

Lesson 13 Molecular Evolution

In this class we learn how to draw molecular evolutionary trees for proteins.

Theory

1. Evolutionary distance: The number of mutations separating two sequences. Sometimes distance is expressed as the total number of mutations and sometimes it is expressed as the number of mutations divided by the number of residues. Evolutionary distance can be obtained from the PAM matrices using the equations:

% difference = 100 - % identity.

$$\% \text{ difference} = 100\% \left(1 - \sum_i f_i M_{ii} \right)$$

f_i = fraction of amino acid i .

M_{ii} = The entry in the diagonal element of the PAM matrix.

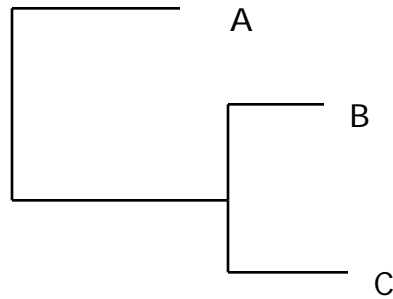
If the M_{ii} from PAM# (where # is the evolutionary distance) gives the right % difference between two sequences, in the above equation, then # is the evolutionary distance between the sequences.

2. Tree notation:

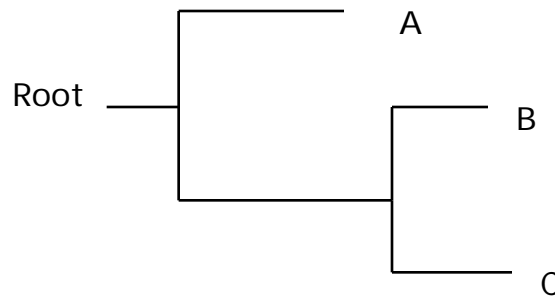
A. Species tree: A tree that gives the evolutionary relationship between species.

B. Gene tree. A tree that gives the evolutionary relationship between molecules that are related and some of which come from different species. A gene tree may correspond to species tree, but will not always do so.

C. Unrooted tree: An unrooted tree does not have a single node from which all other nodes can be reached by moving in the same direction. The direction of evolutionary time is not clear from an unrooted tree. Here is an unrooted tree:



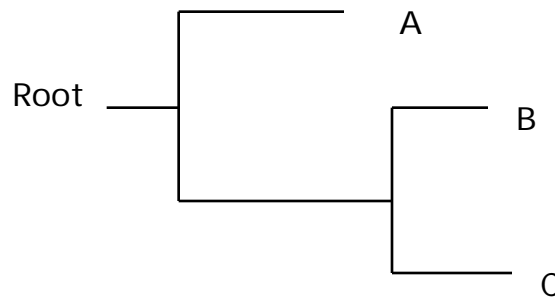
D. A rooted tree is a tree which has a node, called the root, from which all of the other nodes can be reached by moving in the same direction. The root is the common ancestor of the other nodes.



E. Cayley notation:

$(\text{Root}(A,(\text{B},\text{C})))$;

corresponds to the drawn tree:



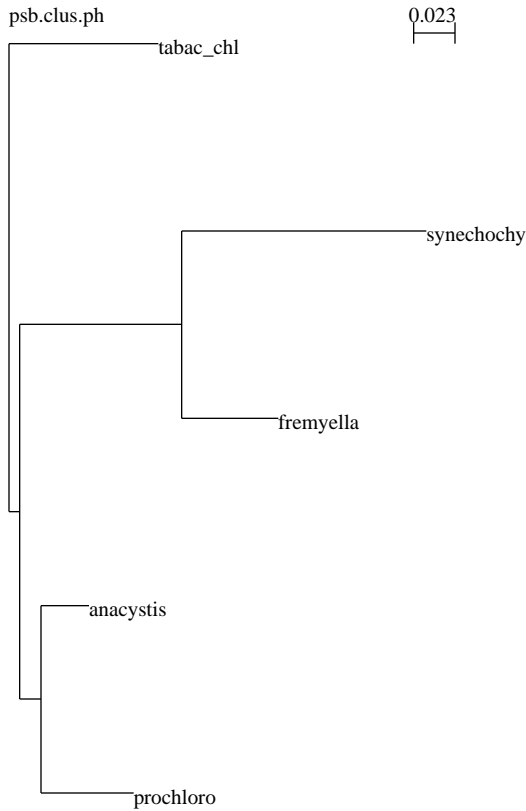
Note: How the Cayley notation diagram is distributed between lines doesn't matter. For example, the above tree can be expressed like this:

```
(Root  
(A,  
(B,C)  
)  
);
```

F. Newick Notation:

Expands Cayley notation to include evolutionary distances:

```
(  
(  
(  
prochloro:0.05132,  
anacystis:0.02636)  
:0.01164,  
tabac_chl:0.08969)  
:0.00809,  
fremyella:0.05377,  
synechochy:0.13585);
```



Distances are reflected in branch lengths.

G. An ultrametric tree is a tree in which all the branches from a given node are equal in length. Ultrametric trees implicitly assume the same rate of molecular evolution in each branch (i.e. a biological clock).

3. Neutral theory of molecular evolution (Motoo Kimura): Nucleic acids and proteins evolve primarily through random genetic drift as opposed to (Neo-Darwinian) selection.

Corollary: By modeling genetic drift one can draw evolutionary trees between molecules.

4. Homologs- Genetically related sequences (I prefer homogenes).

Orthologs - Genes in different organisms that are related by mutation (I prefer orthogenes). In my terminology genes that are homogenetic, homomorphic, and homopractic.

Paralogs - Genes in the same or different organisms that are related by gene duplication as well as speciation (I prefer paragenes). For example, whereas human and mouse hemoglobin are orthologs, human hemoglobin and human myoglobin are paralogs, and human hemoglobin and mouse myoglobin are also paralogs. In my terminology genes that are homogenetic and homomorphic, but not homopractic, are paragenes.

Xenologs - Genes that are related by transfer of DNA between species.

5. Distance methods- Inferring evolutionary trees from the evolutionary distances between the sequences.

A. Exhaustive minimum evolution: An exact method of inferring evolutionary trees from the evolutionary distance of the sequences. However it is subject to the quality of the data and the assumption that the minimum evolutionary distance yields the real tree. Not covered in this lecture.

B. UPGMA (Unweighted Pair Group Method with Arithmetic mean): A simple method of going from distances to trees. Used by Pileup.

C. Neighbor joining: A method of inferring evolutionary trees from the evolutionary distances between the sequences that is more accurate than UPGMA, yet is usually less accurate than exhaustive minimum evolution (see below). Used by Clustalw.

6. Maximum likelihood method - Inferring the most likely evolutionary tree for a group of sequences by considering the probability of all possible mutational paths between them. Used by Protml.

A. Star decomposition maximum likelihood - An approximate way of finding the maximum likelihood tree. Does not always find the most likely tree.

B. Exhaustive maximum likelihood - Finding the maximum likelihood by computing the likelihoods of all possible trees.

7. Reliability of methods: The evolution of protein sequences can in general be traced over longer times than that of nucleic acid sequences.

In decreasing order of reliability:

Exhaustive maximum likelihood star decomposition maximum likelihood Exhaustive minimum evolution neighbor joining UPMGA.

8. Measures of statistical significance:

A. Felsenstein bootstrap - Rerunning tree on randomly chosen subsets of the columns of the alignment and calculating the fraction of times each node that appears in the whole alignment appears in the randomly chosen subsets.

B. Kishino-Hagesawa bootstrap - Computing bootstrap parameters for each possible whole tree based upon the computed likelihood of that tree.

C. Bootstrap reliability criteria

.8- 1.0 - Good

.5-.8 - Acceptable

<. 5 - Questionable

9. Miscellaneous

A. Always remove columns containing gaps from alignments before analyzing.

B. In computing distances correct for back mutations.

Summary of Commands:

Note: In this document different fonts have different meanings:

Times is used to explain commands.

Courier is used to indicate commands and command options.

Courier italics are used to indicate command parameters, for example, filenames.

Courier bold is used to indicate commands that are not displayed.

Courier bold italics are used to indicate computer-generated output.

Helvetica is used to indicate menu items.

<http://www.ncbi.nlm.nih.gov/COG/>

Finds orthologs of a protein sequence.

pileup

Aligns multiple sequences and draws an evolutionary tree according to the distance UPMGA method. Option A sends the tree diagram to a figure file entitled *pileup.figure*.

tk

Sets the graphics output to terminal.

figure *psba.pile.fig*

Outputs the figure file to the graphics output, in this case the terminal.

clustalw

Aligns the multiple sequences (usually) more accurately than Pileup and draws a tree diagram according to the neighbor-joining distance methods. The use of Clustalw to perform a multiple alignment is described in detail in Lesson 8. The following describes the use of Clustalw for evolutionary analysis only: The Clustalw aligned sequences are already in psb.clus.msf.

Type:

clustalw

```
*****  
***** CLUSTAL W (1.7) Multiple Sequence Alignments *****  
*****
```

1. Sequence Input From Disc
 2. Multiple Alignments
 3. Profile / Structure Alignments
 4. Phylogenetic trees
-
- S. Execute a system command
 - H. HELP
 - X. EXIT (leave program)

Your choice: 1

Sequences should all be in 1 file.

7 formats accepted:

NBRF/PIR, EMBL/SwissProt, Pearson (Fasta), GDE, Clustal, GCG/MSF, RSF.

Enter the name of the sequence file: psba.tfa

Sequence format is Pearson

Sequences assumed to be PROTEIN

```
Sequence 1: prochloro      360 aa  
Sequence 2: anacystis     360 aa  
Sequence 3: fremyella     360 aa  
Sequence 4: tabac_chl     360 aa  
Sequence 5: synechochy    360 aa
```

***** CLUSTAL W (1.7) Multiple Sequence Alignments *****

1. Sequence Input From Disc
 2. Multiple Alignments
 3. Profile / Structure Alignments
 4. Phylogenetic trees
-
- S. Execute a system command
 - H. HELP
 - X. EXIT (leave program)

Your choice: 2

***** MULTIPLE ALIGNMENT MENU *****

1. Do complete multiple alignment now (Slow/Accurate)
 2. Produce guide tree file only
 3. Do alignment using old guide tree file
 4. Toggle Slow/Fast pairwise alignments = SLOW
 5. Pairwise alignment parameters
 6. Multiple alignment parameters
 7. Reset gaps between alignments? = OFF
 8. Toggle screen display = ON
 9. Output format options
-
- S. Execute a system command
 - H. HELP
- or press [RETURN] to go back to main menu

Your choice: 9

***** Format of Alignment Output *****

1. Toggle CLUSTAL format output = ON
2. Toggle NBRF/PIR format output = OFF
3. Toggle GCG/MSF format output = OFF
4. Toggle PHYLIP format output = OFF
5. Toggle GDE format output = OFF
6. Toggle GDE output case = LOWER
7. Toggle CLUSTALW sequence numbers = OFF
8. Toggle output order = ALIGNED
9. Create alignment output file(s) now?

0. Toggle parameter output = OFF

H. HELP

Enter number (or [RETURN] to exit): 1

***** Format of Alignment Output *****

1. Toggle CLUSTAL format output = OFF

2. Toggle NBRF/PIR format output = OFF

3. Toggle GCG/MSF format output = OFF

4. Toggle PHYLIP format output = OFF

5. Toggle GDE format output = OFF

6. Toggle GDE output case = LOWER

7. Toggle CLUSTALW sequence numbers = OFF

8. Toggle output order = ALIGNED

9. Create alignment output file(s) now?

0. Toggle parameter output = OFF

H. HELP

Enter number (or [RETURN] to exit): 3

***** Format of Alignment Output *****

1. Toggle CLUSTAL format output = OFF

2. Toggle NBRF/PIR format output = OFF

3. Toggle GCG/MSF format output = ON

4. Toggle PHYLIP format output = OFF

5. Toggle GDE format output = OFF

6. Toggle GDE output case = LOWER

7. Toggle CLUSTALW sequence numbers = OFF

8. Toggle output order = ALIGNED

9. Create alignment output file(s) now?

0. Toggle parameter output = OFF

H. HELP

Enter number (or [RETURN] to exit): 4

***** Format of Alignment Output *****

- 1. Toggle CLUSTAL format output = OFF
- 2. Toggle NBRF/PIR format output = OFF
- 3. Toggle GCG/MSF format output = ON
- 4. Toggle PHYLIP format output = ON
- 5. Toggle GDE format output = OFF

- 6. Toggle GDE output case = LOWER
- 7. Toggle CLUSTALW sequence numbers = OFF
- 8. Toggle output order = ALIGNED

- 9. Create alignment output file(s) now?

- 0. Toggle parameter output = OFF

- H. HELP

Enter number (or [RETURN] to exit):

***** MULTIPLE ALIGNMENT MENU *****

- 1. Do complete multiple alignment now (Slow/Accurate)
 - 2. Produce guide tree file only
 - 3. Do alignment using old guide tree file

 - 4. Toggle Slow/Fast pairwise alignments = SLOW

 - 5. Pairwise alignment parameters
 - 6. Multiple alignment parameters

 - 7. Reset gaps between alignments? = OFF
 - 8. Toggle screen display = ON
 - 9. Output format options

 - S. Execute a system command
 - H. HELP
- or press [RETURN] to go back to main menu

Your choice: 1

WARNING: Output file name is the same as input file.

Enter new name to avoid overwriting [psba.msf]: psba.clus.msf

Enter a name for the PHYLIP output file [psba.phy]: psba.clus.phy

Enter name for new GUIDE TREE file [psba.dnd]: psba.clus.dnd

[Sequences are aligned]

```
*****
***** CLUSTAL W (1.7) Multiple Sequence Alignments *****
*****
```

1. Sequence Input From Disc
2. Multiple Alignments
3. Profile / Structure Alignments
4. Phylogenetic trees

- S. Execute a system command
- H. HELP
- X. EXIT (leave program)

Your choice: 4

```
***** PHYLOGENETIC TREE MENU *****
```

1. Input an alignment
2. Exclude positions with gaps? = OFF
3. Correct for multiple substitutions? = OFF
4. Draw tree now
5. Bootstrap tree
6. Output format options

- S. Execute a system command
 - H. HELP
- or press [RETURN] to go back to main menu

```
***** PHYLOGENETIC TREE MENU *****
```

1. Input an alignment
2. Exclude positions with gaps? = ON
3. Correct for multiple substitutions? = ON
4. Draw tree now
5. Bootstrap tree
6. Output format options

- S. Execute a system command
 - H. HELP
- or press [RETURN] to go back to main menu

Your choice: 1

Sequences should all be in 1 file.

7 formats accepted:

NBRF/PIR, EMBL/SwissProt, Pearson (Fasta), GDE, Clustal, GCG/MSF, RSF.

Enter the name of the sequence file: psba.clus.msf

Sequence format is Pileup/MSF
Sequences assumed to be PROTEIN

Sequence 1: prochloro 367 aa
Sequence 2: anacystis 367 aa
Sequence 3: tabac_chl 367 aa
Sequence 4: fremyella 367 aa
Sequence 5: synechochy 367 aa

***** PHYLOGENETIC TREE MENU *****

1. Input an alignment
 2. Exclude positions with gaps? = ON
 3. Correct for multiple substitutions? = ON
 4. Draw tree now
 5. Bootstrap tree
 6. Output format options
- S. Execute a system command
H. HELP
or press [RETURN] to go back to main menu

Your choice:

Your choice: 5

Enter name for bootstrap output file [psba.clus.phb]:

Note: The default tree diagram is in Phylip bootstrap format.

Enter seed no. for random number generator (1..1000) [111]: 123

Enter number of bootstrap trials (1..10000) [1000]: 100

Each dot represents 10 trials

.....

Bootstrap output file completed [psba.clus.phb]

***** PHYLOGENETIC TREE MENU *****

1. Input an alignment
2. Exclude positions with gaps? = ON
3. Correct for multiple substitutions? = ON


```
protml psba.clus.ng.molphy psba.clus.ng.e.tree >  
psba.clus.ng.e.boot
```

Generates all possible trees for an alignment plus maximum likelihood parameters and bootstrap parameters. Input files are an alignment without gaps in Molphy format and a list of all possible trees in Newick format.

Lab:

This lab refers to the five psbA sequences in /usr2/seq/seqclass/lab13. The question addressed is whether the Prochlorothrix sequence (prochloro) is related to the tobacco chloroplast sequence (tabac_chl) (Kishino et al. J. Mol. Evol. 31, 151-160).

1. Align and generate a tree for these sequences using Pileup. What does Pileup tell you about the relationship between Prochlorothrix and Tobacco Chloroplast? Does Pileup give a statistical measure of the validity of this tree?
2. Align and generate a tree for these sequences using Clustalw. The tree should have bootstrap statistics and recorded in Phylip bootstrap format. If you have a Mac display the tree with NJPLOT. What does Clustalw tell you about the relationship between Prochlorothrix and Tobacco Chloroplast?
3. Align and generate a tree for the Clustalw alignment using the star decomposition and exhaustive options of Protml. Perform a bootstrap analysis on the exhaustive set of trees. What does Protml tell you about the relationship between Prochlorothrix and Tobacco chloroplast? Is the star decomposition tree the same as the maximum likelihood tree?

Bibliography:

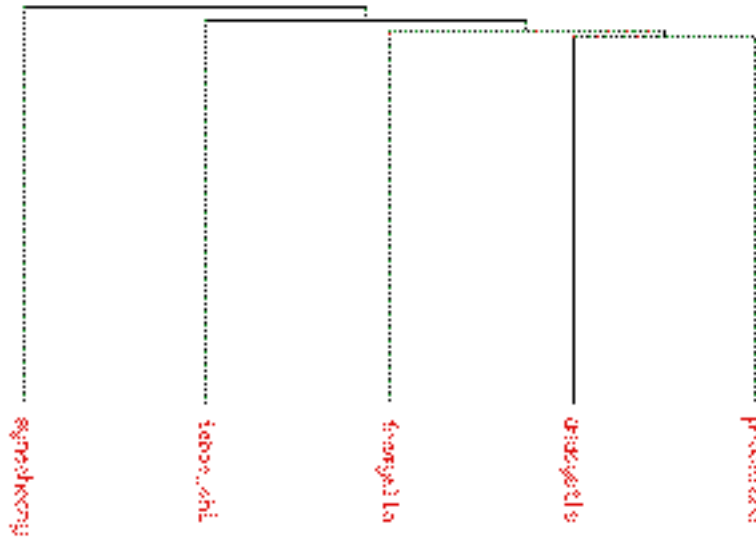
1. GCG Program Manual Pileup.
2. Documentation for Clustalw which can be found in /usr2/seq/doc/clustalw .
3. Documentation for Protml and file conversion programs can be found in /usr2/seq/doc/molphy .
4. An excellent introduction to molecular evolution is: Molecular Evolution by Wen-Hsiung Li, Sinauer, 1997.
5. Other GCG programs for evolutionary analysis are: Paupsearch, Paupdisplay, Distances, Growtree, and Diverge. The first two programs are the GCG Version of PAUP* 4.0 by David Swofford, a widely-used package of evolutionary analysis programs.
6. Phylip is a mammoth collection of evolutionary analysis programs by Joseph Felsenstein. Documentation to Phylip is available on /usr2/seq/doc/phylip . Programs in Phylip can be executed on cuccfa by typing the program name.

Acknowledgement:

I thank Drs. Andrey Rzhetsky of the Genome Center and William Hahn of the Earth and Environmental Science Department for useful discussions pertaining to molecular evolution.

Key to Lab

1. Pileup tree

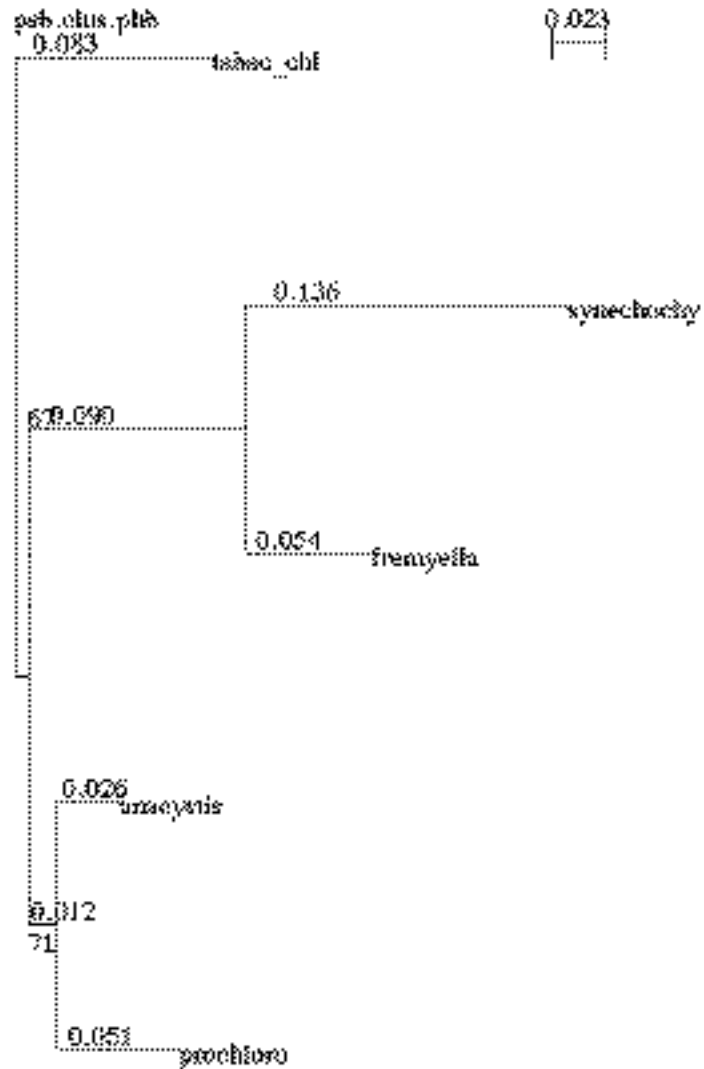


Prochlorothrix does not branch from the same node as Tobacco chloroplast. There is no measure of statistical significance.

2. Clustalw tree in Phylip bootstrap format:

```
(
(
(
prochloro:0.05132,
anacystis:0.02636)
71:0.01164,
tabac_chl:0.08969)
67:0.00809,
fremyella:0.05377,
synechochy:0.13585)TRICHOTOMY;
```

Clustal tree with bootstrap as drawn by NJPlot:



Tobacco chloroplast does not share a node with Prochlorothrix or any of the other photosynthetic bacteria for that matter. According to the bootstrap test the synthetic bacterial nodes are significant.

3. Protml star decomposition tree:

```
((prochloro,anacystis),tabac_chl,fremyella,synechochy);
((prochloro,anacystis),(tabac_chl,fremyella),synechochy);
```

```
      :-----1 prochloro
:-----6
:      :--2 anacystis
:
:      :-----3 tabac_chl
:---7
:      :-----4 fremyella
:
:-----5 synechochy
```

Protml exhaustive maximum likelihood bootstrap values:

Tree	ln L	Diff ln L	S.E.	#Para	AIC	Diff AIC	Boot	P
1	-1790.6	-3.7	6.0	7	3595.1	7.5	0.1610	
2	-1790.7	-3.8	5.8	7	3595.3	7.7	0.1210	
3	-1786.8	0.0	<-best	7	3587.7	0.0	0.4590	
4	-1798.5	-11.6	6.6	7	3610.9	23.3	0.0030	
5	-1793.4	-6.5	8.4	7	3600.8	13.1	0.1610	
6	-1802.6	-15.8	10.0	7	3619.2	31.5	0.0040	
7	-1798.9	-12.1	10.9	7	3611.9	24.2	0.0740	
8	-1803.6	-16.8	10.4	7	3621.3	33.6	0.0070	
9	-1805.7	-18.9	11.4	7	3625.5	37.8	0.0010	
10	-1804.3	-17.5	11.7	7	3622.6	35.0	0.0040	
11	-1804.9	-18.0	10.0	7	3623.7	36.1	0.0010	
12	-1808.6	-21.7	10.8	7	3631.1	43.5	0.0000	
13	-1809.0	-22.1	10.6	7	3631.9	44.3	0.0000	
14	-1804.7	-17.9	11.6	7	3623.4	35.7	0.0040	
15	-1805.7	-18.9	11.4	7	3625.5	37.8	0.0000	

Protml best tree:

```
#3
      :-----1 prochloro
:-----6
:      :--2 anacystis
:
:      :-----4 fremyella
:---7
:      :-----3 tabac_chl
:
:-----5 synechochy
```

No.3	num	length	S.E.	num	length	S.E.
prochloro	1	6.18	1.39	6	2.31	0.93
anacystis	2	1.50	0.74	7	1.13	0.68
tabac_chl	3	9.55	1.76	8		
fremyella	4	5.40	1.33	ln L:	-1786.83	+ - 65.79
synechochy	5	15.26	2.28	AIC :	3587.65	

Tobacco chloroplast doesn't share a node with Prochlorothrix. The best exhaustive tree is the same as the star decomposition tree.