Practice of phylogenetic analysis

Instructor: Kiatichai Faksri, Ph.D.

Objectives: The students should be able to

1. Construct a phylogenetic tree by using basic program for phylogenetic analysis.
2. Understand the process and options for phylogenetic analysis.
3. Interpret the results of the phylogenetic tree.

Introduction

In general, the phylogenetic analysis required many programs (for many steps), i.e. multiple alignments, tree building, tree editing and tree viewing.

SeaView is a simple and multi-platform program designed to facilitate multiple alignment and phylogenetic tree building (in one package program) from molecular sequence data through the use of a graphical user interface.

SeaView contains Clustal, Muscle, PHYLIP and PhyML. The package software is also incorporated a simple tree viewer. This software can be downloaded from http://pbil.univ-lyon1.fr/software/seaview.html

Practices

1. What are your sequences of interest?

Before you perform the phylogenetic analysis, you are the one who know the best what is your hypothesis (and your sequences for analysis). In this practice, the sequences of interest are *16S rRNA* gene sequences (nucleotide sequences) of the following bacteria,

1. Enterobacter cloacae (*E.cloacae* strain LRC85), 1504
2. *Escherichia coli* (*E.coli* strain ATCC 25922*)*, 1532
3. *Helicobacter pylori* (*H. pylori* isolate ATCC 43504), 1503
4. *Klebsiella pneumoniae* (*K. pneumoniae* strain LSRC119), 1504
5. *Proteus vulgaris* (*P. vulgaris* strain CIP103181T*)*, 1505
6. *Salmonella typhi* (*S. enterica* subsp. enterica serovar Typhi), 1542
7. *Staphylococcus aureus* (*S. aureus* partial isolate 30), 1517
8. *Vibrio cholerae* (*V. cholerae* strain VC12-Ogawa), 1491
9. *Yersinia enterocolitica* (*Y.enterocolitica* strain O:3 108 c), 1489

Note: only *Staphylococcus aureus* is a Gram positive bacteria

Step 1: Find the sequence No. 6 *Salmonella typhi* (then save as Styphi 16SrRNA). The last eight sequences above were prepared and can be downloaded from the e-learning. Save them to your PC as **FASTA format**.

SeaView reads and writes various file formats (NEXUS, MSF, CLUSTAL, FASTA, PHYLIP, MASE, Newick) of DNA and protein sequences.

***Question** 1: In general, what are the characteristics or properties of the target sequences or genes for phylogenetic analysis?*

***Question** 2: What are the length of the tested sequence? What should you suggest for the length variation of these sequences before constructing the phylogenetic tree?*

----

2. Multiple sequences alignments

SeaView drives programs muscle or clustalw for multiple sequence alignment

Step2: Installed software SeaView 4.0 (size 2.72 MB)>> by download from e-learning or from http://pbil.univ-lyon1.fr/software/seaview.html. (Extract from zip file to your PC, it ready to launch by double click the icon SE A− directly)

Step 3: Create “multiple sequences FASTA file” >> by copy nucleotide sequences (with FASTA format) of each sequence (9 sequences) into 1 file as the example in the below textbox.

```
>gi|86278349|gb|DQ360844.1| Escherichia coli
AGAGTTTGATCCTGGCT……................................

>gi|333353408|gb|JF772064.1| Enterobacter cloacae
AGAGTTTGATCCTGG……......................................

>gi|333353423|gb|JF772079.1| Klebsiella pneumoniae
AGAGTTTGATCC……...............................................
```

Note: You should change the preferred name of each sequence after “>”
For example “>Escherichia coli”
DO NOT forget to hit the button “ENTER” of the nucleotide sequences after the title name from the first line, so the program will be able to recognize the sequence character after “ENTER”

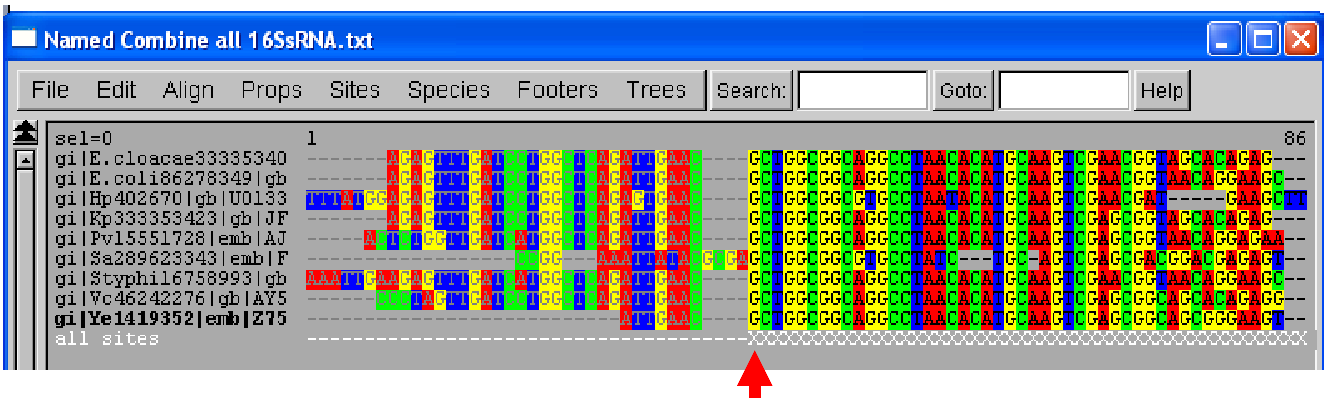
Step 4: Go to Menu of SeaView program>>Click “File”>> Click ”Open” to input the multiple sequence file (FASTA file that you have created).
To open the multi sequence file, you can simply “Drag and Drop” the file into the program.
Then do multiple sequence alignment by >>>click “Align”>> choose “Align all” >>wait a minute >>click “OK”

Now, the aligned sequences are available. You can see that there are gaps inside each sequence.
Note: The multiple sequences alignment is a critical step for phylogenetic analysis. If you start with a bad alignment, the phylogenetic tree will be incorrect. A good alignment should chop off (exclude) the edge regions that are not relate to the actual biology (called artifact) before sending the alignment for phylogenetic analysis. However, the gap may be from indels (true evolution) that should be included into analysis.

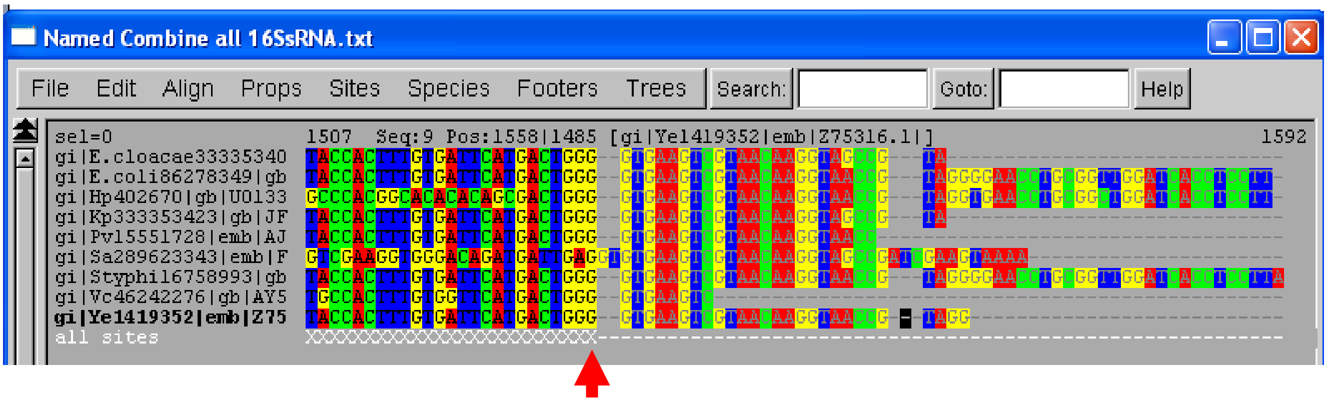
Step 5: Exclude unequal sequences (from the sequencing)>> by click "Sites">>"Create set">>click "OK"
Now, the "XXXXX" (sequence use for analysis) are appear under each nucleotide.

We can exclude the unequal sites (unequal sequences or artifact gap) by left click at "X" under the unwanted nucleotide. Then "X" (selected) will change into "-" (excluded)

To exclude the long sequence, **left clicks** of the mouse to the "X" of one end, then "**right click + Ctr**" in another end as the picture below.



```
sel=0                      1                                                                                     86
gi|E.cloacae33335340       -------AGAGTTTGATCCTGGCTCAGATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAGCACAGAG---
gi|E.coli86278349|gb       -------AGAGTTTGATCCTGGCTCAGATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAAGC--
gi|Hp402670|gb|U0133       TTTATGGAGAGTTTGATCCTGGCTCAGAGTGAAC----GCTGGCGGCGTGCCTAATACATGCAAGTCGAACGAT-----GAAGCTT
gi|Kp333353423|gb|JF       -------AGAGTTTGATCCTGGCTCAGATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAGCACAGAG---
gi|Pv15551728|emb|AJ       -----ACTCTGGTTGATCATGGCTCAGATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGTAACAGGAGAA--
gi|Sa289623343|emb|F       ------------------CCGG---AAATTATACGCGAGCTGGCGGCGTGCCTATC---TGC-AGTCGAGCGACGGACGAGAGT--
gi|Styphi16758993|gb       AAATTGAAGAGTTTGATCATGGCTCAGATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAAGC--
gi|Vc46242276|gb|AY5       ------CCCTAGTTGATCCTGGCTCAGATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGCAGCACAGAGG--
gi|Ye1419352|emb|Z75       ---------------------------ATTGAAC----GCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGCAGCGGGAAGT--
all sites                  --------------------------------------XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```



```
sel=0                      1507  Seq:9 Pos:1558|1485 [gi|Ye1419352|emb|Z75316.1|]                                1592
gi|E.cloacae33335340       TACCACTTTGTGATTCATGACTGGG--GTGAAGTCGTAACAAGGTAGCCCG---TA------------------------------
gi|E.coli86278349|gb       TACCACTTTGTGATTCATGACTGGG--GTGAAGTCGTAACAAGGTAACCCG---TAGGGGAACCTGCGGTTGGATCACCTCCTT-
gi|Hp402670|gb|U0133       GCCCACGGCACACACAGCGACTGGG--GTGAAGTCGTAACAAGGTAACCCG---TAGGTGAACCTGCGGCTGGATCACCTCCTT-
gi|Kp333353423|gb|JF       TACCACTTTGTGATTCATGACTGGG--GTGAAGTCGTAACAAGGTAGCCCG---TA------------------------------
gi|Pv15551728|emb|AJ       TACCACTTTGTGATTCATGACTGGG--GTGAAGTCGTAACAAGGTAACC-------------------------------------
gi|Sa289623343|emb|F       GTCGAAGGTGGGACAGATGATTGAGGTGTGAAGTCGTAACAAGGTAGCCCGATCGAAGTAAAA-----------------------
gi|Styphi16758993|gb       TACCACTTTGTGATTCATGACTGGG--GTGAAGTCGTAACAAGGTAACCCG---TAGGGGAACCTGCGGTTGGATCACCTCCTTA
gi|Vc46242276|gb|AY5       TGCCACTTTGTGGTTCATGACTGGG--GTGAAGTC---------------------------------------------------
gi|Ye1419352|emb|Z75       TACCACTTTGTGATTCATGACTGGG--GTGAAGTCGTAACAAGGTAACCCG---TAGG----------------------------
all sites                  XXXXXXXXXXXXXXXXXXXXXXXXX-------------------------------------------------------------
```

Now, the aligned sequences with already excluded unequal sequences (unwanted sequences) are ready to construct the tree.

Note: the exclusion of head and tail region of sequences will exclude the noise characters (artifact) that irrelevant to the actual biology, especially when the tested sequence is short and the artifact characters (unequal sequence obtained from sequencing machine) might affect the tree topology.

***Question 3**: What do you think if you use the aligned sequences that the lengths are not equal (with unequal sequences at the upstream and/or downstream of the tested region) for building the tree?*

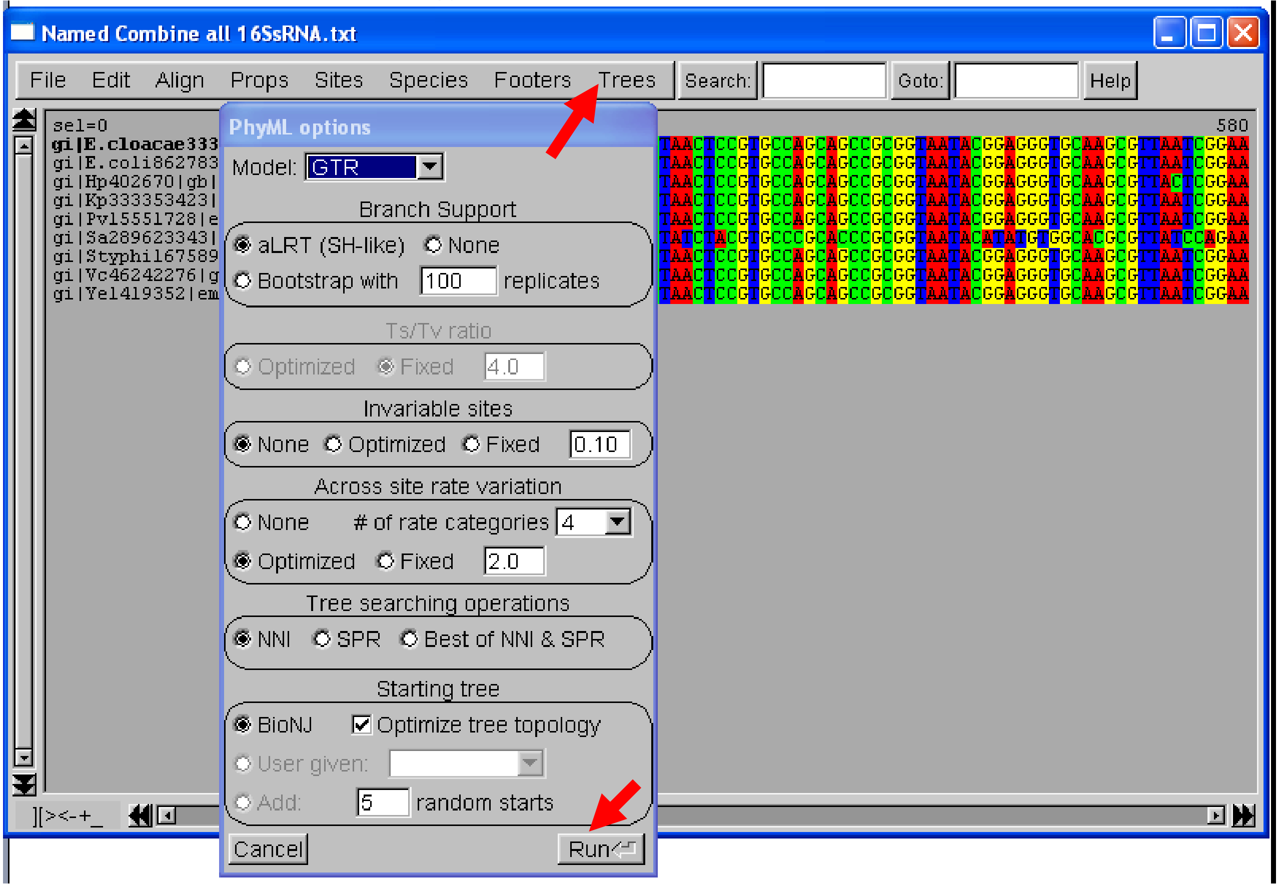
----

3. Construct the tree

SeaView computes phylogenetic trees by
1. Parsimony, using PHYLIP's dnapars/protpars algorithm.
2. Distance, with NJ or BioNJ algorithms on a variety of evolutionary distances.
3. Maximum likelihood, driving program PhyML 3.0.

In the next analysis, we are going to use Maximum likelihood method by PhyML 3.0.

Step 6: Construct the tree by>>go to the top menu of program>>click "Tree">>select "PhyML">>click "Run">> wait a few minutes>>> then click "OK"
Option: You might change the nucleotide substitution model (such as GTR, JC etc.) and compare the tree topology from each model.


Named Combine all 16SsRNA.txt
File Edit Align Props Sites Species Footers Trees
Search: Goto: Help
PhyML options
Cancel
Run

```
sel=0                                                       580
gi|E.cloacae333   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
gi|E.coli862783   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
gi|Hp402670|gb|   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTACTCGGAA
gi|Kp333353423|   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
gi|Pv15551728|e   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
gi|Sa289623343|   TATCTACGTGCCCGCACCCGCGGTAATACATATGTGGCACGCGTTATCCAGAA
gi|Styphi167589   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
gi|Vc46242276|g   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
gi|Ye1419352|em   TAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAA
```

Model: GTR

Branch Support
● aLRT (SH-like) ○ None
○ Bootstrap with 100 replicates

Ts/Tv ratio
○ Optimized ● Fixed 4.0

Invariable sites
● None ○ Optimized ○ Fixed 0.10

Across site rate variation
○ None # of rate categories 4
● Optimized ○ Fixed 2.0

Tree searching operations
● NNI ○ SPR ○ Best of NNI & SPR

Starting tree
● BioNJ ☑ Optimize tree topology
○ User given:
○ Add: 5 random starts

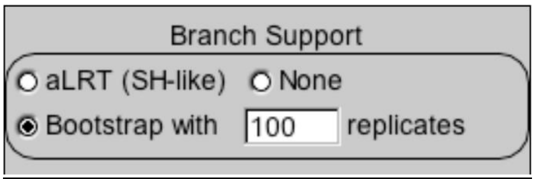
Now, the tree by ML method is revealed (i.e. **this is the Tree 1**)

Actually, you can save the tree as pdf format, but the there is some defect in the software, **so please use "Print screen" option as a solution** or you can save the output file from the tree construction step and open the file in other Tree viewing software for viewing/editing and save as pdf, JPEG or preferred format.

Question 4: *What does the output tree (**Tree 1**) called (between cladogram and phylogram) and why is that?*

Step 7: Evaluate the tree by >>go to window with aligned sequences>>click "Tree">>select "PhyML">>select "Bootstrap with 100 replicates" (from Branch support menu) >> click "Run">> click "OK"

Option: you might increase the boostraping replica into 1000, but it might take hours to finish it.



Branch Support

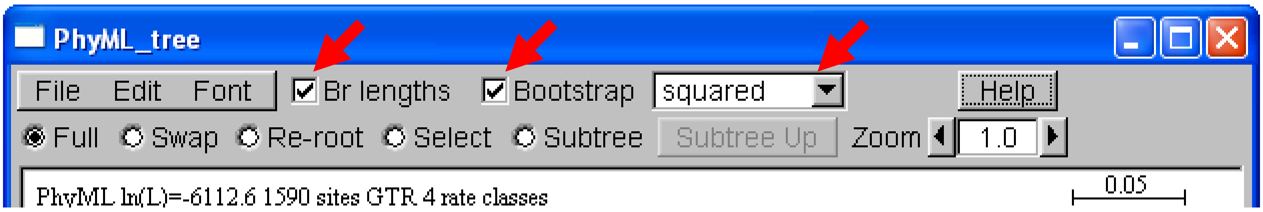
○ aLRT (SH-like) ○ None

● Bootstrap with 100 replicates

It takes around 3 minutes for bootstrapping analysis (100 replicates).
It takes even longer time when select several hundred (e.g. 1,000) replicates.
(In general study, bootstrapping with 1,000 replicates is preferred)

The tree by ML method with bootstrapping is now revealed.

Step 8: At the top of the tree window, select "Br lengths" , "Bootstrap" and change "squared into "cladogram".



PhyML_tree

File Edit Font ☑ Br lengths ☑ Bootstrap squared Help

● Full ○ Swap ○ Re-root ○ Select ○ Subtree Subtree Up Zoom 1.0

PhyML ln(L)=-6112.6 1590 sites GTR 4 rate classes

When you select "Br lengths" , "Bootstrap", (NOT change into "cladogram") save the display tree (Yes!, by print screen) (i.e., this is the **"Tree 2"**)

Question 5: *What happen when you click "Br lengths", "Boot strap" or change "squared" into "cladogram" from the analyzed tree and what does each of the results from these options mean?*

Question 6: *Where is the root of the obtained tree? Why do you think it is that root?*

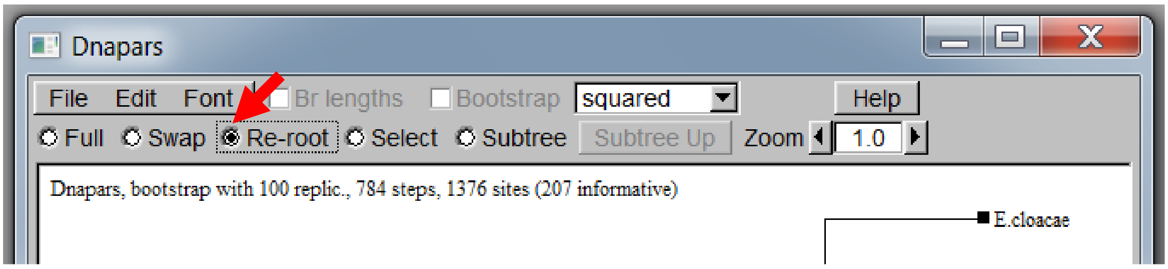
Step 8: Construct the tree by different methods>>go back to the alignment window>>click "Tree">>click "Parsimony" (i.e. **Tree 3**) (and then "Distance methods", (i.e. **Tree 4**)), don't forget to select "bootstrap with 100 replica" (or 1,000 replica).

The trees by Parsimony and Distance methods are now revealed.

Note: When you save the tree figure, it should show Bootstrap value (and branch length if available) to provide more information of the tree.

Question 7*: What is the root of "Tree 3" and "Tree 4". Is that correct? , What should you do next?*

Step: 9 Re-root the tree of Tree 3 >> go to Parsimony tree (Tree 3)>>click Re-root at top menu >>Click the node (small black box) of the right root (Gram positive one) to re-root the tree.



Dnapars, bootstrap with 100 replic., 784 steps, 1376 sites (207 informative)

Now, the Tree 3 with the corrected root is available.

Note: if you want to display bootstrap value (and branch length if available), select "Full", then "Bootstrap" menu (and Br lengths if available) will be active and selectable.

Save the re-rooted version of Tree 3 with bootstrap value for the assignments

Question 8*: What are the differences among the trees by ML (Tree 2), Parsimony (Tree 3) and Distance methods (Tree 4)? Please discuss the detail of differences among the obtained trees.*

Question 9*: Do you think ML method is a good method for the given data or not, why is that? Can you construct the consensus tree from Tree 2 and Tree 4? After the concensus tree construction, what is the node that is changed?*

Note: you can save the Tree as Newick format, and then it can be opened in other Tree viewing software such as FigTree (http://mac.softpedia.com/get/Graphics/FigTree.shtml). This program is more versatile, allowing you to colour branches and taxa, redraw the tree in a number of ways, collapse branches, etc.

(The tree figure can be saved as SVG or pdf format (or just print screen and crop))

Interpretation of Phylogeny

We are going to concentrate only "Tree 2" (with ML method)
Now you got the tree that constructed from the dataset. What is the tree mean?

***Question10* (Final question):** *What are the interpretation of the tree in term of phylogeny, relationships among species, robustness and association to phenotypes (you can google for the detail of characteristics of these bacteria)?*

After the class:
Send the lab report to instructor within 2 weeks after the class by

- Answers of 10 questions
- Show the trees that you get from analysis
  1. "Tree 2" (ML) that show branch length and bootstrap values
  2. "Tree 3" (Parsimony) that was re-rooted and show bootstrap values
  3. "Tree 4" (Distance) that show branch length and bootstrap values
- Interpretation of the "Tree 2 " (ML method) in the "Final question"

The students who copy the answers from others will be punished by 50% reduction of the total score for particular students.

Note: there are several programs that can detect plagiarism (copy/paste issue).

References

1. Gouy, M., S. Guindon, and O. Gascuel. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol **27:**221-4.
2. Saeview verson 4.0 retrieved from http://pbil.univ-lyon1.fr/software/seaview.html
3. Susan Holmes, Bootstrapping Phylogenetic Trees Theory and Methods, Statistical Science, 2003, Vol. 18, No. 2, 241–255.
4. Training manual, Working with Pathogen Genomes,19-24 February 2012, Wellcome Trust-Mahidol University-Oxford Tropical Medicine Research Programme Bangkok, Thailand