Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 24

1

TODAY'S TOPICS

- Phylogenetic analysis/visualization software
  - Transfer bootstrap expection
  - Software and visualization
- Protein structure databases
  - Protein structures
  - Secondary structure assignment
  - Protein structure databases (PDB)

2

Renewing Felsenstein's phylogenetic bootstrap in the era of big data

F. Lemoine[1,2], J.-B. Domelevo Entfellner[3,4], E. Wilkinson[5], D. Correia[1], M. Dávila Felipe[1], T. De Oliveira[5,6] & O. Gascuel[1,7]*

452 | NATURE | VOL 556 | 26 APRIL 2018

Transfer distance (or R distance):

The minimum number of elements to be transferred (or removed) to transform on partition into the other

(S1, S2), (S3, S4, S5) — Distance=0 → (S1, S2), (S3, S4, S5)

(S1, S2, S3), (S4, S5) — Distance=1 → (S1, S2), (S3, S4, S5)

Reference (original) tree T

Bootstrap tree T*

(S1, S2, S3), (S4, S5) $\delta(b,b1^*)=1$ (S1, S2), (S3,S4, S5) (S1, S2, S4), (S3,S5) $\delta(b,b2^*)=2$

$\delta(b,b^*)$: Distance between branch b in tree T and branch b^* in BS tree T^*:
- Compare the two partitions defined by branch b in tree T and all branches in BS tree T*
- Find the b^* with the minimum $\delta(b,b^*)$

Transfer Bootstrap Expectation (TBE) for branch b:

$$\mathrm{TBE}(b) = 1 - \frac{\phi(b, T^*)}{p-1}$$

Average from all BS trees

p: number of taxa from the smaller of the two clusters

$\phi(b,T^*)=\min_{b^* \in T^*} \delta(b,b^*)$

BS: count only identical partitions

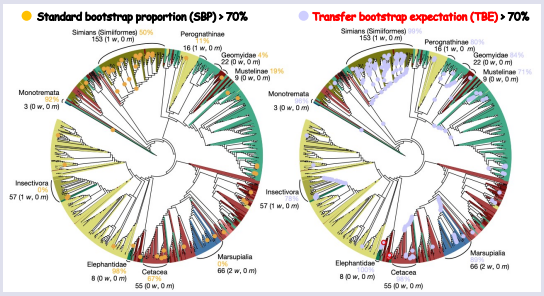
TBE: include most similar partitions (no need to be identical)

3

Transfer bootstrap for big phylogenies

Renewing Felsenstein's phylogenetic bootstrap in the era of big data

Robustness of Felsenstein's Versus Transfer Bootstrap Supports With Respect to Taxon Sampling



- Deep branches in large phylogenies are often not supported by SBP
- TBE supports are higher without inducing falsely supported branches

4

PHYLIP format for input alignment

PHYLIP

https://phylipweb.github.io/phylip/

Phylip tutorial on Canvas

Sequence number

Alignment length

(Interleaved alignment format)

```
 4    125
1c2r        -----GDAAK GEKEFN-KCK TCHSIIAPDG TEIVKGAKTG PNLYGVVGRT
1ycc        TEFKAGSAKK GATLFKTRCL QCHTVEKGG- -----PHKVG PNLHGIFGRH
3c2c        ----EGDAAA GEKVSK-KCL ACHTFDQGG- -----ANKVG PNLFGVFENT
1etp        ----AGDAEA GQGKVA-VCG ACHGVDGNS- --------PA PN----FPKL

            AGTYPEFKYK DSIVALGASG FAWTEEDIAT YVKDPGAFLK EKLDDKKAKT
            SGQAEGYSYT DAN---IKKN VLWDENNMSE YLTNPKKYIP G--------T
            AAHKDNYAYS ESYTEMKAKG LTWTEANLAA YVKNPKAFVL EKSGDPKAKS
            AGQGERYLLK QLQ-DIKAGS TPGAPEGVGR KVLEMTGMLD P---------

            GMAFK-LAKG GE--DVAAYL ASVVK
            KMAFGGLKKE KDRNDLITYL KKACE
            KMTFK-LTKD DEIENVIAYL KTLK-
            ---LS----D QDLEDIAAYF SSQKG
```

This space can be kept empty. Or the 2nd, 3rd, ... alignment blocks can start without keeping this empty space.

But no sequence names.

Sequence names should be ten characters in length.
Add spaces AFTER the sequence name if names are too short.
Do not include space or "()[]:;," in the sequence names.

5

Strict *vs.* relaxed PHYLIP alignment format

Strict Phylip format

```
 4    20
1c2r        -----GDAAK GEKEFN-KCK
1ycc        TEFKAGSAKK GATLFKTRCL
3c2c        ----EGDAAA GEKVSK-KCL
1etp        ----AGDAEA GQGKVA-VCG
```

```
 4    20
1c2r567890-----GDAAK GEKEFN-KCK
1yccXXXXXXTEFKAGSAKK GATLFKTRCL
3c2cYYYYYY----EGDAAA GEKVSK-KCL
1etpZZZZZZ----AGDAEA GQGKVA-VCG
```

- Only up to 10 characters are used as the sequence names.
- No space is needed between names and sequences.

In the strict format: Names > 10 letters will be truncated!

Relaxed phylip format

```
 4    20
1c2rABCDEFGHI -----GDAAK GEKEFN-KCK
1c2rABCDEFGH  TEFKAGSAKK GATLFKTRCL
1c2rABCDEFG   ----EGDAAA GEKVSK-KCL
1c2rABCDE     ----AGDAEA GQGKVA-VCG
```

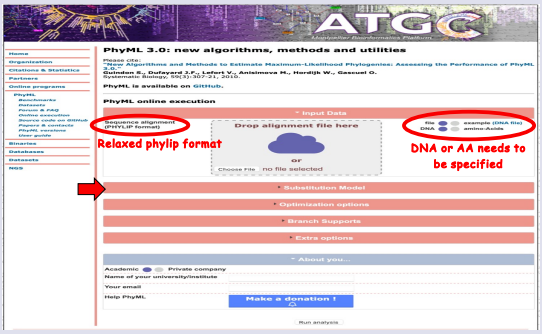
- Relaxed format allows to have longer sequence names.
- At lease one space is needed between names and sequences.
- Used in PhyML, RAxML, *etc.*

- Relaxed format is not supported by PHYLIP (sequence data start at 11th character)

6

Maximum likelihood phylogeny: PhyML

http://www.atgc-montpellier.fr/phyml/

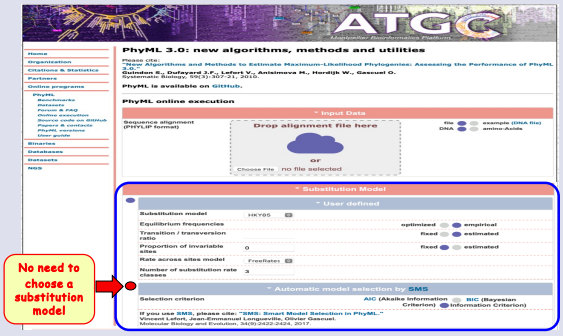


BIOS477/877 L24 - 7

7

Maximum likelihood phylogeny: PhyML

http://www.atgc-montpellier.fr/phyml/

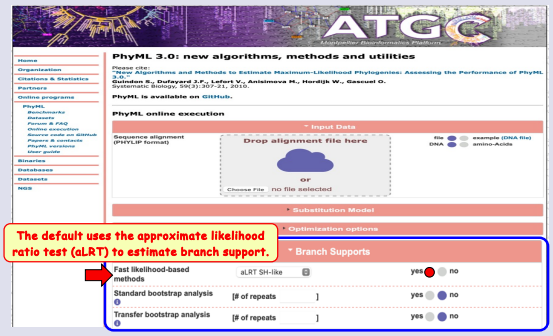


BIOS477/877 L24 - 8

8

Maximum likelihood phylogeny: PhyML

http://www.atgc-montpellier.fr/phyml/

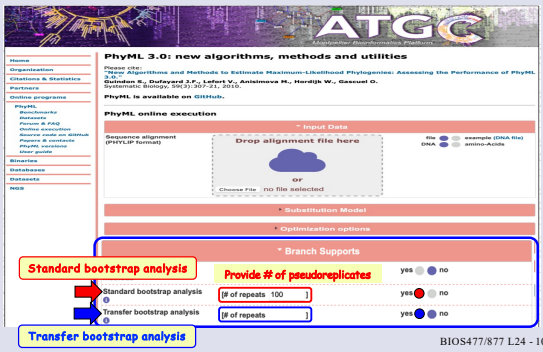


BIOS477/877 L24 - 9

9

Maximum likelihood phylogeny: PhyML

http://www.atgc-montpellier.fr/phyml/



BIOS477/877 L24 - 10

10

Maximum likelihood phylogeny: PhyML

```
                        ---  PhyML 3.3.20190909  ---
                http://www.atgc-montpellier.fr/phyml
                Copyright CNRS - Universite Montpellier

. Sequence filename:                          bglobin_phy
. Data set:                                   #1
. Initial tree:                               BioNJ
. Model of amino acids substitution:          LG
. Number of taxa:                             7
. Log-likelihood:                             -1985.89283
. Unconstrained log-likelihood:  -849.15979   -3252.62416
. Composite log-likelihood:                   386
. Parsimony:                                  6.04806
. Tree size:                                  Yes
. Discrete gamma model:                       4
  - Number of classes:
  - Gamma shape parameter:                    3.801
  - Relative rate in class 1:                 0.44308 [freq=0.250000]
  - Relative rate in class 2:                 0.76803 [freq=0.250000]
  - Relative rate in class 3:                 1.08438 [freq=0.250000]
  - Relative rate in class 4:                 1.70451 [freq=0.250000]

. Run ID:                                     none
. Random seed:                                1586748614
. Subtree patterns aliasing:                  no
. Version:                                    3.3.20190909
. Time used:                                  0h0m2s (2 seconds)
```

Substitution model selected

Gamma parameters estimated

Approximate Likelihood-Ratio Test (aLRT) values

```
(HBB_HUMAN:0.01369099,HBB_HORSE:0.18953198,((HBA_HUMAN:0.03417792,HBA_HORSE:0.11210715)0.877000:0.21433432,(GLB5_PETMA:0.67451297,(MYG_PHYCA:1.35997426,LGB2_LUPLU:1.98802500)0.066000:0.17694980)0.968000:0.53663911)1.000000:0.75082181);
```

Standard bootstrap values [10 pseudoreplicates] (bootstrap values are shown in absolute # (not in proportion)

```
(HBB_HUMAN:0.01364377,HBB_HORSE:0.18957711,((GLB5_PETMA:0.67450654,(LGB2_LUPLU:1.98820683,MYG_PHYCA:1.36008247)4:0.17680914)8:0.53665495,(HBA_HUMAN:0.03417817,HBA_HORSE:0.11210788)7:0.21432437)9:0.75087832);
```

Transfer bootstrap values [10 pseudoreplicates]

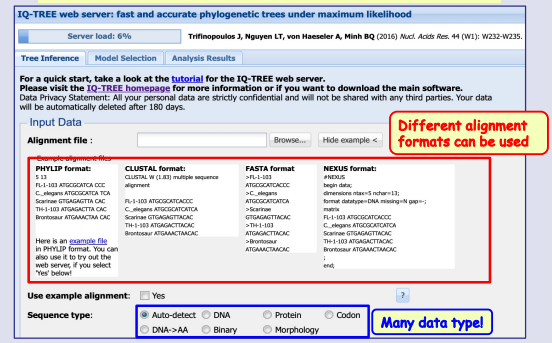
```
(HBB_HUMAN:0.01334429,HBB_HORSE:0.18931947,((MYG_PHYCA:1.64464609,(GLB5_PETMA:
0.63323591,LGB2_LUPLU:2.28443966)0.600000:0.15967461)0.950000:0.49889968,(HBA_HUMAN:
0.03507174,HBA_HORSE:0.11042045)0.900000:0.18778014)0.900000:0.79016985);
```

L24 - 11

11

Maximum likelihood phylogeny: IQ-TREE

http://www.iqtree.org; http://iqtree.cibiv.univie.ac.at (other websites available)

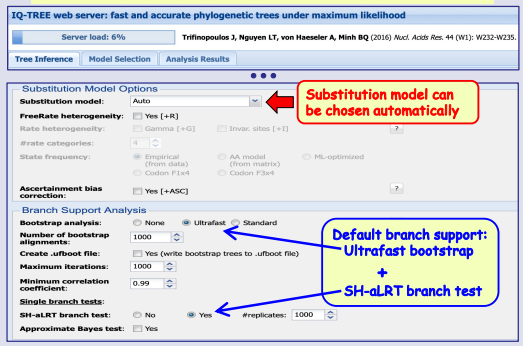


BIOS477/877 L24 - 12

12

Maximum likelihood phylogeny: IQ-TREE

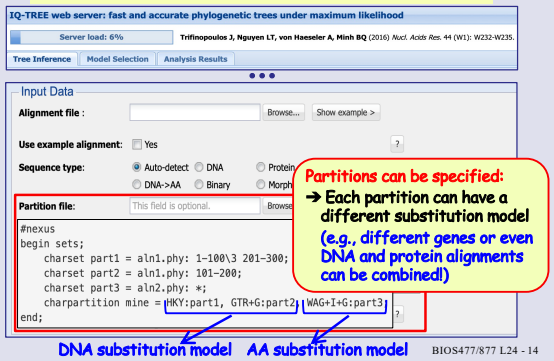
http://www.iqtree.org; http://iqtree.cibiv.univie.ac.at (other websites available)



13

Maximum likelihood phylogeny: IQ-TREE

http://www.iqtree.org; http://iqtree.cibiv.univie.ac.at (other websites available)



14

Phylogenetic method programs/websites

- PhyML http://www.atgc-montpellier.fr/phyml/ (3.0)
  - ➜ Smart Model Selection is included (no need to choose a substitution model).
  - ➜ Standalone or web interface (aLRT SH-like, standard bootstrap, and transfer bootstrap)
- RAxML http://sco.h-its.org/exelixis/web/software/raxml/index.html
  - ➜ Randomized Axelerated Maximum Likelihood (includes rapid bootstrap)
- RAxML-NG https://github.com/amkozlov/raxml-ng/releases
  - ➜ Faster and more accurate than RAxML, includes transfer bootstrap
- IQ-TREE http://www.iqtree.org/ (IQ-TREE 2 also available)
  - ➜ A fast and effective stochastic algorithm for estimating ML phylogenies
  - ➜ Includes model selection and ultrafast bootstrap; IQ-TREE 2 includes transfer bootstrap
- FastTree 2 http://microbesonline.org/fasttree/
  - ➜ Approximately-maximum likelihood phylogenetic trees (standalone only)
- Booster https://booster.pasteur.fr/ (transfer bootstrap with PhyML and FastTree)
- Los Alamos databases and tools
  https://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html (PhyML, IQ-TREE)
- MrBayes Bayesian Inference of Phylogeny https://nbisweden.github.io/MrBayes/index.html
- CIPRES (Cyberinfrastructure for Phylogenetic Research) http://www.phylo.org/
  - ➜ High performance parallel codes: RAxML-NG, MrBayes, GARLI, Bali-Phy, PAUP, IQ-Tree, etc.
- NGPhylogeny.fr https://ngphylogeny.fr/
  - ➜ includes PhyML, MrBayes, FastME, TNT, etc., also BMGE, Gblocks, etc. for filtering
- Phylogeny.fr http://www.phylogeny.fr

15

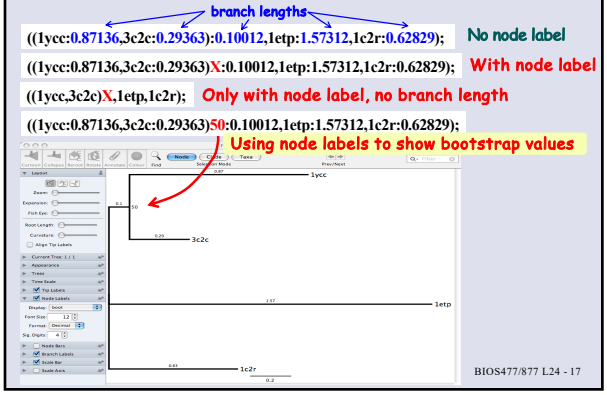
More phylogeny programs/websites

- MEGA X http://www.megasoftware.net/
  - ➜ UPGMA, NJ, ME, MP, ML (include visualization)
- Phylip3.698 https://phylipweb.github.io/phylip/
  - ➜ Includes: FM, UPGMA, NJ, ME, MP, ML
  - ➜ Rphylip, an R interface for Phylip: http://www.phytools.org/Rphylip/

  Phylip tutorial on Canvas
  Youtube tutorials available
- Phylemon http://phylemon.bioinfo.cipf.es
  - ➜ includes Phylip, PhyML, MrBayes; also ModelTree
- SeaView http://doua.prabi.fr/software/seaview
  - ➜ MSA (Muscle, Clustal Ω) and phylogeny (NJ, BioNJ, MP, PhyML, including transfer bootstrap)
- MAFFT https://mafft.cbrc.jp/alignment/server/
  - ➜ Includes NJ (distance: JTT, WAG, etc.)
- DAMBE http://dambe.bio.uottawa.ca/DAMBE/dambe.aspx
  - ➜ UPGMA, NJ, FastME, FM, MP, ML (many substitution models; visualization)
- CRAN (R projects for phylogenetics) https://cran.r-project.org/web/views/Phylogenetics.html
- Ape (R package for Analysis of Phylogenetics and Evolution)
  https://cran.r-project.org/web/packages/ape/index.html

16

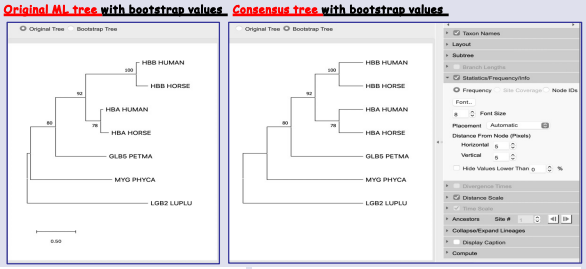
Newick tree format with node labels



17

MEGA Molecular Evolutionary Genetics Analysis (https://www.megasoftware.net)

Original ML tree with bootstrap values Consensus tree with bootstrap values



((((HBB_HUMAN:0.04045989,HBB_HORSE:0.15320710)1.0000:0.64439084,(HBA_HUMAN:0.02935227,HBA_HORSE:0.10205076)0.7800:0.19434686)0.9200:0.38828004,GLB5_PETMA:0.65024328)0.8000:0.31847481,MYG_PHYCA:1.04418589,LGB2_LUPLU:1.59298882);

((((HBB_HUMAN,HBB_HORSE)1.0000,(HBA_HUMAN,HBA_HORSE)0.7800)0.9200,GLB5_PETMA)0.8000,MYG_PHYCA,LGB2_LUPLU);

MEGA includes bootstrap values as node labels.
No branch lengths are included in the consensus tree.

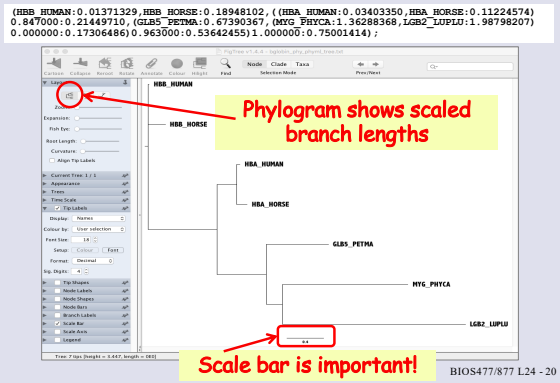
18

Visualization of phylogeny

- FigTree (Macintosh, Windows, Linux/Unix) http://tree.bio.ed.ac.uk/software/figtree/
- iTOL (Web service) http://itol.embl.de/
- TreeViewer (Macintosh, Windows, Linux/Unix) https://treeviewer.org/
- IcyTree(Web service) https://icytree.org/
- Phylo.io (Web service) http://phylo.io/index.html
- PRESTO (Web service) http://www.atgc-montpellier.fr/presto/
- EvolView v2 (Web service) http://www.evolgenius.info/evolview/
- Phylogenetic tree (newick) viewer (Web service) http://etetoolkit.org/treeview/

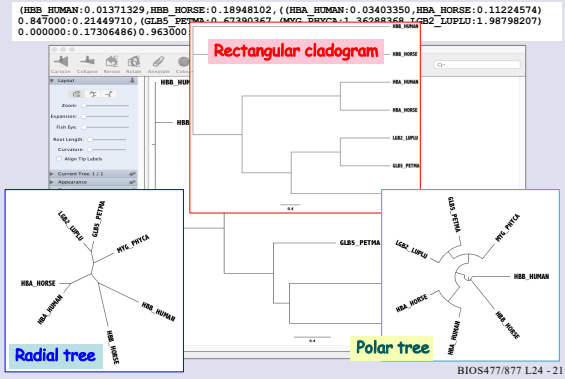
19

FigTree (http://tree.bio.ed.ac.uk/software/figtree/)



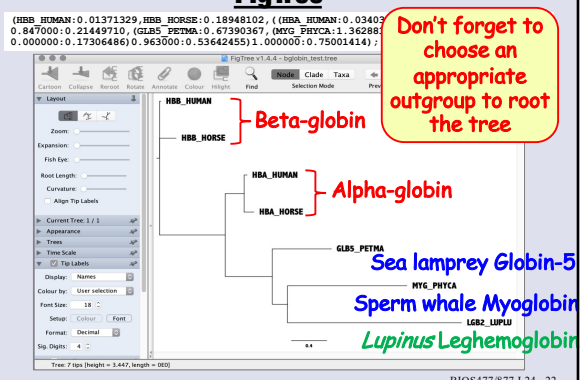
20

FigTree



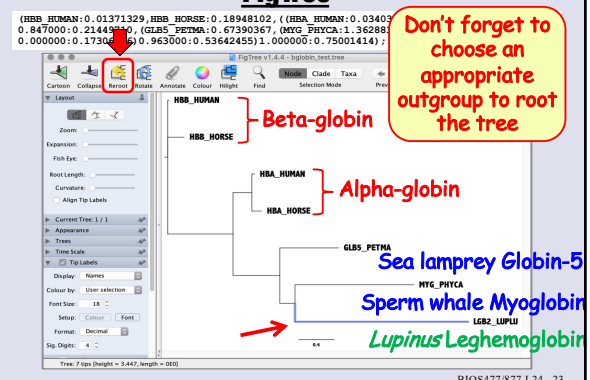
21

FigTree



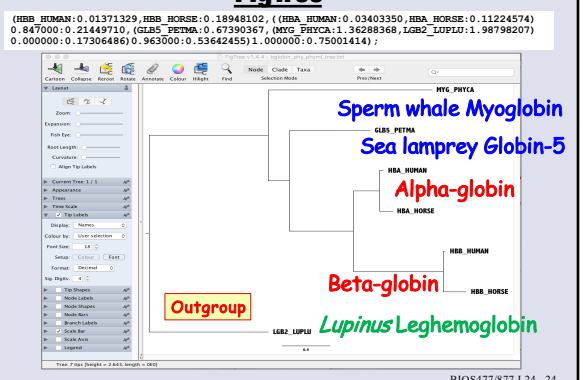
22

FigTree



23

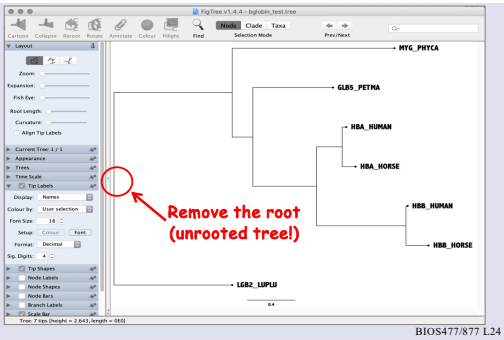
FigTree



24

FigTree

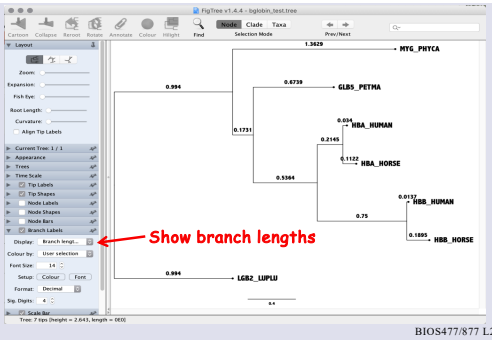
(HBB_HUMAN:0.01371329,HBB_HORSE:0.18948102,((HBA_HUMAN:0.03403350,HBA_HORSE:0.11224574)
0.847000:0.21449710,(GLB5_PETMA:0.67390367,(MYG_PHYCA:1.36288368,LGB2_LUPLU:1.98798207)
0.000000:0.17306486)0.963000:0.53642455)1.000000:0.75001414);



25

FigTree

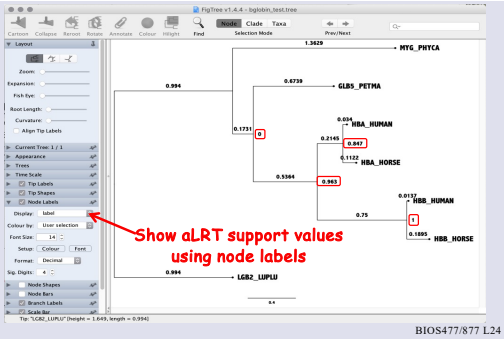
(HBB_HUMAN:0.01371329,HBB_HORSE:0.18948102,((HBA_HUMAN:0.03403350,HBA_HORSE:0.11224574)
0.847000:0.21449710,(GLB5_PETMA:0.67390367,(MYG_PHYCA:1.36288368,LGB2_LUPLU:1.98798207)
0.000000:0.17306486)0.963000:0.53642455)1.000000:0.75001414);



26

FigTree: aLRT support as node labels

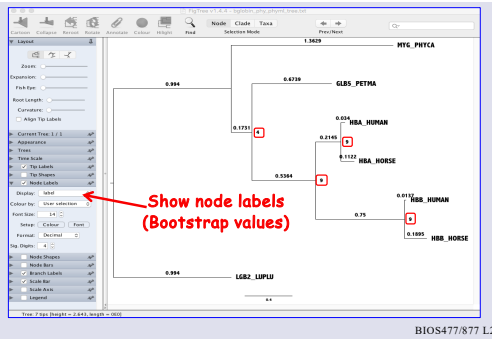
(HBB_HUMAN:0.01371329,HBB_HORSE:0.18948102,((HBA_HUMAN:0.03403350,HBA_HORSE:0.11224574)
0.847000:0.21449710,(GLB5_PETMA:0.67390367,(MYG_PHYCA:1.36288368,LGB2_LUPLU:1.98798207)
0.000000:0.17306486)0.963000:0.53642455)1.000000:0.75001414);



27

FigTree: bootstrap support as node labels

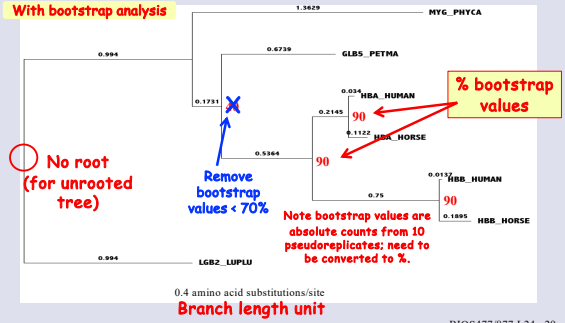
(HBB_HUMAN:0.01371329,HBB_HORSE:0.18948102,((HBA_HUMAN:0.03403350,HBA_HORSE:0.11224574)9:
0.21449710,(GLB5_PETMA:0.67390367,(MYG_PHYCA:1.36288368,LGB2_LUPLU:1.98798207)4:
0.17306486)9:0.53642455)9:0.75001414);



28

Bootstrap values as node labels

(HBB_HUMAN:0.01371329,HBB_HORSE:0.18948102,((HBA_HUMAN:0.03403350,HBA_HORSE:0.11224574)9:
0.21449710,(GLB5_PETMA:0.67390367,(MYG_PHYCA:1.36288368,LGB2_LUPLU:1.98798207)4:
0.17306486)9:0.53642455)9:0.75001414);



29

IQ-Tree and iTOL (http://itol.embl.de/)

IQ-Tree Newick format tree file:

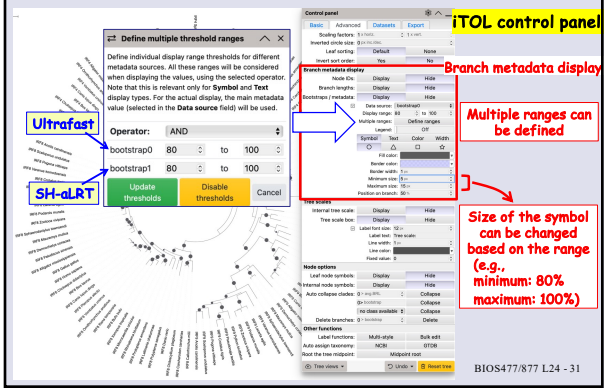
(IRF4 Homo sapiens:0.0261131273,IRF4 Pteropus alecto:0.0178822599,(((IRF4 Canis lupus dingo:0.0220230929,IRF4 Bos taurus:0.0459977745)88.5/100:0.0079686246,(IRF4 Vombatus ursinus:0.0699729028,(IRF4 Ornithorhynchus anatinus:0.0336223152,(((IRF4 Gallus gallus:0.0482000904,IRF4 Alligator mississippiensis:0.0658074429)92.4/99:0.0181052721,((IRF4 Mauremys mutica:0.0000026997,IRF4 Dermochelys coriacea:0.0103795045)42.4/95:0.0058138194,IRF4 Pelodiscus sinensis:0.0125162652)94.9/100:0.0190368037)92/99:0.0244166567,((((IRF4 Varanus komodoensis:0.0302789819,(((IRF4 Zootoca vivipara:0.0024867736, ...

IQ-Tree can provide two branch support values (e.g., ultrafast bootstrap/SH-aLRT)

iTOL can use both information

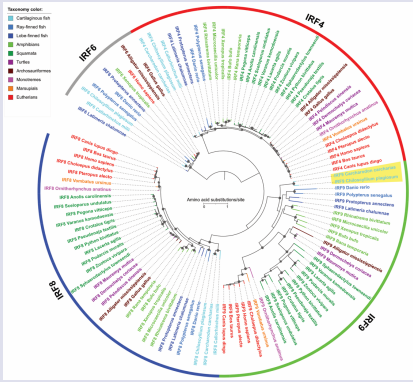
30

IQ-Tree and iTOL (http://itol.embl.de/)



31

IQ-Tree and iTOL (http://itol.embl.de/)

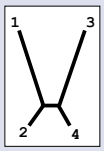


More examples available on the tree gallery

32

Phylogenetic methods: pros and cons

- Criticisms to distance methods
  - Depend on distance estimation method
  - Summarizing a set of sequences by a pairwise distance matrix loses information
- Criticisms to maximum parsimony methods
  - "Long branch attraction" problem
    - → If the internal branch is short relative to the terminal branches, by chance 1 and 3 may acquire the same nucleotide independently
    - → ((1,3),(2,4)) may become the MP tree
  - No multiple hit correction
- Criticisms to maximum likelihood methods
  - Require an explicit model of evolution
    - → Which model to use?

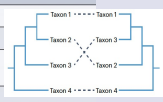


33

What causes phylogenetic incongruence

Steensyk *et al.* (2023); also see Kapli *et al.* (2021)

Driver of incongruence	Factor
Incomplete lineage sorting	Biological
Horizontal gene transfer	Biological
Hybridization or introgression and recombination	Biological
Natural selection	Biological
Sampling (taxon and locus)	Analytical, stochastic error
Insufficient number of genes or divergent sites	Analytical, stochastic error
Erroneous orthologue detection	Analytical, systematic error
Model misspecification	Analytical, systematic error
Multiple sequence alignment errors	Analytical, treatment error
Excessive trimming	Analytical, treatment error
Inappropriate character recoding	Analytical, treatment error



Incongruent gene vs. species trees

Insufficient taxon sampling
Fast evolving lineage
Outgroup selection

Long-branch attraction
Inadequate model complexity

34

Guidelines for phylogenetic analysis

Anisimova *et al.* (2013); Kapli *et al.* (2020)

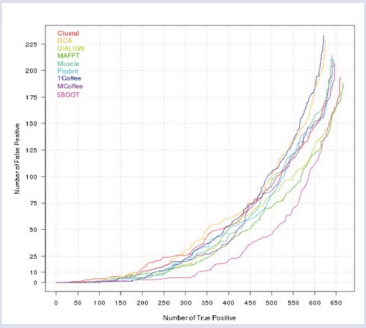
- Data
  - DNA or protein? Coding or noncoding?
  - Quality of data (*e.g.*, sequencing error, gene annotation)
  - Identification of homologues (orthologues only? paralogues?)
  - Carefully choose the outgroup sequence(s)
- Reconstruct multiple sequence alignment
  - Which method?
  - Data trimming/filtering (filter out unreliable alignment regions)
- Selection of substitution model
  - Simplified vs. over-fitting (model testing can be done)
- Phylogenetic inference
  - Analyze data combined (with partitions) or individually
  - Methods (ML, Bayesian, or NJ for speed)
  - Branch support statistics (bootstrap, *etc.*)

But with caution!

35

Incorporating alignment uncertainty into Felsenstein's phylogenetic bootstrap to improve its reliability

Jia-Ming Chang [1,*,†], Evan W. Floden [2], Javier Herrero [1,‡], Olivier Gascuel[3], Paolo Di Tommaso [2] and Cedric Notredame [2,4,*]



Super-MSA: Concatenation of 7 alignments (ClustalW, DCA, Dialign2, Mafft, Muscle, ProbCons, and T-Coffee)

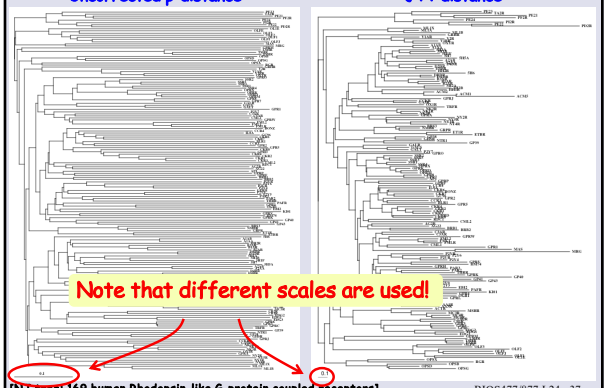
SBOOT: Bootstrap analysis is done using Super-MSA

SBOOT outperformed individual alignments for phylogenetic accuracy.

36

Effect of distance estimation on phylogeny

Uncorrected p-distance | JTT distance



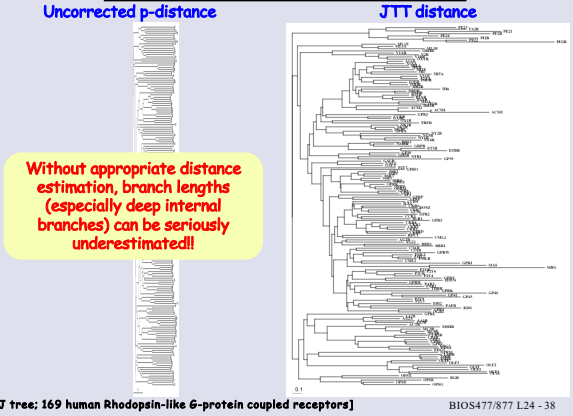
[NJ tree; 169 human Rhodopsin-like G-protein coupled receptors]

BIOS477/877 L24 - 37

37

Effect of distance estimation on phylogeny

Uncorrected p-distance | JTT distance

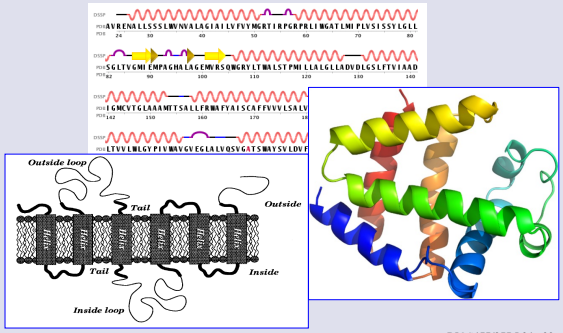


[NJ tree; 169 human Rhodopsin-like G-protein coupled receptors]

BIOS477/877 L24 - 38

38

Prediction of protein structures

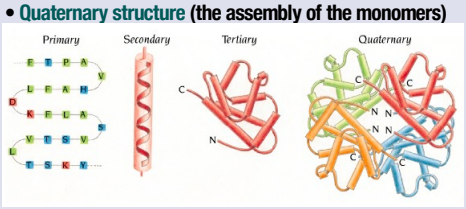


BIOS477/877 L24 - 39

39

Hierarchical nature of protein architecture

- Primary structure: amino acid sequence
- Secondary structure: alpha helices, beta sheet, etc. (hydrogen bonding pattern of main chain):
- Tertiary structure (the assembly and interactions of the helices and sheets)
- Quaternary structure (the assembly of the monomers)

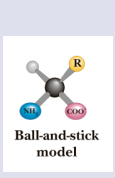


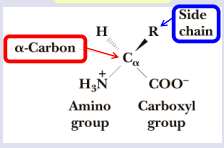
BIOS477/877 L24 - 40

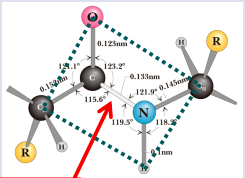
40

Proteins = polypeptide chains

Amino acid

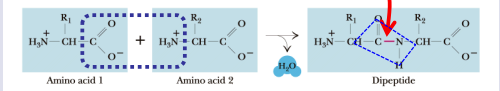






Peptide bond

6 atoms exist in a plane

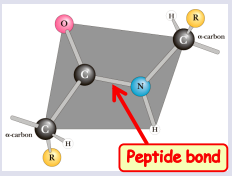


BIOS477/877 L24 - 41

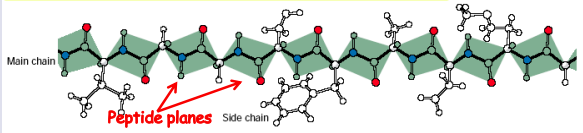
41

Proteins = polypeptide chains

- Backbone (main chain): atoms that participate in peptide bonds
  - → ignores the side chains
  - → 6 atoms (O, C, N, H, and C_α) of the peptide group lie in a plane



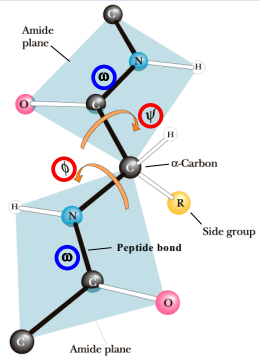
Backbone is a linked sequence of rigid planar peptide group



BIOS477/877 L24 - 42

42

Torsion angles



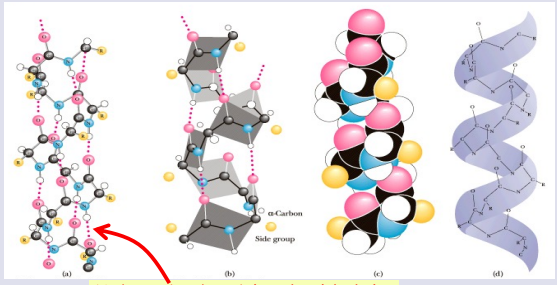
- The conformation of the backbone can be described by the torsion angles (or dihedral angles or rotation angles): ϕ and ψ
  - → The entire path of the backbone is known if ϕ and ψ are specified
- Some values of ϕ and ψ are more likely than others
  - → Due to steric interference between non-bonded atoms

BIOS477/877 L24 - 43

43

Secondary structure

➢ Alpha helix: ϕ = -60 and ψ = -45

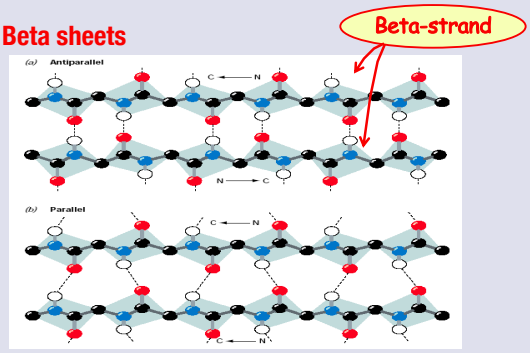


BIOS477/877 L24 - 44

44

Secondary structure

➢ Beta sheets



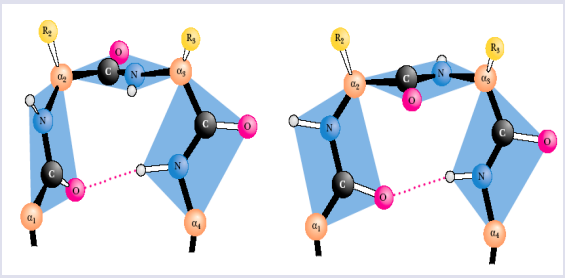
BIOS477/877 L24 - 45

45

Secondary structure

➢ Beta turns

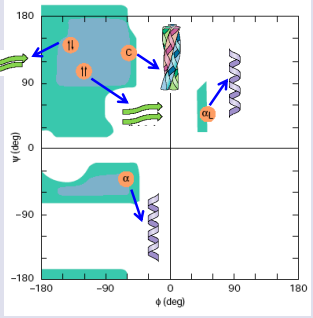
- Proline and glycine occur frequently in beta turns



BIOS477/877 L24 - 46

46

Ramachandran plot



- The sterically allowed values for ϕ and ψ can be calculated (colored regions)

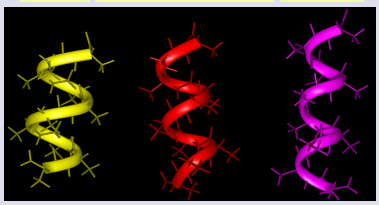
α: right-handed alpha helix
α_L: left-handed alpha helix
↑↑: parallel beta sheet
↑↓: antiparallel beta sheet
C: collagen helix

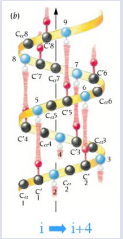
BIOS477/877 L24 - 47

47

Secondary structure conformation

π-helix Righ-handed α-helix 3_{10}-helix





	Φ	Ψ	H-bond pattern
Right handed α-helix	-57	-47	i + 4
π-helix	-57	-70	i + 5
3_{10} helix	-49	-26	i + 3
Parallel β-sheet	-119	113	
Antiparallel β-sheet	-139	135	

BIOS477/877 L24 - 48

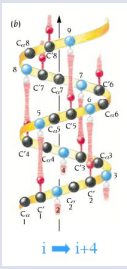
48

Secondary structure assignment

- **DSSP: Dictionary of Secondary Structure of Proteins**
  https://pdb-redo.eu/dssp
  - to standardize secondary structure assignment
  - a database of secondary structure assignments for all protein entries in PDB

  Kabsch and Sander (1983)

  Based on the backbone-backbone hydrogen bonds (*e.g.,* α if two consecutive amino acids have i → i+4 hydrogen bonds) and geometrical features extracted from x-ray coordinates



- **STRIDE: Secondary Structural Identification method**
  https://github.com/MDAnalysis/stride Frishman & Argos (1995)

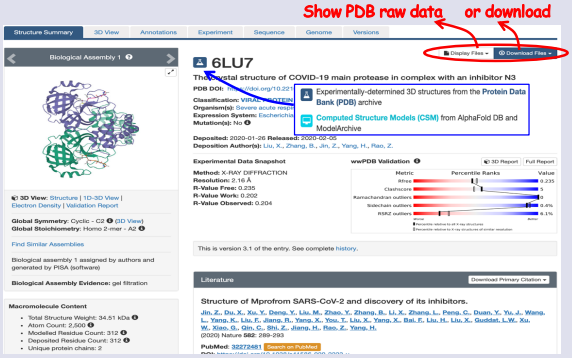
  ➡ uses empirically derived hydrogen bond energy and ϕ - ψ torsion angle criteria

49

PDB: Protein Data Bank

50

PDB: Protein Data Bank

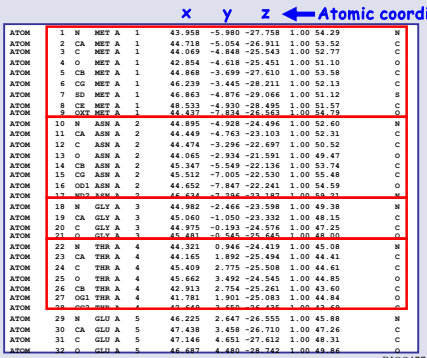


51

PDB format: http://www.wwpdb.org/docs.html

52

PDB format: http://www.wwpdb.org/docs.html



http://www.rcsb.org/pdb/

53