Bioinformatics tools for phylogeny and visualization

Yanbin Yin

Spring 2013

Homework #3

- Under “A la Carte” mode, unselect multiple alignment (default is MUSCLE)

Homework assignment 6

1. Take the MAFFT alignment http://cys.bios.niu.edu/yyin/teach/PBB/purdue.cellwall.list.lignin.fa.aln as input and use MEGA5 to build a phylogenetic tree

2. Try maximum likelihood (ML), neighbor-joining (NJ) and maximum parsimony (MP) algorithms with 100 bootstrap replications and compare the running time and the topology of the resulting trees. If encounter errors, try to use the HELP link to find out and solve it

3. Color the branches and leafs in the resulting ML tree graph using different colors for different gene subfamilies

Homework assignment 6 Cont.

4. Export the tree as a newick format file

5. Use the original sequence file in http://cys.bios.niu.edu/yyin/teach/PBB/purdue.cellwall.list.lignin.fa to calculate the lengths of C3H/C4H/F5H proteins (try to search "length" in galaxy server) and identify the Pfam domains in the C3H/C4H/F5H protein sequences; with the two results, prepare a domain definition file

6. Prepare a color definition file for different gene subfamilies (see step 3); upload the newick tree file, the color definition file and the domain definition file to iTOL to color display the tree

Write a report in word or ppt to include all the operations and screen shots.

Due on March 12 (send by email)

Office hour:
Tue, Thu and Fri 2-4pm, MO325A
Or email: yyin@niu.edu

Outline

- Introduction to phylogenetic analysis

- Hands on practice of Clustalx and MEGA 5

Putty: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html

Secure SSH client: http://cys.bios.niu.edu/yyin/teach/SSHSecureShellClient-3.2.9.exe

ClustaX: http://cys.bios.niu.edu/yyin/teach/clustalx-2.1-win.msi

MEGA: http://cys.bios.niu.edu/yyin/teach/MEGA5.10_Setup.exe

PHYLIP: http://evolution.genetics.washington.edu/phylip.html

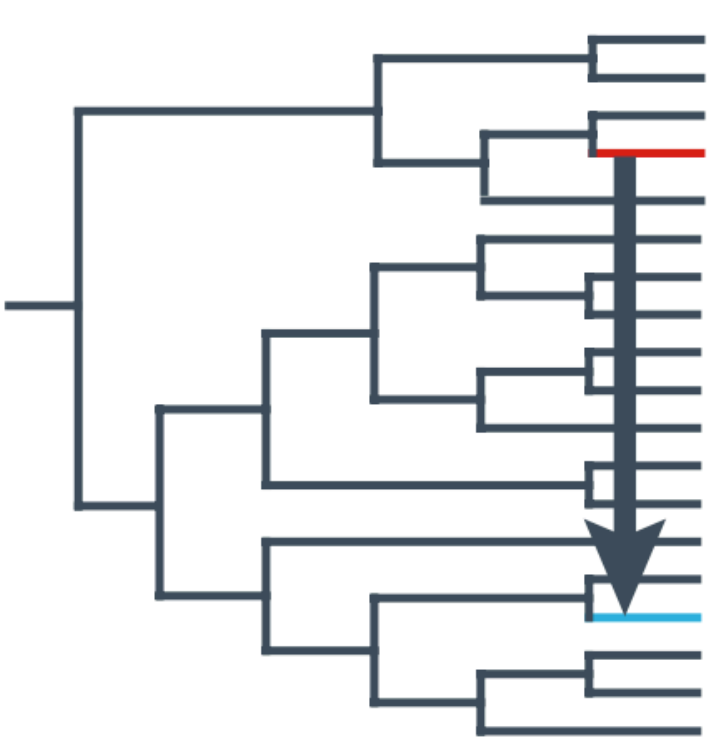
Jalview: http://www.jalview.org/

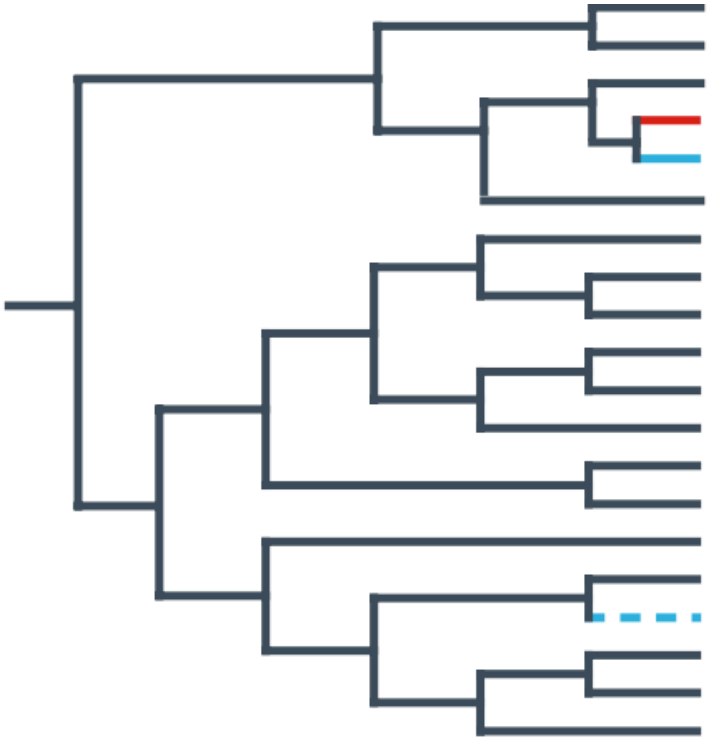
ITOL: http://itol.embl.de/

Phylogenetics is the science of estimating the evolutionary past, in the case of molecular phylogeny, based on the comparison of DNA or protein sequences:

- Study the evolution of genomes and gene families (duplication and transfer)
- Study the diversity of genes or fragments
- Cluster homologous sequences into subfamilies based on evolutionary history
- Infer functions for unknown genes

A simple case of horizontal gene transfer





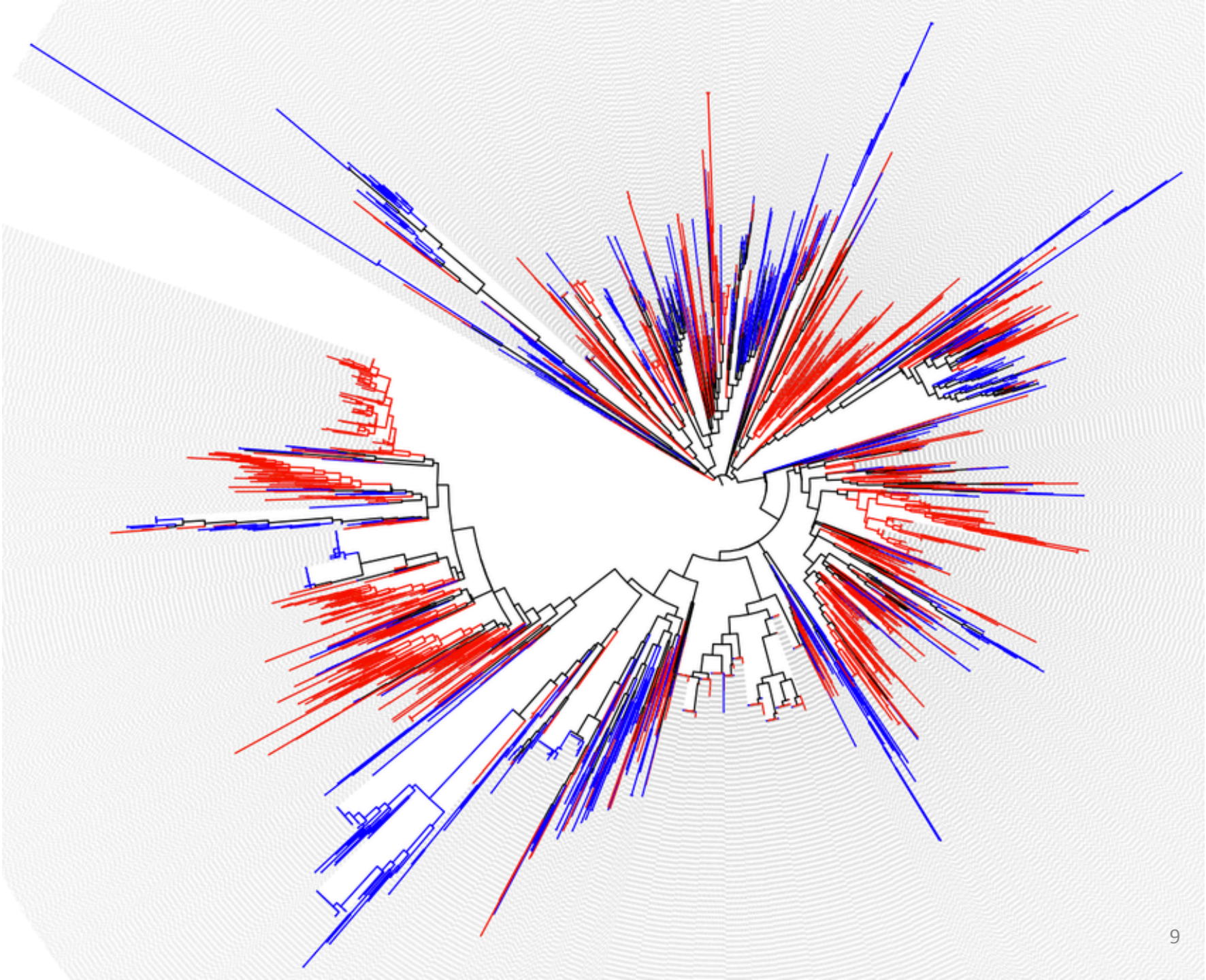
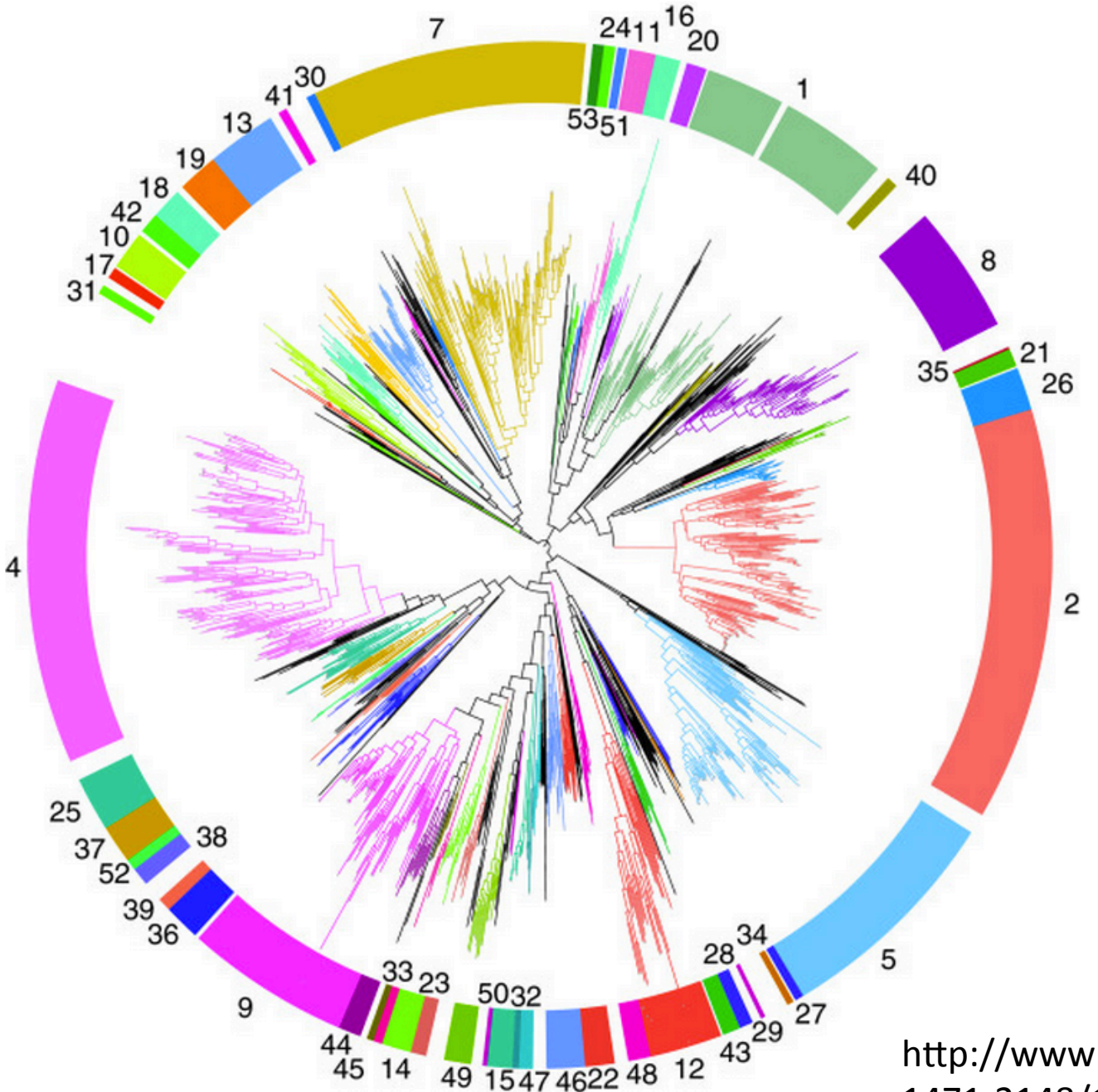


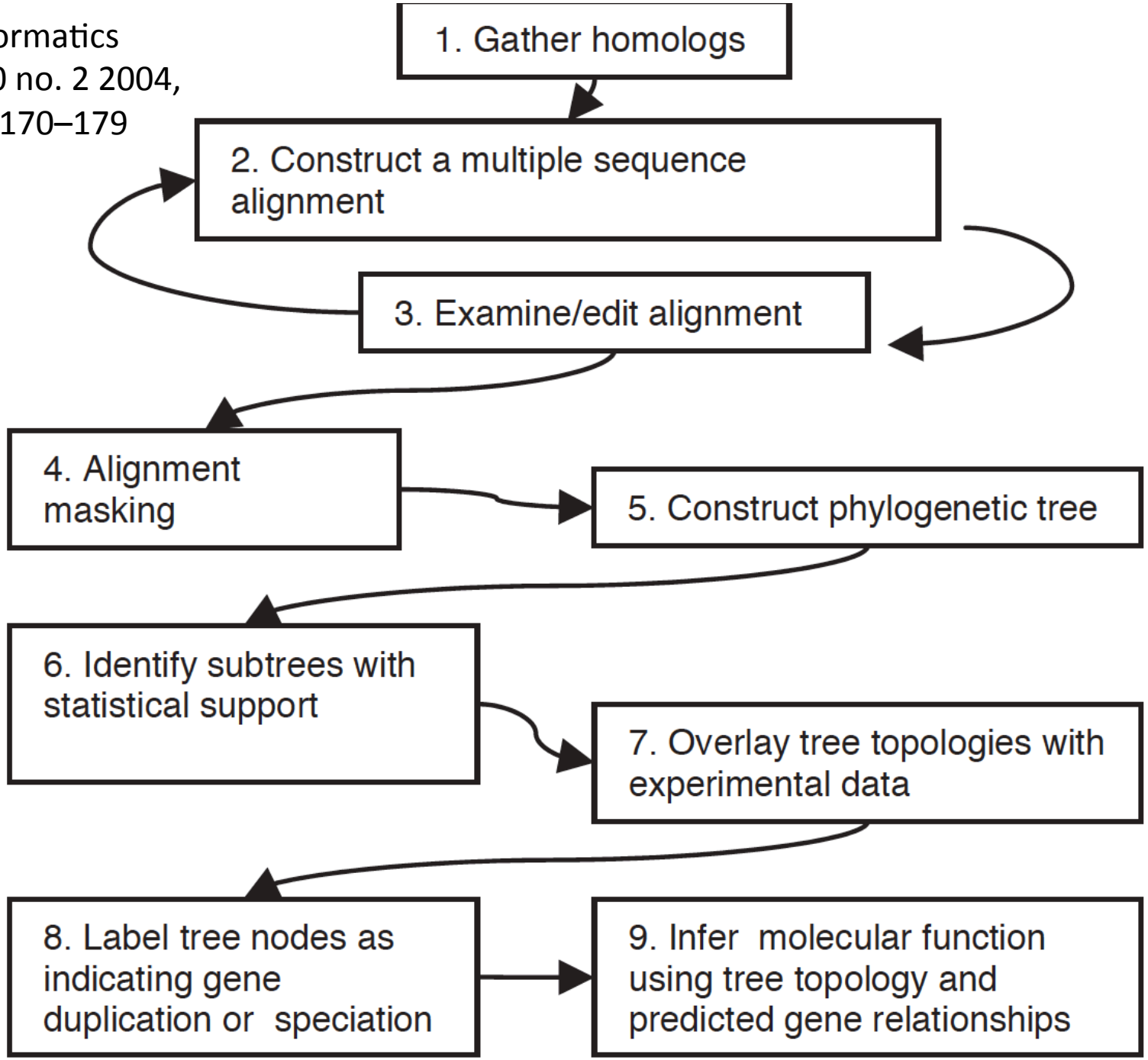
Figure 1.



http://www.biomedcentral.com/1471-2148/12/186

Phylogenetic tree of family GH5. In this circular phylogram, the branches corresponding to subfamilies 1–53

Bioinformatics
Vol. 20 no. 2 2004,
pages 170–179



Step 1. Assembling a dataset

BLAST, FASTA, domain/family based (HMMER)

Step 2. Multiple sequence alignment

MAFFT, MUSCLE, Clustal Omega

Step 3. Phylogeny reconstruction

MEGA5, PHYML, RAxML, GARLI, MrBayes, FastTree

Step 4. Tree visualization

TreeView, TreeDyn, MEGA5, iTOL

Phylogenetic trees are calculated by applying mathematical models to infer evolutionary relationships between molecules or organisms (here sequences), based on a set of characters that describe their differences.

Four main categories of phylogenetic reconstruction methods:

1. Maximum parsimony approaches create trees using the minimum number of ancestors needed to explain the observed characters

2. Distance matrix methods, such as neighbor joining, allow more sophisticated evolutionary models than parsimony

3. Maximum likelihood methods search a set of tree and evolutionary models to find the ones most likely to generate the observed characters

4. Bayesian approaches offer more flexibility, as they allow optimization of all aspects of a tree (model, topology, branch length)

Syst. Biol. 55(2):314–328, 2006

Maximum likelihood and Bayesian, in general, outperformed neighbor joining and maximum parsimony in terms of tree reconstruction accuracy.

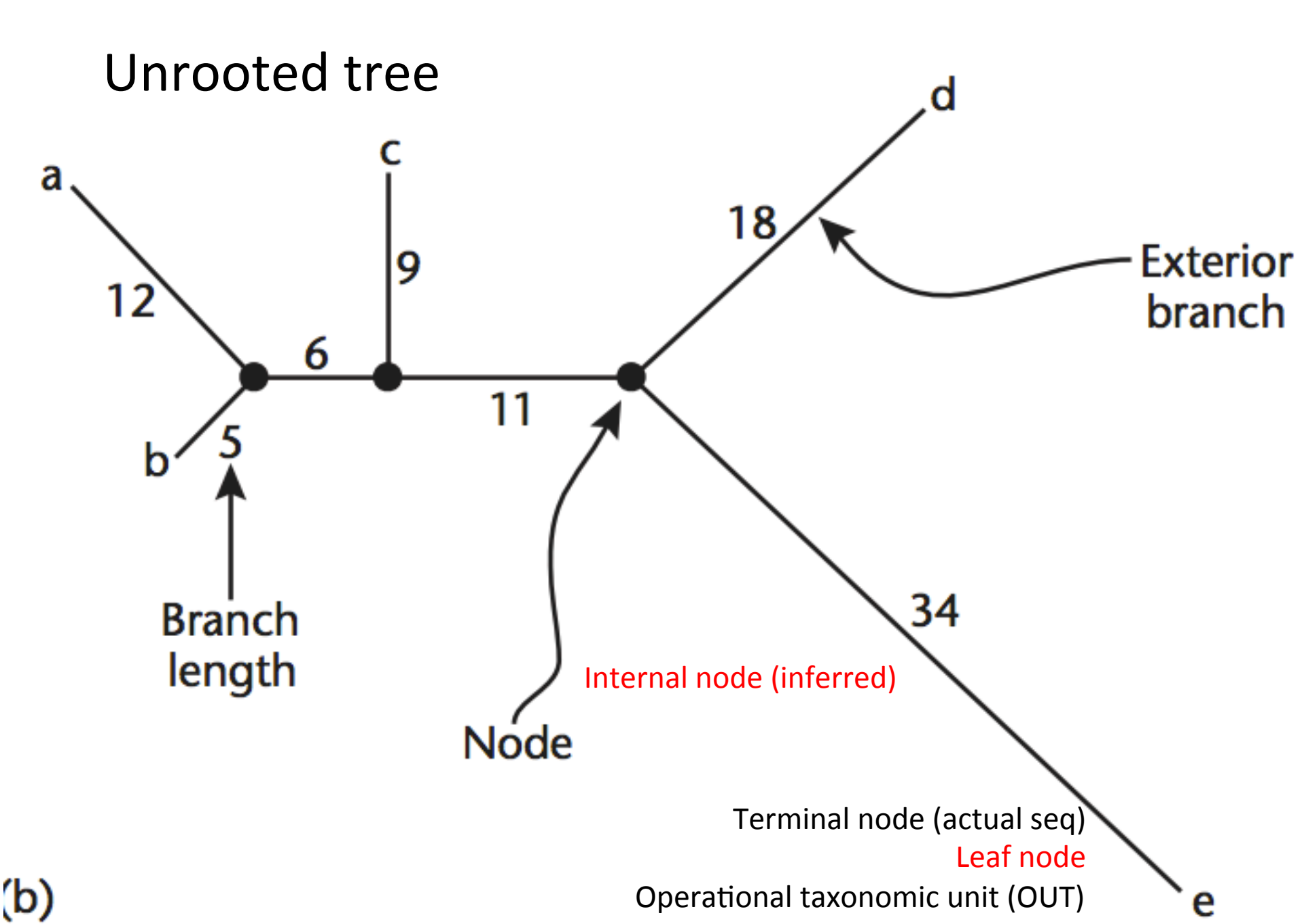
In general, our results indicate that as alignment error increases, topological accuracy decreases.

Results also indicated that as the length of the branch and of the neighboring branches increase, alignment accuracy decreases, and the length of the neighboring branches is the major factor in topological accuracy.

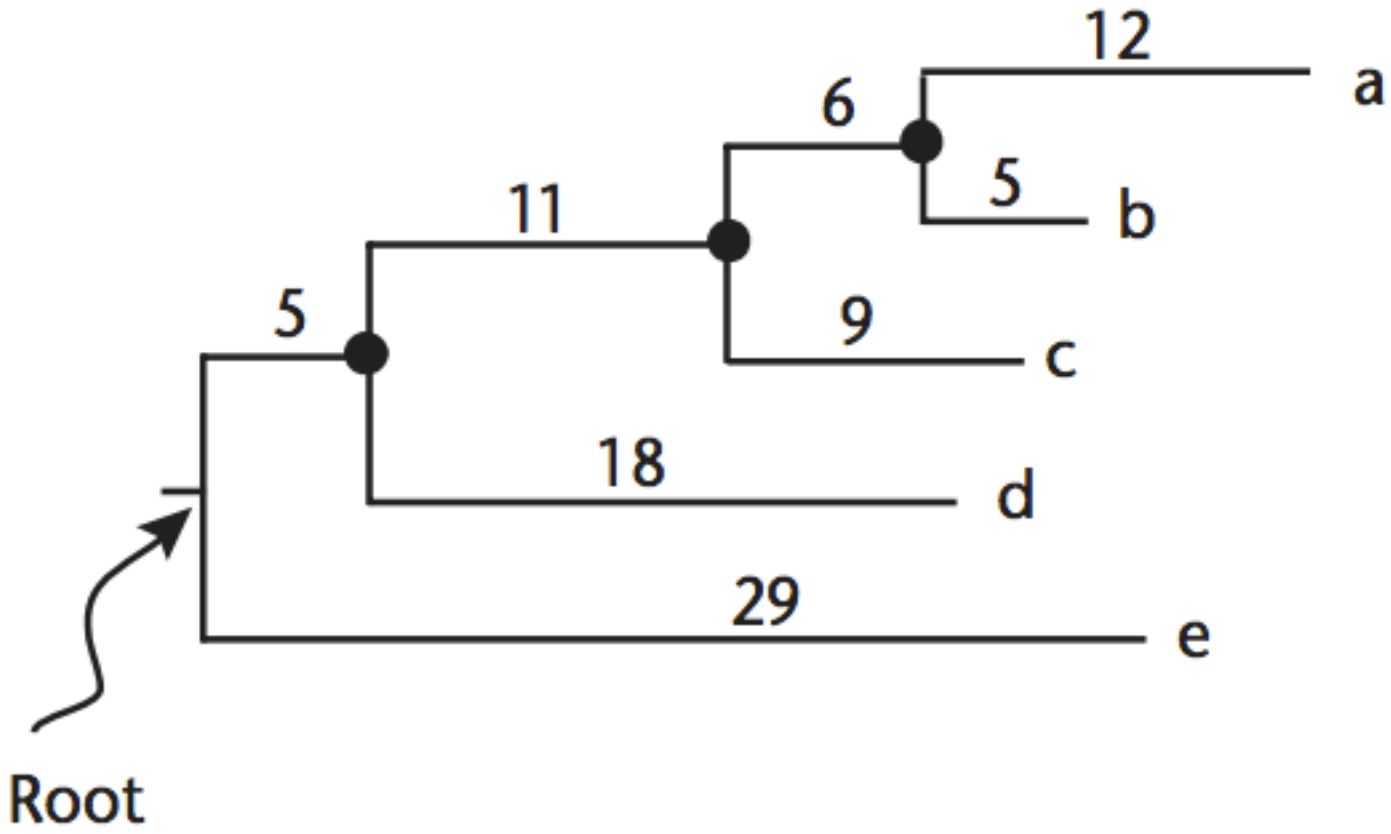
Mol Biol Evol (2005) 22 (3): 792-802.

Over the variety of conditions tested, Bayesian trees estimated from DNA sequences that had been aligned according to the alignment of the corresponding protein sequences were the most accurate, followed by Maximum Likelihood trees estimated from DNA sequences and Parsimony trees estimated from protein sequences

Unrooted tree

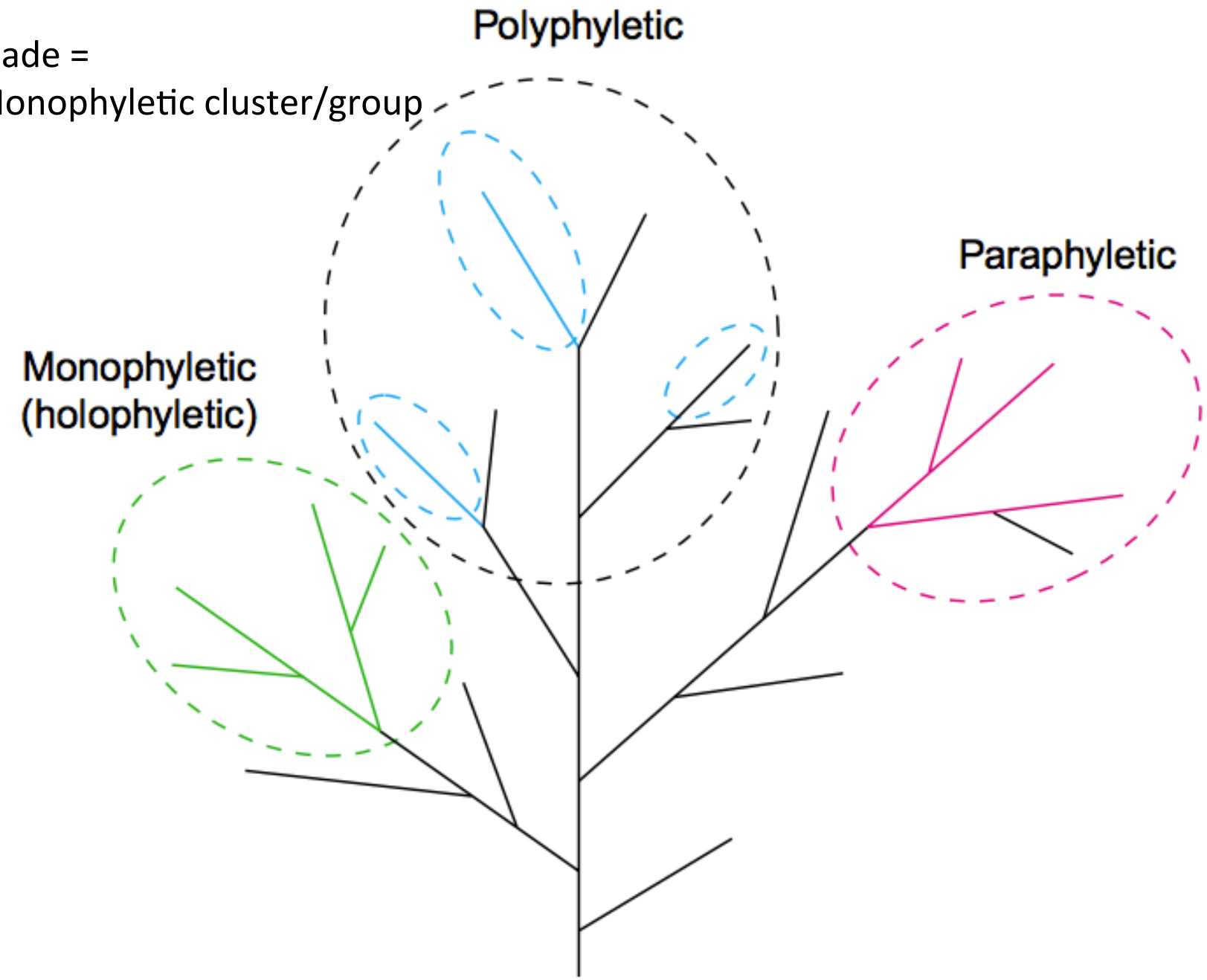


Rooted tree
Root is often selected based on prior knowledge

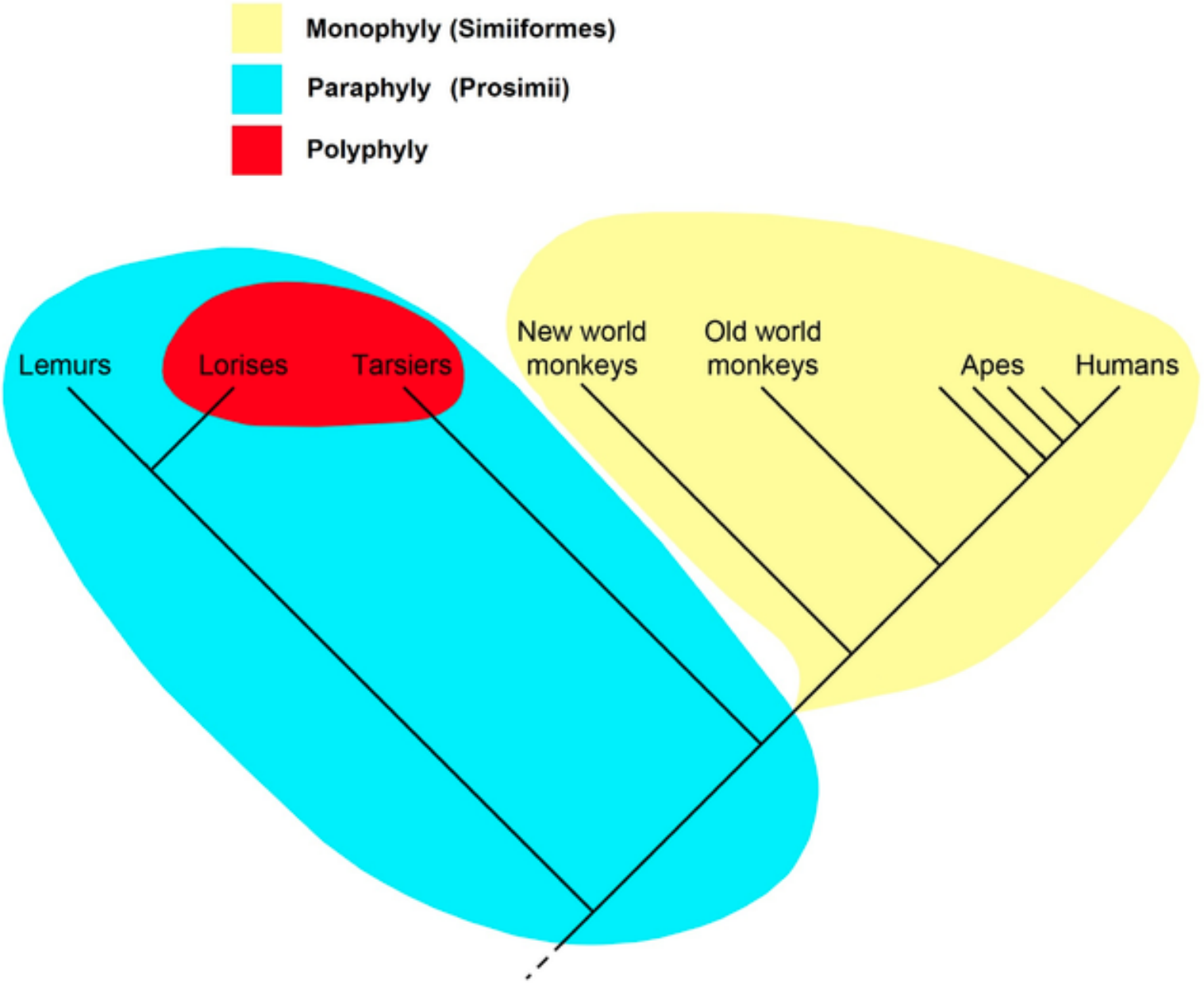


Branches are drawn with lengths proportional to the divergence (difference) between two nodes

Clade =
Monophyletic cluster/group



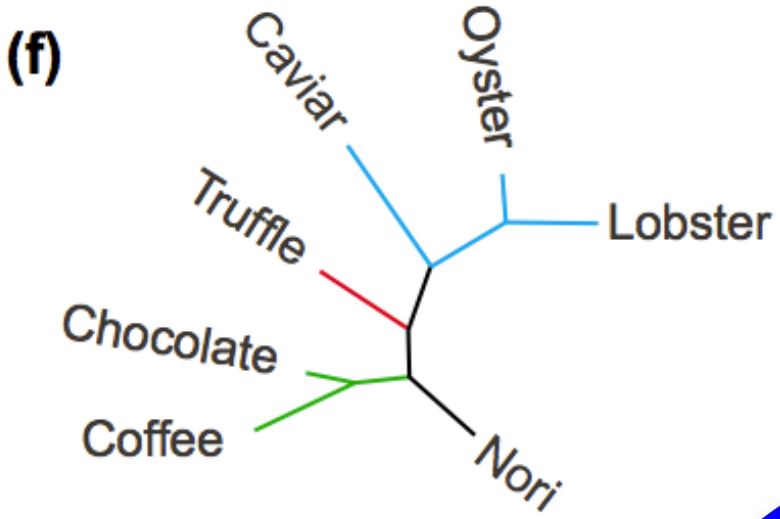
TRENDS in Genetics

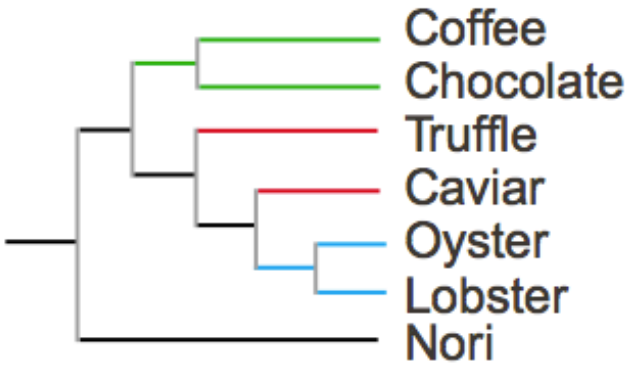


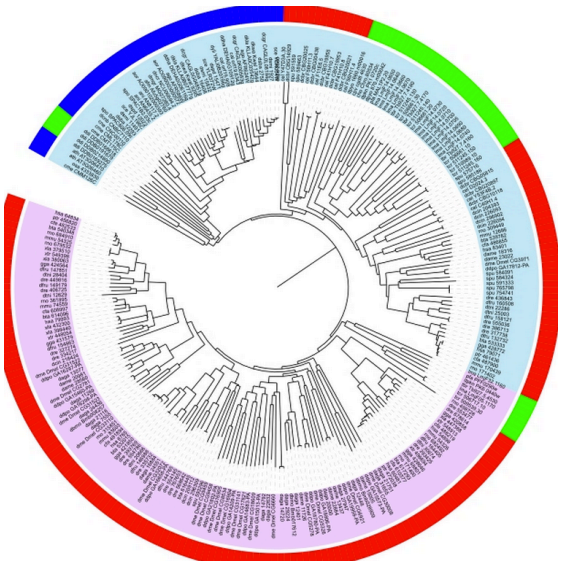
http://en.wikipedia.org/wiki/File:Monophyly,_paraphyly,_polyphyly.png

Radial view

Rectangular view

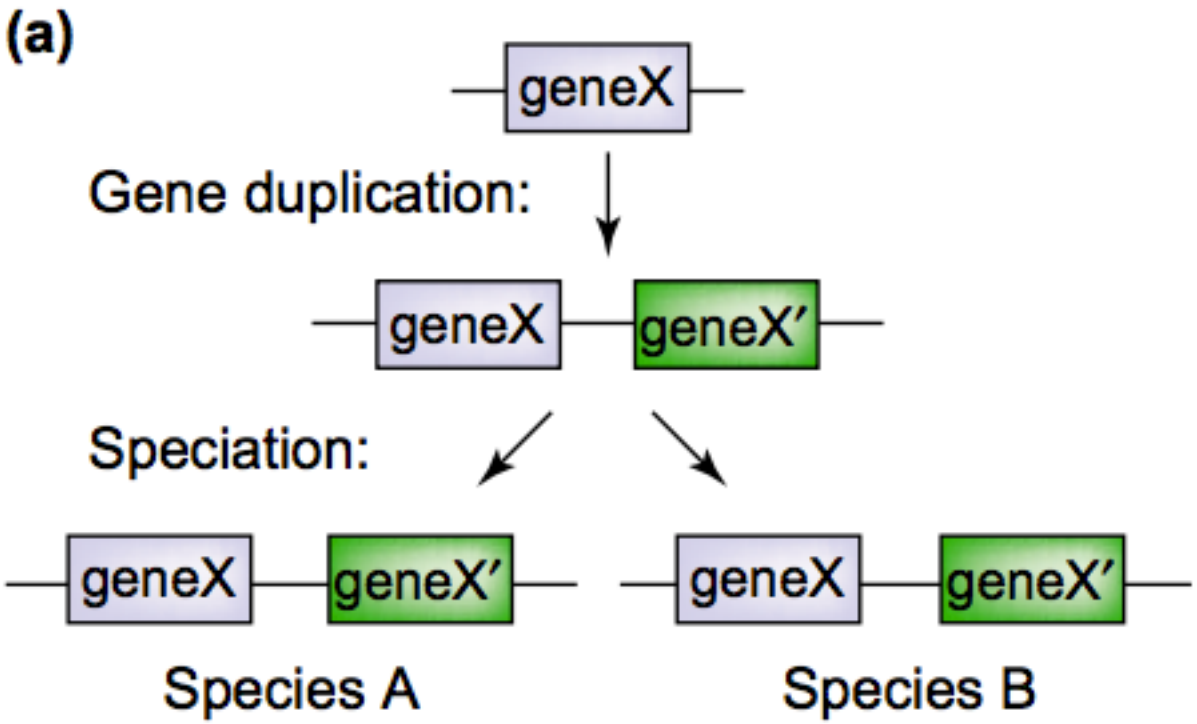


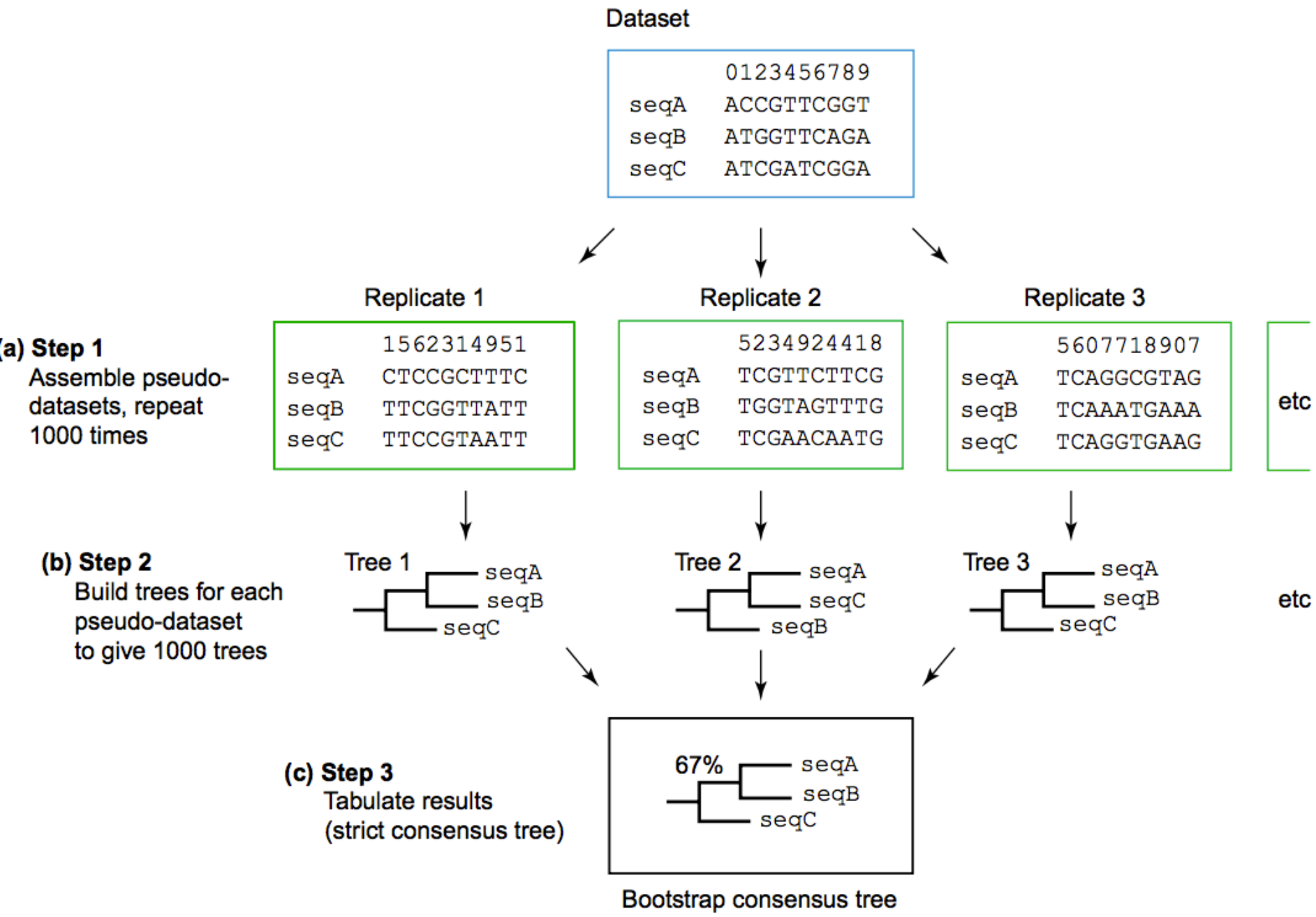




TRENDS in Genetics

Circular view


(a)
geneX
Gene duplication:
geneX
geneX′
Speciation:
geneX
geneX′
geneX
geneX′
Species A
Species B


Dataset
0123456789
seqA ACCGTTCGGT
seqB ATGGTTCAGA
seqC ATCGATCGGA
Replicate 1
Replicate 2
Replicate 3
(a) Step 1
Assemble pseudo-datasets, repeat 1000 times
1562314951
seqA CTCCGCTTTC
seqB TTCGGTTATT
seqC TTCCGTAATT
5234924418
seqA TCGTTCTTCG
seqB TGGTAGTTTG
seqC TCGAACAATG
5607718907
seqA TCAGGCGTAG
seqB TCAAATGAAA
seqC TCAGGTGAAG
etc
(b) Step 2
Build trees for each pseudo-dataset to give 1000 trees
Tree 1
seqA
seqB
seqC
Tree 2
seqA
seqC
seqB
Tree 3
seqA
seqB
seqC
etc
(c) Step 3
Tabulate results (strict consensus tree)
67%
seqA
seqB
seqC
Bootstrap consensus tree

Hands on practice: see the Feb26-2013-2.pdf

Next class: advanced manipulation and visualization of tree graphs