you to understand, the basic of transformation not the nitty-gritty microbiological detail, but enough for you to understand how to apply it in environmental problems.

And then after that we will go in and see how we make a taxonomy trees and have a annotate, how we make sense sort of a sequences and how we make sense of the f genetic evolution, how it helps us in making trees. And even addressing the primary question of biology how did life start. So, if you make trees ideally we can trace back evolution to the first ancestor which we called as last universal common ancestor Luca. So, the how we make a trees how we make sense out of biological data that is why we are going together today. Let us get started, all right.

(Refer Slide Time: 01:25)



(Refer Slide Time: 01:27)

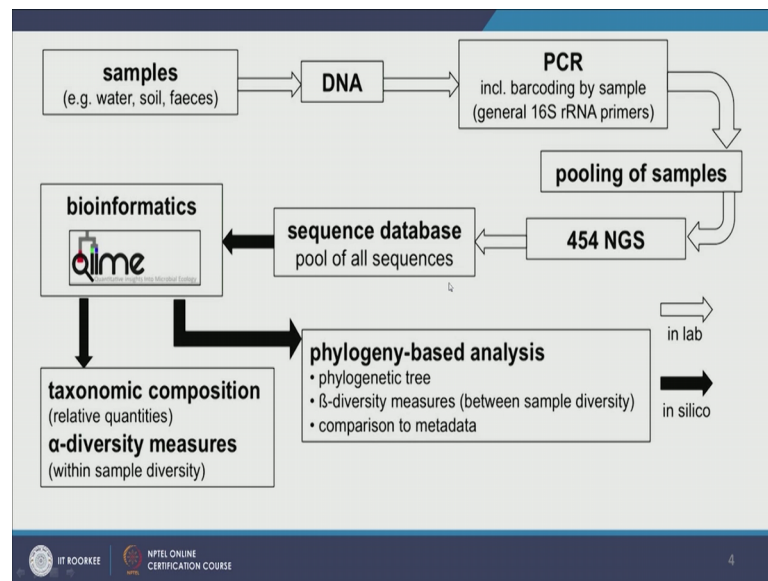


So, let look at transformation and conjugation in details. So, in transformation, we have this DNA it is lying outside the cell we have DNA binding protein on the cell membrane which binds which sticks the extracellular DNA to the microwave cell and then a nuclear when one of the protein will allow, the extracellular DNA to enter in a form of single strand here inside the cell. So, one it is nucleated, so one strand is destroyed outside and notice RecA a protein is present here. So, the RecA will cause recombination to happen here and next thing you know you have recombination that has occurred. It is a homologous recombination because the entire extracellular DNA as been incorporated and this is your transformation typical transformation that is used in conventional cloning techniques, and it is also very important in terms of antimicrobial resistance.

Now, let us look at transduction. Usually when a bacteriophage attaches to a virus cell that bacteriophage viruses that attacks a cell, it secretes its DNA or RNA into the cell. Forces the it hijacks the cell forces it to make multiple copies of its own proteins and then we have multiple virus left, they lies a cell and they come out. Now, at times what happens is that some bacteriophage when they attack a cell not only do they force a cell to make its protein and its structure, its RNA, DNA, but at times they also pick up the DNA of cell fragments of DNA of the cell. So, these phages that are that are made after this cells was lysed and high after hijacking and lysing this particular cell, and now have the fragments of its DNA.

Now, at times when they attack another cell which is the recipient they can inject these fragments of DNA, which can undergo homologous recombination into the chromosome and then become the part of the bacterial chromosome thus this is transduced bacterium. So, transduction is also very common in RecA and bacteria both and it is very very important when we are studying horizontal gene transfer.

(Refer Slide Time: 03:38)



Now, let us move on and let us try to see how we make sense of all this information, how we can use these techniques and understanding to take a environmental sample and understand what is going on. So, let say you take your sample it is either water sample soil sample fecal matter or some other things, if your environmental engineer like I am you are typically working with either water, fecal matters like or activated sludge or other biomass that is a byproduct of fecal matter or waste or you are working with soils. So, what we do we first, first step is we extract DNA first step is to extract nucleic acid.

In this particular diagram, we are extracting only DNA, but you might also want to extract RNA. So, all nucleic acids depending on what your interest is and then what we do is we use PCR. So, we do polymerase chain reaction, usually we the one that we are most interesting in is 16S rRNA gene albification. So, this is a ribosomal genes 16S rRNA 16S small sub unit rRNA and the I have mentioned previous lectures about what we use 16S rRNA. So, if you are confused go ahead and take a look at those lectures.

And PCR is basically a chemical reaction through which we amplify the gene of interest which 16S rRNA here.

Now, once this gene has been amplified, we pull all the samples. So, how we pull it, this is very very important because this diagrams is not only telling you how we make sense out of data, but is actually giving you an example of the second and third generations sequencing techniques. So, I think it is a good idea for me to go ahead and explain to you more in detail. So, earlier what we would do once we have amplified 16S rRNA, ideally it is from one bacteria not from a pool, environmental pool like it is here.

So, between extraction of DNA and PCR here this another step of cloning. So, we extract DNA, we amplify the PCR we clone it and then we do sequencing on it, but here we are doing different work. What we are doing here is now in this 16S rRNA, we have the 16S rRNA gene from different microbes. We might have it from proteobacteria, we might have it from firmicutes, we might have it from very very different microbes we might even have some 16S rRNA then basically 16S rRNA from different microbes that are present in our sample.

Now, what we do when we usually we have 100 of sample let say I have 100 samples, I can add a barcode to each of the sample. So, barcode is a very short nucleotide gene. So, it is an oligonucleotide; and its sequence acts like an identifier. So, whenever I get a sequence that matches to my barcode, I know which sample it came from. So, a for 100 samples, sample 1, sample 2, sample 3, sample 4, I will have hundred barcodes and each of one barcode will go to one sample, and it will attach itself to the amplicon to the gene that I have amplified within this case is 16S rRNA. So, I amplified by including barcode.

Now, I pull all the samples I can now take the 16S rRNA amplicon from all samples put them in one tube, the reason I can do is because amplicon from each sample has a barcode. So, once a sequence I know which sample it came from. And then this is 454 NGS which is basically 454 pyrosequencing, I have talked about it in one of the previous lectures. So, go ahead and take a look what 454 next generation sequencing is.

So, after doing pyrosequencing, I create first two files, the first two files I pull all the data I do bio informatics on it. So, I find a quality score, I find out first I will check for camera, which are artificial sequence that is not really there. And I look I do different kinds of analysis. And now what I can do it, I can annotate, I can align the sequences that

I have which a good quality sequences. And by the way in bioinformatics portion, we in bioinformatics portion, we also get rid of the barcodes.

So, we get rid of the barcodes and we instead of writing barcode atgc whatever the barcode is we write sample one, these are the sequences sample; two these are the thousands of sequences or millions of sequences. Now, all the sequences can be aligned to the database that we have well established, well theoretical databases. And then for each sequence, you can get some taxonomic information, this sequence matches proteobacteria, betaproteobacteria within proteobacteria. This probably is betaproteobacteria sometimes some might get even more information. So, not only do I know it is firmicutes, but also no it is clostridium within firmicutes. I might also needs clostridium similar to clostridium thermocellum. So, now I know all right (Refer Time: 08:00) thermocellum may be cellulose degradation. So, this kind of information I get from after aligning and annotating my sequences.

Now, using the taxonomic composition I can do two kind of analysis. I can do alpha diversity analysis; I can see how diverse a sample is within itself. Let say I take water sample; I get 100 sequences from water sample. So, maybe in 100 sequences, I have let us say 30 unique sequences. So, basically there are 30 species that are present and there are 100 that I have sequence in water. And I get 100 sequence from fecal matter, but instead of 30 unique sequences I have 85 unique sequences. So, I can say the alpha diversity of water in fecal matter such as that fecal matter is more diverse than so water. So, this is a alpha diversity. I can also look at beta diversity when I want to know ok.

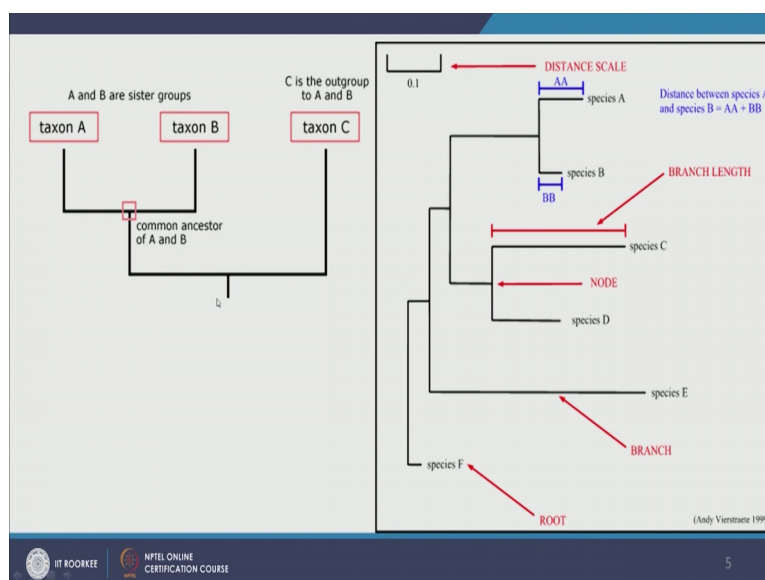
Fecal matter of termite verses fecal matter of human being and I want to know which is more diverse. So, I want to locate beta diversity. I want to look at water in the influent of wastewater treatment plant, and water in the effluent, and then look at the diversity right. So, these kind of intra sample analysis are known as beta diversity analysis. And what I can do is not so for example, I am and comparing I am doing beta diversity analysis between the influent of wastewater treatment plant and effluent of wastewater treatment plant.

When I compare them, I can locate metadata what was the BOD in the influence. What was the total heterotrophic plate count, what was the total 16S rRNA number that I got from my quantitative polymerase chain reaction QPCR verses what where these

attributes for my effluent what was the heavy metals in influent what does the heavy metals in effluent. So, I can compare all this I say (Refer Time: 09:46) when I compare I can see how biomass impact a diversity, so when biomass reduces 100 times between influent and effluent what is a reduction in diversity.

So, I can do this kind of analysis meaning for the analysis using the meta data. I can also make phylogenetic trees. So, what are these phylogenetic trees, we will get that we will next of the lectures rest of the lecture is dedicated to that. So, a phylogenetic trees will look like this to you.

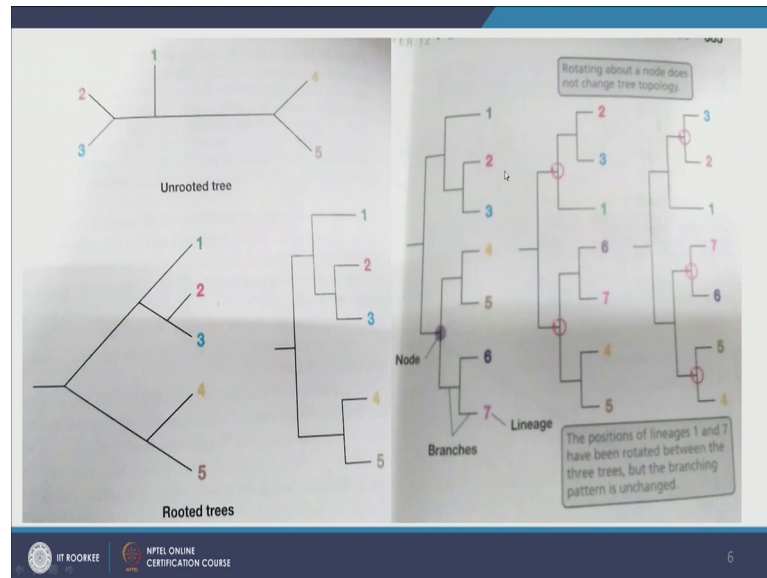
(Refer Slide Time: 10:16)



Here usually these nodes they are linked to your ancestor. So, here must be an ancestor. And from this ancestor they where two speciation between taxonomy group C and some other taxonomy group which is common ancestor taxon A and taxon B and here there was another speciation A and B. So, A and B are sister groups they came from same common ancestor. C is the outgroup of A and B outgroup means it is not belonging to the sibling family of A and B. So, this is how you understand.

Now, often what happens is the longer these lines are so they are two ways of writing making these dendrograms, these are called dendrograms. One of the ways is that the further two species are from each other the longer the lines would be. For example, the distance between A and B is summation of this length and this length distance between species A and B is equal to A A plus B B. So, branch length and this is the node all right.

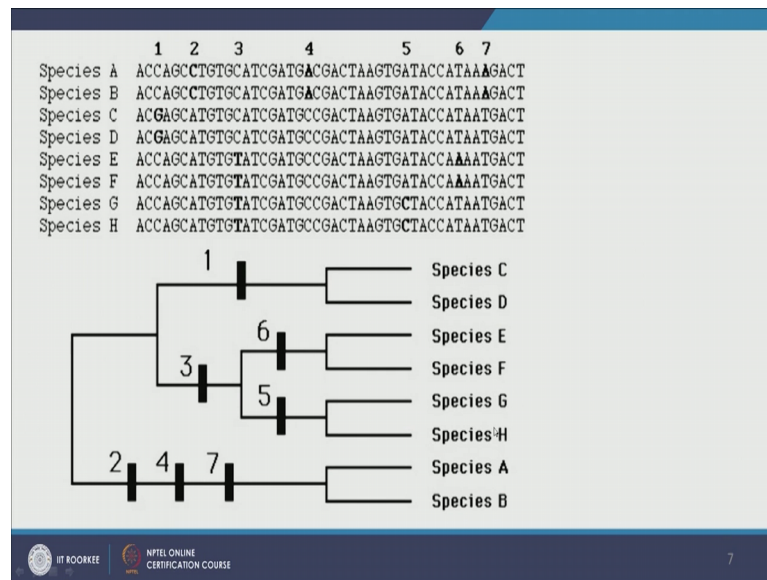
(Refer Slide Time: 11:17)



Now, let us look at these two pictures from your textbook. This particular dendrograms does not have any common ancestor; we do not know if this is a common ancestor, if this is the common ancestor, this is common ancestor. So, this kind of tree is called unrooted tree. basically here I have five different sequences, and I looked at the similarity and I made the trees. Oh, yeah, by the way, I also want to tell you how to make this tree, but let us start from here now.

Now, in these cases they give a direction and they tell you all this is the common ancestor from which two sisters branches came and then speciation happen further, 4, 5 are sisters; 1, 2, 3 are out group to it; 2, 3 are sisters 1; 4, 5 are outgroup to it. This can also be represented in this way. So, these are the nodes, these are the branches, these are the lineages.

(Refer Slide Time: 12:02)



Now, let us see how we make these trees. So, from a first queue file, you will first of all you will get species A, species B. So, you can get detailed information as gamma proteobacteria, what gamma proteobacteria, delta proteobacteria, what kind of delta proteobacteria or you can just get OT 1, OT 2, OTO 2, OTO 3 or species A, B, C, D so on and so forth. And here you have sequence usually they are very long, but in this case for represented a purposes, this is enough. So, the first step is alignment we align these sequences to each other. So, look here A is aligned at so happens that all of this species have A in this place. In the next place all of them have C; third all, but two have G. So, we have highlighted it in bold species C and D have a base pair difference here. In the fourth, all of them have A; fifth G and so on and so forth, in this one that there are two differences here. So, all the differences are highlighted in bold.

Now, once we have highlighted the differences, we can make a matrix of similarity. So, matrix of similarity, so we can find out nucleating distance between the samples or we can find breakout distance or many statistical parameters, basically they will tell you how similar or how dissimilar each species is to each other. And usually they are represented in matrix form. So, you will have species A, B, C, D written here species A, B, C, D written here and you will have similarities numbers written here.

Now, once you have your similarity or dissimilarity matrix, by the way I say similarity or dissimilarity matrix, because similarity in this case is equal to 1 minus dissimilarity. So,

once you have your similarity matrix or dissimilarity matrix, you can do clustering and this is hierarchical clustering by the way. And I use for my analysis, I use software R, it is open source free of cost software. And I find it very user friendly, but it does have a steep learning curve. I encourage you to learn and there is actually an NPTEL course on R, so I encouraged you to go and find out about it.

(Refer Slide Time: 14:14)

Handwritten notes: R, open source statistical software; 'vegan' package; B-C, E; pvclust.

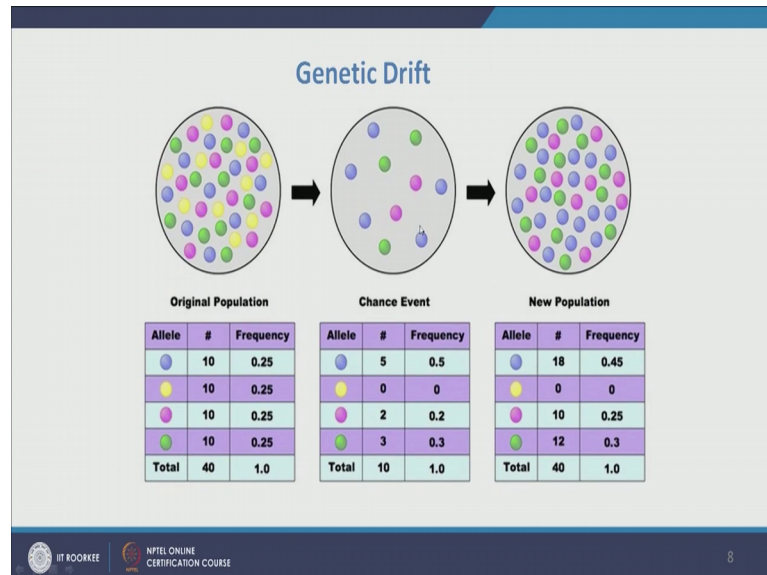
	1	2	3
1	25	100	
2	26		100
3	84		

So, in R there is a very nice package called vegan. So, within the R, which is an open source software a statistical software which I use a lot, there is a package called vegan. And this vegan package allows you to calculate different kinds of distances, for example, you can calculate break-width distance, you can calculate Euclidean distance and all of these will help you make matrices which look like this. So, you have your, so these are your species 1, 2, 3, species 2, 3, 4. And then you can have similarity how similar they are maybe species 1 and 2 are 25 percent similar; 1 and 3 are 26 percent similar; 1 and 4 are 84 percent similar; and then you can have obviously, 2 and 2 are 100 percent similar; and 3 and 3 are 100, so that diagonal usually runs 100.

So, this way you this similarity matrix you can use vegan package. And this another package that I used it is called pv clust. So, this package as to be make hierarchical dendrograms, do hierarchical clustering. And the beauty is that it uses the similarity to group the similar species together and when it does that it also tells me if the grouping is significantly statistically significant or not. So, in summary the way to make this

dendrograms is by aligning the sequences, calculating their similarities - species similarity, and then clustering the similar ones together.

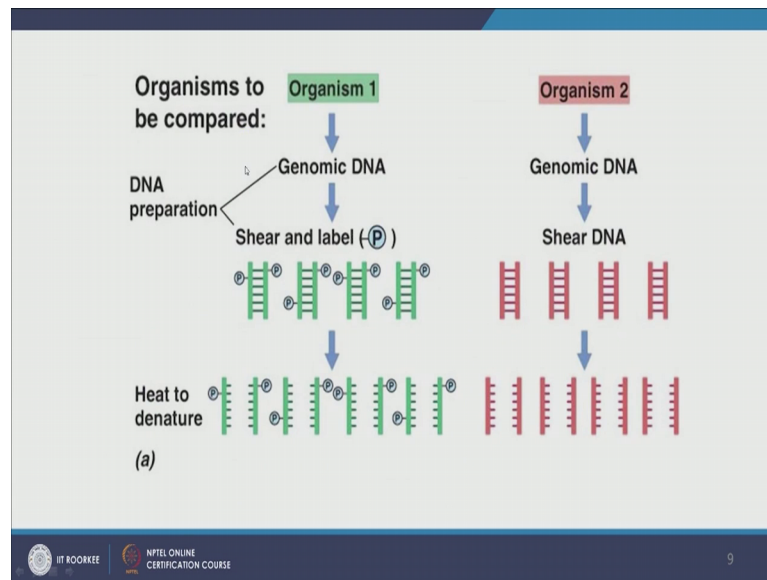
(Refer Slide Time: 16:20)



The other way in which the genetic evolution happens which we did not cover in the previous lecture apart from vertical mutations and transfer of genes, horizontal or horizontal gene transfer and recombination is genetic drift. So, genetic drift is natural drift in the genetic signature of a population or of a community because of different reasons, one of them is bottleneck. So, for example, here in the original population we have obviously, four different colors of balls or four different colors of genes and all of them are equally distributed. So, their frequency is 10, 10, 10, out of 40 balls. And then overtime for someone chance after some chance event, you know there was no induced mutation nothing, the yellow balls just disappeared, some of the greens disappeared, some of the purples and blue disappeared and the frequency varied. So, the distribution varied.

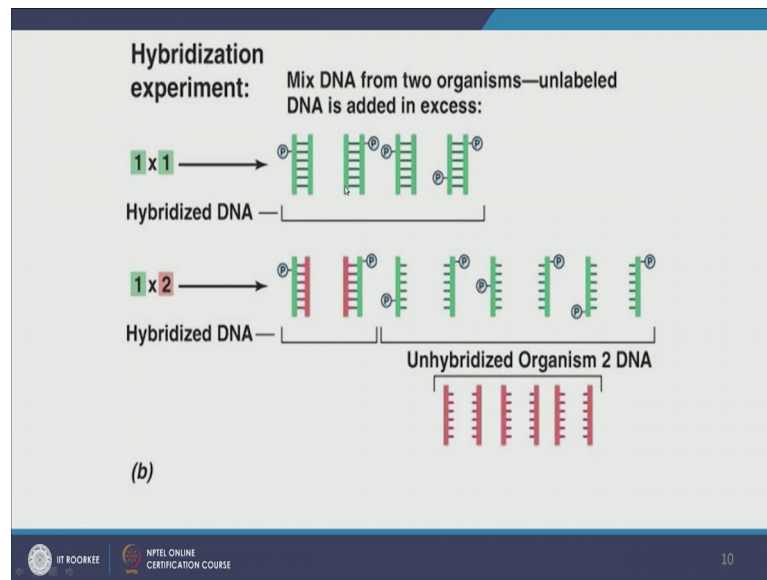
And then when again repopulation happened I got a very different community. So, the genetic pool has drifted. So, the drift suggest that it is not induced intentionally, it is not a drastic change, but happens over time and it is almost natural we cannot stop it. So, now after talking about genetic drift, now what I want to understand is like if I have these microbes blue, yellow, purple, green, how do I tell which of them are similar which of them are not similar so that I can create hierarchical clusters like this.

(Refer Slide Time: 17:52)



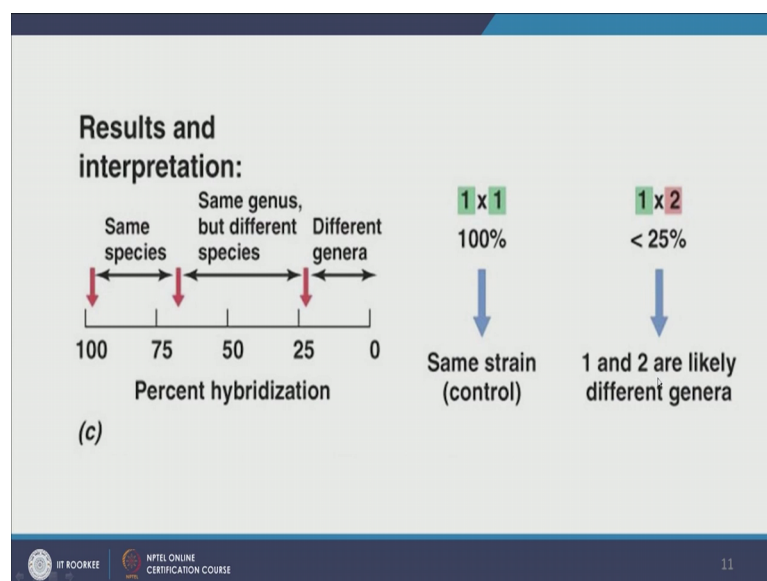
So, for that the most through technique is DNA-DNA hybridization now how does DNA-DNA hybridization work just take a look. So, let us say I know what organism 1 is I have grown it in lab, I have cultured it, I have fully sequenced it. And now I have isolated organism 2 and I suspect that it is similar to organism 1 or it may be organism 1. So, in order to understand whether it is organism 1 or how different or how similar it is to organism 1, what I do is I extract DNA from both genomic DNA from organism 1 and organism 2. I infer organism 1; I will shear the DNA. So, I made small fragments of the DNA and I have labored it with a tag. And in the this, I have just shear the DNA using similar process. Now, here I am heating now with heat the DNA will denature. So, the double standard fragment will becomes singles standard fragment.

(Refer Slide Time: 18:44)



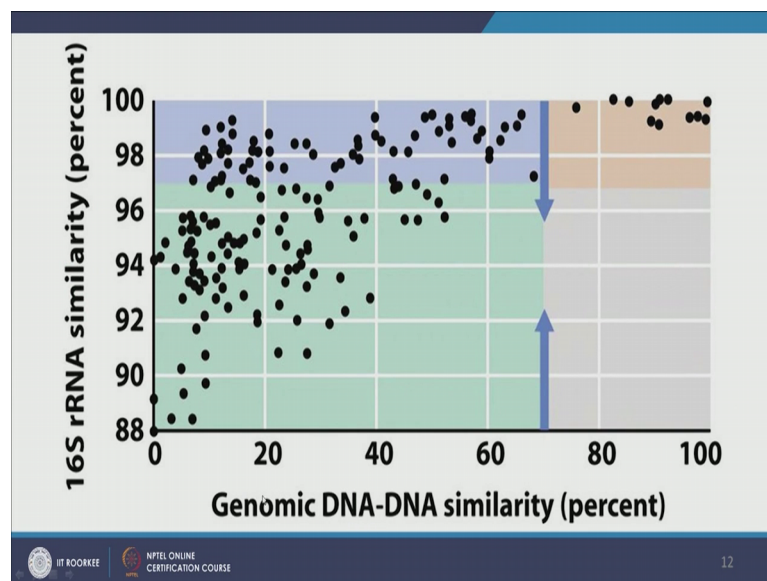
Now, the next step is what I will do is I will mix the DNA from two organism and I will make sure that the unlabeled one is in excess. So, I will have two kinds of I will have. So, here the unlabeled one is excess this is in excess. So, I will have two different combinations at the end. The green-green will known-known will attached to each other with 100 percent complement. The green and red will attach to each other they might be some imperfect matches. And then I will have unhybridized organisms here, which I have not hybridized with each other for various reasons.

(Refer Slide Time: 19:32)



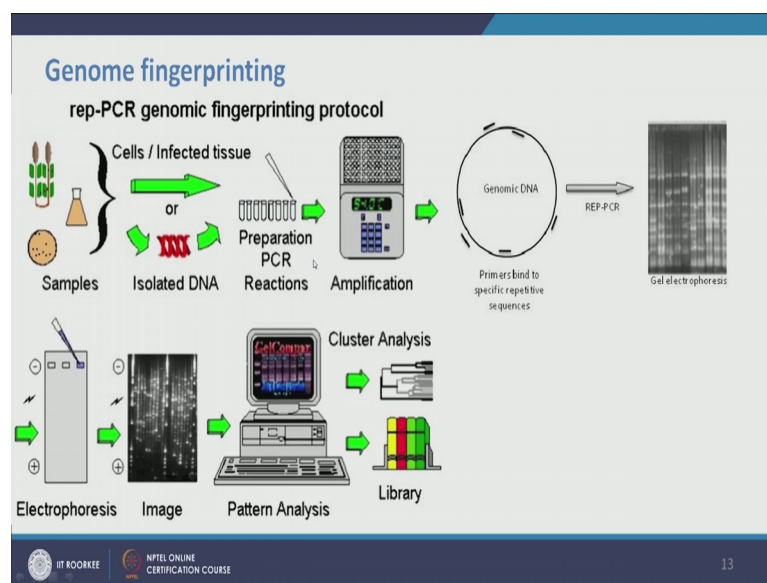
Now, I can look here what percentage have are same. So, these are 100 percent similarity. And what percentage are of from different genera, the known and unknown. If I get up to 97, 98 percent similarity, I say they are same species; if it is less than 70 percent similarity, I say they are same genus, but different species; and if they are less than 20 percent, then I say they are different genera. So, depending on how similar there are let say they have 97 percent similarities, so it was nearly perfect hybridization I will say they are same species.

(Refer Slide Time: 20:04)



Now, the question is I have mentioned before we often use 16S rRNA similarity to make phylogenetic trees. Now, here I am talking about genomic DNA-DNA similarity. Now, note 16S rRNA is a very small portion of your entire genome. And when I measure entire genome on x-axis and 16S rRNA similarity on y-axis, what I can say is that when remember here what we have talked about if it is more than 70 percent similarity, they are same species. So, if the genomic DNA has more than 70 percent similarity which is here top right corner, and that would be that would correspond to 97 percent similarity in 16S rRNA similarity. So, in last week homework, the organism that had more than 97 percent similarity, we can say they are from same species by one basis of DNA-DNA hybridization.

(Refer Slide Time: 20:58)



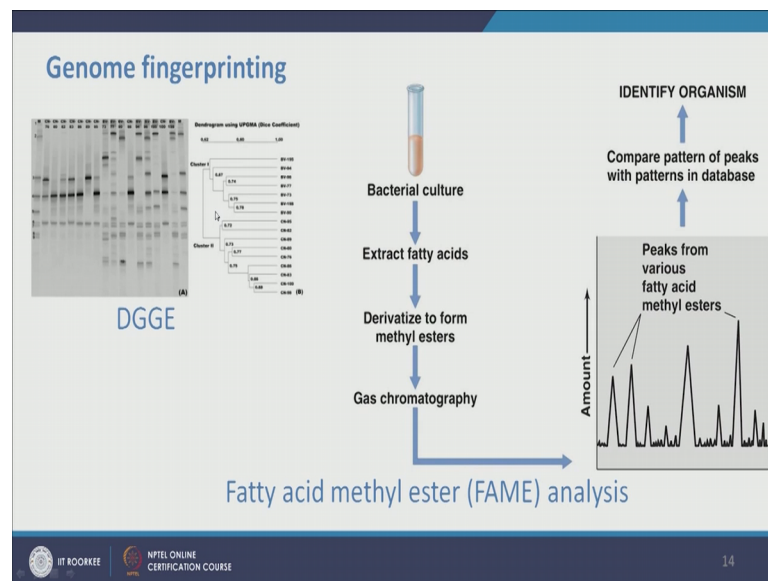
Another way in which we understand what microbes are present is rep-PCR and rep-PCR is a wonderful technique in which we use the repetitive sequences palindromic sequences within a genomics DNA of a microbe to classify different kinds of microbes. Let us say I have different forms of different strains of pseudomonas in my samples, I know all of them are pseudomonas, but I want to know how many different kinds of pseudomonas I have, and what are their relative abundances. So, I can design primers that will bind to the repetitive sequences in the pseudomonas, and these are the repeater sequences wherever they are the primers will bind and they will amplify.

So, because they are at they are unequally distributed along the genome and among different microbes, they will make unequal amplicons. And one unequal amplicons made I can run them on gel I can do gel electrophoresis which is basically separation of nucleic acids on bases of their size. Because the ways gel electrophoresis is work is that on where we load our samples, which is here on the top, we apply negative charge; and DNA being negatively charged molecule is pushed away from it. So, and the speed at which it escapes away from the negative pool is dependent on its size. The smaller particles will run faster, the larger one will be slow to go.

So, the long ones for example, this amplicons would be somewhere here, the shorter one should be somewhere here. So, depending on the signatures, I can say all right, the similar ones for example, these three look similar. So, they might be the same

pseudomonas. This one, this one and this one looks similar, so they might be same pseudomonas. So, I can tell count how many different kinds of microbes I have. So, basically rep-PCR will look like this I take my samples, isolate DNA, I prepare my PCR reaction for rep-PCR and do my PCR. And I done my gel electrophoresis, I get my image, I do analysis of my image and I create my cluster diagram. I see how similar or different the genomes are.

(Refer Slide Time: 22:59)



The other technique that is used a lot was used a lot actually nowadays we call it stone age techniques at least I called it stone age technique is DGGE. So, this is a denaturing gradient gel electrophoresis. So, the gel is laid in such a way that it has a gradient of denaturing agent. So, urea or other denaturing agent are put in this gel, but it has a gradient it might have 40 percent to 60 percent. So, here at the bottom, there is 60 percent denaturing agent. So, by the time DNA comes here, it will denature more here, it would not. So, as that your amplicon travel, they get denature. And according to what their sizes is and how fast how much what their GC content is they get separated along the one of the axis.

Now, the samples that have similar fingerprints are likely to have similar microbes. And the beauty of DGGE is I can actually pick up, for example, look here this band and this band look similar instance that they have three prominent bands. Now, this one is dominant in this, but this is not very bright in this. So, I can actually pick it up and see

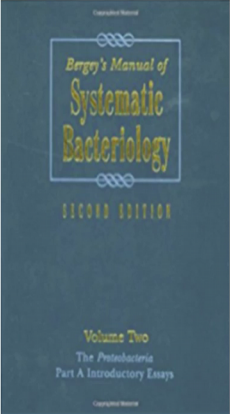
what even though these sample look like they have similar composition, but this one is dominant in this conditions. So, let us find out what this is. So, I can actually sequenced. And even if I do not sequence it, I can cluster these samples together. So, the similar ones will be clustered together. For example, these two will be nearer to each other than this and this would be. Now, this one and this one are very similar, so they were likely to be clustered together.

But visually we cannot analyze really well. So, we use software for it and where we identity the bands, and we analyses the similarity in create dendrograms. I must say that DGGE is quite outdated now; we have better techniques. We just sequence them instead of being fingerprinting and then we use the sequences to calculate the similarity matrix and then to do hierarchical clustering.

Another analysis that has been used a lot is FAME or fatty acid methyl ester analysis where we basically extract the fatty acid from bacterial culture. We derivatize them to form methyl esters, and then we do gas chromatography and then we get the peaks. And then on basis of the peaks we tell what how many different kinds of microbes that we have. Now, again because this is mostly based on the fatty acids and not necessary the nucleic acid; this is not as reliable as other DNA based techniques are in definitely not as reliable as next generation sequencing techniques. And thus it is not very popularly in used any more.

(Refer Slide Time: 25:33)

Rank or level	Example
Species	<i>E. coli</i>
Genus	<i>Escherichia</i>
Family	Enterobacteriaceae
Order	Enterobacteriales
Class	γ-Proteobacteria
Phylum	Proteobacteria
Domain	Bacteria



The image shows the cover of 'Bergey's Manual of Systematic Bacteriology, Second Edition, Volume Two: The Proteobacteria. Part A: Introductory Essays'. The cover is dark blue with gold lettering. It features the title 'Bergey's Manual of Systematic Bacteriology' in a large, stylized font, with 'SECOND EDITION' below it. At the bottom, it specifies 'Volume Two' and 'The Proteobacteria. Part A: Introductory Essays'.

IT ROOKIEE | NPTEL ONLINE CERTIFICATION COURSE

15

So, I would like to end this lecture here by giving you a very important information on the taxonomic ranks of microbes. So, we have been using these words very casually until now in all lectures, but I think it is a very good idea to go through them. So, we have three domains in microbiology – eukaryote, bacteria and archaea, so these are called domain, these are the first broad classification we do. So, remember we have Luca the least the last universal common ancestor displayed into three domains. I have been using the word kingdom also casually, but the technical word is domain.

Now, domain are further divided into different phylum, which will be like proteobacteria or acidobacteria or firmicutes. Now, they are further divided into classes like gamma proteobacteria, delta proteobacteria, alpha, beta, epsilon data proteobacteria. Now, even just knowing that this is delta proteobacteria, I can get some information from it. Delta proteobacteria might have sulfate reducers for example, because many sulfate reducers are found to be in delta proteobacteria; or if phylum was firmicutes and class was closterium, I can say alright quite possible that cellulose degraders are how is in closterium because that is the information we have.

Now, we know that this is not necessarily true, because some cellulose degraders are necessarily closterium anyway, but this is informative the class is further divided into orders. So, entro so this is now I am going towards from broad classification to finer classification. So, enterobacteriales this is the order. From order, if you have more sequence, we can get more information; then we have families like enterobacteriaceae and then genus *Escherichia*, and then species such as *Escherichia coli* or *E coli*, which is the model organism for studying bacteria.

So, we go from domain to phylum to class to order to family to genus to species. And my dear students I promise you in this class you do not have to memorize biodot, but this taxonomic ranks you do need to memorize by biodot. The other important information I want to give you is of this Bergey's manual of systematic bacteriology there are two manuals of Bergey, one is based on morphology, the other base based on genetics. So, this is the volume 2, and it is based on genetics. So, what they have done is they have talked about different phylum and classes in fact, all known phylum classes, order, family, genus and species of bacteria. So, very very important manual to have. If possible try to get a softcopy or a hardcopy of this manual and this will be very helpful for you

throughout your life. So, Bergey's manual of systematic bacteriology is very important. And I encourage you to keep it with you.

Dear students, this is all for today in this class. In next class, we will go ahead and we will talk about more about genomics; and now is the time that we will dive straight into the actual environmental problems, and how we apply everything you have studied until now to solve them. Also we will take a brief subway to understand about latest microbiological tools that are used in microbiology especially in applied environmental microbiology. So, that is all for today.

Thank you very much.