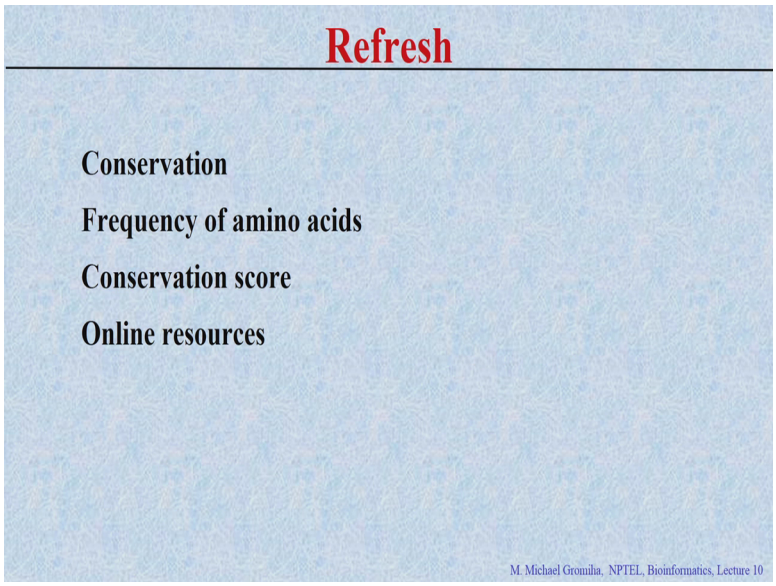
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 10a
Phylogenetic Trees

In today's lecture, we will discuss about the construction of phylogenetic trees, yesterday's class we discussed various aspects on conservation.

(Refer Slide Time: 00:26)



So, could you please remember something what we discussed in the last lecture?

Student: Frequency.

Yeah right. So, what is the meaning of conservation?

Student: So, how far a residue is conserved in a …

Right in any protein sequence, how far a specific residue maintains at the same position in different organisms, in different homologous sequences. Whether it maintains the same residue at the same position or it changes with respect to different organisms, or with respect to different time.

So, if it is same then we call that this residue is conserved at the particular position. So, how to understand or how to derive, how to qualitatively measure, the conservation of the residues at different positions in a sequence, in this case it is a 2-step procedure. What is the first step?

Student: Frequency.

Yeah we try to calculate the frequency of occurrence of amino acid residues at any particular position, then we convert these frequencies into scores. For getting the frequency of occurrence there are different ways to get the frequency of occurrence. What are the different ways to get the frequency of occurrence? Unweighted frequencies; in this case we do not give any weightage to any sequence or no weightage any amino acids; all residues and all sequences are treated equally. So, it is useful if you compare closely related sequences, in this case most of them are from homologous families and several sequences, residues are the same.

So, is useful for the closely related sequences. What is the other method to calculate the frequency?

Student: Unweighted.

Weighted frequencies. What you do with the weighted frequencies?

Student: different weightages

Now, if you give different weightage, any specific weightage for any specific residue which like to be maintained the same position, then we give weightage for any particular sequence. For example, if we have many sequences from closely related ones, homologous sequences, and some of them from distant related ones, then if you give weightage to include the information from distant related ones, then we give weightage. So, the third one we discussed about the independent counts, how randomly distributed, with respect to your original distribution, that you can get independent counts to get the frequency.

Then when frequency is calculated then we convert these frequencies into scores. So, we discussed about different types of methods to get scores, what are different methods to calculate the score?

Student: Entropy based.

Entropy based.

Student: Variation

Variant based method.

Student: Sum of.

A sum of pairs method, right. So, in the case of the entropy based method how to work? So, you get a probability of the particular amino acid residue.

Student: Hum.

That is frequency of residues at any particular position multiplied with this.

Student: log

Logarithmic of ones. So, you will give the entropy based method. In the case of variants based method, it gets the information for the same amino acid residues at different positions. How far they changed, how far they maintain at different positions in amino acid sequence to the get the information. In the case of sum of pairs method what we do? We take these pairs and then we give its any matrix. So, we can normalize any matrix and you can use the any matrix to see to give weightage to similar residues compared with the residues of the different kinds.

So, we get the conservation, then we discussed about the normalization. So, we can use different normalization procedures to normalize the scores. In the case of entropy based method if your particular position is occupied with the same residue what is its score?

Student: 0.

Zero, because the frequency is one. So, one in logarithmic of one, that is equal to 0. So, we get the numbers, negative values for the variable ones. So, you can also normalize based on 0 to 1 or based on this average value of 1 and so on. So, we discussed about few online resources, what are the online resources we discussed to calculate the conservation score?

Student: AL2CO

AL2CO right.

Student: Yes sir.

And then.

Student: Consurf.

Consurf - what is the input for AL2CO server?

Student: Alignment

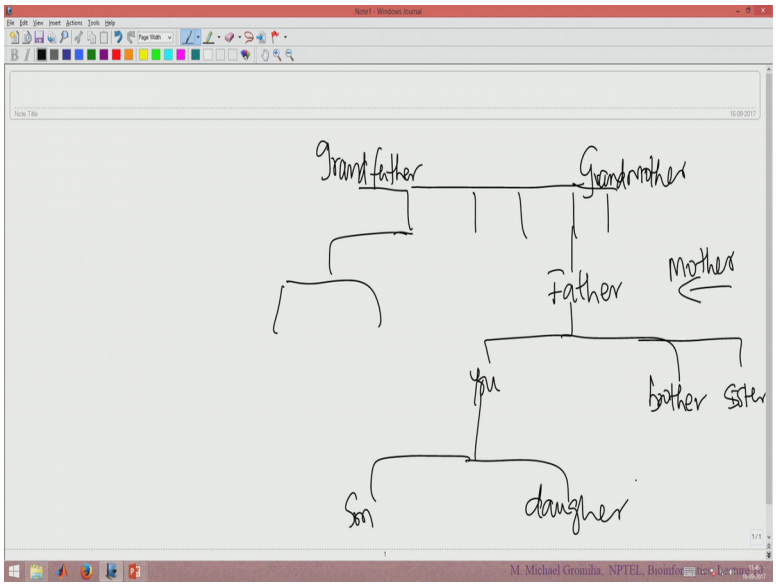
Alignment. We get the multiple sequence alignment. So, it takes a multiple sequence alignment and we give the score. We give all the options, you can choose any of the options; you get the score and for the Consurf?

Student: PDB

You can give the PDB ID. It will automatically get the sequences and do all the alignment and look at the coservation score. So, today if you have these sequences, we check discussed about the multiple sequence alignment; there are many sequences and which sequences are close to each other. And how far it takes to change one sequence to other sequences with respect to different organisms. So, in this case you can draw a tree type of method right, make a tree to show which residues are similar to each other.

So, phylogeny, if you define, this a description of the relationship, any biological relationship, expressed as in the form of tree. For example, if you take the human genealogy. So, the grandfather here and a grandmother.

(Refer Slide Time: 05:29)



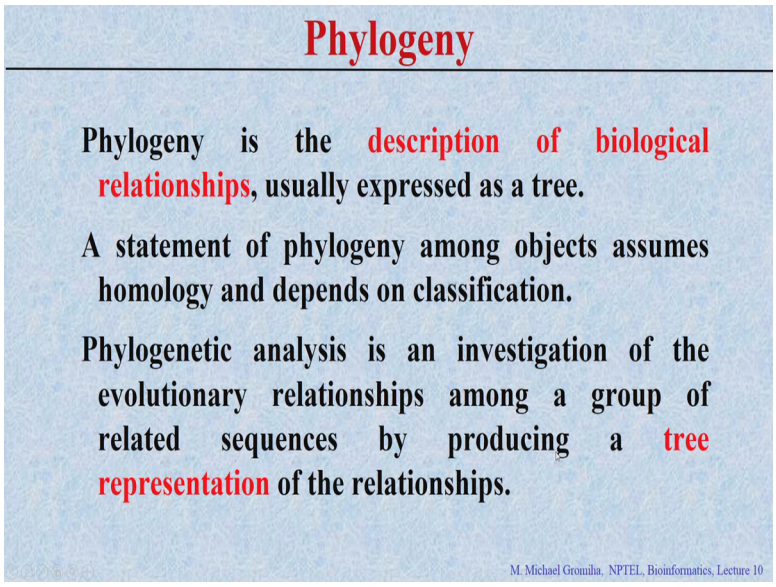
So, they have several children. So, now, if you see here is the father. Likewise from the mother tree you get the mother. So, here we have, several children right? So, we have this is may be you, this is your sibling, brother, sister so on.

Here also you can have different things. So, like this for the different son, sons and daughters of these grandparents then here, you are here. So, you have your son your daughter and so on right. So, we have these relationships, to understand the relationship for example, whether you can see any similarities in the shape or any of these similar organs you know if it look similar, then you can say they may be close to each other. From this tree to this you can easily, you can know that and from here, you see you have the relationship and here to grandmother-daughter, daughter to the grandmother some possibilities, then if you go for the more and more generations then we do not know.

So, but you can listen that he looks like his grandfather, he looks like his aunt. So, how to do this? There are different ways, one is look like the shape, or you have to look at the any of the organs, or that then it depends on time. How long it took from one generation to other generation. So, you can understand two different aspects, one is you can see the time from one generation to other generation, and also you can see the relationship how far the two people are related to each other on what aspects.

So, if we look the closely related ones, you can see much similarities; when it goes further and further you can see similarities, but less compared with the closely related ones.

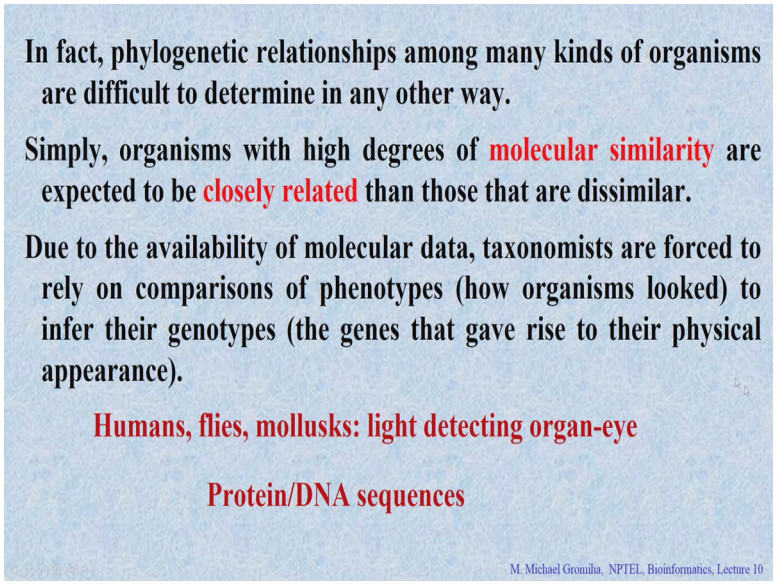
(Refer Slide Time: 07:39)



Likewise you can see the phylogeny. So, this is among different objects, which assumes the homology because they are similar to each other and it also depends upon how they classify. So then, these analyses, what is a phylogenetic analysis? It is an investigation of this evolutionary relationship, from one organ of different organisms among the group of related sequences by giving a tree-like for representation.

Just we discussed earlier. So, we give a tree-like representation to discuss the relationship, to understand the evolutionary relationship. In fact, there are various ways to explain the relationship, but is a easiest way for example, if we say A and B are related to each other, B and C are related to each other, A and D are related to each other. So, we can explain different ways, but the tree-wise representation is the easiest possible way to understand the relationship among different organisms.

(Refer Slide Time: 08:31)



So, we see the organisms with high similarity like it, just I discussed earlier, they are closely related.

So, then they are this similar. So, in this case the ones which are closely related they put nearby each other. So, we can easily understand these are close to each other. So, due to the availability of data, the taxonomists first they start, they started to construct trees right. So, they rely on the comparison of phenotypes, like how the organisms look like and you infer a genotype, what gives this type of appearance. Likewise in the family tree as I showed earlier. So, you can see look at this how they look like, then they will try to compare okay; which is look like. So, these are similar to each other, right.

Then they try to do this, but when it happens to the various systems for example, if they take the light-detecting organ, eye. To see the common behavior of these different organisms then, sometimes it fails because this is, eyes you can see from humans, flies, mollusk and so on. So, then they try to use different types of information to infer the sequences which are similar to each other, the organisms which are similar to each other. So, currently due to the availability of different sequences or protein sequence or DNA sequences, it is easy and it is reliable to compare the sequences and identify the lineage between different organisms, because we have the different sequences for DNA. Where you can get the DNA sequences?

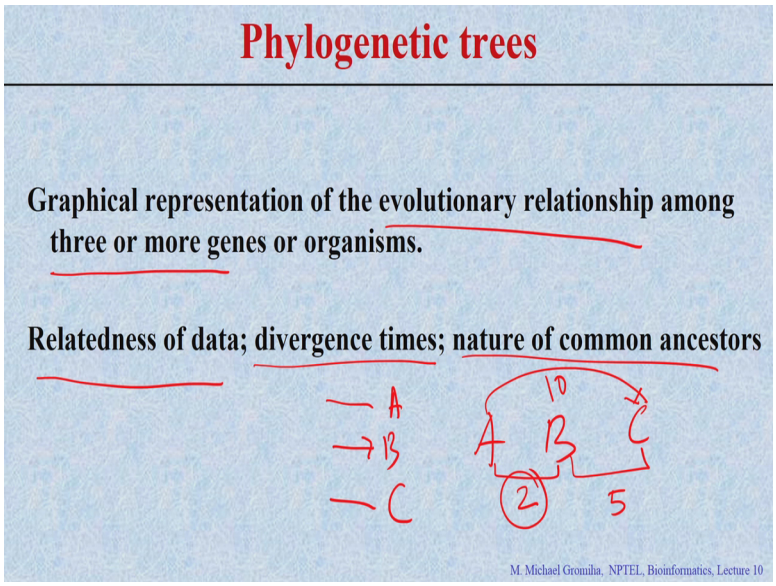
Student: GenBank.

GenBank or EMBL or DDBJ and so on. So, where should we get the protein sequences?

Student: UniProt.

UniProt is a unique resource, we get the protein sequences. So currently, also we get different sequences from several organisms and the sequences are also reliable. So, easily you can compare these sequences to understand the evolutionary relationship and how far it took from one organism to different organism, and or how about the different variations in the sequences and so on.

(Refer Slide Time: 10:28)

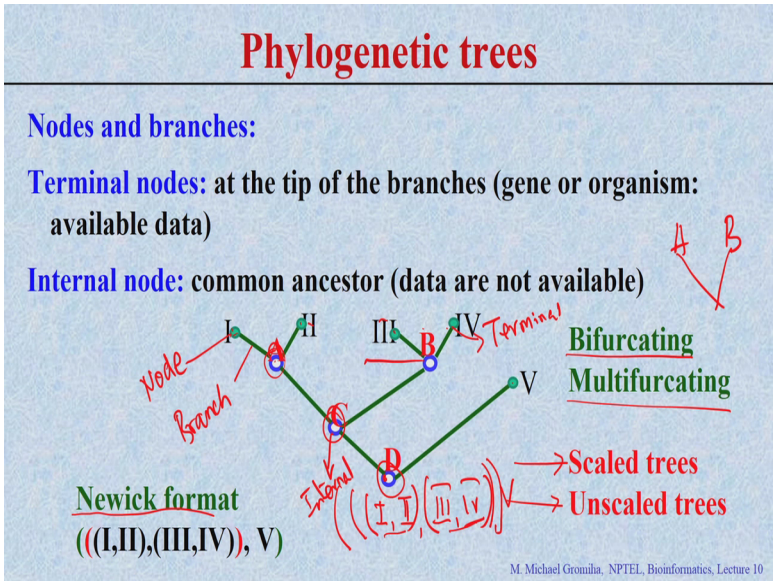


So, now if we have the sequences, we can construct trees to show the evolutionary relationship of any genes or organism. At least 3 or more genes or any organisms. For example, if you have 3 organisms A B C. If A and B are similar to each other with minor change for example, if there is a two amino acid change and B and C are also similar to each other, which are 5 amino acid difference and you can see A and C are similar to each other, for example you can see the 10 amino acid difference. Then if you have 3 cases, one is A, one is B, one is C, which are close to each other? Here A and B are close to each other because we have the mismatch of only two residues. Likewise if you have more number of data, 3 or more sequences from the differences in the nucleotides or in the amino acids, you can see how they are related, which two are close to each other and how long it takes to get this divergence; depending upon the variations you can estimate

the time - on this much time that it takes to get this convergence, to get the divergence times for one organisms to another organisms.

Then you can see the nature of common ancestors, alright if you make a tree you can see which will be the common ancestor. For example, here A and B are close to each other. So, they may have a common ancestor like this, you can see the common ancestor when you construct phylogenetic trees.

(Refer Slide Time: 11:59)



So, how to represent trees, how to draw trees? So, I will show one example here. So, if we have a common ancestor here if this is the one D is here. So, then from here I take the trees. So, this goes from number 5 through C you can see 3 and 4 and through A you can see I and II.

When you make a tree for example, this is A this is B both are common to each other you can connect these two. So, if you see this figure. So, which ones are close to each other.

Student: I and II, III and IV

I and II are close to each other, and III and IV are close to each other and this I and II and then III and IV it takes time, but they have some similarity this is the common ancestor C and then all these things V. So, they will be connected together at the point D.

So, when you draw this there are two different aspects. So, only we can see the dots and we can see the lines. So, lines represent branches and the dots represent nodes, you can see this is the node and this line they are branches and if you see the nodes, there are different types of dots here. You can see the two different types of circles. One is a closed one here, II, III, IV and V.
And also you can see some of the nodes which are in the interior here, there's open circles here this, this and this.

So, here these two circles, they represent different aspects, the one with the ends, tip of the branches they are called the terminal nodes. These are called terminal nodes, like you can see the gene or any organism or any sequence, that is we know the data, that information is available. And we can see other dots, inside ones and these are the internal nodes for example, these are internal nodes, which represent common ancestor.

But this information we do not know because if we look into the UniProt database or we look into the DDBJ database, you look at the sequences that information we know then which two are common and which two are emerged in specific period of time that we do not know. So, when you make these type of trees you can try to understand which organisms are close to each other and which organism emerges first and how long it takes to have the next one and so on. So, if we have this type of trees, there are two types of trees for example, we see this node it has two branches.

Some cases you have 3, some case you have 4. So if there is 2 then this is called the bifurcating ones, the nodes. So, there can be more branches. So, this one we call as multifurcating. So, here we have two. For example, if we have another one here there will be 3 or 4. So, in this case you can distinguish the bifurcating as well as the multifurcating. And another aspect is, in some cases you know how long it takes for one organism to another organism, some cases we do not know. So, this will tell you the scaled trees and the unscaled trees.

If you take 100 years or 10,000 years. So, if you can scale it then this is called a scaled trees and if you do not know the time then that case you can call this as a unscaled trees. So, now, if you have a trees to making large trees there like space. In this case you can simplify the trees in the form of specific notations like you can use some parentheses. If

you write the parentheses this will tell you how the tree looks like and this format is called the Newick format.

How to write this Newick format? For example, if you see this tree, I and II are close to each other. So, I and II are close to each other and then III and IV are close to each other and these two are close to each other, this is this node A and here these are close to each other this node B, but these are close to each other. So, now, we can say these two are close to each other. Now the next one is the number V, V is alone. So, V is here and this is and this 5 are connected with each other.

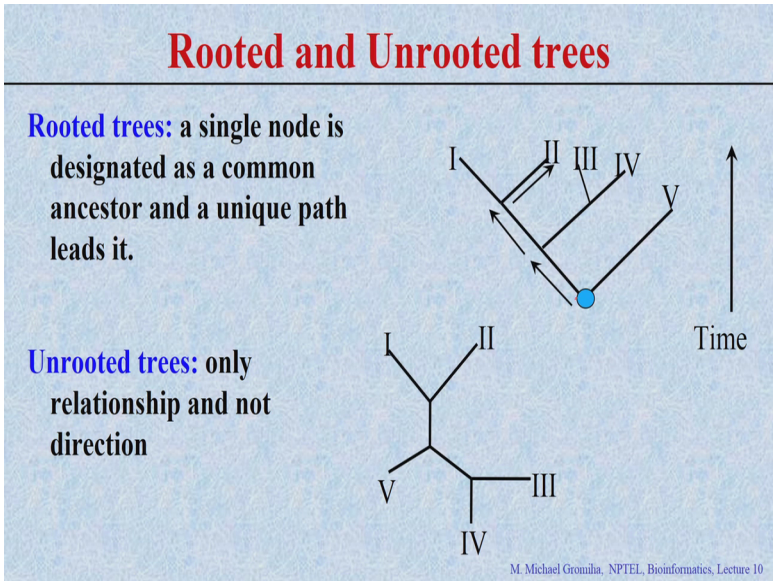
So, you can see another bracket here then we can say this is the format. So, go to the internal brackets, then you can construct the tree for example, I and II is connected III and IV connected and these two are connected with each other and V is connected with everyone. So, there is a format which you can write, represent trees using this parentheses; this format is called the Newick format. We have a trees we have nodes you have branches, 2 types of nodes, terminal nodes and internal nodes. Terminal nodes represent any genes or any organisms. So, data are available internal node provides the information on common ancestor that we do not know when you construct the trees then we can have the information.

And there are scale trees and unscaled tress. So, what is the difference between scaled and unscaled trees?

Student: They

Time we know the time or we do not know time, then you have the depending upon these branches we have bifurcating and multi-furcating and we can write these trees in a form of this specific format this is called the Newick format. Now there is another type of trees there is called a rooted trees and the unrooted trees; what is rooted trees and what's called unrooted trees.

(Refer Slide Time: 17:39)

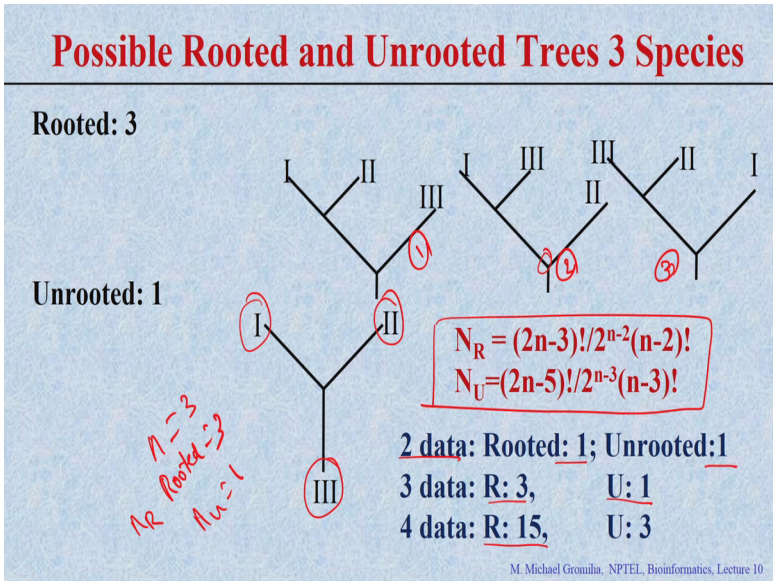


The rooted trees means you can see a single node, which is designated as a common ancestor. So, in this case we know the ancestor, but the lineage we know, but we do not know the relationship.

But we have an ancestor and you have the path to lead to this tree, all this I II III IV V; finally, comes to this point through a specific period of time; that's called a rooted tree. So, rooted tree a single node is designated as a common ancestor. In the case of unrooted trees we know the relationship, but we do not know their direction for example, I and II are related and IV and V are related. So, all they are related, but where they starts where they ends we do not know.

Now, but they are related to each other. For example, if you see a group of people who can say they look similar. So, may be related to each other, but I do not know where we have ancestor maybe fourth generation fifth generation, that we do not know. Some cases we know you look similar, but we know ancestor, your forefather or your grandparent, great grandparent was he. So, you both from the just different genealogy; this is how we can see the rooted trees as well as the unrooted trees ok.

(Refer Slide Time: 18:47)



Now, so for example, if you have the 3 species, I II III. How many times you can have the rooted trees and how many times we can have the unrooted trees right.

If you have 3 species I II III, if the unrooted tree, we do not know where the common ancestor. In this case you can draw in any way whatever way you draw, so, you will get only one tree. Because we do not know about ancestor, but the case of rooted ones ancestor could be I, ancestor could be II, ancestor could be III and in this case you will have 3 different ways to construct these rooted trees. So, this is 1, this is 2. So, you have the third one, right here this is I, here II is here and III is here and then it emerges I and II.

Here first I is here, and second and third; here first is the second one, this is the common one and first and third. There are 3 different ways you can have the rooted trees if you think about the ancestor. Likewise if we have two data, how many rooted trees and how many unrooted trees? One rooted trees and one unrooted trees if you have 3 data 3 rooted trees and one unrooted tree. If you have 4 data depending upon the first one we'll get 15 rooted trees and 3 unrooted trees. So, this is the equation used to calculate the number of rooted and unrooted trees depending upon the organism.

So, n is the number of species and we can use this equation 2 n minus 3 factorial, divided by 2 n minus 2, into n minus 2 factorial. If this is the equation who can fit these numbers

here. So, number of rooted trees and the number of unrooted trees for example, if we have 3 data how many number rooted trees?

Student: 6 minus

6 minus 3 that is equal to 3 factorial, that is equal to 6 divided by?

Student: 2

2 right. So, 3 minus 2 equal to 1, here equal to 1 into 1, 1. So, this equal to 3 by 1 that is equal to 1 right. So, your n equal to 3 then this case N_R equal to 3 right. So, n equal to 3. So, number of rooted trees equal to 3, you can use this equation. Likewise, if you say unrooted trees this is n equal to 3 then N_U equal to 1, we can fit this equation likewise you can use 4 data or 5 data, you can get the number of rooted and unrooted trees and this question I think once they asked in the GATE exam. So, how to relate number of rooted or unrooted trees.