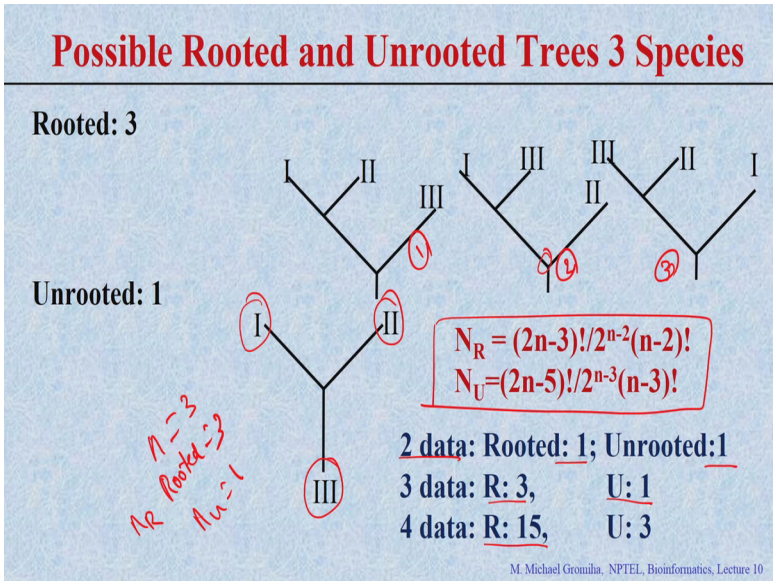
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

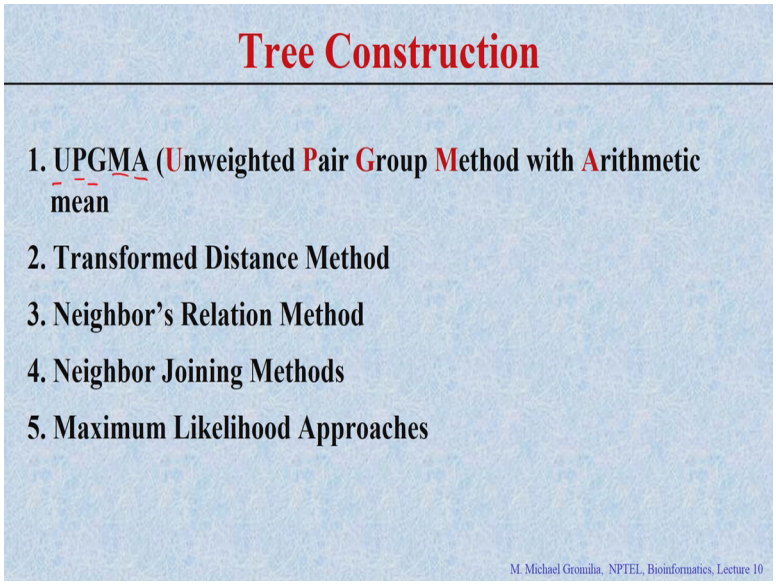
Lecture - 10b
Phylogenetic Trees II

(Refer Slide Time: 00:18)



Then how to construct the trees? Now for example, if we have, let us say sequences, right and how to construct the trees. So, there are various ways, there are several ways to construct the trees. So, one of the foremost method and the popular methods you see UPGMA.

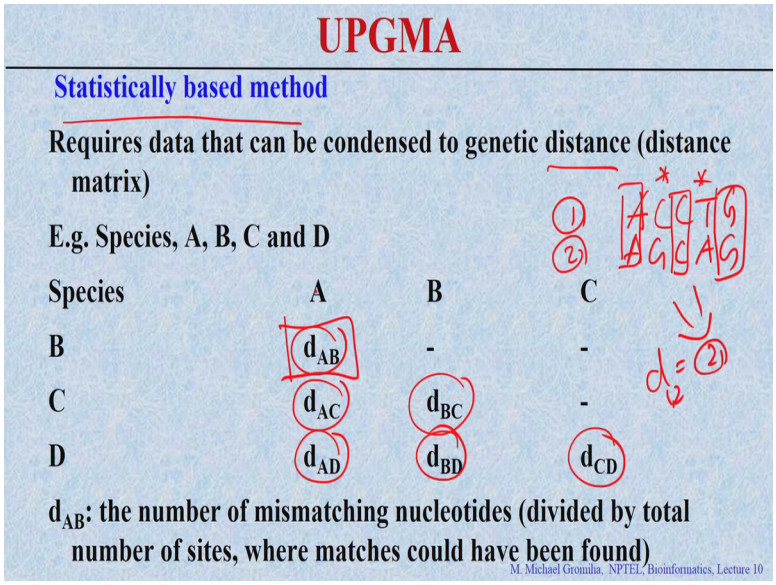
(Refer Slide Time: 00:32)



This is Unweighted Pair Group with Arithmetic Mean right, this is simplified as UPGMA method. It is very common method, and you can easily understand how to construct the trees using UPGMA method. It has a good understanding, but it has some issues some disadvantages. So, rectify that one, that are several other methods have been proposed, that is transformed distance method, neighbor's relation method, neighbor joining method, maximum likelihood approaches and so on.

So, the developed several approaches to construct trees. So, we will see how to construct trees based on UPGMA and what are the principles used in the other types of methods. So, it is a statistical-based method right. So, it requires the data to be connected or to be condensed with genetic distance. For example, we can use we can use DNA sequences or you can use protein sequences. So, they look at these sequences and see how they are different from each other.

(Refer Slide Time: 01:37)



They calculate the distance right, it is a statistically based method; they use statistics to analyze the data to construct the trees based on the distance.

So, how do you think about the distance here? In this case we have, we know only the sequences ACCTG this is your sequence number one, you can sequence number two you can see AGCAG. So, how to calculate the distance between these two? It is in the two dimensional case, this is not the three dimensional one. So, in this case, if you see how far they are different from each other? So how far each different from each other; so what is the difference, how many nucleotides are different?

Student: 2.

This is same.

Student: (Refer Time: 02:20).

This is different, this is same, this is different this is same. So, here you can see that distance is two. Distance difference between A and B or 1 and 2, one comma two, that is equal to 2. So, for example, if you take the species ABCD. 4 species right, we have put ABC, ABCD.

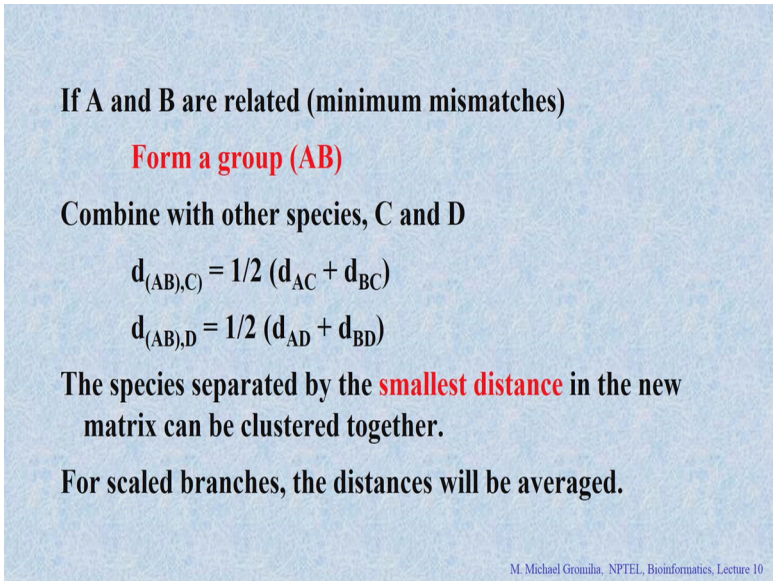
So, you calculate the distance between A and B, let us put d_{AB} and A and C is d_{AC} and A and D is d_{AD}. Right this is same, because B and B are the same. So, that is 0, and B and C

already we include here right. So, this why they put dash here likewise B and C you can get here, B and D, and C and D. So, we get the distance among all possibilities. Among the all possibilities you can get the number of mismatching sites from this which two are close to each other? The one with…

Student: Less.

Less number of mismatching, right. So, we will show some of the examples and if then they are related for example, if A and B are these are closest one for example, then you can say they are very close, they will have the similarities, they are close to each other and then you can combine this group with C and D. They use this equation.

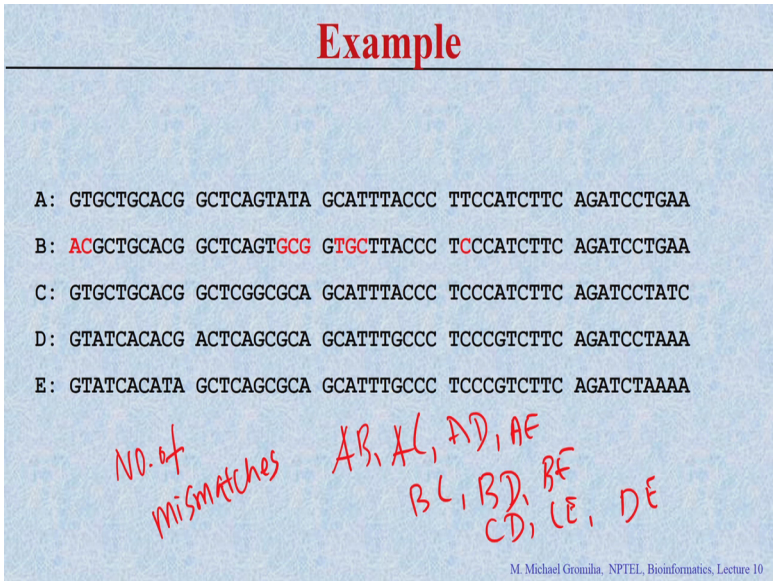
(Refer Slide Time: 03:36)



To combine A B with C they take the AC plus BC by 2 because this is a C you have to combine. So, you say C is common because A and B you have to combine. So, take this A and this B and take the average. So, then we will get the AB into C. Likewise you can do A B with D right. So, you choose A with D and B with D, then we take the average. So, you get the ABD.

Now, you can see the smallest ones then see which one, which two are close to each other. Likewise, construct for everything and we based on the number information based on the smallest distance we can construct a tree.

(Refer Slide Time: 04:21)



So, now you have the five sequences A B C D E, right. These are DNA sequences for the five species A B C D and E. Now we need to construct a tree. Which parameter you have to calculate?

Student: Distance.

Distance; So how many distances you have to calculate? Distance between?

Student: Between all possible.

A and B, A and C, A and D, A and E. Likewise BC.

Student: (Refer Time: 04:46).

BD, BE, CD, and CE right, all the possibilities right AB, BC.

Student: AB, AC.

AC.

Student: AD.

AD.

Student: AE.

AE, BC, BD, BE, CD, CE and DE. We get the numbers from this number which one or which two which pair is close to each other?

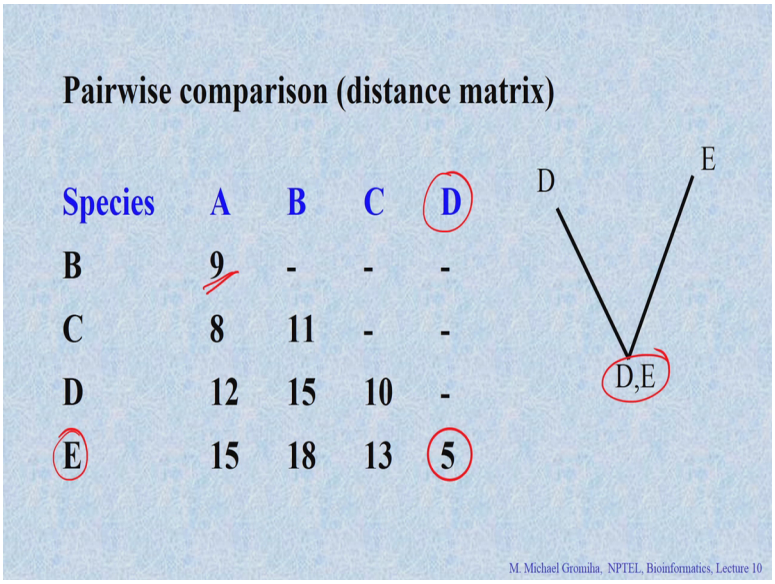
Student: That one.

Based on the distance, right. For example if you take A and B. So, we take the sequence A and B and compare these two sequences, and you can see number of mismatches right. How many number of mismatches right. So, how many mismatches between A and B?

Student: 9.

9? Ok, I put 9.

(Refer Slide Time: 05:38)



So, if you see this you can make this matrix A B C D and here B C D E because five organisms. So, A and B there are 9, likewise A and C mismatch is 8, and A and D is 12, A and E is 15. So, like BC, BD, BE and CD and CE right, then DE. Between the D and E if you see D and E. So, there are five mismatches; 1, 2, 3, 4, 5. So, five mismatches. So, this is the closest, this is the lowest value. So, from this what can we infer D and E are?

Student: Related, closely related.

Close to each other. Among all the combinations, this is very low, the lowest number of mismatch. So, you can see D and E are close to each other. So, you put the D is here E is

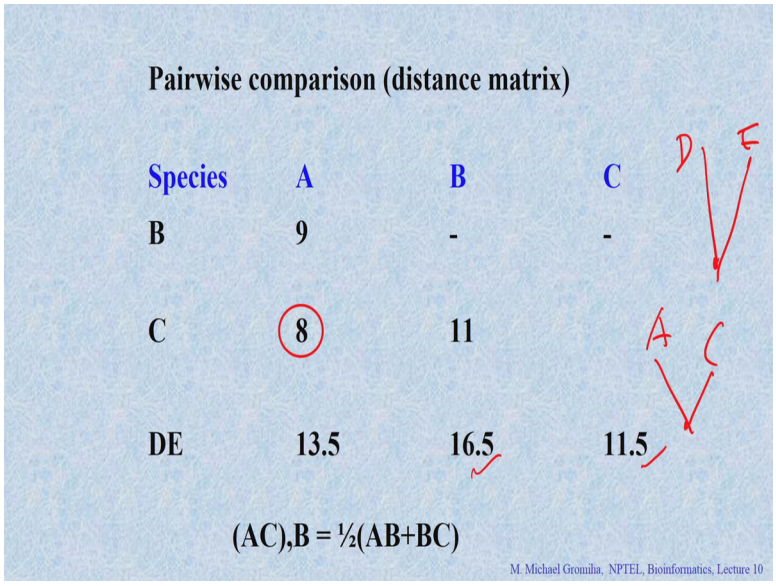
here and D and E are close to each other, we make first branch you can make and first node also you can make, this is the common node. So, D and E are close to each other.

Now, what to do?

Student: (Refer Time: 06:36).

Combine this DE with all other species, A B and C. So, how to do this? We take the average right. For example, if you see, B and A, we do not make any changes, because we need to combine D and E.

(Refer Slide Time: 06:49)



So, here you put the 9 as it is, and you have to put the C right and the B and DE we have to calculate because B and C we do not touch. So, if we take the D and E 12 plus 15 what is the average?

Student: 13.5.

13.5 right with respect to A. and D and E with respect to B.

Student: 16.5.

16.5 and here.

Student: 11.5.

11.5 right because we combined this D and DE with respect to A, with respect to B and with respect to C. So, now, I get this matrix; from this matrix which one is the lowest number?

Student: AD.

AD is the lowest one.

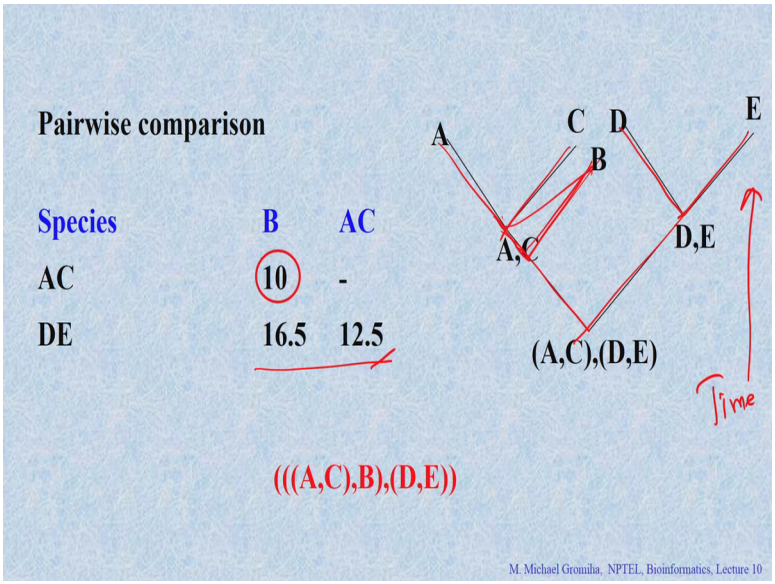
Student: (Refer Time: 07:28) 8.

It is a number 9, 8, 11, 13.5, 16.5 and 11.5 from this 8 is the lowest. So, what do I what do you infer from this one?

Student: A C are close.

A and C are close to each other, right.

(Refer Slide Time: 07:40)



So, you can see A and C are close to each other, earlier we did this D this E, now from here we can see A and C are close to each other, then what next what you have to do?

Student: combine.

You have to combine this AC with others. So, you put the AC here and the B and AC, D over here 3, B, AC and DE because A and C already we merged, D merged already then

B is there. So, you have the B, AC and D. So, combine this AC with B right this is equal to 10 because 9 plus 11 equal to 20 right. A plus B C right. So, 10 by 2 this is equal to 10. Then with the D to E. So, the 12.5 and 16.5 if you see here this is C to A, I combine A and C. So, take the average 13.5 and 11.5. So, this will be 12.5. Here you will touch because with B and because we are doing with AC, so here this equal to 6 minus 5.

So, from this matrix now we constructed the next matrix. from this which is the lowest one?

Student: 10.

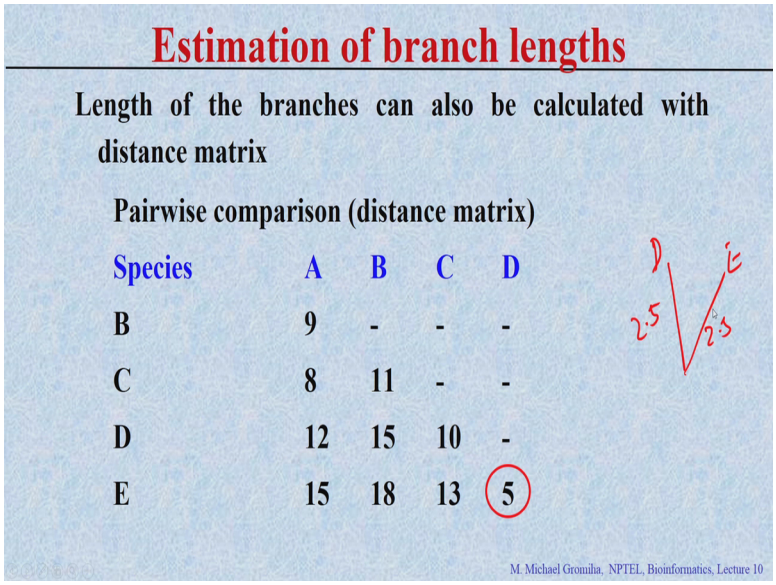
Ten is the lowest one, so AC with B. So, we have the AC here common AC right with the AC we can connect with B right we can go with this one, with if you put like this, then you can see all the three are the same line. This way I put this line and then this is related with B, and then if you this is out, then the other these two will be the remaining. So, then DE and A C, the DE and A C are common to each other finally, you can they get the tree.

So, from this one we can construct the tree right. D and E are close to each other, E and C are close to each other and this AC is close to B, and this is close to D and DE. So, we can construct this graph. So, when you have this graph now we can easily tell. So, which organisms they are close to each other.

The next question is how long it takes to evolve from one to other. So, you have time frame. So, can we able to estimate the time, right because we have some numbers. These numbers tell the number of mismatches, see with the number of mismatches you can see with less mismatches it took less time. If more number of mismatches it takes time right to go from one organism to other organism.

So, now we have, okay now you see the length of the branches, we can calculate from the distance matrix.

(Refer Slide Time: 10:14)

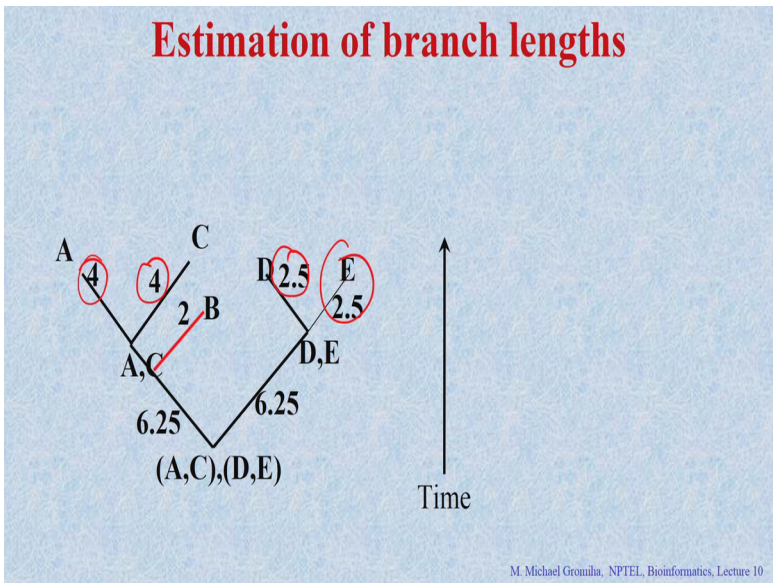


This is how to calculate. This is a pairwise comparison, we construct the matrix right, this matrix is same as the one we derived here, this one, right the same matrix. Now this is E and D is five. We assume they evolve the same time, with this branch is 5 then we divide this to 2. So, each one will get?

Student: 2.5

2.5, if it is D is here and the E is here.

(Refer Slide Time: 10:45)



This will take 2.5 and this will take 2.5. So, made this, is 2.5 and here this is 2.5. Now, the next one if you take AC is equal to?

Student: 8.

8. So, if you can draw this A and C this is equal to 8 you divided by 2. So, you put 4 and 4, 8. Then AC with B, right AC with B what is AC with B? It is 10. So, already here we put 4 and 4, 8. So, remaining 2, we give here for the B. So, total will be 10. Then AC with the B equal to 10 and then totally if you see the A to E, at least 15. So, you give this is 2.5, the rest we have put here this is 12.5 plus 2.5 this equal to 15.

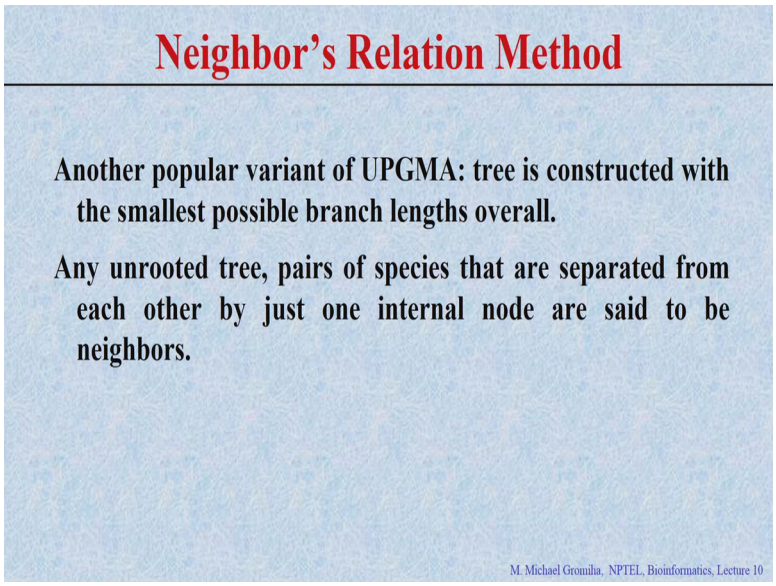
So, made these number, then you from this you can tell. This evolved first because it is very close, then this will takes 4 and here you take 6.5 from AC to D right. From this you can estimate the time approximately from one organism to other organisms. This is a method you can easily construct trees right, with simple statistics.

So, what the principle used in the UPGMA method?

Student: Distance

Distance between the two sequences. How far they are different, based on that we can construct a tree.

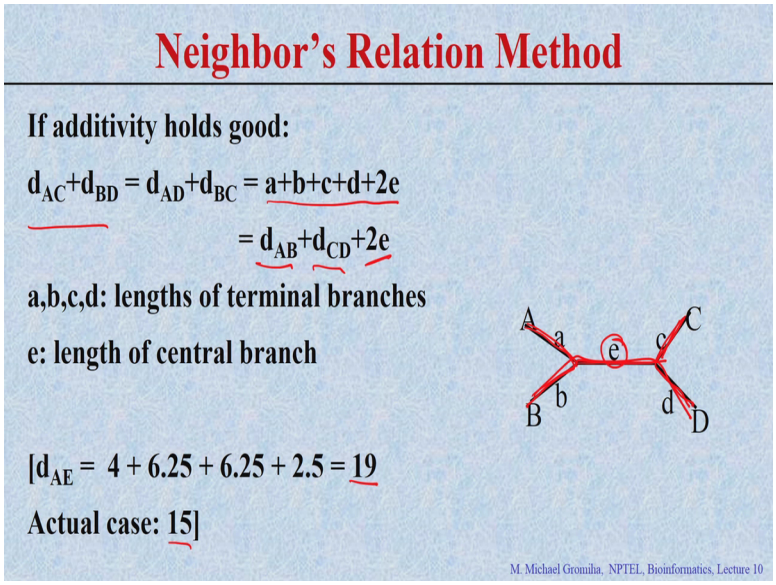
(Refer Slide Time: 12:07)



So, this is another method, this is called neighbor relation method, here also this is similar to UPGMA method, but this is a unrooted tree right, and you can see in the UPGMA method, sometimes the number is not the equal. For example, if you go from here to here, if you add up these numbers as well as if you add the values here for example, A to E this is 15, but A to E if you add from here to here, you will give the different number.

In this case it is not able to exactly account some numbers. So, for that one to make in correction in this methods, they put few more conditions right. They try to join all the neighbors not just joining one by one, they are try to join different numbers and see the closest one which one is the minimum. So, they use that criteria to develop this methods.

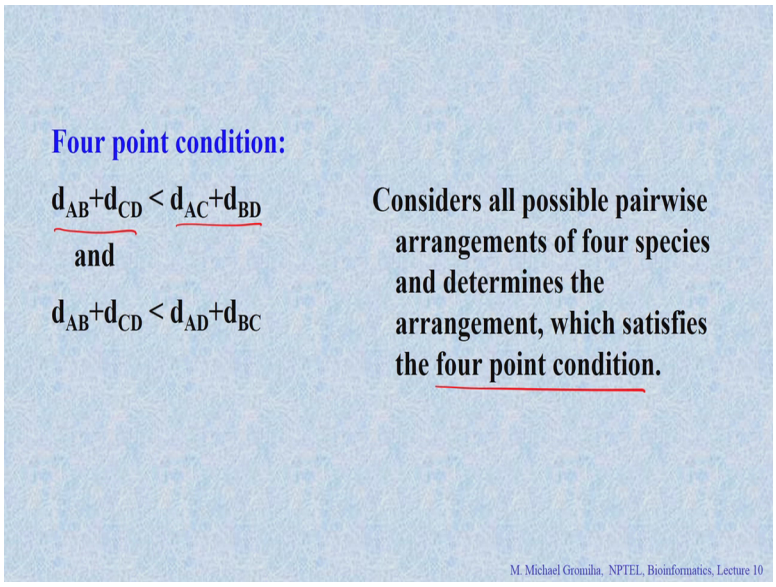
(Refer Slide Time: 13:02)



How to do this, it is a unrooted tree? So, A B; A here, B here, C and D. So, here from A to B and C to D this is not equal as A to C and B to D. So, added another line between these two, that is E for example, here this length is A and this length is B, and this length is C and this length is D and they put additional length as E. So, if you see AC plus BD this one, and this one right this is similar to AD plus BC you can give as a plus b plus c plus d plus 2e; that means, AB plus CD and 2e and here you can see the discrepancy between this UPGMA method and this method they considered this also in the node.

So, because of that this is the value you get from the UPGMA method. So, if you add up AE the 19, but actually case it is 15 because this is, this missing. To take care of this conditions.

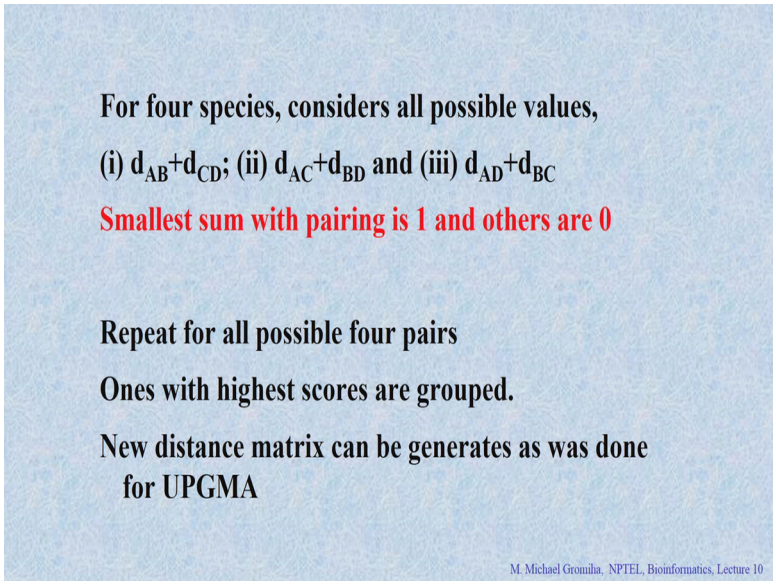
(Refer Slide Time: 14:06)



They have made four point conditions. If you put AB plus CD that should be less than AC plus BD, because if you see here A B plus C D, there should be less than AC plus BD because we have this value E as well as A B plus C D. This also less than AD plus BC right here. Either you take this or you take these right that should be less okay.

Now, if you have different species, they consider all the pairwise arrangements and put in such a way that they should satisfy this four point condition right.

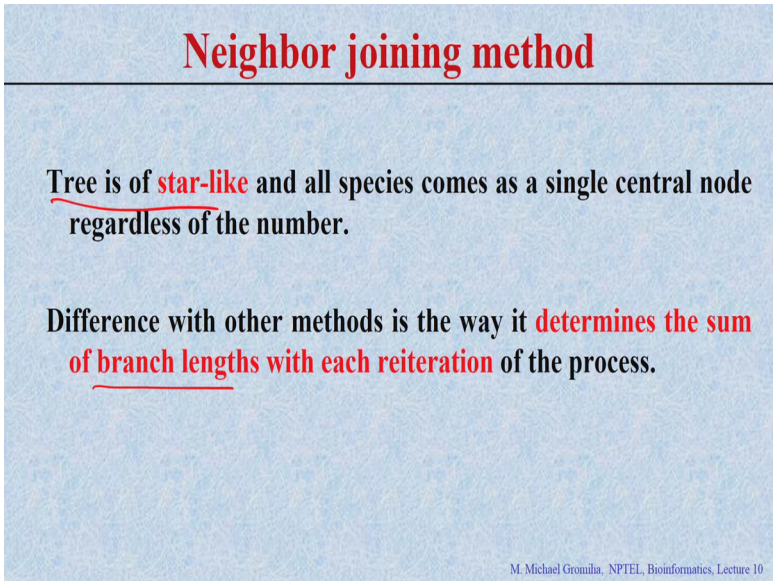
(Refer Slide Time: 14:47)



Now, I have this one, for example I can have four species, you have different values you can calculate A B plus C D and AC plus BD and AD plus BC, and the smallest sum, which can come close to each other that is one and others put 0.

Then we repeat for all possible pairs and take the closest ones. Once with the highest score or the group and then from that groups you can calculate the UPGMA method to get the distance. Then along with the UPGMA method, they considered special conditions to derive these trees right.

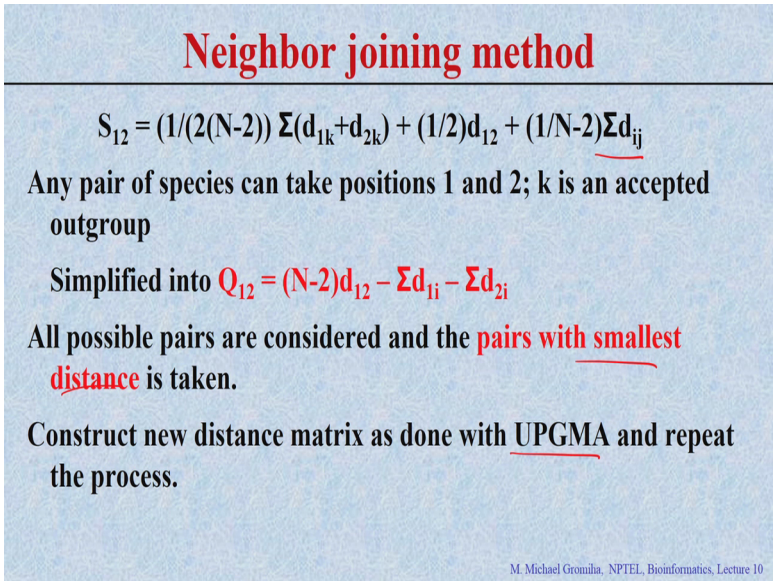
(Refer Slide Time: 15:18)



There is another method that is called neighbor joining method; in this case instead of going one by one they make a star like tree.

So, each species connected and then see how far they can connect with each other regardless of this any of these numbers. So, the difference with the other methods here you can see the sum of the branch lengths with each reiteration process, because we considered each one separately. And finally, they try to see how which one has the closest distance.

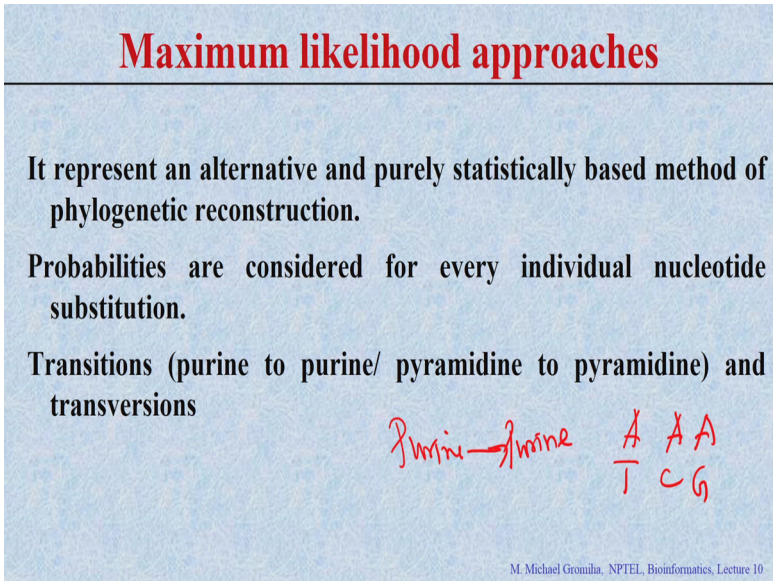
(Refer Slide Time: 15:48)



Based on that they derive this equation, this is a complicated equation with the depending upon the distance to get what are the possible pairs, which is connected to each other with respect to the smallest distance.

Once the smallest distance could find, then they can use this standard method to get the distance matrix as well as to get the trees. The simplest one is they try to utilize the information from different species together.

(Refer Slide Time: 16:18)



Than doing one by one. Then recently they had developed another method right, whatever we use, the UPGMA method, we consider all the nucleotides all the amino acids with equal weightage, and because what is the value, number we use from UPGMA method? The distance.

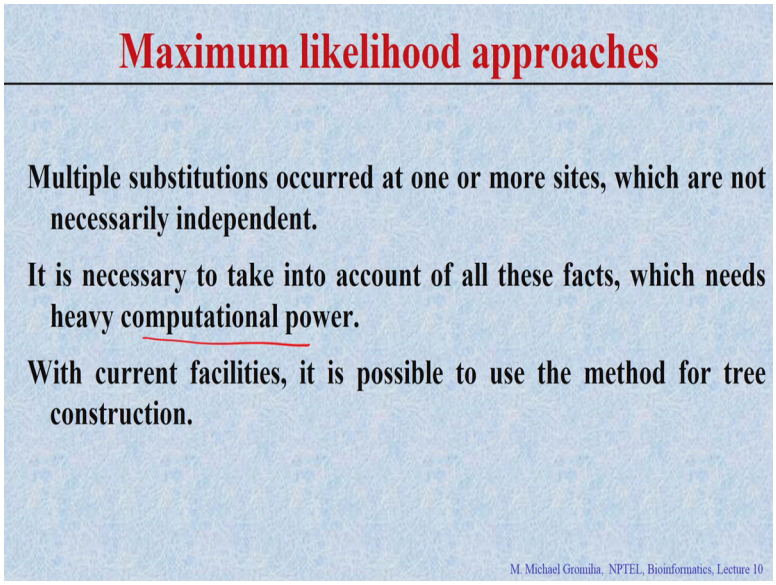
So, it is ATA is, A change to T or A change to C or A change to G, we take this one because we take only the mismatches. So, in the maximum likelihood method these are the statistical based method, but they give weightage. For example, the case of nucleotides, right what are the weightage they give usually? Purine to purine and pyrimidine to pyrimidine right they give some weightages and this is a transition and the transversions or transversion.

Student: (Refer Time: 17:07).

Purine to pyrimidine or vice versa. In this case we give more penalty right. So, in the in the maximum likelihood method, they try to give weightage to transitions and transversions. Likewise if you take the amino acids, when we construct trees, right they also use some sort of information for example, they can use a popular matrix, which matrix? PAM matrix or BLOSUM matrix so you can see the similarities. Also they can also design a matrix right based on the physicochemical properties or the molecular weights, right size.

So, if you have the misalignment, the alignment with mismatches, see similar residues or completely different residues, they give weightage, accordingly they can also develop a tree right, so different ways to construct phylogenetic trees. So, in this case, if there are multiple substitutions right may be independent, or sometimes dependent right, that also we can you can take into consideration right.

(Refer Slide Time: 18:04)

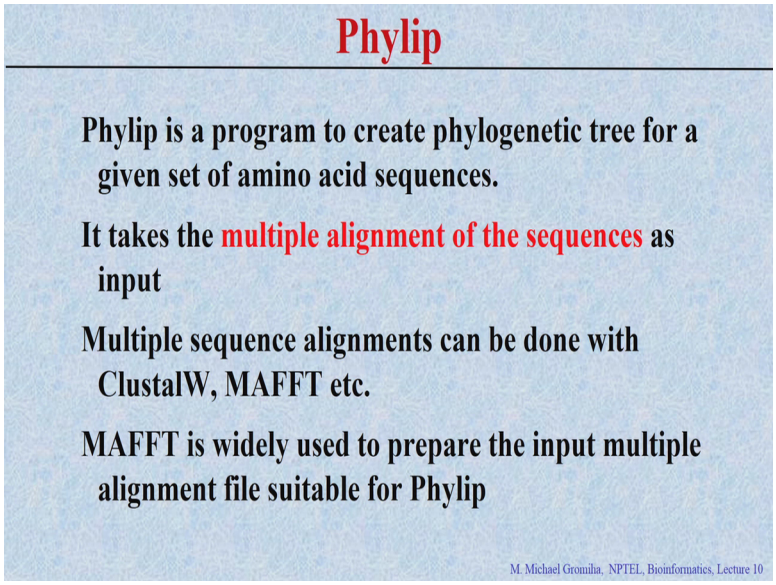


But to take all these aspects it requires high computational power, with the current facility it is possible to use all the information. This is a reason initially they tried to develop a method with the simplest possible, that is UPGMA method, and with the availability of computational power they try to increase the complexities right. So, that we can, if you if you increase the complexity, you can get better performance. But performance you need to sacrifice in terms of time right. If you have more time, you can have more time to analyze and will better results.

(Refer Slide Time: 18:47)

So, now is it possible to construct trees by considering all these aspects? If you look into the literature there are so many methods available, here I list of each set of methods. So, on PHYLIP is one of the most popular methods, even currently it is a widely accepted method right for constructing trees and it will take few minutes to just demonstrate the functioning of these PHYLIP how to do this.

(Refer Slide Time: 19:08)



It Is a program for constructing a phylogenetic tree for any given set of sequences, if you get a DNA sequences or amino acid sequences it will creates the phylogenetic tree. So, construct a phylogenetic tree. So, what is the input they acquire?

Student: multiple sequence.

In this case they require multiple sequences alignment, we can we need at least three sequences. We take the sequences and make the alignment right, and use the alignment as the input for constructing tree. How to get the multiple sequence alignment? What are the methods commonly available to align the sequences using multiple sequence alignment? ClustalW currently Clustal Omega, right MAFFT.

Student: (Refer Time: 19:43).

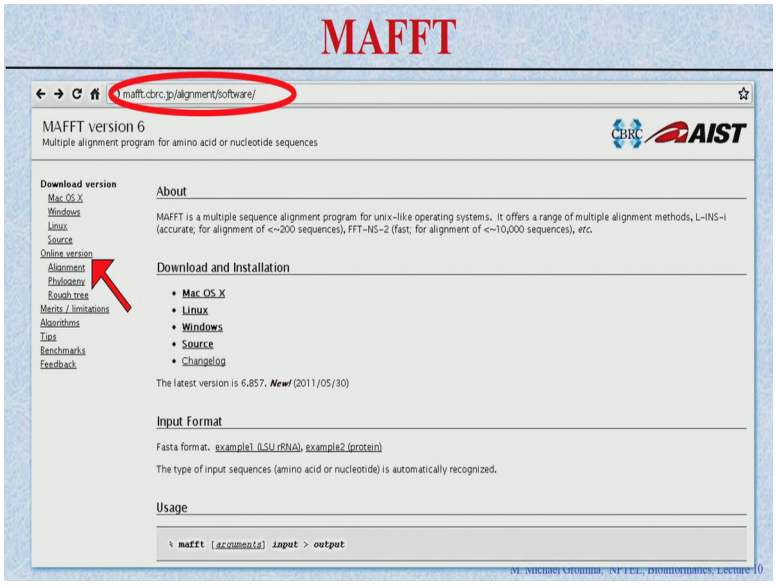
Promols.

Student: (Refer Time: 19:44).

MUSCLEe and so, on right; so Phylip automatically gets the information from MAFFT, if you give the MAFFT alignment it will automatically take the alignment and then give the tree. It is very easy.

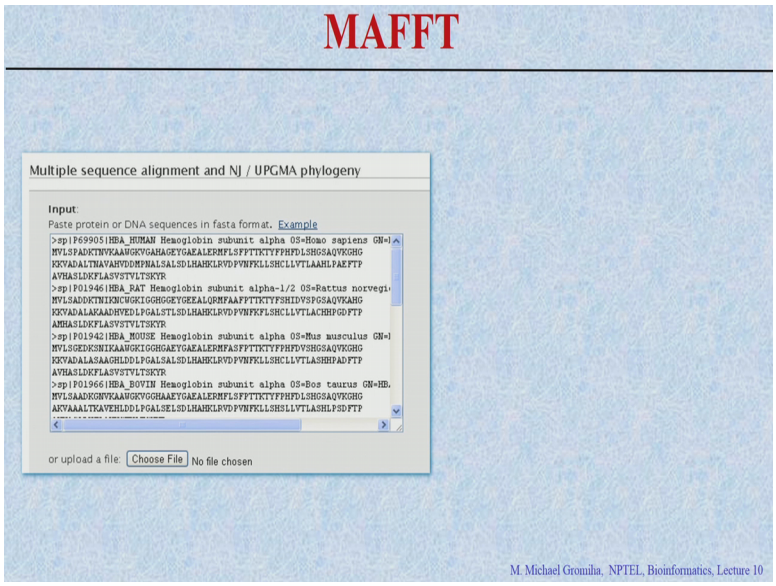
So, it is widely used to prepare the input file suitable for Phylip. We use MAFFT there is an option to save the file in Phylip format. So, you do not have to worry about formatting. Run MAFFT, save the multiple sequence alignment in Phylip format, and you can give this as input to the to run Phylip.

(Refer Slide Time: 20:15)



So, this is the home page for MAFFT right. So, they have a download version as well as online version, you can go to this website and then you can access MAFFT. If you like to use the online version just you go to the online version and give your sequence.

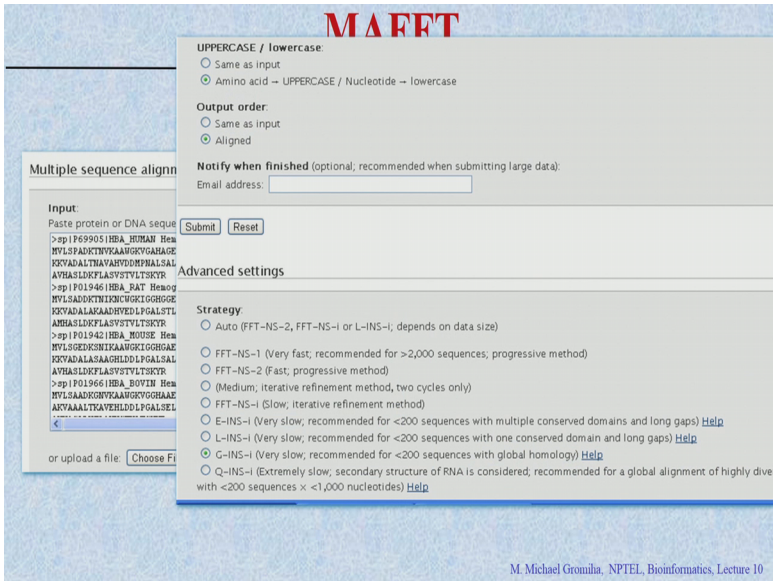
(Refer Slide Time: 20:28)



What these are your sequences, but you will auto it will create the your multiple sequence alignment right.

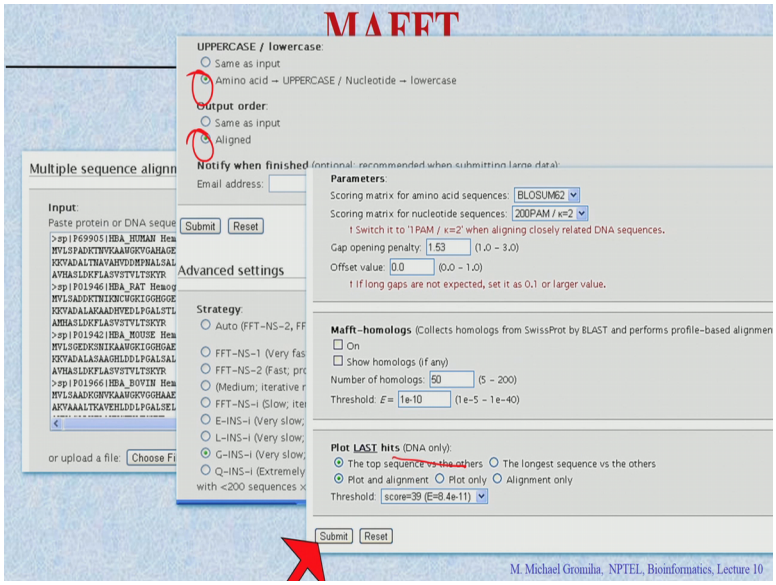
So, if you do this, it will ask for the conditions for the parameters.

(Refer Slide Time: 20:36)



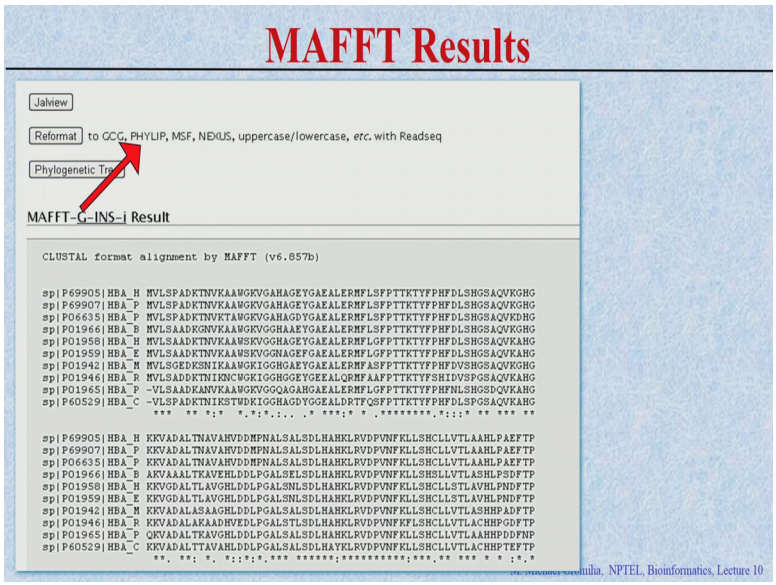
Which parameters you want right, you have the aligned one, you need the aligned ones. So, we need the aligned one, you click here right and these amino acid sequence, also here this is a, this is recommended if less than 2200 sequences; we click that one. Depending upon your sequences and the different data you need, you can choose any of these settings.

(Refer Slide Time: 21:02)



If you do this. So, now, you can submit their data right, is asking for this alignment and the plot right. So, you ask for the matrix. So, we can use the BLOSUM matrix right. So, and if you click on submit.

(Refer Slide Time: 21:14)



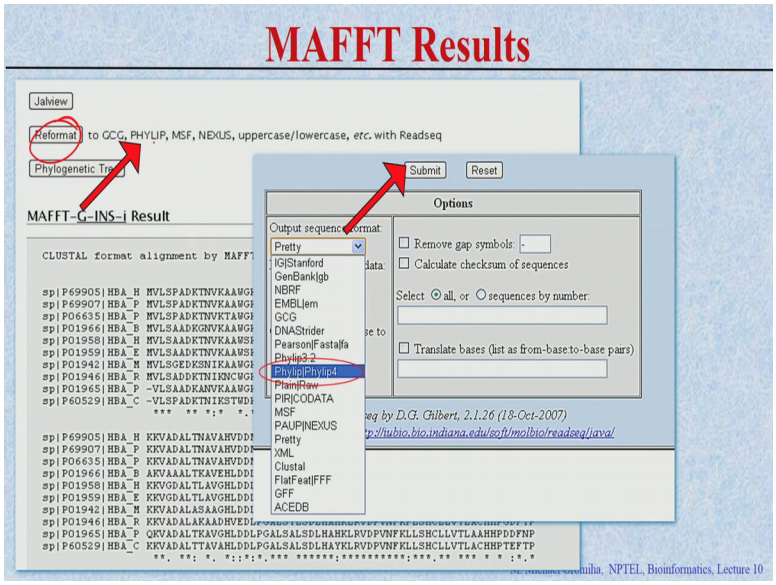
So, you will a get the result. This is a multiple sequence alignment.

Now, the question is whether you need to reformat this are not? Right, do you want to reformat it? Yes, because we had to, want to use these for?

Student: Phylip

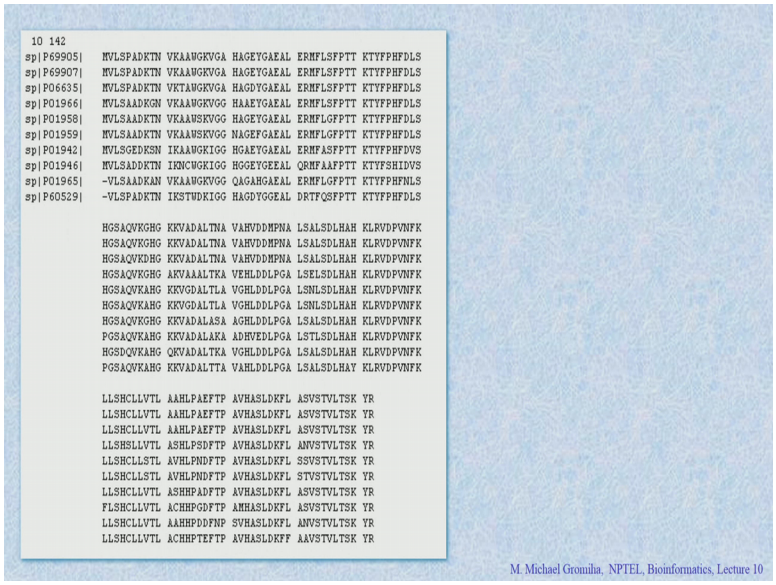
Phylip right. So, you have a reformat option here right and you can use different format. So, if you use want to Phylip you can use Phylip. Currently MAFFT also includes to construct trees directly, that is also possible, but if you use Phylip you format with the Phylip format.

(Refer Slide Time: 21:46)



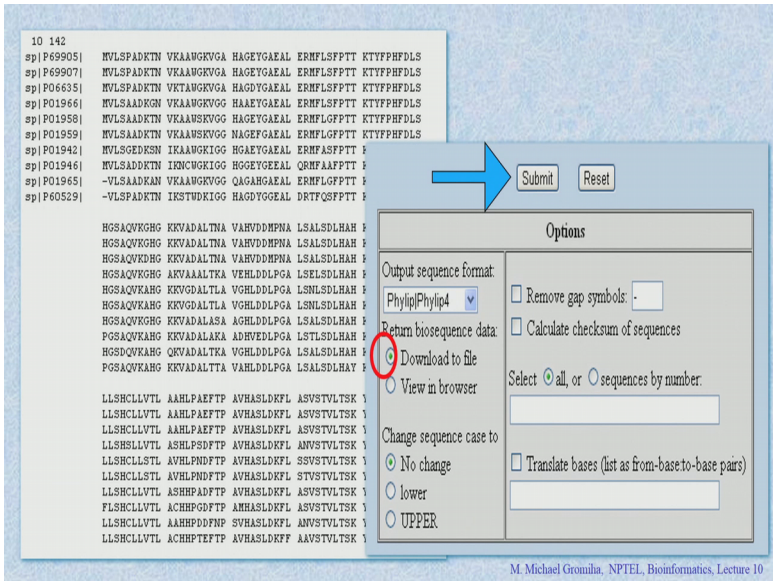
So, we do this right, it will ask for the which Phylip version you want, based on the the version if you submit then you can save the file.

(Refer Slide Time: 21:55)



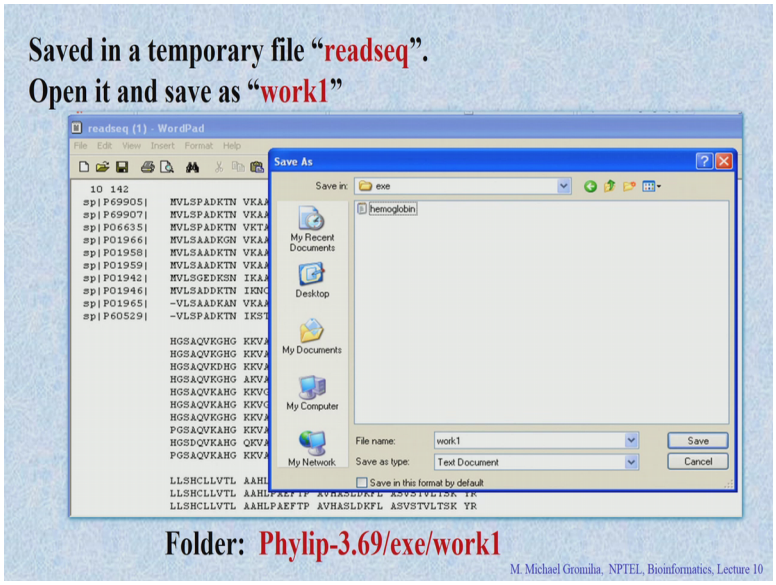
This is the Phylip format right, this different from the MAFFT format. So, these are your sequences, they are aligned for the Phylip right.

(Refer Slide Time: 22:05)



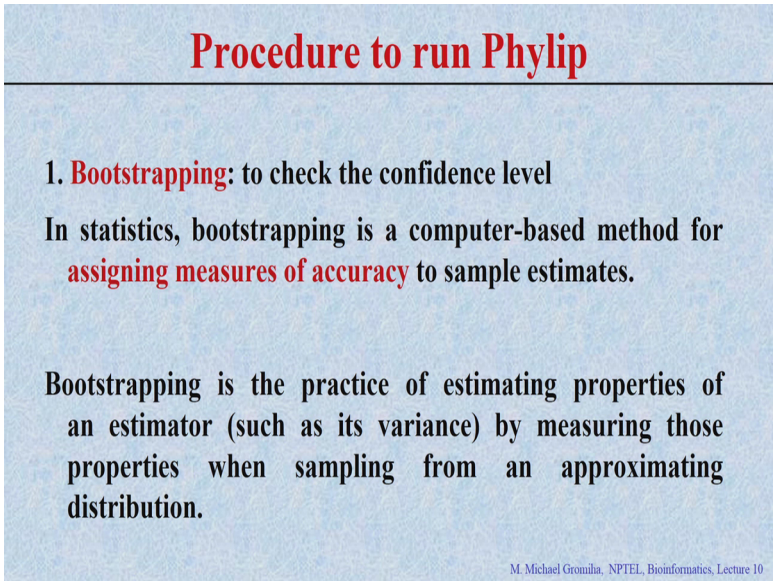
And you can save this right, you can download the file right, you can save the file now. So, you have the input file for Phylip now.

(Refer Slide Time: 22:08)



Now, next one is you need to run Phylip to construct the trees.

(Refer Slide Time: 22:15)



So, to run any of these programs and if you have to check whether your results are significant or not. For example, if you give 10 sequences, it will construct a tree right and what will happen if there is a change right? Whether the program depends upon these sequences are also, this different from the new set of sequences or completely randomized sequences. Because you had 10 sequences you will get a tree, if you completely randomize a sequence, there also you get a tree right. Your tree is the same as randomized tree or it is unique for your sequences. If it is unique then what will you infer?

Student: significant.

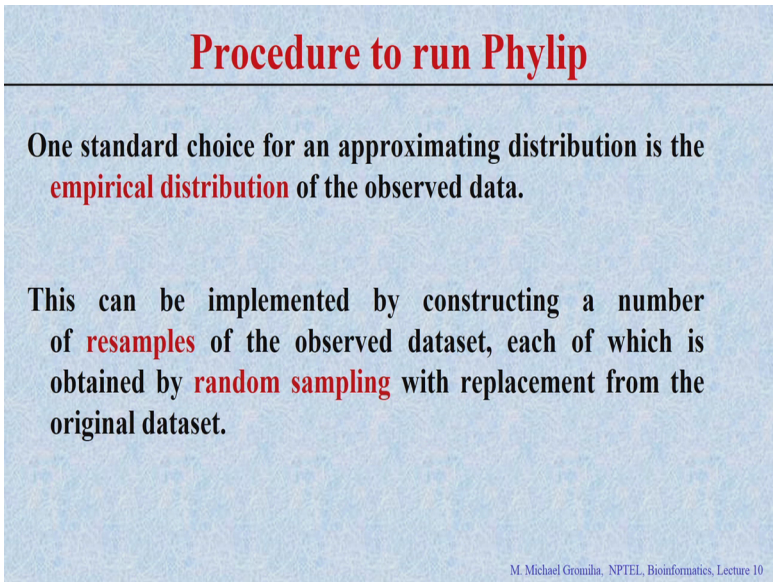
You need significant because you will get a unique tree. So, that is good for only your sequences. If you have tree and the randomized tree are the same, then what will you infer?

Student: not significant

Is not significant because it could be possible by random, and exactly one and two are close to each other, even if you take any random sequences one and two will be close to each other right. For in this case they use a method called bootstrapping, to increase the confidence level, whether your tree is confident or not.

So, in statistics we say computer based method, to assess the measure of accuracy for any your analysis. So, what to do? This is the practice of estimating the properties of the estimator, because here we want to have the proper alignment, such as its variance and so on. From the sampling of independent data right, you can have the various several different data right, and from this sampling, you see whether your data is significant or not.

(Refer Slide Time: 23:55)

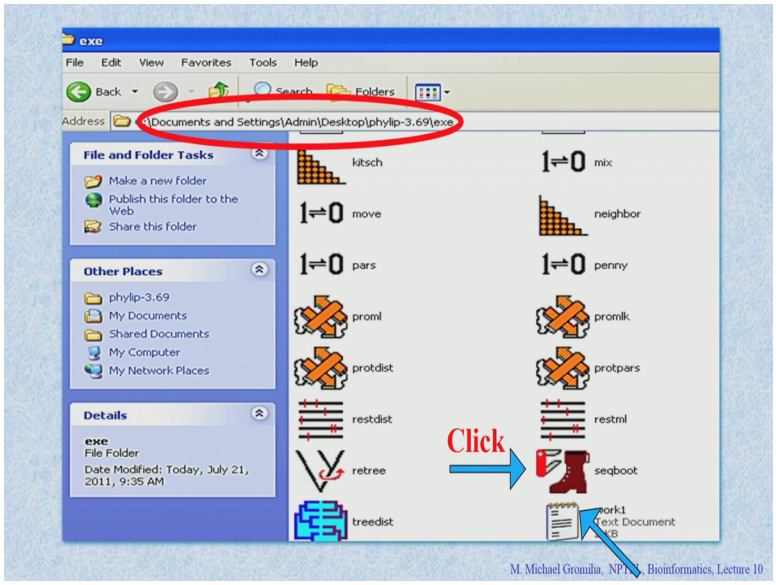


How to do this? See how we do various empirical distribution for example, you can resample. You construct large number of resampling for example, 100 times, 1,000 times, 10,000 times you can resample the data, and from this sample you construct the trees and compare.

For example if you had 10 sequences, if you align you will get 10. Each sequence you can sample many times, 100 times, 1,000 times you can take. Now from the pool you take any 10 and then again you construct. Do it for 1,000 times and 10,000 times and then see how many times you get the same two sequences are aligned together right. If it is completely random you get very random distribution.

If two are really close related, then always you get this close to each other. I will show you how to do this.

(Refer Slide Time: 24:41)



So, in this case first we have to do the bootstrapping. So, that is the when you download Phylip and when you install Phylip, you will get all these files. So, one is the bootstraping method here, this is your input file work on we saved in the previous one.

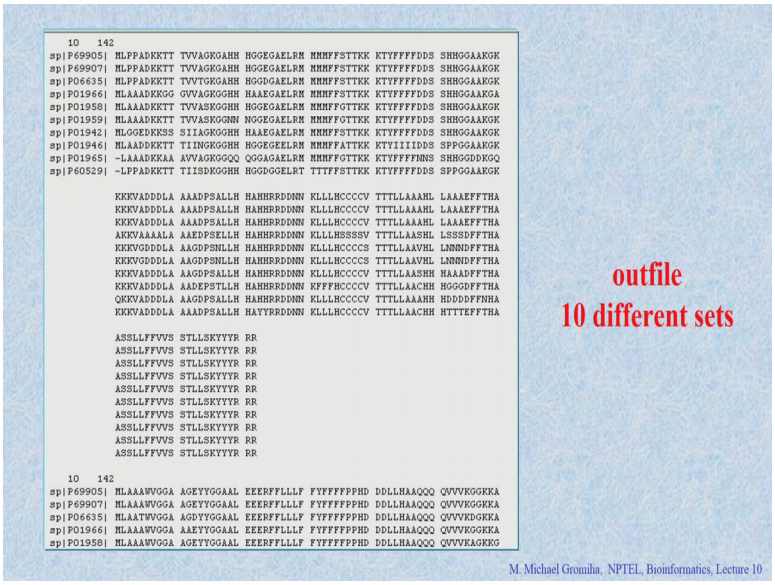
(Refer Slide Time: 25:02)

So, you get this bootstrapping right, it will ask for the input. So, how many replicates do you want? We gave here 10 sequences, how many replicates you need to get for each sequence? In 100 or 10,000 1,000 anything you write right. If you want also it ask for the different options. So, you have give any weight or any characters or you can see the

sequences, which type of settings you want, bootstrap or jackknife or whatever, right here you put the bootstrap. So, what are the sampling procedure did you want right this is a regular sampling procedure, and here replicates right. If you want to accept you put Y, but if you want to change just you can change right. You have Y to accept right and type the letter for one to change anything any letter you can change. If you want to change replicate R you put R right then you will you can change you put R then you have to change the number of replicates. Whatever you want to change, put this letter then accordingly you can change fine.

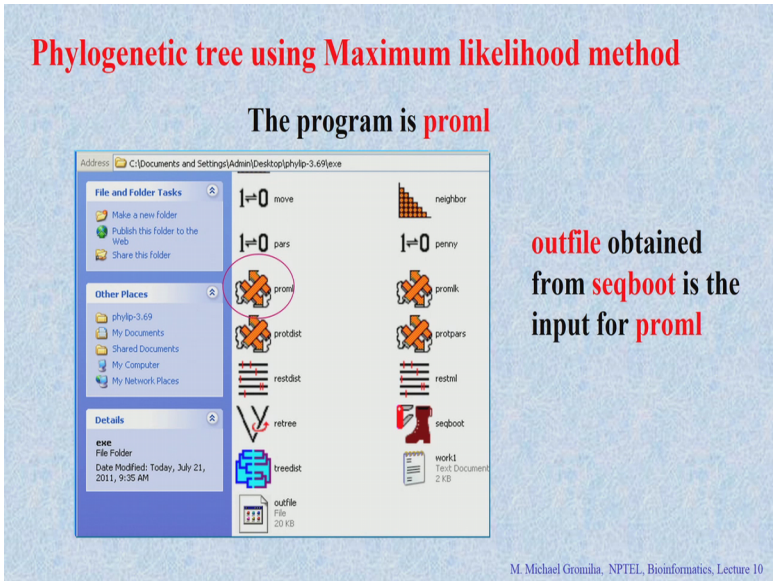
So, and if you change the replicates then accept right, then the they ask for a random seed that is for the programming purpose right and finally, it will get if you put 10 replicates, it generate 10 replicates. Then output is written this file right you can see this now the outfile if you open the outfile it contains 10 replicates okay.

(Refer Slide Time: 26:14)



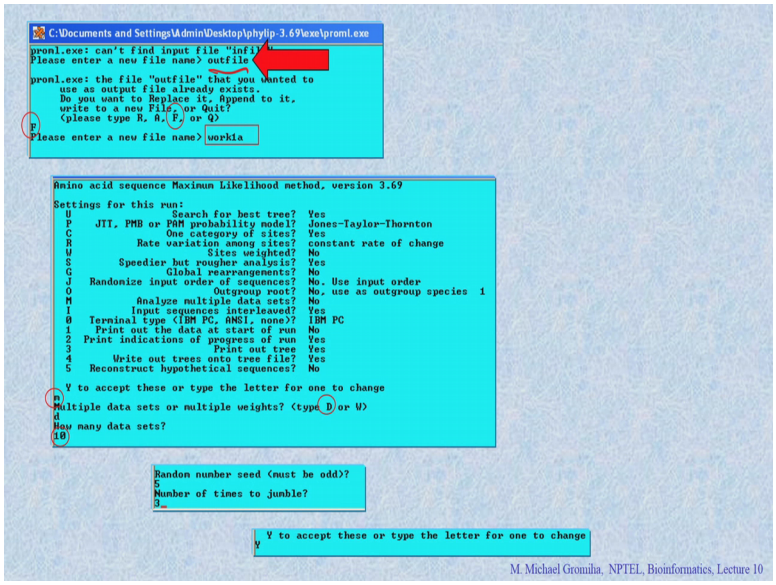
So, here you can see a 10 replicates right, the outfile contains 10 different sets of or each of your sequences right. Now you have to use the method, now you have the bootstrap for the bootstrapping you did sampling and you get a lot of your replicates.

(Refer Slide Time: 26:21)



Now, there different ways to get the tree right. So, one is the maximum likelihood, this is the one which considers the substitutions. So, this is the proml, this is a program which can run for the maximum likelihood method. So, go with this one. So, the out the outfile you obtain from the bootstrapping, this can be the input to the proml. So, you do not have to do anything.

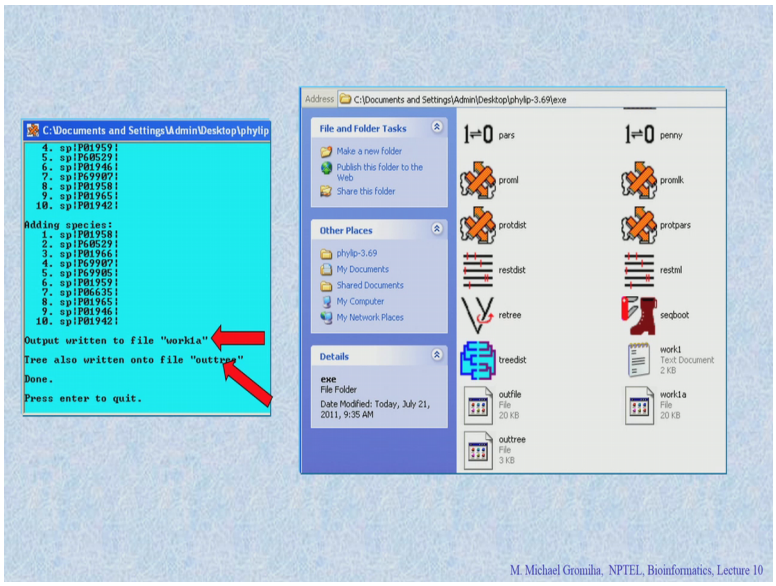
(Refer Slide Time: 26:54)



So, run this one, giving the input as this output out file. So, you can see this outfile as a input, then you can write the what the file name which one we need to save.

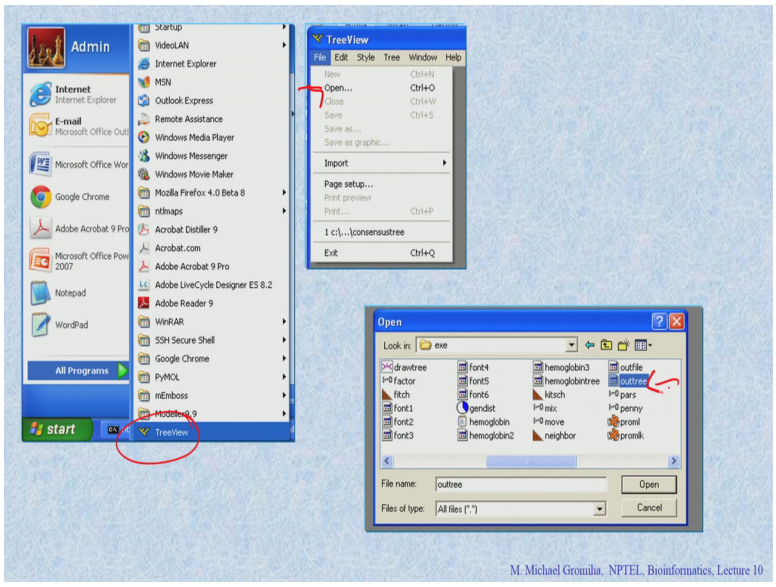
So, you give there a file name, right then again they ask for the same question, you that you want to in change anything; if you do not want to change just you put to Y. So, if you want to change, you can give the letters appropriately and you can change. How many times you want to jumble this sequence again right. So, you can also do that.

(Refer Slide Time: 27:19)



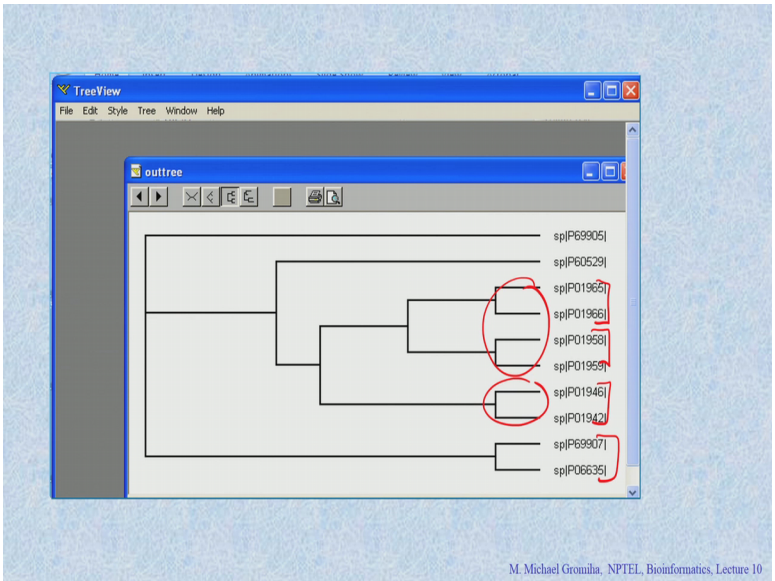
And finally, you can get these things. Okay with there are they got the trees and they wrote in this outtree this output we can see here right, file it is saved ok.

(Refer Slide Time: 27:32)



Now, you can have the tree and you can view the tree using this program called TreeView right. This you can work with the Windows system. So, you go to TreeView, open it, then open the file or go to the open here right, and here open this file name, here outtree is the filename, if you open this you will get the trees.

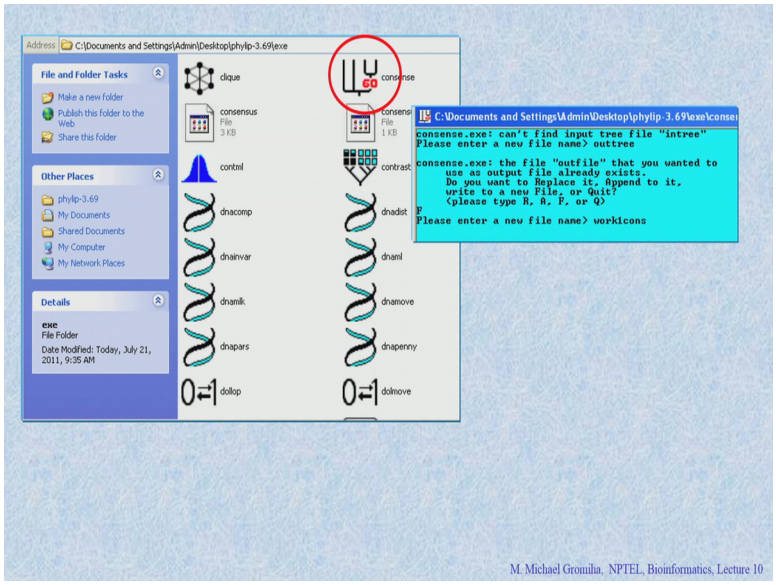
(Refer Slide Time: 27:51)



So, the 10 different cases. So, you can see the trees. So, there is which two are close to each other? See these two are close to each other, and these two are close to each other, these two are close to each other, these two are close to each other, and these two, these two and these two are again these are close to each other, this line right. So, you can see the lineage between among these different sequences.

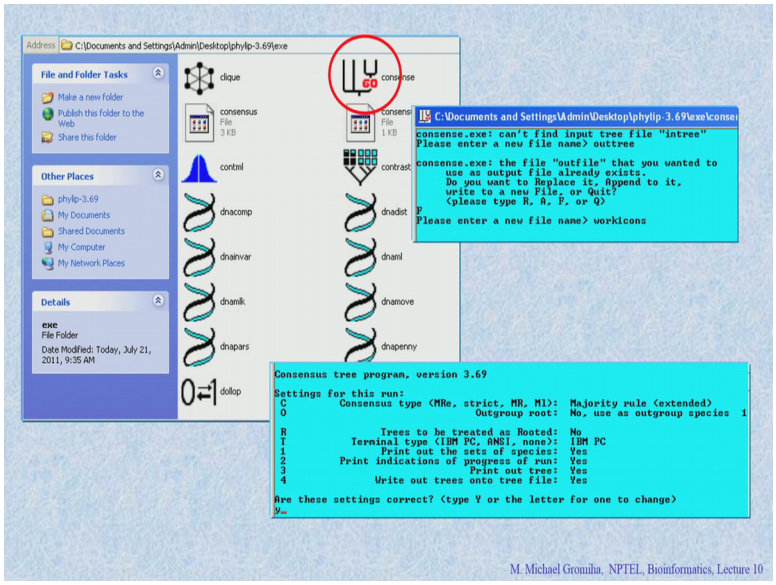
Now, the question is how far you are confident that these two are close to each other.

(Refer Slide Time: 28:22)

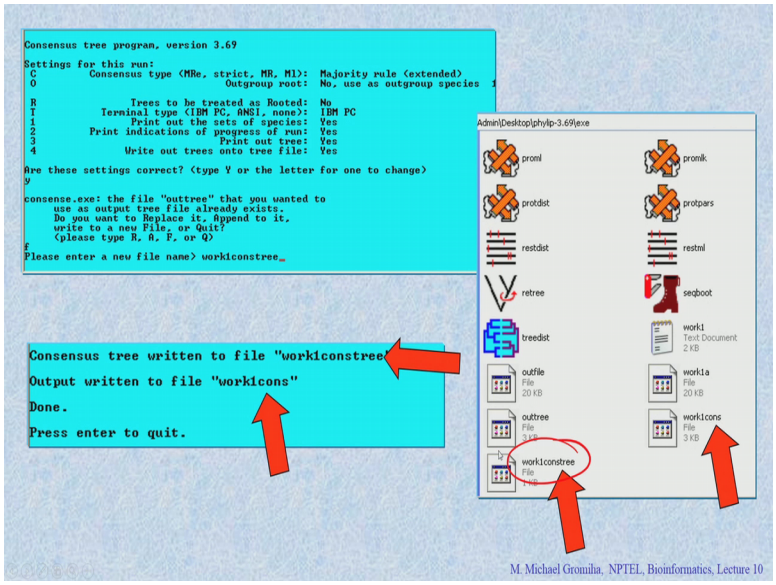


So, for this case you can go to consense tree right. So, go with this a work1consensus, this is your file.

(Refer Slide Time: 28:29)

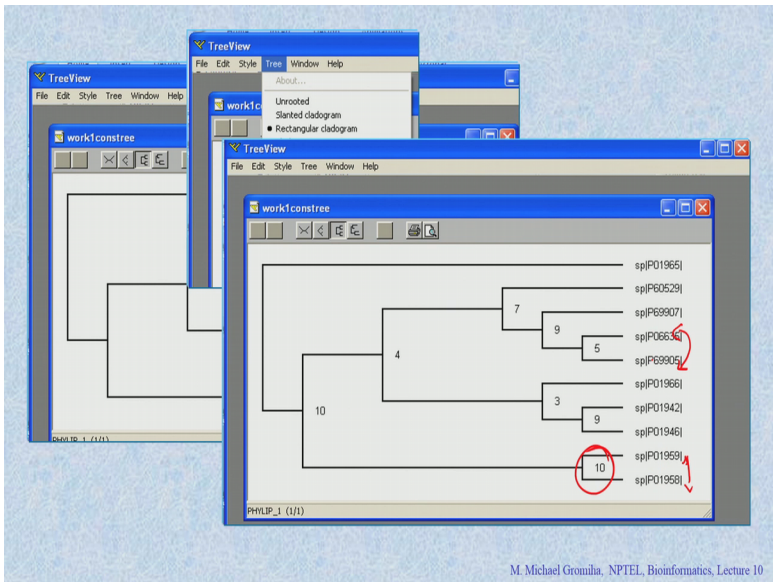


(Refer Slide Time: 28:30)



So, finally, if you give the files right finally, you can get this a file work1cons right.

(Refer Slide Time: 28:37)



Now, if you get the tree this show as tree. So, here is that option called rectangular cladogram, if you click on that then we will get this with numbers. From this one you can what is meaning of these different numbers?

Student: (Refer Time: 28:51).

Right, what is the significance? If this is a 10, these are 10 options out of 10 you all the 10 you these two are identical to each other. In this case between these two the possibility is only 50 percent, between these two, the possibility is 100 percent. So, here is more confident that these two are close to each other, compared with these two are close to each other. Likewise you have the numbers this will tell you the closest sequences as well as how confident you can see that these two sequences are close to each other.

So, in summarize, what did we discuss in this class?

Student: Phylogenetic

Construct your trees. What is the meaning of tree, it will give you the information regarding?

Student: Phylogenetic relationship.

Right, relationship among the different sequences and how far the time taken to evolve from one organism to different organisms right. There are different ways to construct the trees; what is the most common method right?

Student: UPGMA.

UPGMA method right. What is the input for the UPGMA method?

Student: Sequence

Sequences; sequences, who which information they obtain from the sequence?

Student: Distance.

Distance right. They take the mismatches and using the mismatches they will construct a trees. A lot of other methods also available for constructing trees right and the maximum likelihood method is one of the most widely used methods, because that uses the information regarding the they.

Student: (Refer Time: 30:10).

Characteristic of nucleotides or amino acids; What is the program we discuss to construct a trees?

Student: Phylip.

Phylip, like what is the input for the Phylip?

Student: Multiple sequence.

Multiple sequence alignment you can use MAFFT to get the multiple sequence alignment right and then you can construct the trees right, and you can also validate using bootstrapping method right fine right.

So, far we discussed various aspects for example, sequence alignment, conservation and the trees and so on. Next classes, we will discuss about the different parameters or different properties or different features, which can be derived from this amino acid sequences and how these features or the properties will be useful to understand this structure and function.

Thank you very much.