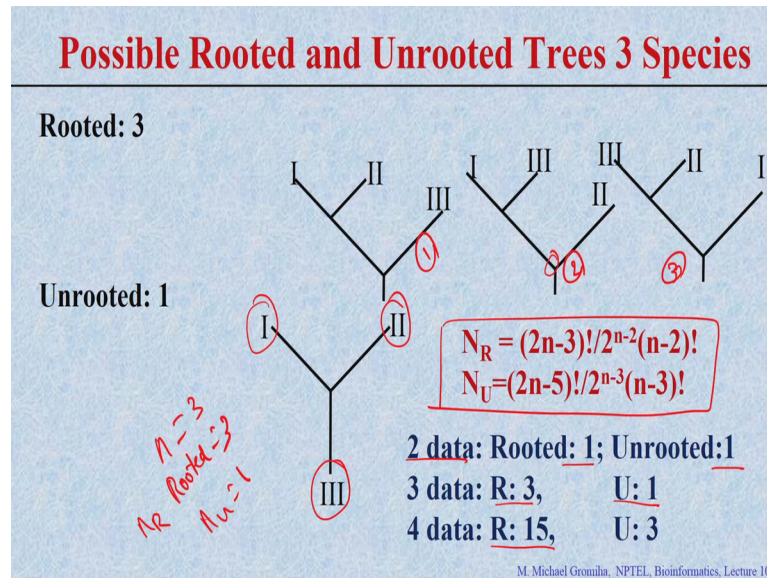


Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 10b
Phylogenetic Trees II

(Refer Slide Time: 00:18)



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

(Refer Slide Time: 00:32)

Tree Construction

1. UPGMA (Unweighted Pair Group Method with Arithmetic mean)
2. Transformed Distance Method
3. Neighbor's Relation Method
4. Neighbor Joining Methods
5. Maximum Likelihood Approaches

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

This is a unweighted pair group with arithmetic mean right this is where they simplified as UPGMA method it is very common method, and even easily understand how to construct the trees using UPGMA method. It is a good understanding, but it has some way some issues some disadvantages. So, rectify that one, that are several other methods have been proposed that is transformed distance method, neighbors 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 use we can use DNA sequences or proteins protein sequences. So, they look at these sequences and see how they are different from each other.

(Refer Slide Time: 01:37)

UPGMA

Statistically based method

Requires data that can be condensed to genetic distance (distance matrix)

E.g. Species, A, B, C and D

Species	A	B	C
B	d_{AB}	-	-
C	d_{AC}	d_{BC}	-
D	d_{AD}	d_{BD}	d_{CD}

d_{AB} : the number of mismatching nucleotides (divided by total number of sites, where matches could have been found)

① ACCTG

② AGCAG

$d_2 = 2$

M. Michael Gronulla, NPTEL, Bioinformatics, Lecture 10

They calculate the distance right it is a statistically based method; they use a 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 the we know the only the sequences ACCTG this is you are at a consequence number one, you can sequence number two you can see AGCAG . So, got 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 true distance difference between A and B or one and two one comma two that is equal to 2. So, for example, if you take the species ABCDE four species right we have put ABC, ABCD.

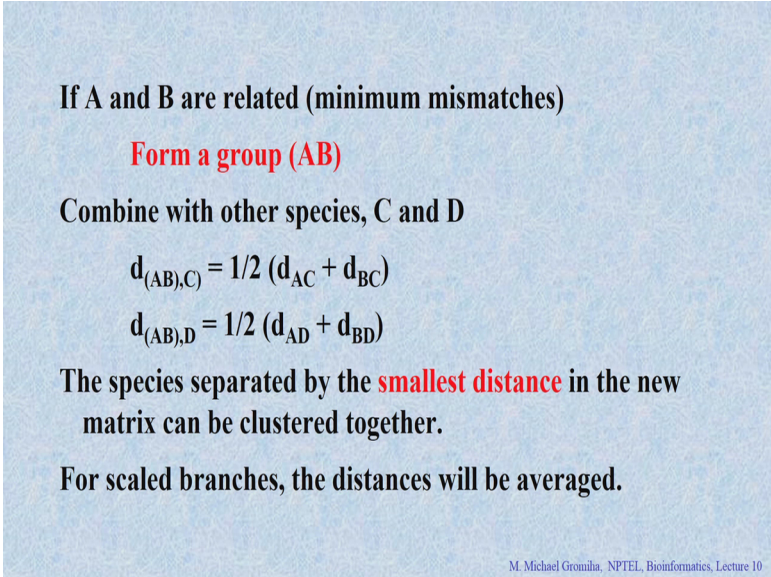
So, you calculate the distance between A and B let us put d_{AB} and A and C is the AC and A and D is AD right this is a same because B and B are the same. So, there is a 0 and B

and C already include here right. So, this way they put dash here likewise B and C you can get here B and D and C and d. So, 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)



If A and B are related (minimum mismatches)

Form a group (AB)

Combine with other species, C and D

$$d_{(AB),C} = 1/2 (d_{AC} + d_{BC})$$
$$d_{(AB),D} = 1/2 (d_{AD} + d_{BD})$$

The species separated by the **smallest distance** in the new matrix can be clustered together.

For scaled branches, the distances will be averaged.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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 equal to a B with D right. So, you consists 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 C which are which two are close to each other likewise you 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)

Example

A:	GTGCTGCACG	GCTCAGTATA	GCATTACCC	TCCATCTTC	AGATCCTGAA
B:	ACGCTGCACG	GCTCAGTGGG	GTGCTTACCC	TCCATCTTC	AGATCCTGAA
C:	GTGCTGCACG	GCTCGGCGCA	GCATTACCC	TCCATCTTC	AGATCCTATC
D:	GTATCACACG	ACTCAGCGCA	GCATTGCCC	TCCGTCTTC	AGATCCTAAA
E:	GTATCACATA	GCTCAGCGCA	GCATTGCCC	TCCGTCTTC	AGATCTAAA

Handwritten notes:
No. of mismatches
AB, AC, AD, AE
BC, BD, BE
CD, CE, DE

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

So, now you have a these five sequences ABCDE right this is DNA sequences for the five species ABCD and E. Now we need to construct a tree which parameter you have to calculate?

Student: Distance.

Distance; so how many distance you have to calculate? Distance between?

Student: Between all possible.

A and B, A and C, A and D A and E likewise B C.

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.

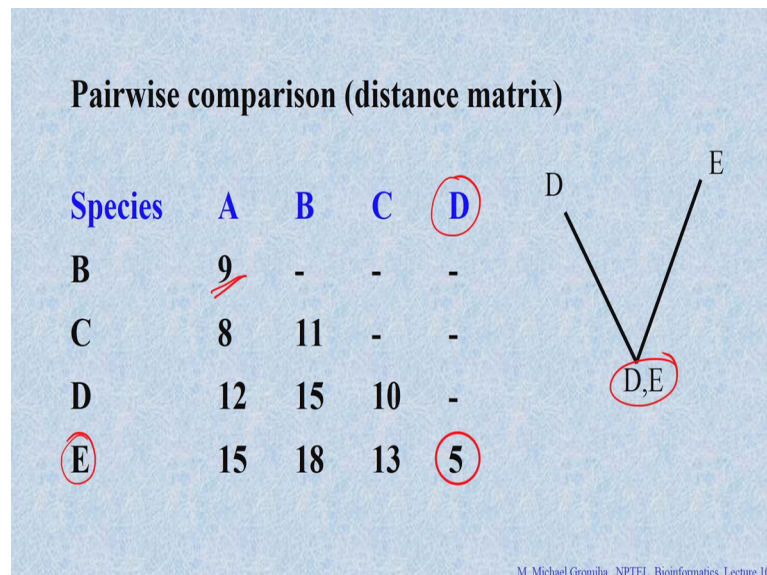
Student: (Refer Time: 05:15).

Based on the distance right for actually if it 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.

9I put 9.

(Refer Slide Time: 05:38)



So, if you see these 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 de if you see D and DE. So, there are five mismatches; 1, 2, 3, 4, 5. So, first 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 all the combinations, this is the 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).

A combine this DE with all other species A B and C. So, what to do this? We take the average right for actual if you C B and D a we do not make any changes, because we need to compare D and E.

(Refer Slide Time: 06:49)

Pairwise comparison (distance matrix)

Species	A	B	C	D	E
B	9	-	-		
C	8	11	-		
DE	13.5	16.5	11.5		

$(AC), B = \frac{1}{2}(AB+BC)$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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

Student: 13.5.

13.5 right with respect to A and D and de 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.

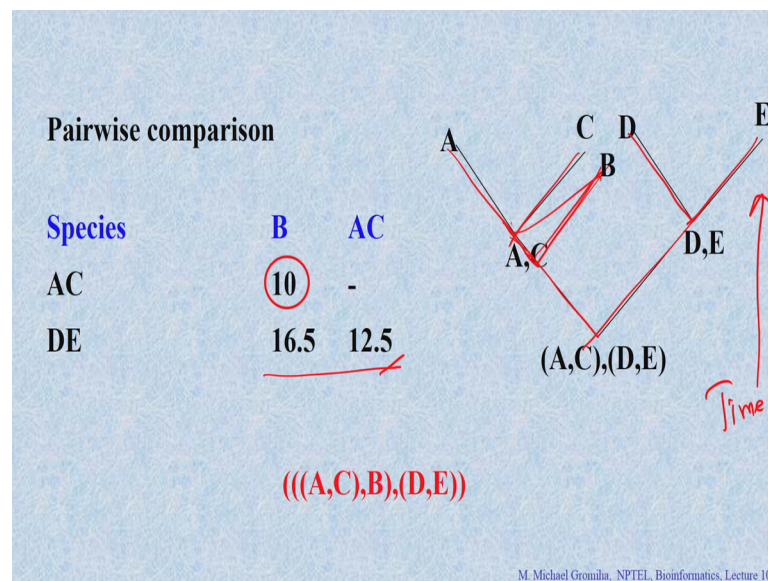
Student: (Refer Time: 07:28) 8.

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 Care 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: (Refer Time: 07:57).

You have to combine this AC with others. So, you put the AC here and the B and D AC D (Refer Time: 08:08) 3 B AC and de because A and C are 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 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 1.

Student: 10.

Ten is the lowest one, so AC with the B. So, we have the AC here common AC right with the AC we can connect with the 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 for 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 organism.

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

(Refer Slide Time: 10:14)

Estimation of branch lengths

Length of the branches can also be calculated with distance matrix

Pairwise comparison (distance matrix)

Species	A	B	C	D
B	9	-	-	-
C	8	11	-	-
D	12	15	10	-
E	15	18	13	5

Handwritten red notes:
A vertical line with 'D' at the top and 'E' at the bottom. The segment from D to a point is labeled '2.5', and the segment from that point to E is labeled '2.5'.

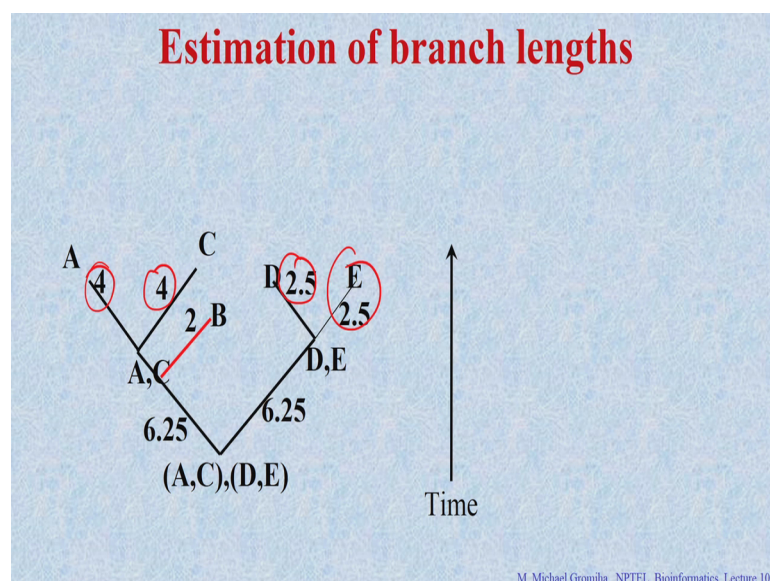
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

This is how to calculate this is a pairwise comparison, we can check 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 were the same time, with this is branch is 5 then we do add 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, that 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, totally be 10 then AC with the B equal to 10 and then totally if you see the A to E at least 15. So, they give this is 2.5, if you are 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, at the principle used in the UPGMA method?

Student: Distance (Refer Time: 12:00).

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

(Refer Slide Time: 12:07)

Neighbor's Relation Method

Another popular variant of UPGMA: tree is constructed with the smallest possible branch lengths overall.

Any unrooted tree, pairs of species that are separated from each other by just one internal node are said to be neighbors.

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 that 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)

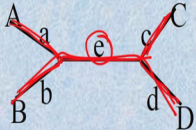
Neighbor's Relation Method

If additivity holds good:

$$\underline{d_{AC} + d_{BD}} = d_{AD} + d_{BC} = \underline{a + b + c + d + 2e}$$

$$= \underline{d_{AB} + d_{CD} + 2e}$$

a, b, c, d: lengths of terminal branches
e: length of central branch



$[d_{AE} = 4 + 6.25 + 6.25 + 2.5 = \underline{19}]$

Actual case: 15

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

How to do this it is a unrooted tree? So, A B; A here to 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 2 e; that means, A B plus C D and 2 e and here you can see the discrepancy between this UPGMA method and this method they considered this also in the cord.

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

(Refer Slide Time: 14:06)

Four point condition:

$$\underline{d_{AB} + d_{CD}} < \underline{d_{AC} + d_{BD}}$$

and

$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

Considers all possible pairwise arrangements of four species and determines the arrangement, which satisfies the four point condition.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

They have made a four point conditions. If you put AB plus C D that should be less than AC plus B D A 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 ok.

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)

For four species, considers all possible values,
(i) $d_{AB} + d_{CD}$; (ii) $d_{AC} + d_{BD}$ and (iii) $d_{AD} + d_{BC}$
Smallest sum with pairing is 1 and others are 0

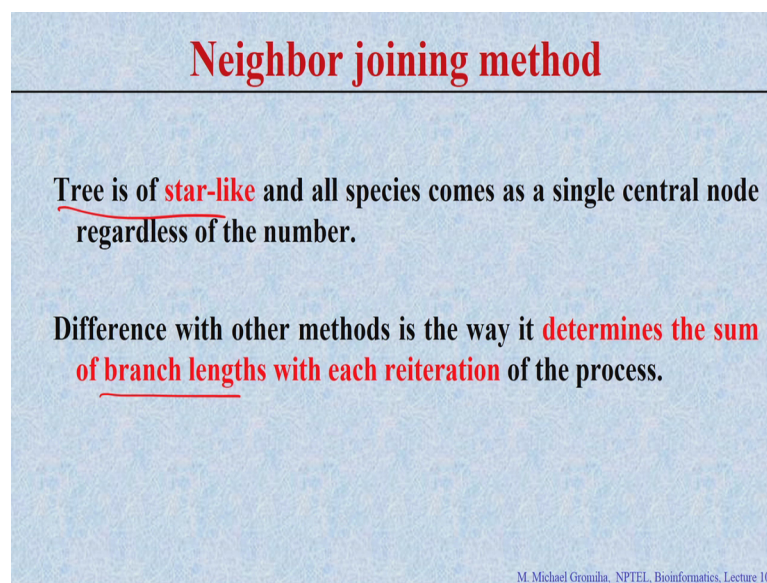
Repeat for all possible four pairs
Ones with highest scores are grouped.
New distance matrix can be generated as was done for UPGMA

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

Now, I have this one for I can have four species, you have different values you can calculate $AB + CD$ and $AC + BD$ and $AD + BC$ 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 UGPMA method to get the distance. Then along with the UGPMA method, they considered special conditions to derive these trees right.

(Refer Slide Time: 15:18)



Neighbor joining method

Tree is of star-like and all species comes as a single central node regardless of the number.

Difference with other methods is the way it determines the sum of branch lengths with each reiteration of the process.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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 because less of this any of these numbers. So, they 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)

Neighbor joining method

$$S_{12} = (1/(2(N-2)) \sum (d_{1k} + d_{2k}) + (1/2)d_{12} + (1/(N-2)) \sum d_{ij}$$

Any pair of species can take positions 1 and 2; k is an accepted outgroup

$$\text{Simplified into } Q_{12} = (N-2)d_{12} - \sum d_{1i} - \sum d_{2i}$$

All possible pairs are considered and the pairs with smallest distance is taken.

Construct new distance matrix as done with UPGMA and repeat the process.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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)

Maximum likelihood approaches

It represent an alternative and purely statistically based method of phylogenetic reconstruction.

Probabilities are considered for every individual nucleotide substitution.

Transitions (purine to purine/ pyrimidine to pyrimidine) and transversions

Purine → Purine A A A
 T C G

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

Then doing one by one then recently they had developed another method right what would we use UGPMA 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 that 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 methods which were 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 transitions 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 say 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 are commonly different residues they give weightage, accordingly they can also develop a tree right, different ways to construct for phylogenetic trees. So, in this case if there are multiple substitutions right may be independent or sometimes independent right that also we can you can take into cons consideration right.

(Refer Slide Time: 18:04)

Maximum likelihood approaches

Multiple substitutions occurred at one or more sites, which are not necessarily independent.

It is necessary to take into account of all these facts, which needs heavy computational power.

With current facilities, it is possible to use the method for tree construction.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

But to take all these aspects it requires high computational power, what is 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 a 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)

Program to construct trees

Phylip

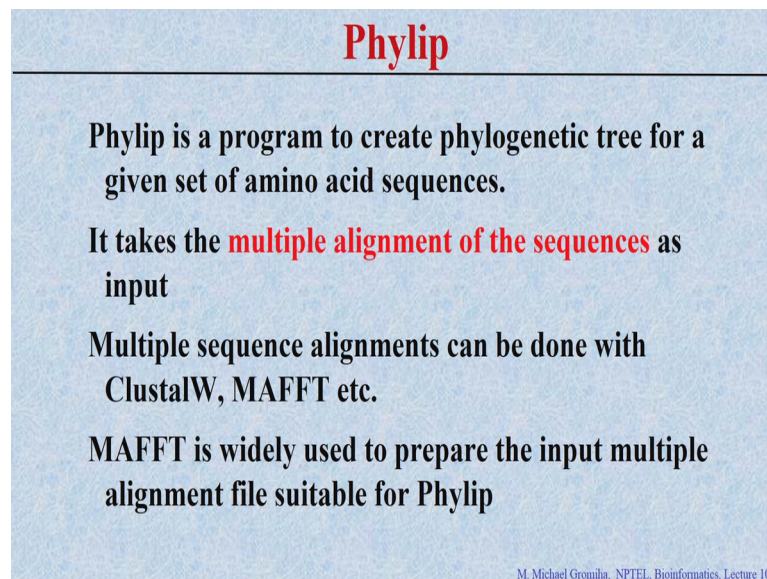
as Windows executables (not counting executing in a "DOS box"). Programs available as source code which is Windows-specific are listed below under Unix. (Note that compilers available on Windows systems, particularly the free Cygwin and MinGW compilers, can also be used to compile generic source code). Programs run in interpreted environments such as Perl, Python, R or MATLAB can also be run under Windows if the programs are listed above under Unix.

• PHYLIP	• DNASIS	• Mesquite	• MrModeltest	• MESA
• PAUP*	• MINSIPNET	• PhylEdit	• SymmeTREE	• MultiPhd
• TREECON	• BioEdit	• SYN-TAX	• TreeJuxtaposer	• NnmbTree
• GDA	• ProSeq	• FFP	• Network	• ArboDraw
• SeqPup	• PAL	• DIVA	• Spectronet	• SPAGeDi
• MOLPHY	• WINCLADA	• TreeFitter	• Phylogen	• CRCAAnalyzer
• GeneDoc	• NONA	• Phylo_win	• Phylap	• DualBrothers
• COMPONENT	• Phylogenetic Independence	• SplitTree	• Dnatrex	• PaupUp
• TREEMAP	• PERLE	• PAST	• Ima2	• Ntng
• COMPARE	• HY-PHY	• GeneStudio_Pro	• ProtTest	• SSA
• RAEDistance	• TreeExplorer	• Treefinder	• GEODIS	• Multidivtime
• TreeView	• Genie	• PPH	• TreeSetVis	• ParaFit
• Phylo dendron	• Vanilla	• MetaPIGA	• TreeMe	• IDIC
• POPGENE	• MEGA	• Phyltools	• ModelGenerator	• TreeMaker
• TFPGA	• TNT	• MSA	• Simplot	• CodonRates
• GeneTree	• GelCompar II	• Mgenome	• PHYLORGE	• EakPhy
• MVSP	• BioNumerics	• APE	• ProDist	• CoMET
• ESTCALC	• TCS	• PHASE	• START2	• TreeDm
• Genetic	• FORESTER	• PHYLML	• IQ-TREE	• DigTree
• NJplot	• Populations	• YCDMA	• STC	• Genious
• unrooted	• T-REX	• NSA	• TreeSAAP	• Brownie
• Arlequin	• MrBayes	• BEAST	• Swap	• Mac5
• DAMBE	• EDIBLE	• Clann	• Swap PH	• BayesPhylogenies
• DnaSP	• Winboot	• Jevtrace	• TreeGraph.2	• BayesTraits
• PAML	• r8s	• MMTau	• DIVERGE	• MrEnt

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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 a 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 they will take few minutes to just demonstrate the functioning of these PHYLIP how to do this.

(Refer Slide Time: 19:08)



Phylip

- Phylip is a program to create phylogenetic tree for a given set of amino acid sequences.
- It takes the **multiple alignment of the sequences** as input
- Multiple sequence alignments can be done with ClustalW, MAFFT etc.
- MAFFT is widely used to prepare the input multiple alignment file suitable for Phylip

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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: (Refer Time: 19:21).

In this case in (Refer Time: 19:22 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).

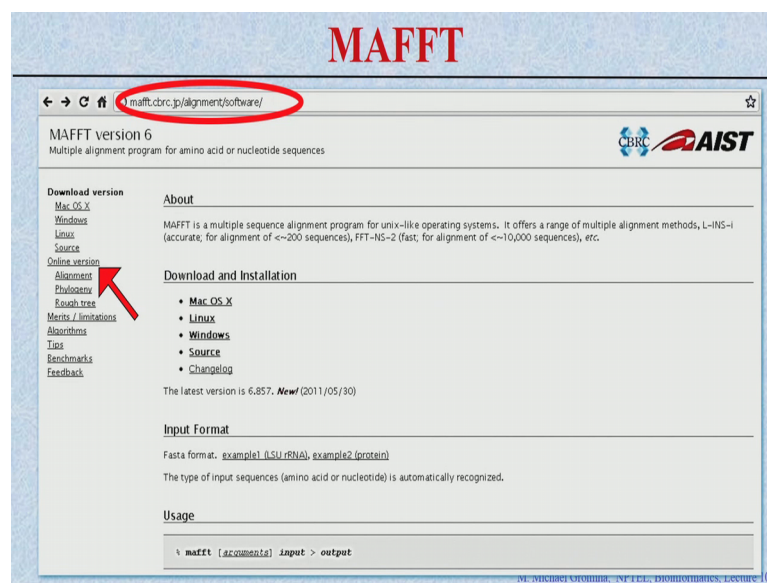
Promols.

Student: (Refer Time: 19:44).

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

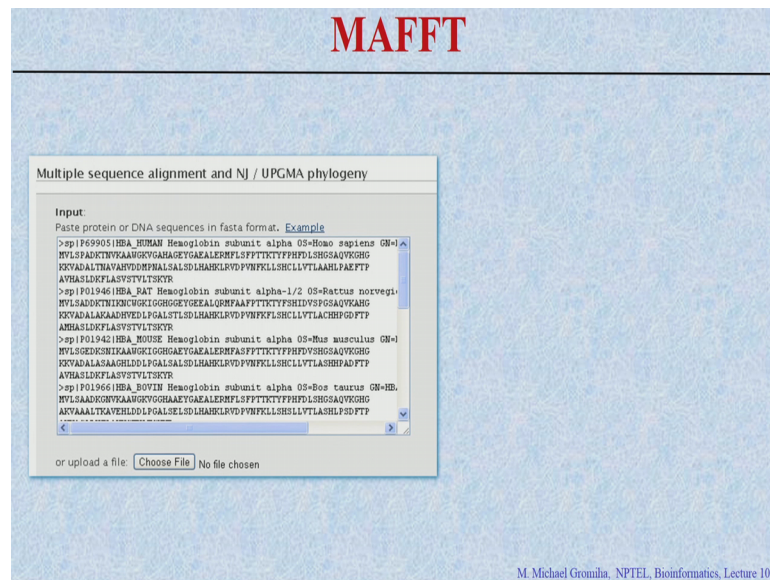
So, it is widely used to prepare the input file suitable for phylip we use MAFFT there is an option to save the file in phylip format. So, you do not have to worry about formatting run MAFFT save the multiple sequence alignment in phylip format and you can give this as input to the to run phylip.

(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.

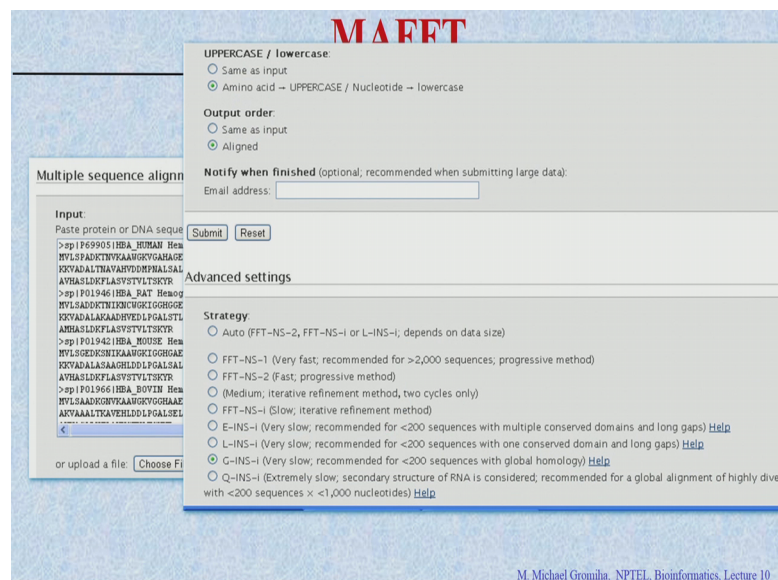
(Refer Slide Time: 20:28)



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 that 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)

MAFFT

UPPERCASE / lowercase:
☐ Same as input
☒ Amino acid → UPPERCASE / Nucleotide → lowercase

Output order:
☐ Same as input
☒ Aligned

Multiple sequence alignment

Input:
Paste protein or DNA sequence
>sp|P69905|HBA_HUMAN HBA
MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
AYHASLDKFLASVSTVLTSTR
>sp|P01942|HBA_MOUSE HBA
MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
AYHASLDKFLASVSTVLTSTR
>sp|P01946|HBA_BOVIN HBA
MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
AYHASLDKFLASVSTVLTSTR
or upload a file:

Notify when finished (optional: recommended when submitting large data):
Email address:

Submit

Advanced settings

Strategy:
☐ Auto (FFT-NS-2, FFT-NS-1, FFT-NS-2 (Fast), FFT-NS-2 (Medium, iterative), FFT-NS-1 (Slow, iterative), E-INS-i (Very slow), L-INS-i (Very slow), G-INS-i (Very slow), Q-INS-i (Extremely slow) with <200 sequences x

Parameters:
Scoring matrix for amino acid sequences:
Scoring matrix for nucleotide sequences:
! Switch it to '1PAM / k=2' when aligning closely related DNA sequences.
Gap opening penalty: (1.0 - 3.0)
Offset value: (0.0 - 1.0)
! If long gaps are not expected, set it as 0.1 or larger value.

MAFFT-homologs (Collects homologs from SwissProt by BLAST and performs profile-based alignment)
☐ On
☐ Show homologs (if any)
Number of homologs: (5 - 200)
Threshold: (1e-5 - 1e-40)

Plot LAST hits (DNA only):
☒ The top sequence vs the others ☐ The longest sequence vs the others
☒ Plot and alignment ☐ Plot only ☐ Alignment only
Threshold:

Submit

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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)

MAFFT Results

to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, etc. with Readseq

MAFFT-G-INS-i Result

CLUSTAL format alignment by MAFFT (v6.857b)

```
sp|P69905|HBA_H MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P69907|HBA_P MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P06635|HBA_P MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P01966|HBA_B MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P01959|HBA_H MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P01959|HBA_E MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P01942|HBA_H MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P01946|HBA_B MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P01965|HBA_P MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
sp|P060529|HBA_C MVLSPADKTNVKAAGKVGAGHAGEYGAELERFLSFTTKTYFPHFDSLHSGSAQVKGHG
*** ** *
sp|P69905|HBA_H KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P69907|HBA_P KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P06635|HBA_P KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P01966|HBA_B KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P01959|HBA_H KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P01942|HBA_H KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P01946|HBA_B KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P01965|HBA_P KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
sp|P060529|HBA_C KRVADALTNVAHVDDMPNALSALSDLHAHKLKLVDPVNFKLLSHCLLVTLAAHLP AEFTP
*** ** *
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

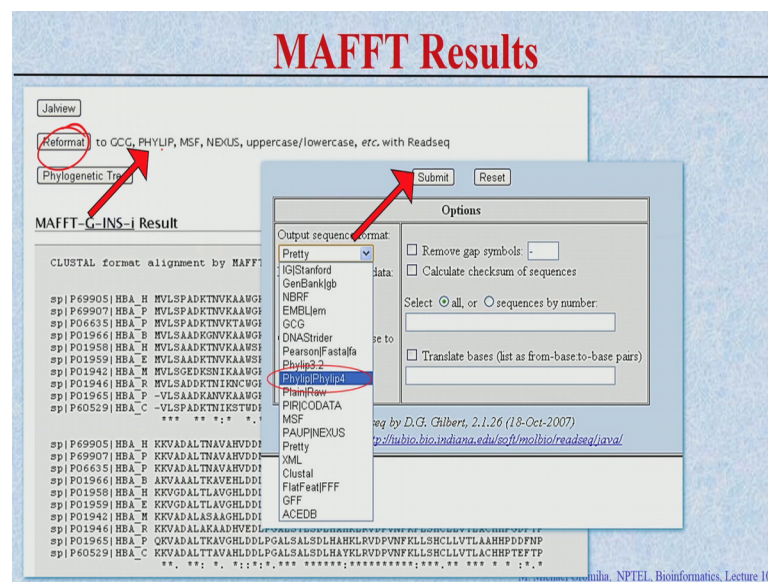
So, you will get the result this is a multiple sequence alignment.

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

Student: Philip

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

(Refer Slide Time: 21:46)



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

(Refer Slide Time: 21:55)

```

10 142
sp|P69905| NVLSPADKTN VKLAAGKVGGA HAGEYGAEL EMFLSFPTT KTYFPHFDSL
sp|P69907| NVLSPADKTN VKLAAGKVGGA HAGEYGAEL EMFLSFPTT KTYFPHFDSL
sp|P06635| NVLSPADKTN VKTAGKVGGA HAGDYGAEL EMFLSFPTT KTYFPHFDSL
sp|P01966| NVLSAADKGN VKAAGKVGQG HAEGYGAEL EMFLSFPTT KTYFPHFDSL
sp|P01958| NVLSAADKTN VKAAGKVGQG HAGDYGAEL EMFLGFFT KTYFPHFDSL
sp|P01959| NVLSAADKTN VKAAGKVGQG HAGEYGAEL EMFLGFFT KTYFPHFDSL
sp|P01942| NVLSGEDENK IKAAGKIGG HAGEYGAEL EMFASFTT KTYFPHFVDS
sp|P01946| NVLSADKTN IENCKGKIGG HGGEGEAL QRHAAFTT KTYFSHIVDS
sp|P01965| -VLSAADKAN VKAAGKVGQG QAGAGHAEL EMFLGFFT KTYFPHFMLS
sp|P60529| -VLSADKTN ISTWDKIGG HAGDYGEAL DRTQSFPTT KTYFPHFMLS

HGSQAQVRGSG KKVADALTHA VARVDMPNHA LSAISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVADALTHA VARVDMPNHA LSAISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVADALTHA VARVDMPNHA LSAISDLHAH KLRVDPNPDK
HGSQAQVRGSG ARVAAALTEA VEHLDDLPGA LSEISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVGDALTLA VOHLDDLPGA LNSISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVGDALTLA VOHLDDLPGA LNSISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVADALASA AGHLDDLPGA LSAISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVADALAKA ADHVEDLPGA LSTISDLHAH KLRVDPNPDK
HGSQAQVRGSG QKVADALTKA VOHLDDLPGA LSAISDLHAH KLRVDPNPDK
HGSQAQVRGSG KKVADALTTA VAHLDDLPGA LSAISDLHAY KLRVDPNPDK

LLSRLCLVTL AARLPAEFTF AVHASLDRFL ASVSTVLTSK YR
LLSRLCLVTL AARLPAEFTF AVHASLDRFL ASVSTVLTSK YR
LLSRLCLVTL AARLPAEFTF AVHASLDRFL ASVSTVLTSK YR
LLSRLCLVTL ASHLPEQDFT AVHASLDRFL ANVSTVLTSK YR
LLSRLCLVTL AVHLPNDPFT AVHASLDRFL SVSSTVLTSK YR
LLSRLCLVTL AVHLPNDPFT AVHASLDRFL STVSTVLTSK YR
LLSRLCLVTL ASHNPADFT AVHASLDRFL ASVSTVLTSK YR
FLSRLCLVTL ACRHPGDFT AVHASLDRFL ASVSTVLTSK YR
LLSRLCLVTL AARHPDQFNP SVHASLDRFL ANVSTVLTSK YR
LLSRLCLVTL ACRHPTFTF AVHASLDKFF AAVSTVLTSK YR

```

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

(Refer Slide Time: 22:05)

```

10 142
spt|P69905| NVLSPADKTN VKAANGKVG A HAGEYGAEL EMFLSFPTT KTYFPFDLS
spt|P69907| NVLSPADKTN VKAANGKVG A HAGEYGAEL EMFLSFPTT KTYFPFDLS
spt|P06635| NVLSPADKTN VKTANGKVG A HAGDYGAEL EMFLSFPTT KTYFPFDLS
spt|P01966| NVLSAADKNK VKAANGKVG G HAAYGAEL EMFLSFPTT KTYFPFDLS
spt|P01958| NVLSAADKTN VKAANGKVG G HAGEYGAEL EMFLGFFT KTYFPFDLS
spt|P01959| NVLSAADKTN VKAANGKVG G HAGEYGAEL EMFLGFFT KTYFPFDLS
spt|P01942| NVLSGEDSKN IKANGKIGG G HAGEYGAEL EMFLASFTT F
spt|P01946| NVLSADKTN IKNCWKIGG G HGEYGEAL QRMFAFFT F
spt|P01965| -VLSAADKN VKAANGKVG G QAGAGAEAL EMFLGFFT F
spt|P60529| -VLSPADKTN IKSTWDKIGG HAGDYGEAL DRTQSFPTT F

```

Submit

Reset

Options

Output sequence format:

Phyipl/Phyipl4 ▼

Return biosequence data:

☒ Download to file

☐ View in browser

☐ Remove gap symbols: -

☐ Calculate checksum of sequences

Select ☒ all, or ☐ sequences by number:

Change sequence case to

☒ No change

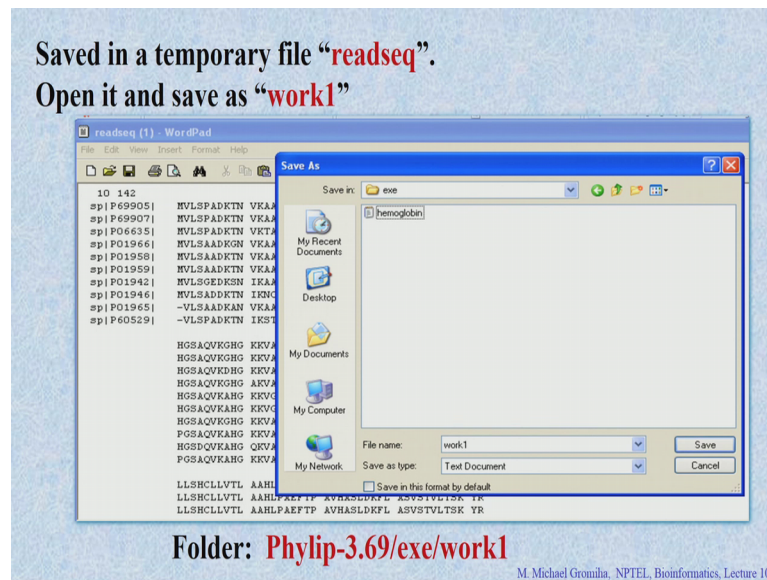
☐ lower

☐ UPPER

☐ Translate bases (list as from-base-to-base pairs)

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 a trees.

(Refer Slide Time: 22:15)

Procedure to run Phylip

1. **Bootstrapping:** to check the confidence level

In statistics, bootstrapping is a computer-based method for **assigning measures of accuracy** to sample estimates.

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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 they are 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: (Refer Time: 22:57) significant.

You need significant because you will get a Unix 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: (Refer Time: 23:10).

Is nothing 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 suggests 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)

Procedure to run Phylip

One standard choice for an approximating distribution is the **empirical distribution** of the observed data.

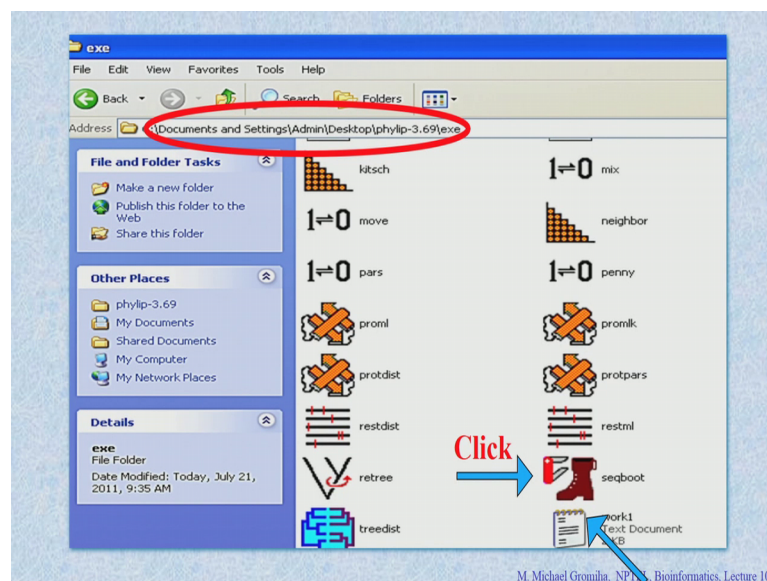
This can be implemented by constructing a number of **resamples** of the observed dataset, each of which is obtained by **random sampling** with replacement from the original dataset.

How to do this? See how we do various empirical distribution for example, you can resamples you construct large number resampling for example, 100 times, 1000 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, 1000 times you can take. Now from the pool you take any 10 and then again you construct. Do it for thousand times and 10000 times and then see how many times you get the same two sequences are aligned together right. Even if it is completely random you do not get you do not get any random distribution.

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

(Refer Slide Time: 24:41)



So, in this case first we have to do the bootstrapping. So, there is the here you can when you download phylip and when you insert phylip, you will get all these files. So, one is the C boot bootstrap ing method here, this is your input file work on we saved in the previous one.

(Refer Slide Time: 25:02)

The screenshot shows the command-line interface of the seqboot.exe program. At the top, a message indicates that the program cannot find the input file 'infile' and prompts the user to enter a new file name, which is 'work1.txt'. The main menu displays the 'Bootstrapping algorithm, version 3.69' and lists various settings for the run. A red arrow points to the 'Block size for block-bootstrapping?' option, which is set to '1 (regular boot)'. Another red arrow points to the 'Number of replicates?' option, which is set to '10'. A third red arrow points to the 'Output written to file "outfile"' message. The bottom of the screen shows the progress of the run, with a list of completed replicates from 1 to 10. A red arrow points to the 'Random number seed (must be odd)?' option, which is set to '5'. The bottom right corner of the slide contains the text 'outfile contains 10 replications'.

```

C:\Documents and Settings\Admin\Desktop\phylop-3.69\seqboot.exe
seqboot.exe: can't find input file "infile"
Please enter a new file name> work1.txt

Bootstrapping algorithm, version 3.69

Settings for this run:
D Sequence, Morph, Rest., Gene Freqs? Molecular sequences
J Bootstrap, Jackknife, Permute, Reurite? Bootstrap
X Regular or altered sampling fraction? regular
B Block size for block-bootstrapping? 1 (regular boot)
R How many replicates? 100
U Read weights of characters? No
C Read categories of sites? No
S Write out data sets or just weights? Data sets
I Input sequences interleaved? Yes
O Terminal type (IBM PC, ANSI, none)? IBM PC
1 Print out the data at start of run No
2 Print indications of progress of run Yes

Y to accept these or type the letter for one to change

Output written to file "outfile"
Done.
Press enter to quit.

Bootstrapping algorithm, version 3.69

Settings for this run:
D Sequence, Morph, Rest., Gene Freqs? Molecular
J Bootstrap, Jackknife, Permute, Reurite? Bootstrap
X Regular or altered sampling fraction? regular
B Block size for block-bootstrapping? 1 (regular boot)
R How many replicates? 10
U Read weights of characters? No
C Read categories of sites? No
S Write out data sets or just weights? Data sets
I Input sequences interleaved? Yes
O Terminal type (IBM PC, ANSI, none)? IBM PC
1 Print out the data at start of run No
2 Print indications of progress of run Yes

Y to accept these or type the letter for one to change
R Number of replicates?
10

Y to accept these or type the letter for one to change
S Random number seed (must be odd)?
5
  
```

outfile contains 10 replications

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

So, you get it 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 1000 anything you write right. If you want all also it ask for the different options. So, you have they 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 if a y to accept right and type the letter for one to change anything any letter you can change. If you want to change replicates you put r right then you will you can change you put r then you have 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 keep the accept right then the they ask for a random sheet that is for the programming purpose right and finally, it will get if you put 10 replicates, it generate 10 replicates. Then output is return this file right you can see this now the outfile if you open the outfile it contains 10 replication ok.

(Refer Slide Time: 26:14)

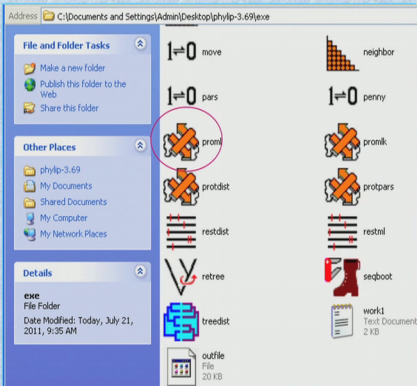
10 142	
sp P69905	MLPFDKTKT TVVAGGAHH HGGEGAELRN HNNFSTTEK KTTTTFFDS SHHGGAARKG
sp P69907	MLPFDKTKT TVVAGGAHH HGGEGAELRN HNNFSTTEK KTTTTFFDS SHHGGAARKG
sp P06635	MLPFDKTKT TVVAGGAHH HGGEGAELRN HNNFSTTEK KTTTTFFDS SHHGGAARKG
sp P01946	MLAADKKG CVVAGGGGHH HAAGAELRN HNNFSTTEK KTTTTFFDS SHHGGAARKG
sp P01958	MLAADKTKT TVVAGGGGHH HGGEGAELRN HNNFGTTEK KTTTTFFDS SHHGGAARKG
sp P01959	MLAADKTKT TVVAGGGGHH HGGEGAELRN HNNFGTTEK KTTTTFFDS SHHGGAARKG
sp P01942	MLGDEKES SIAGGGGHH HAAGAELRN HNNFSTTEK KTTTTFFDS SHHGGAARKG
sp P01946	MLAADKTKT TIHNGGGHH HGGEGAELRN HNNFATTEK KTTTTIDS SPGGGAARKG
sp P01945	-LAADKGA AVVAGGGGCG CGAGAELRN HNNFGTTEK KTTTTFMS SHHGDDGQ
sp P60529	-LPADKTKT TIIDGGGHH HGGGGELST TTTTSTTEK KTTTTFFDS SPGGGAARKG
KKKVADDLA	LAAPFSALLH HAHHRDDNN KLLNLCCCV TTTLAAHL LAAEFFFTHA
KKKVADDLA	LAAPFSALLH HAHHRDDNN KLLNLCCCV TTTLAAHL LAAEFFFTHA
KKKVADDLA	LAAPFSALLH HAHHRDDNN KLLNLCCCV TTTLAAHL LAAEFFFTHA
AKVAAALAA	LAEPFSLLH HAHHRDDNN KLLNLSSSV TTTLAAHL LSSSFFTHA
KKKVGGDLA	LAGPFSNLLH HAHHRDDNN KLLNLCCCV TTTLAAHL LNNDDFFTHA
KKKVGGDLA	LAGPFSNLLH HAHHRDDNN KLLNLCCCV TTTLAAHL LNNDDFFTHA
KKKVADDLA	LAGPFSALLH HAHHRDDNN KLLNLCCCV TTTLAAHL LAAEFFFTHA
KKKVADDLA	LAGPFTLLH HAHHRDDNN FFFPRCTC TTTLAAHL HGGGDDG
QKKVADDLA	LAGPFSALLH HAHHRDDNN KLLNLCCCV TTTLAAHL HDDDFFFTHA
KKKVADDLA	LAGPFSALLH HATYRDDNN KLLNLCCCV TTTLAAHL TTTFFFFTHA
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
ASSLFFVVS	STLKSYYTR RR
10 142	
sp P69905	MLAAAVWGA AGETYGGAAL EEEFFLLFL FYFFFFPPHD DLLLHAAQQQ QVVKGKGA
sp P69907	MLAAAVWGA AGETYGGAAL EEEFFLLFL FYFFFFPPHD DLLLHAAQQQ QVVKGKGA
sp P06635	MLAAAVWGA AGETYGGAAL EEEFFLLFL FYFFFFPPHD DLLLHAAQQQ QVVKGKGA
sp P01966	MLAAAVWGA LAETYGGAAL EEEFFLLFL FYFFFFPPHD DLLLHAAQQQ QVVKGKGA
sp P01958	MLAAAVWGA AGETYGGAAL EEEFFLLFL FYFFFFPPHD DLLLHAAQQQ QVVKGKGG

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)

Phylogenetic tree using Maximum likelihood method

The program is **proml**



The screenshot shows a Windows Explorer window with the address bar set to "C:\Documents and Settings\Admin\Desktop\phylop-3.69\exe". The left sidebar shows the "File and Folder Tasks" and "Other Places" sections. The "Other Places" section lists "phylop-3.69", "My Documents", "Shared Documents", "My Computer", and "My Network Places". The "Details" section shows the "exe" file type, "File Folder", and "Date Modified: Today, July 21, 2011, 9:35 AM". The main pane displays a list of files and folders, including "move", "pars", "proml", "protolist", "restdist", "retree", "treedist", "outfile", "neighbor", "penny", "promk", "protpars", "restml", "seqboot", "work1", and "Text Document". The "proml" file is circled in red.

outfile obtained from **seqboot** is the input for **proml**

M. Michael Gromiha | NPTEL Bioinformatics, Lecture 10

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)

```
C:\Documents and Settings\Admin\Desktop\phylip 3.69\exe\proml.exe
proml.exe: can't find input file "infile"
Please enter a new file name? outfile
proml.exe: the file "outfile" that you wanted to
use as output file already exists.
Do you want to Replace it, Append it,
write to a new file, or Quit?
(please type R, A, F, or Q)
R
Please enter a new file name? workfile

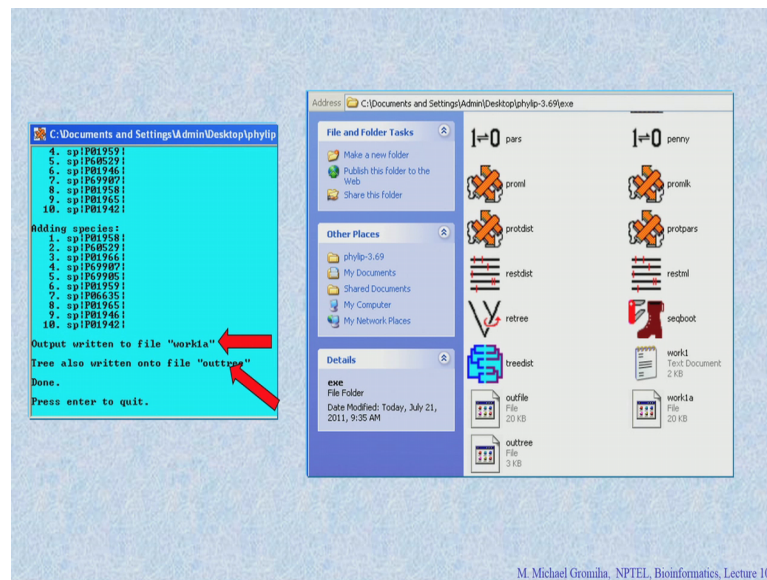
Amino acid sequence Maximum Likelihood method, version 3.69
Settings for this run:
U Search for best tree? Yes
P JTT, PAM or PMB probability model? Jones-Taylor-Thornton
C One category of sites? Yes
R Rate variation among sites? constant rate of change
V Sites weighted? No
S Speedier but rougher analysis? Yes
G Global rearrangements? No
J Randomize input order of sequences? No. Use input order
O Outgroup root? No. use as outgroup species 1
M Analyze multiple data sets? No
I Input sequences interleaved? Yes
0 Terminal type (IBM PC, ANSI, none)? IBM PC
1 Print out the data at start of run? No
2 Print indications of progress of run? Yes
3 Print out tree? Yes
4 Write out trees onto tree file? Yes
5 Reconstruct hypothetical sequences? No
Y to accept these or type the letter for one to change
D Multiple data sets or multiple weights? (type D or U)
10 How many data sets?
Random number seed (must be odd)?
5
Number of times to jumble?
3
Y to accept these or type the letter for one to change
Y
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 10

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.

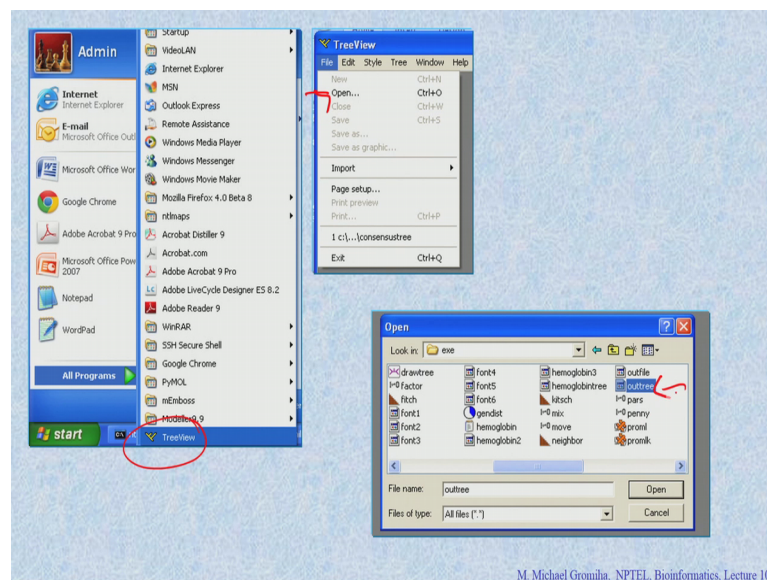
So, you give there a file name all right then again they ask for the same question you that you want to in change anything; if you do not want to change just you put to y. So, if you want to change, you can give the letters appropriately and you can change. How many times you want to jumble this sequence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

(Refer Slide Time: 27:19)



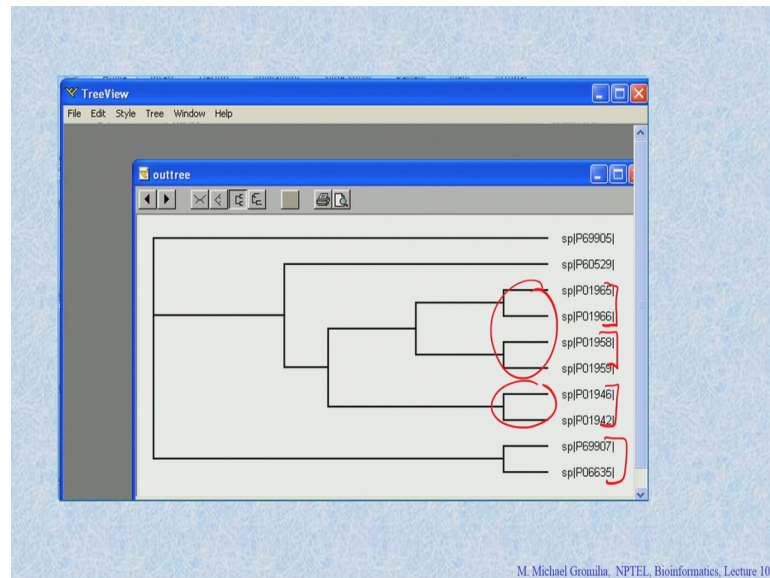
And finally, you can get these things with there are they got the trees and they wrote in this out tree this output we can see here right file it is saved ok.

(Refer Slide Time: 27:32)



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

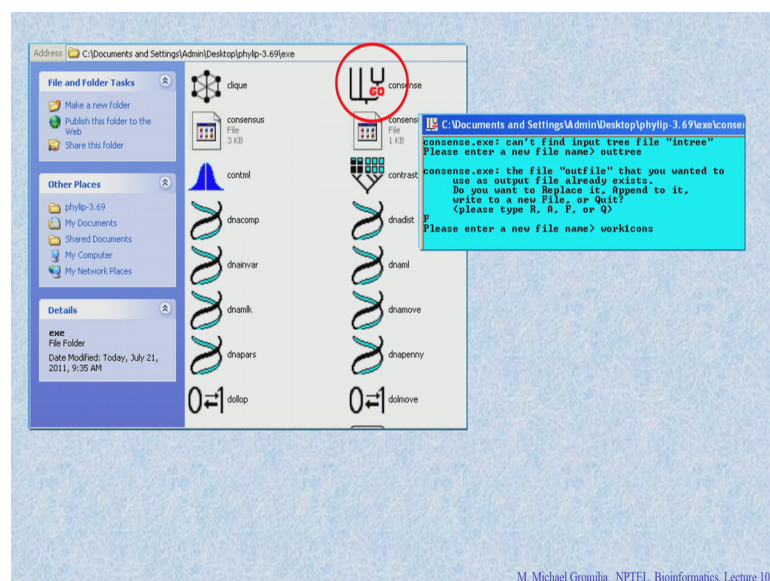
(Refer Slide Time: 27:51)



So, the 10 different cases. So, you can see the trees. So, there is which two are close to each other, say 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 linery between among these different sequences.

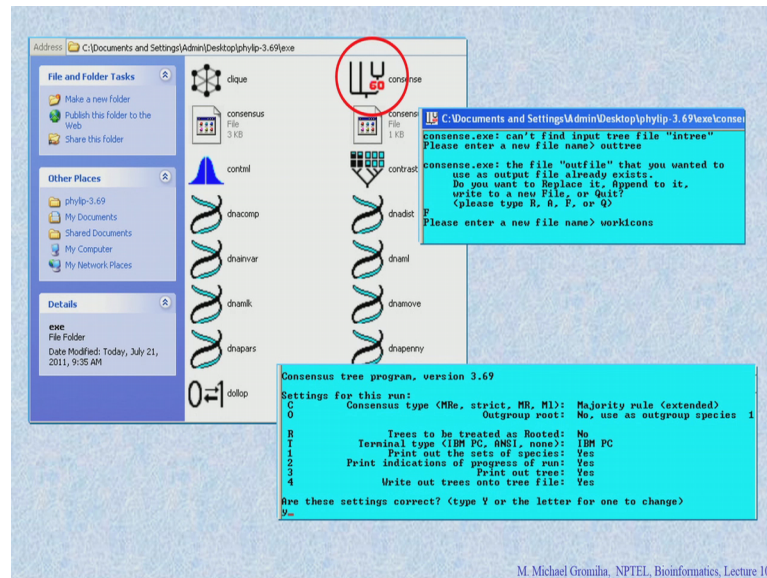
Now, the question is how far you are confident that these two are close to each other.

(Refer Slide Time: 28:22)

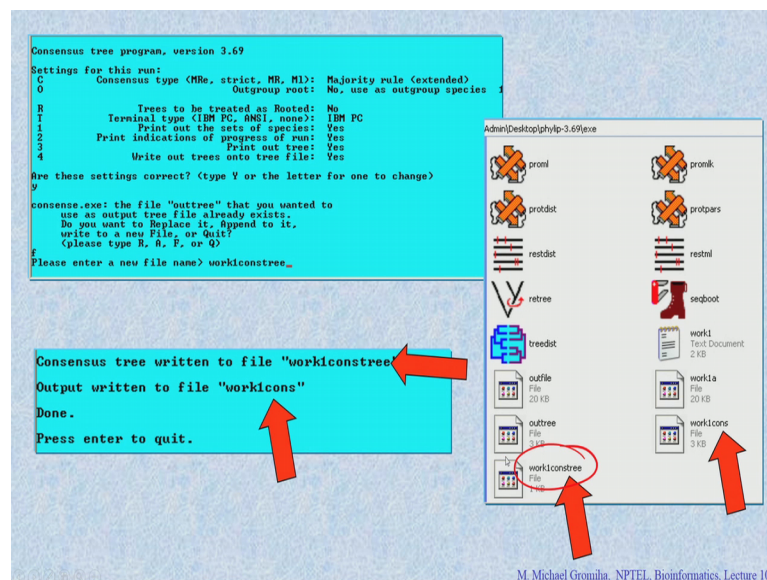


So, for this case you can go to consensus tree right. So, go with this a work on consensus this is your file.

(Refer Slide Time: 28:29)

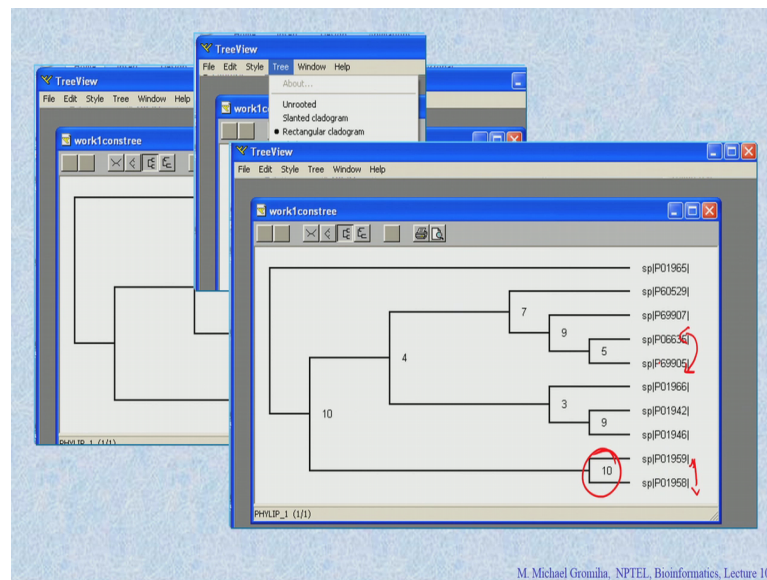


(Refer Slide Time: 28:30)



So, finally, if you give the files right finally, you can get this a file work on consensus right.

(Refer Slide Time: 28:37)



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

Student: (Refer Time: 28:51).

(Refer Time: 28:52) right what is a significant if this is a 10, these are 10 options out of 10 you all the 10 you these two are indent 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: (Refer Time: 29:25).

Construct your trees what is a means of tree it will give you the information regarding?

Student: (Refer Time: 29:29) 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 (Refer Time: 29:47).

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 construct a 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 tree 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 structural function.

Thank you very much.