Molecular Evolution and Phylogeny in Microbiology Research

Sishuo WANG

CONTENT

1 Background

2 Preparation

3 Tree building

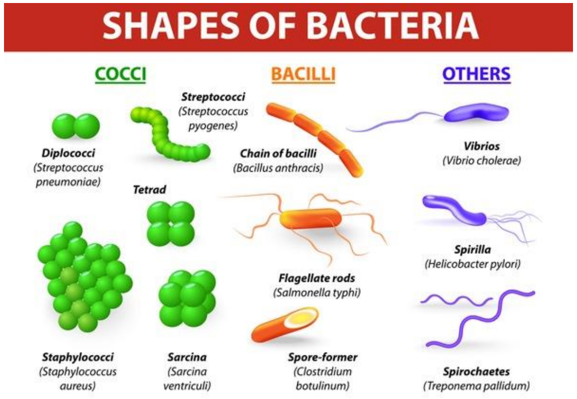
4 Additional topics

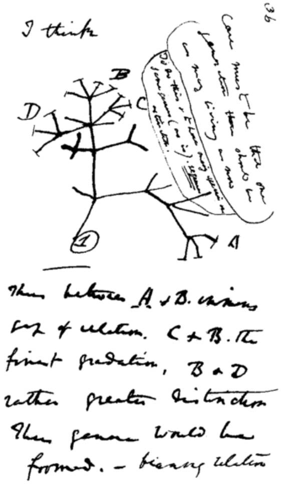
Background: classification vs. phylogenetics

- Classification: the process of organizing and categorizing anything
  - Gram staining: Gram-negative, gram-positive
  - Shape: rod-shaped, spiral, coccoid
  - Pathogenic vs. non-pathogenic

- Phylogeny: purely based on the evolutionary relationships

; (from Greek φῦλον phylon "tribe" + γένεσις genesis "origin")


SHAPES OF BACTERIA
COCCI
BACILLI
OTHERS
Diplococci (Streptococcus pneumoniae)
Streptococci (Streptococcus pyogenes)
Tetrad
Staphylococci (Staphylococcus aureus)
Sarcina (Sarcina ventriculi)
Chain of bacilli (Bacillus anthracis)
Flagellate rods (Salmonella typhi)
Spore-former (Clostridium botulinum)
Vibrios (Vibrio cholerae)
Spirilla (Helicobacter pylori)
Spirochaetes (Treponema pallidum)

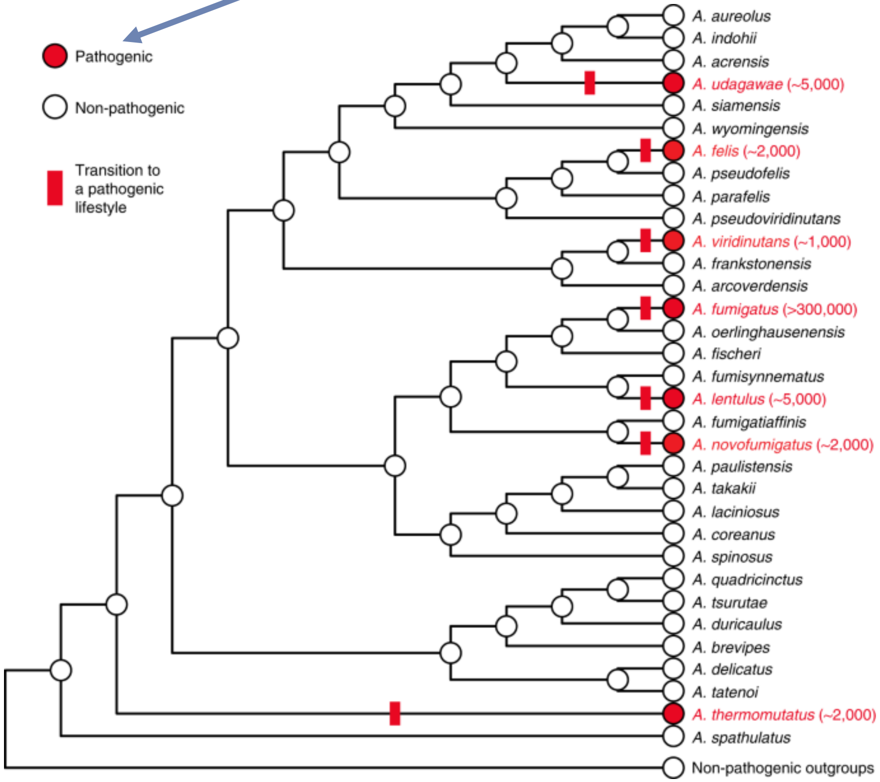

36
I think
1
A
B
C
D
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extinction.
Thus between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction. Thus genera would be formed. — bearing relation

Charles Darwin's 1837 sketch

https://www.shutterstock.com/image-vector/common-bacteria-infecting-human-vector-illustration-160035755

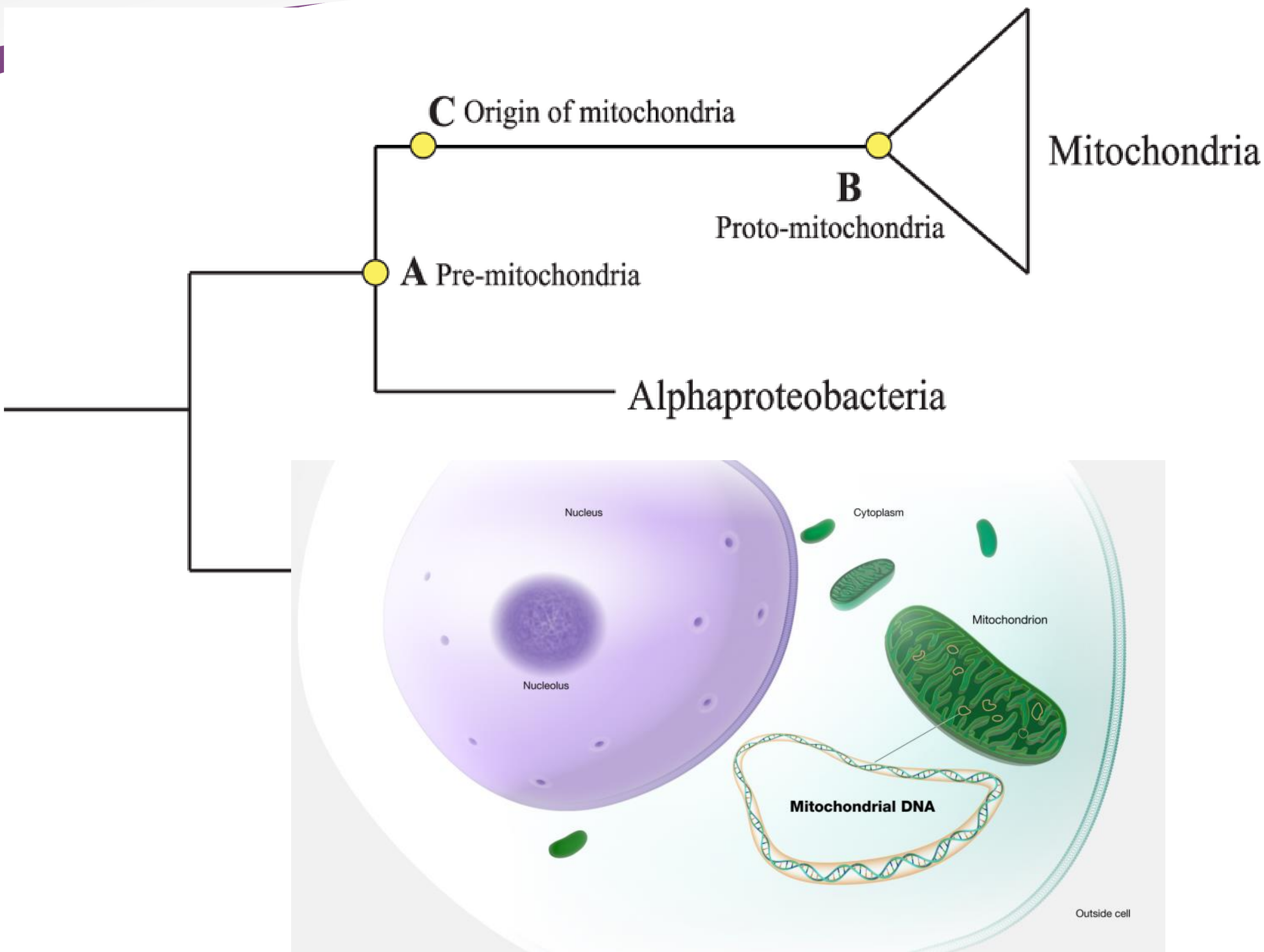
Why study phylogeny:
mapping traits on the phylogeny

Independent transitions to pathogenic in the fungal genus *Aspergillus.*



Rokas, 2022

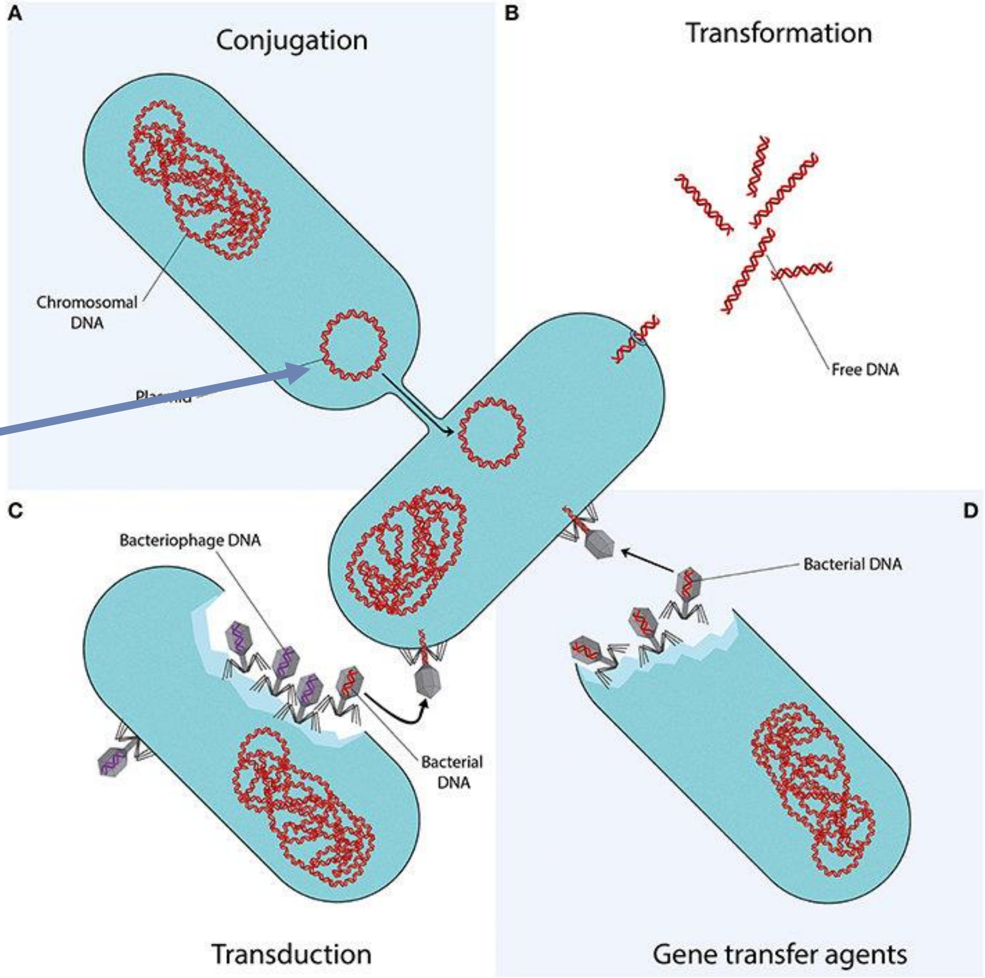
Why study phylogeny: reveal intrinsic connections



https://www.genome.gov/genetics-glossary/Mitochondrial-DNA

What’s horizontal gene transfer (HGT)

Under antibiotics pressure, strains w/ AMR genes on the plasmid may have a higher opportunity to survive!



DOI:10.3389/fmicb.2016.00173

Can you identify “horizontally transferred” words?

English: I eat dim sum.

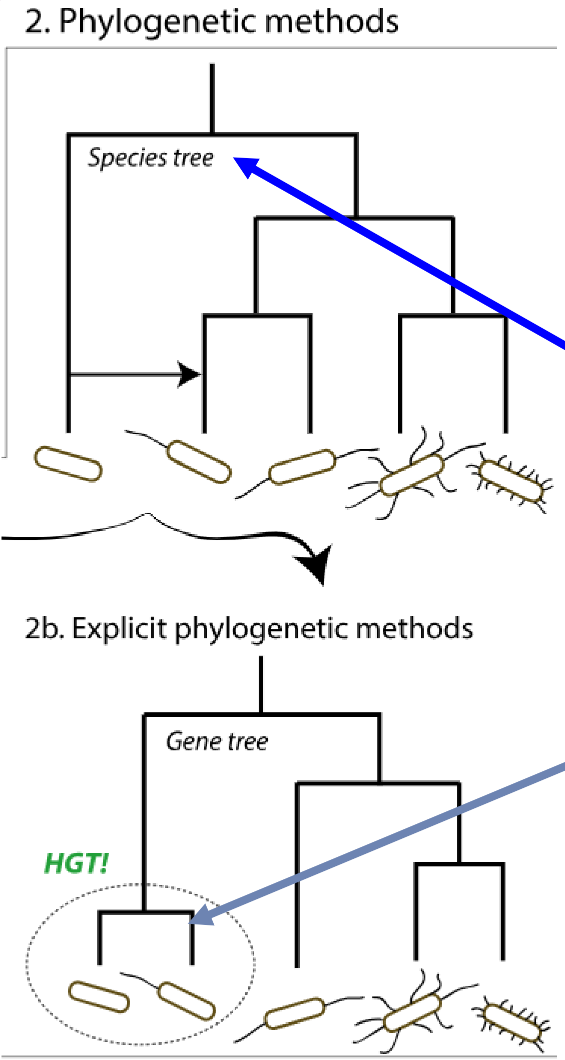
English: He enjoys sushi.

Can you identify "horizontally transferred" words?

- English: I eat dim sum.
- German: Ich esse dim sum.
  - Source: Cantonese

- English: He enjoys sushi.
- German: Er genießt sushi.
  - Source: Japanese

Why studying phylogeny: Horizontal gene transfer (HGT)

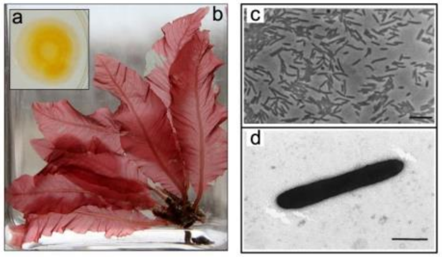


Conflict btwn gene tree and species tree!

Ravenhall et al. 2015

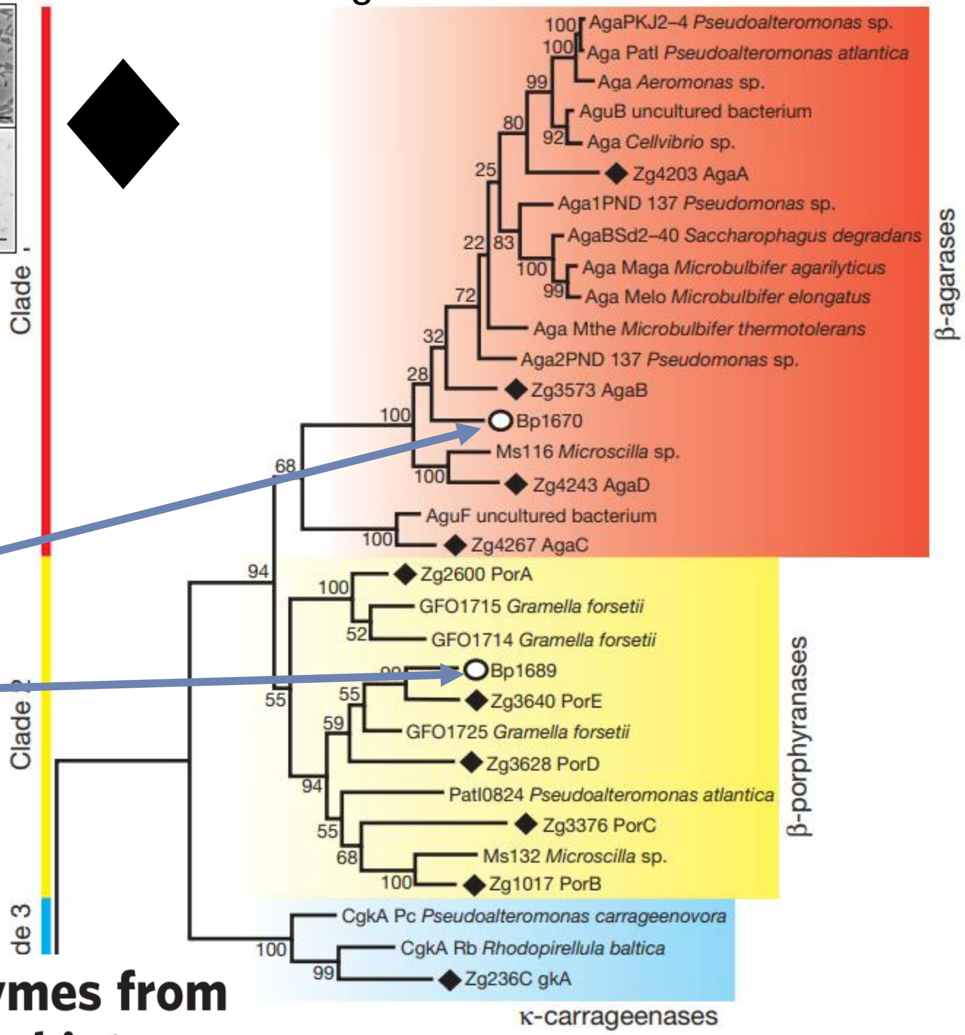
HGT: a real example

Z. galactanivorans



◆

Phylogenetic analysis of GH16 galactanases

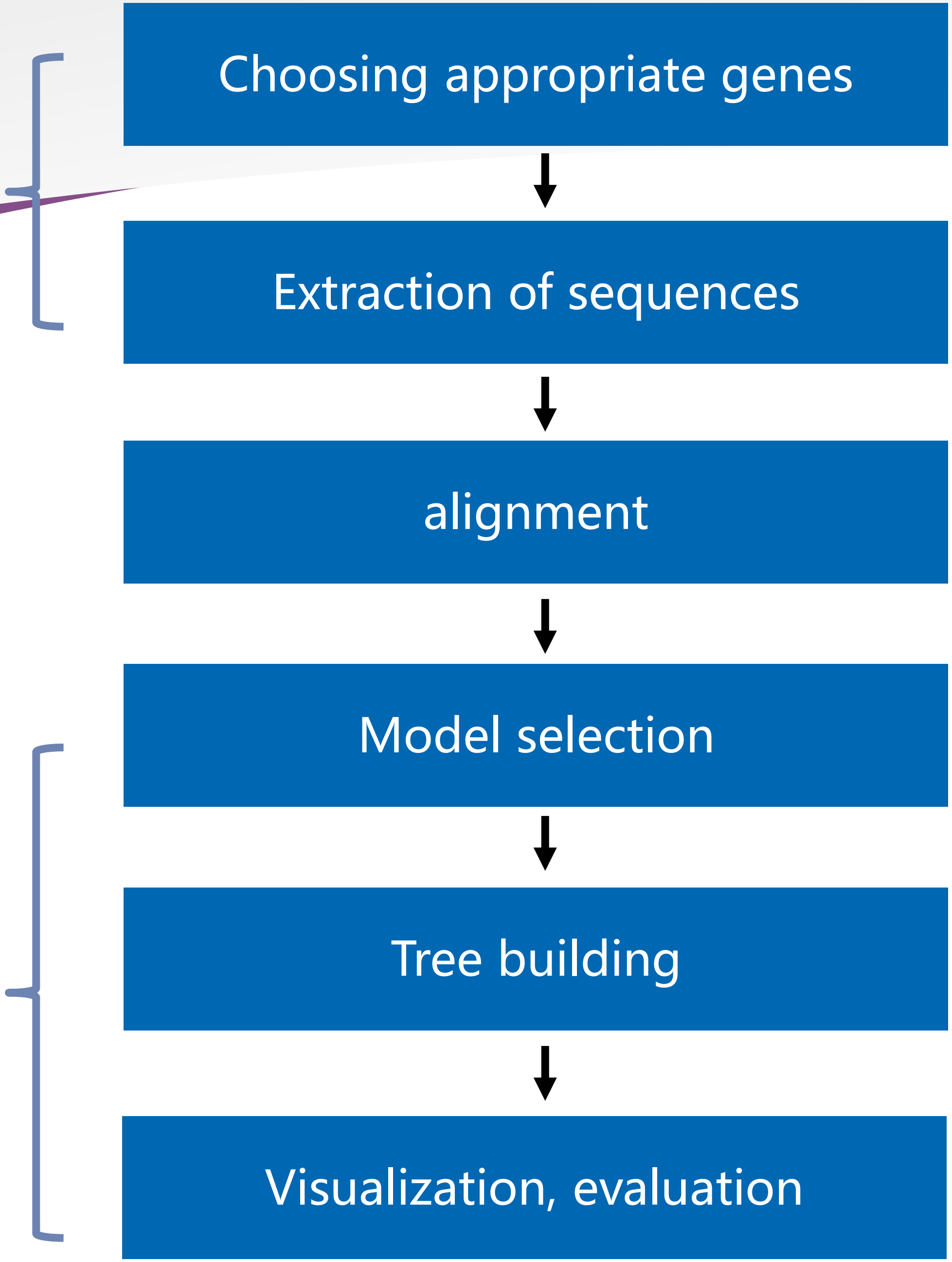

Clade 1
Clade 2
Clade 3
100 AgaPKJ2–4 Pseudoalteromonas sp.
100 Aga PatI Pseudoalteromonas atlantica
99 Aga Aeromonas sp.
80 AguB uncultured bacterium
92 Aga Cellvibrio sp.
25 Zg4203 AgaA
Aga1PND 137 Pseudomonas sp.
22 83 AgaBSd2–40 Saccharophagus degradans
100 Aga Maga Microbulbifer agarilyticus
99 Aga Melo Microbulbifer elongatus
72 Aga Mthe Microbulbifer thermotolerans
32 Aga2PND 137 Pseudomonas sp.
28 Zg3573 AgaB
100 Bp1670
Ms116 Microscilla sp.
68 100 Zg4243 AgaD
AguF uncultured bacterium
100 Zg4267 AgaC
β-agarases
94 Zg2600 PorA
100 GFO1715 Gramella forsetii
52 GFO1714 Gramella forsetii
Bp1689
55 55 Zg3640 PorE
59 GFO1725 Gramella forsetii
Zg3628 PorD
94 PatI0824 Pseudoalteromonas atlantica
55 Zg3376 PorC
68 Ms132 Microscilla sp.
100 Zg1017 PorB
β-porphyranases
CgkA Pc Pseudoalteromonas carrageenovora
100 CgkA Rb Rhodopirellula baltica
99 Zg236C gkA
κ-carrageenases

○

gut Bacteroidetes

Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota

Hehemann, et al., 2010, Nature



Choosing appropriate genes

Extraction of sequences

alignment

Model selection

Tree building

Visualization, evaluation

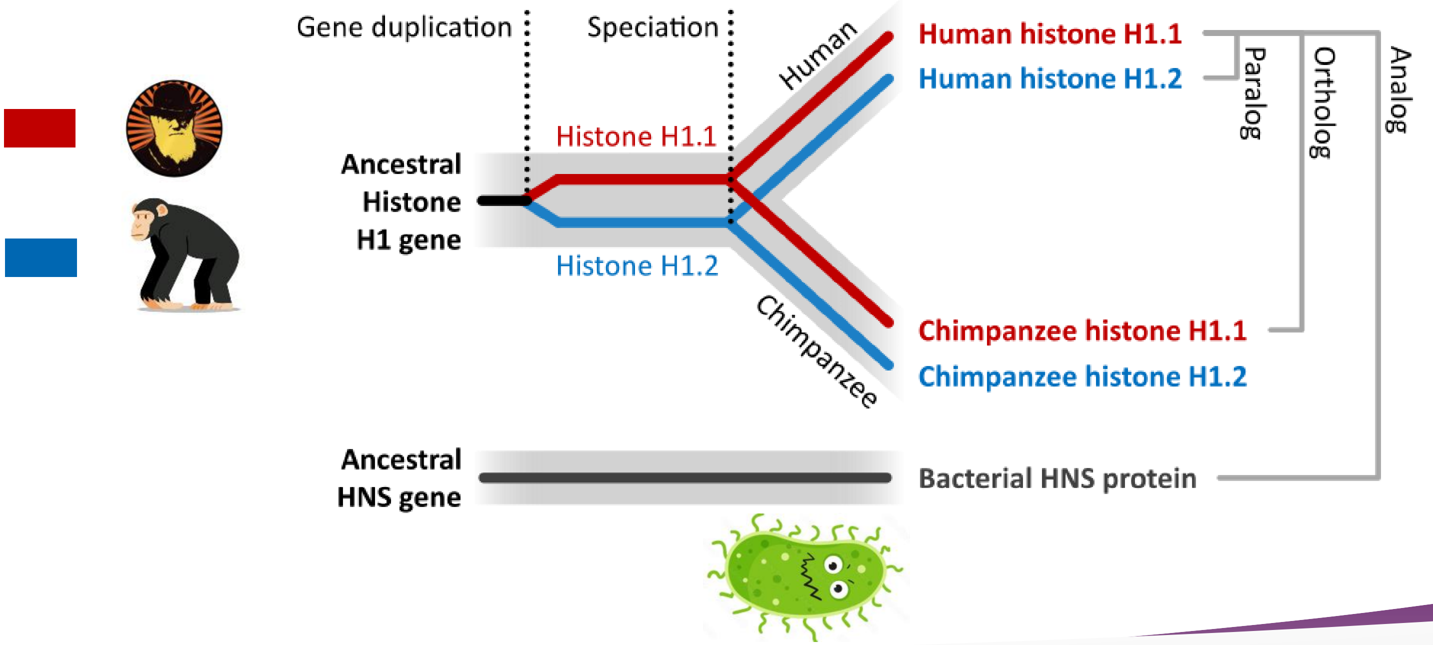
Preparation: what data to use to build a tree?

- For inferring species relationship: orthologs (resistant to HGT)
  - 16S
  - Other marker genes (house-keeping)
  - Single-copy genes

- For inferring gene trees: all genes within the same family
  - ❑ Gene family: homologs, or genes that are evolutionarily related (share a common origin)

Orthologs vs. Paralogs

- Ortholog: genes separated by **speciation**
- Paralog: genes separated by gene **duplication**.
- Analog: genes not evolutionarily related but with similar functions



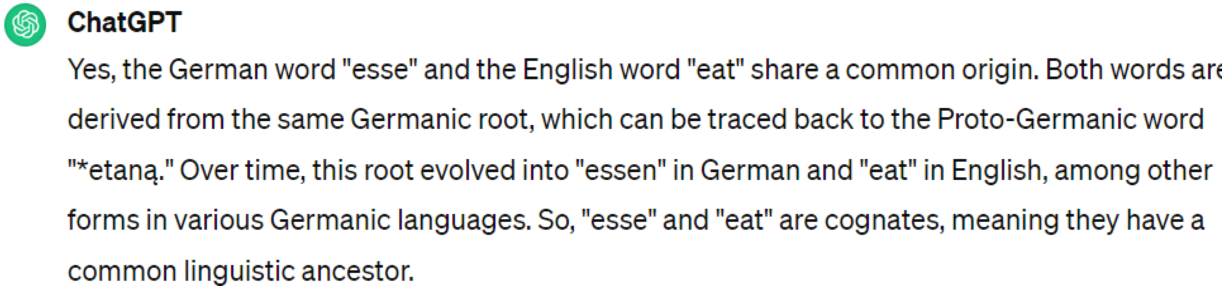
https://en.wikipedia.org/wiki/Sequence_homology

Why using Orthologs for species tree

- English: I eat dim sum.
- German: Ich esse dim sum.

- Ortholog:
  - I – Ich
  - eat – esse

- HGT:
  - dim sum

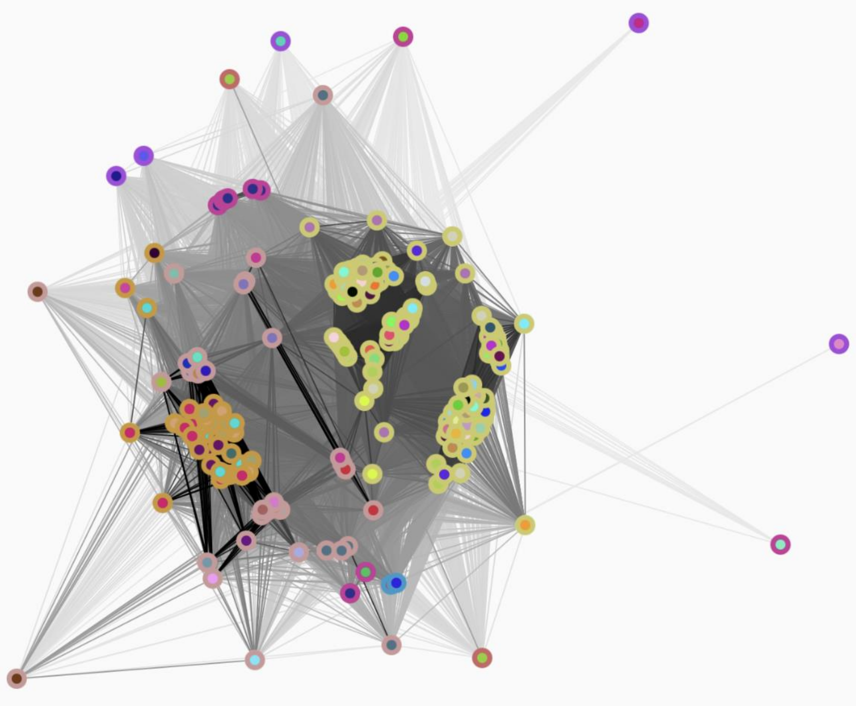


ChatGPT

Yes, the German word "esse" and the English word "eat" share a common origin. Both words are derived from the same Germanic root, which can be traced back to the Proto-Germanic word "*etaną." Over time, this root evolved into "essen" in German and "eat" in English, among other forms in various Germanic languages. So, "esse" and "eat" are cognates, meaning they have a common linguistic ancestor.

How to identify orthologs (and paralogs)

- Clustering based on sequence similarity (BLAST) into families
  - Inparanoid
  - OrthoMCL
  - OrthoFinder
  - MMseqs2



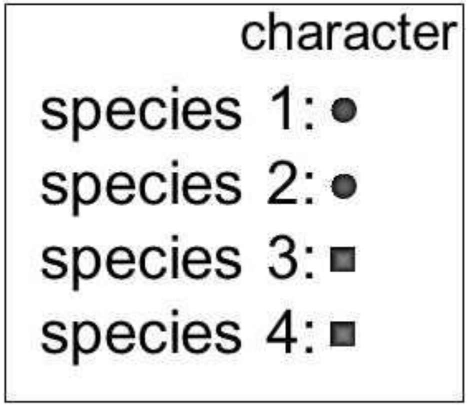
- Typically, you can take **single-copy** genes as orthologs.
- If higher resolution needed, SNP or genome-wide nucleotide tree is needed.

https://orthomcl.org/orthomcl/app

Building a tree

- Most parsimony (MP)
- Distance-based methods (not introduced)
- Maximum likelihood (ML)
- Bayesian approaches

Tree building: maximum parsimony (MP)



character

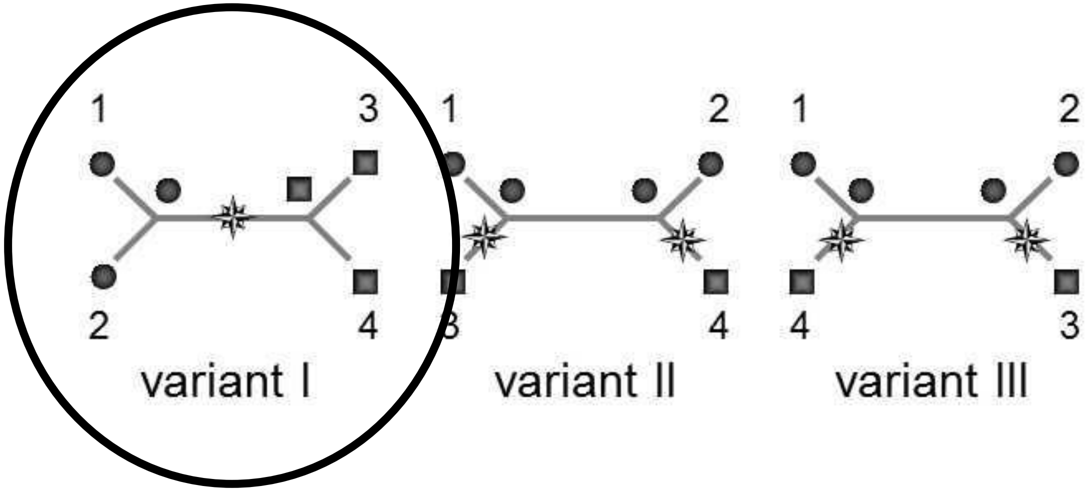
species 1: ●

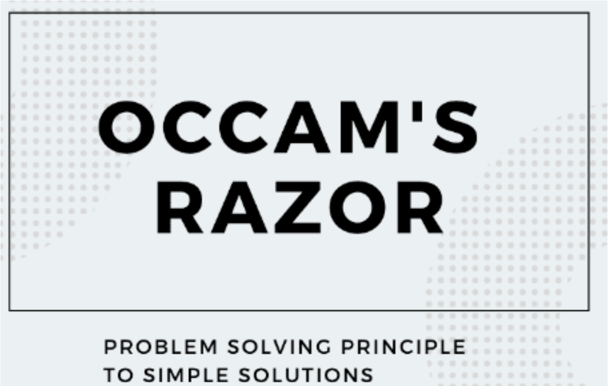
species 2: ●

species 3: ■

species 4: ■

Occam's razor: the tree that involves the least changes in the character (nucleotides, amino acids, etc.), is the best.

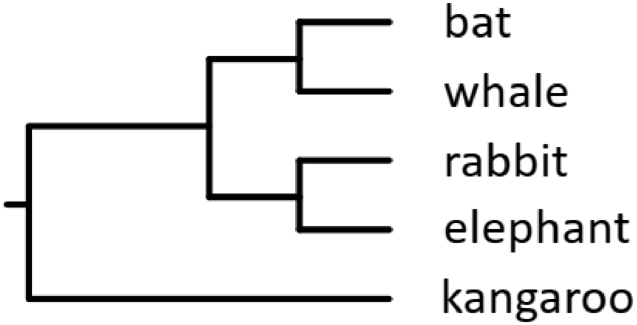




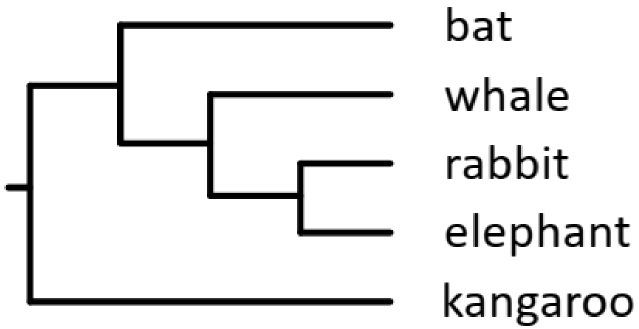
https://www.frozenevolution.com/xxiii52-homologies-and-homoplasies-can-also-be-distinguished-using-maximum-parsimony-principle

Maximum parsimony

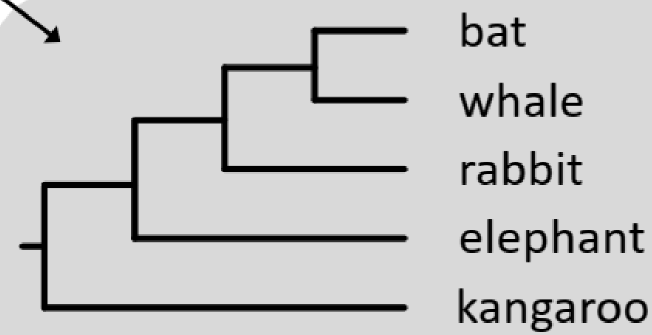
E. coli	CCGTTAGTAACT
Rhizobium	CCGTTAGTAACT
Rhizobium	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



7 steps



8 steps



6 steps

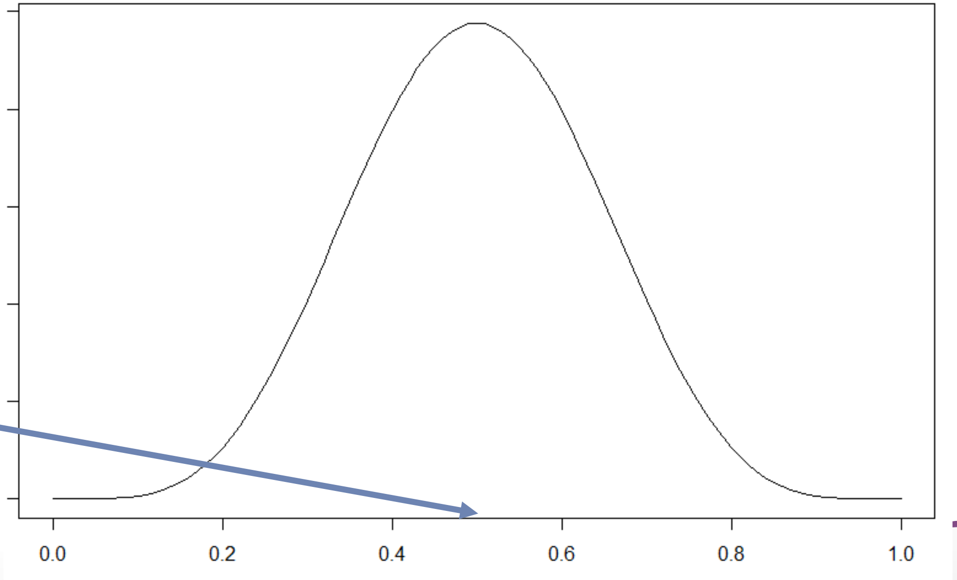
Taken from Simon Ho’s slides

Maximum likelihood estimation

- Maximum likelihood estimate (MLE) maximizes the likelihood.
- Consider you flip a coin for 10 times and 5 heads and 5 tails. What's the MLE of the probability the coin landing heads?

$$f(\theta) = \theta^5 (1-\theta)^5$$

$$\hat{\theta} = 0.5$$

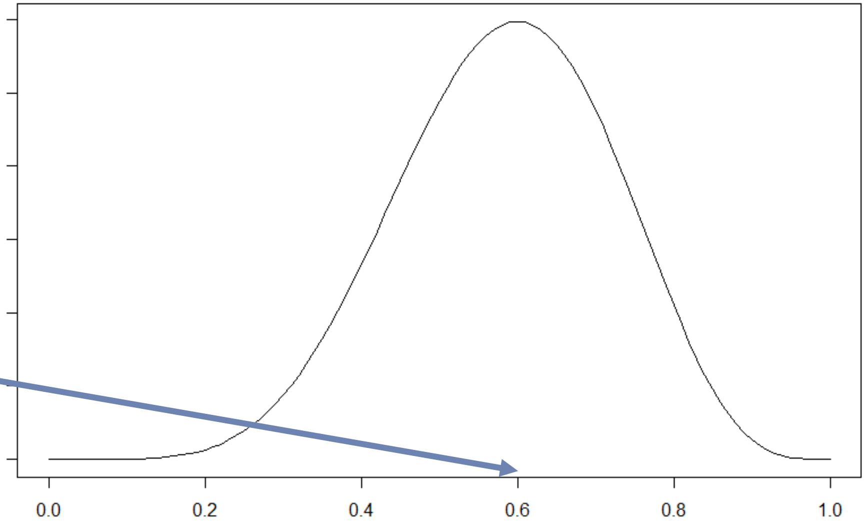


Maximum likelihood estimation

- Maximum likelihood estimate (MLE) maximizes the likelihood.
- Consider you flip a coin for 10 times and 5 heads and 5 tails. What's the MLE of the probability the coin landing heads?
- What's if you 6 heads and 4 tails?

$$f(\theta) = \theta^6(1-\theta)^4$$

$$\hat{\theta} = 0.6$$

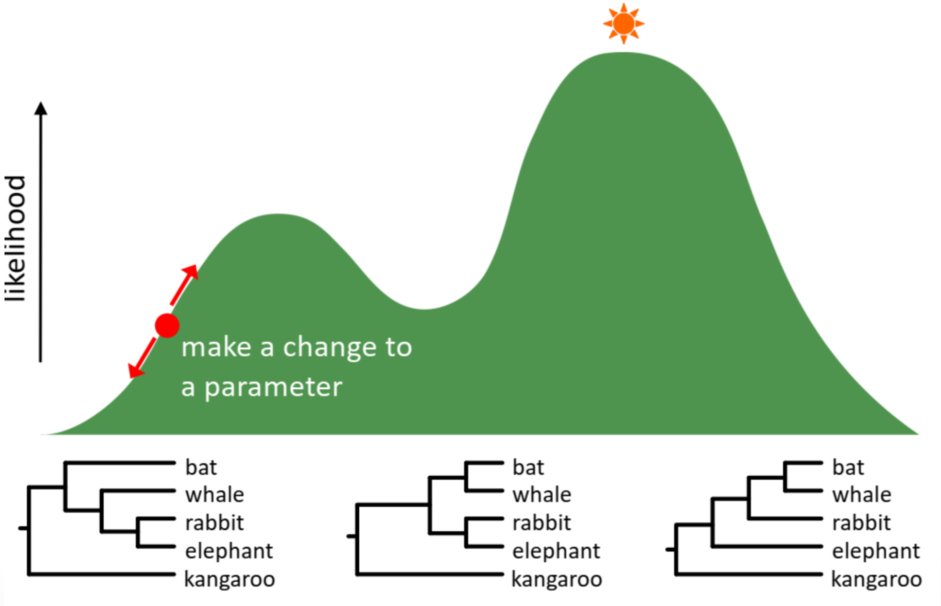


Maximum likelihood tree

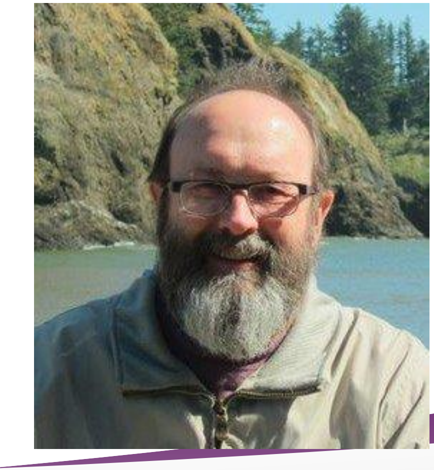
Two key points for a tree:

- Branch length
- Topology

So the ML tree consists of **topology + branch** lengths that maximize the likelihood given the observed DNA/protein sequences.

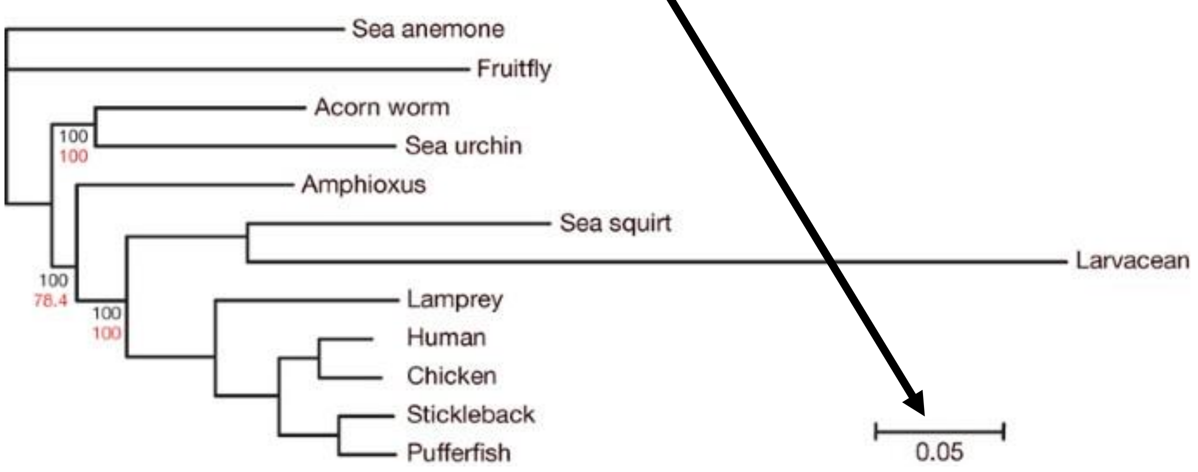


Simon Ho's slides



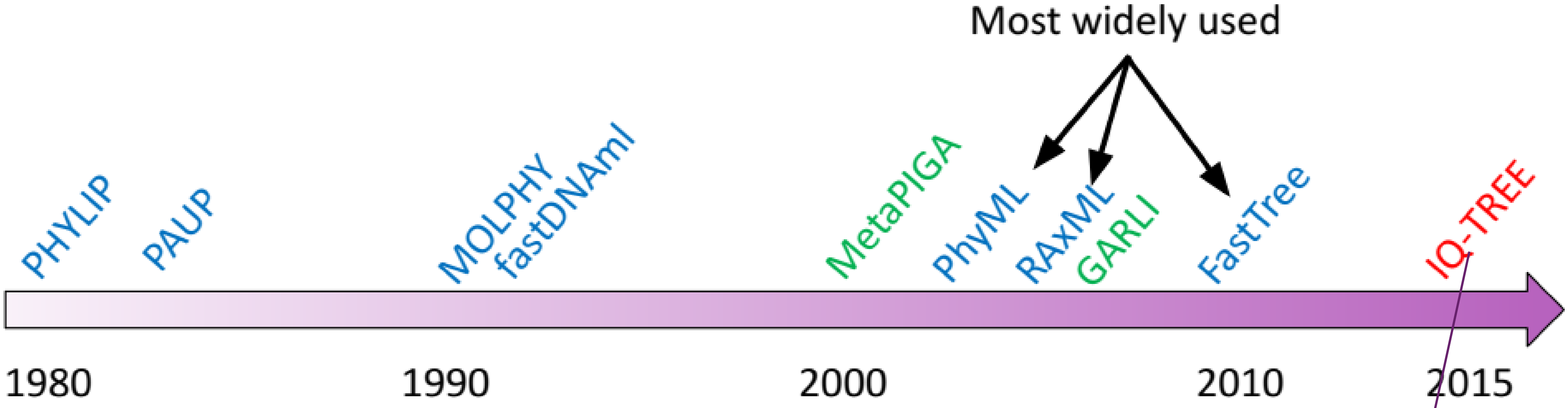
Felsenstein, J. 1981

In ML and Bayesian phylogenies, **branch length** represent substitutions/site (has nothing to do with time). So a longer branch means more changes compared to the ancestor.



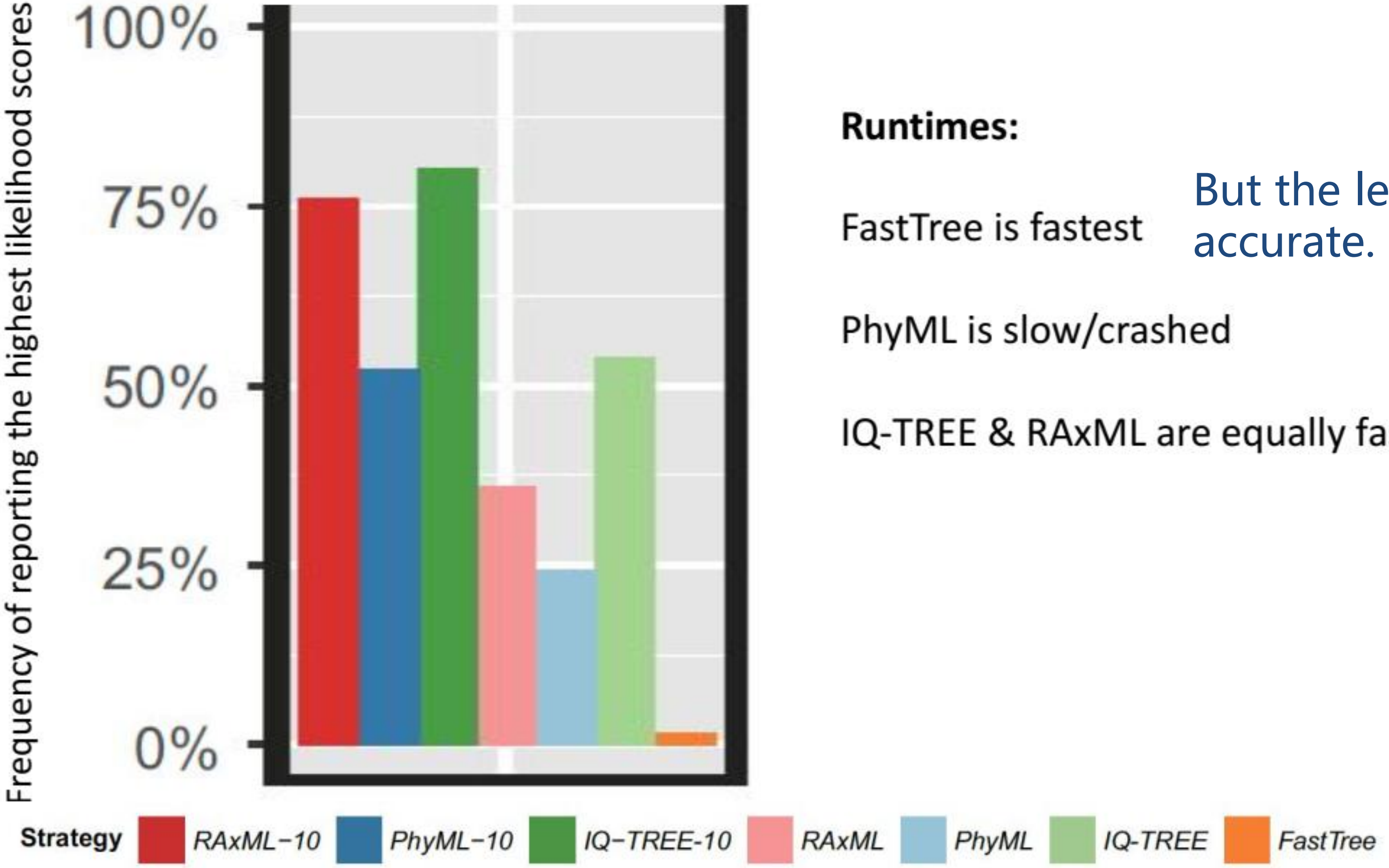
https://www.nature.com/articles/nature06967

maximum likelihood phylogenetics software



Taken from Minh’s slides

An independent benchmark by Zhou et al. (2018)



Runtimes:

FastTree is fastest

But the least accurate.

PhyML is slow/crashed

IQ-TREE & RAxML are equally fast

Modified on Minh’s slides

Bayesian inference

- Set priors for each parameter: branch length, topology, and others.
- The result will be an update of your priors based on the data.

Yang, ZH & Rannala, B., 1997

A toy example

We can publish the work in Nature!!!

If a professor says confidently he can publish your study in Nature, how likely will the study be published in Nature?

The probability of your study being published in Nature, conditioned on your boss saying so.

Pr(your study being published in Nature | your boss saying so)

Let’s write it in this way:
A: really publish in Nature
B: prof says “we’ll publish the study in Nature”
P(A|B)

A toy example

$$\underbrace{\boxed{P(A|B)}}_{\text{posterior}} = \underbrace{\boxed{P(A)}}_{\text{prior}} \times \frac{\boxed{P(B|A)}\ \text{likelihood}}{\boxed{P(B)}\ \text{marginal}}$$

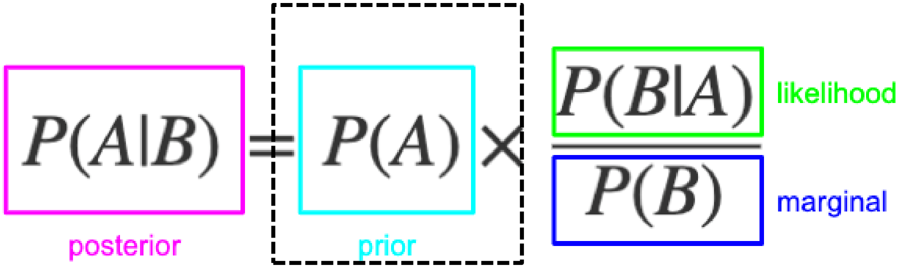
A: really publish in Nature

B: **prof says** “we’ll publish the study in Nature”

- Prior: $P(A) = 0.1$ → $P(\bar{A}) = 1 - 0.1 = 0.9$
- Cond prob:
  - $P(B|A) = 1.0$
  - $P(B|\bar{A}) = 0.5$
- $P(B) = P(A)P(B|A) + P(B|\bar{A})P(\bar{A}) = 0.1 * 1 + 0.5 * 0.9 = 0.55$

$P(A|B) = 0.1 * 1.0 / 0.55 = 0.18$

What if the journal has a higher acceptance rate?



$$\boxed{P(A|B)} = \boxed{P(A)} \times \frac{\boxed{P(B|A)}}{\boxed{P(B)}}$$

posterior

prior

likelihood

marginal

A: really publish in Nature

nature
At last – a computer program that can beat a champion Go player
ALL SYSTEMS GO

B: **prof says** “we’ll publish the study in Nature”

- Prior: P(A) = 0.2 → $P(\bar{A}) = 1 - 0.2 = 0.8$
- Cond prob:
  - $P(B|A) = 1.0$
  - $P(B|\bar{A}) = 0.5$
- $P(B) = P(A)P(B|A) + P(B|\bar{A})P(\bar{A}) = 0.2 * 1 + 0.8 * 0.5 = 0.6$

$$P(A|B) = 1.0 * 0.2 / 0.6 = 1/3$$

What if your boss is more conservative?

$$\underbrace{P(A|B)}_{\text{posterior}} = \underbrace{P(A)}_{\text{prior}} \times \frac{\overset{\text{likelihood}}{P(B|A)}}{\underset{\text{marginal}}{P(B)}}$$

A: really publish in Nature

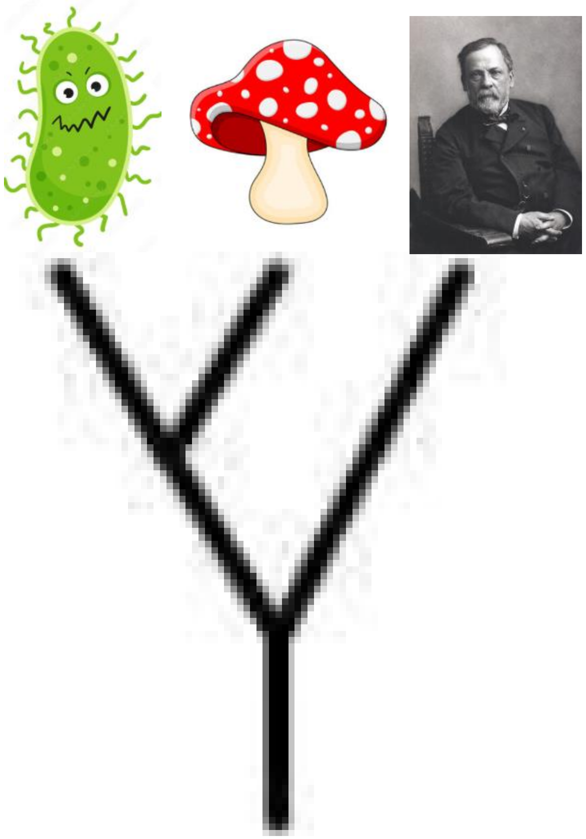
nature
At last – a computer program that can beat a champion Go player
ALL SYSTEMS GO

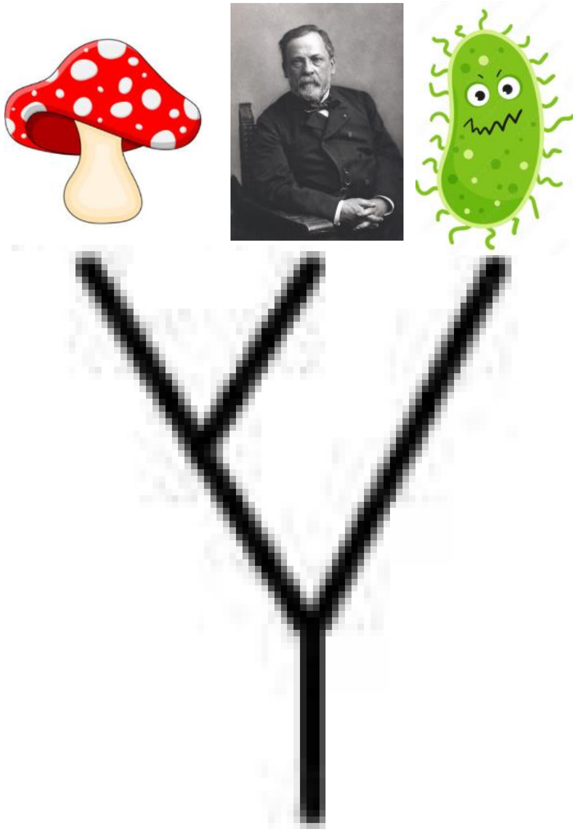
B: **prof says** “we’ll publish the study in Nature”

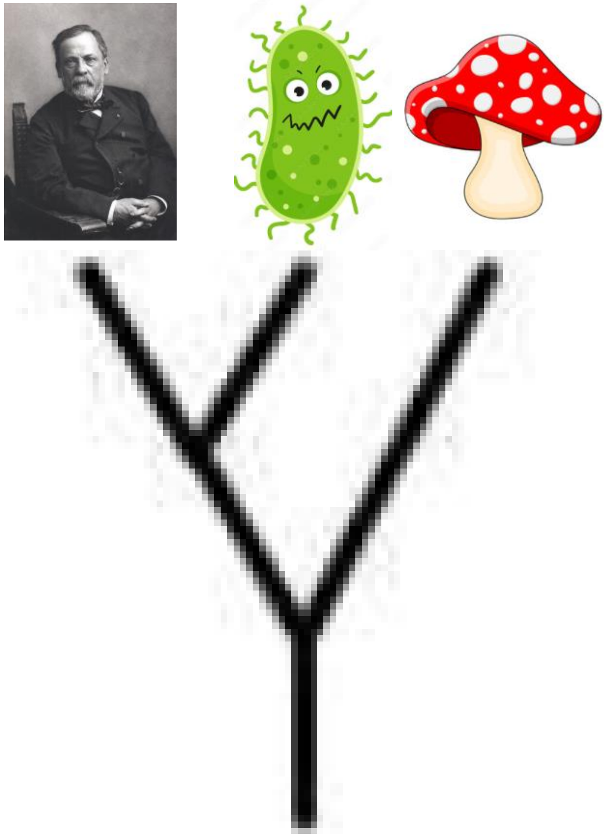
- Prior: $P(A) = 0.1$ → $P(\bar{A}) = 1 - 0.1 = 0.9$
- Cond prob:
  - $P(B|A) = 1.0$
  - $P(B|\bar{A}) = 0.2$
- $P(B) = P(A)P(B|A) + P(B|\bar{A})P(\bar{A}) = 0.1 * 1 + 0.9 * 0.2 = 0.28$

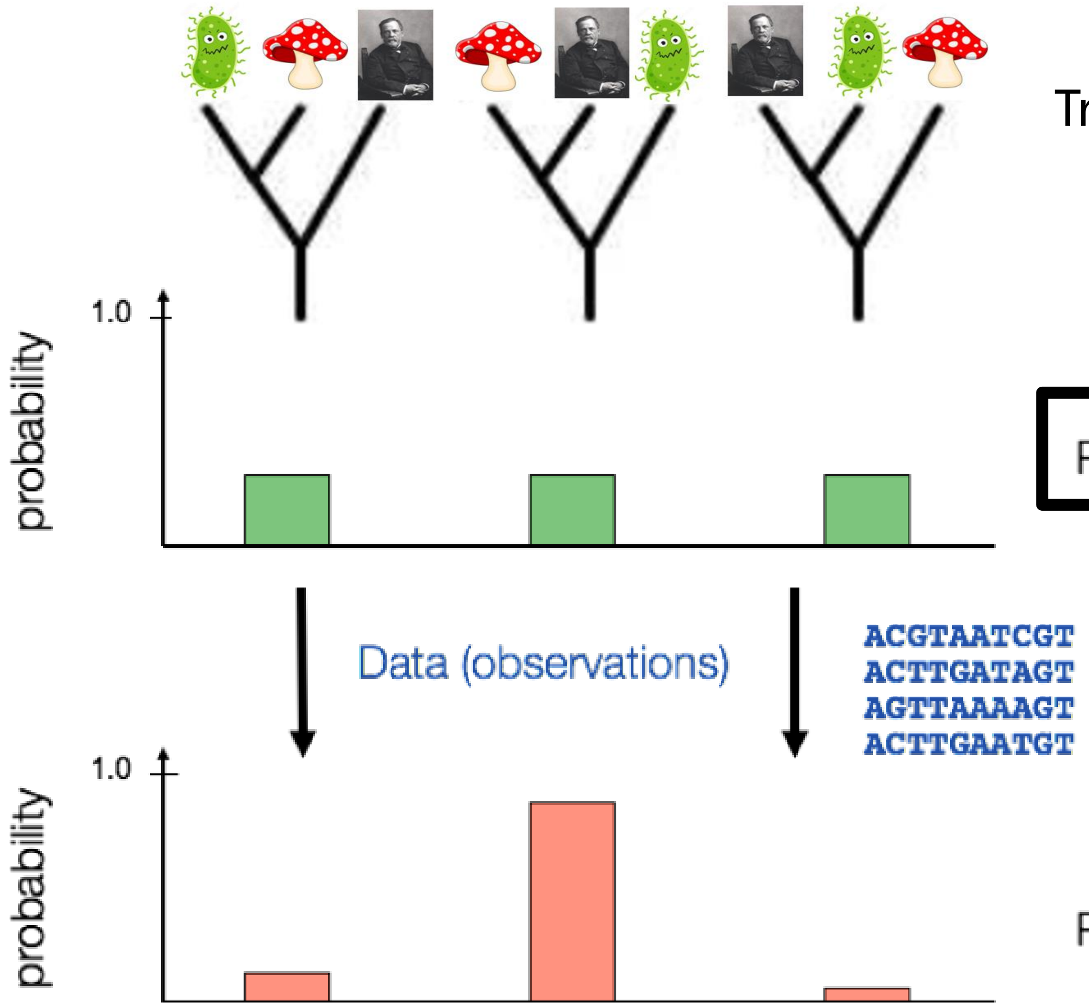
$P(A|B) = 1.0*0.1/0.28 = 0.36$

Which of the following tree topologies is the most likely?









Tree topology

Prior distribution

Data (observations)

ACGTAATCGT
ACTTGATAGT
AGTTAAAAGT
ACTTGAATGT

Posterior distribution

Will give you the posterior prob of each tree instead of showing a single “best” tree.

modified on Suchard’s slides

Bayesian methods

- PhyloBayes
  - complex models

- BEAST

BEAST
Bayesian Evolutionary Analysis Sampling Trees

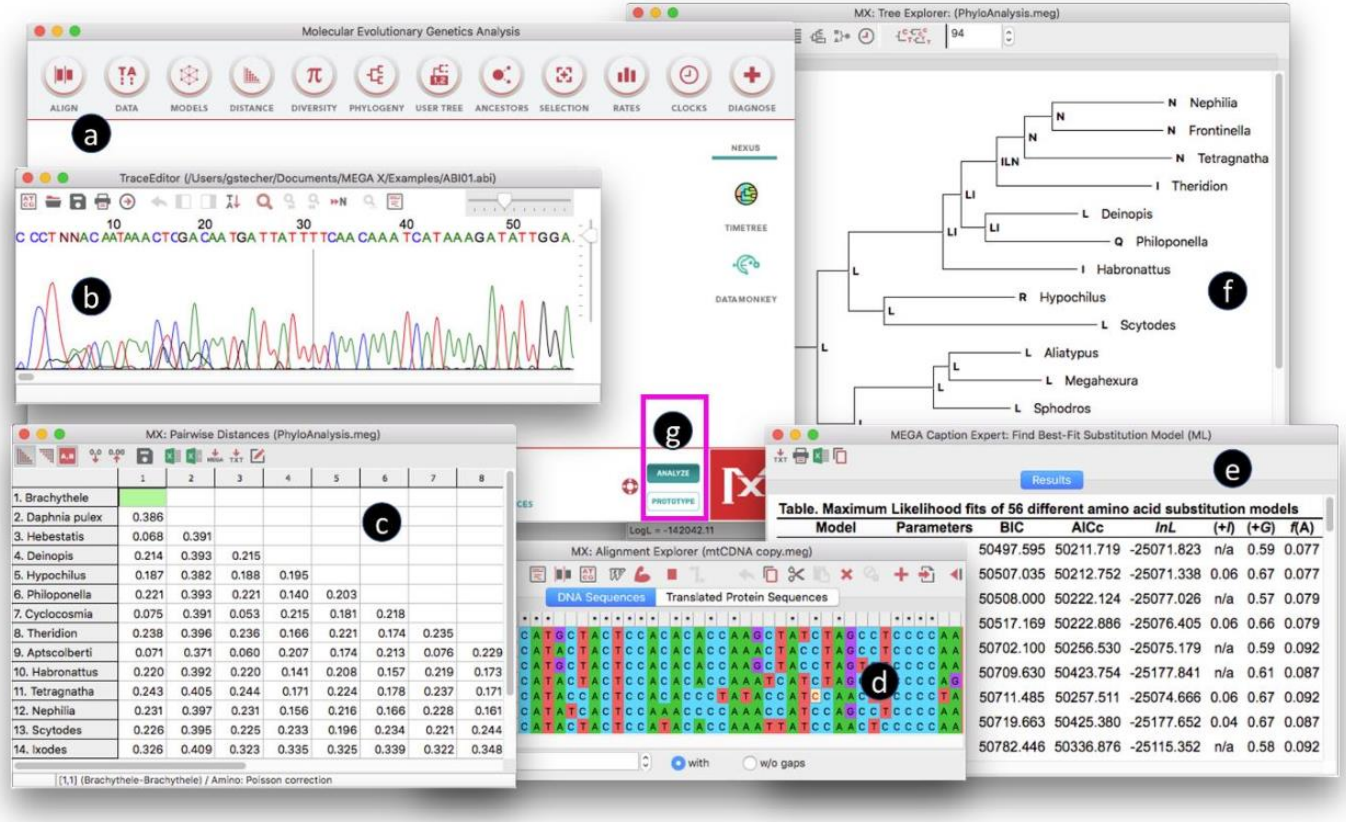
  - often combined with molecular clock
  - GUI but still not easy for beginners
  - GPU accelerated

- MrBayes
  - Not updated any more (but still useful)

- RevBayes (inheritance of MrBayes)
  - community efforts
  - so many functions
  - difficult to handle for beginners

MEGA Molecular Evolutionary Genetics Analysis

Very user-friendly phylogenetics software with GUI, involving MP, distance-based, and ML phylogeny building, as well as many other functions.



MX: Pairwise Distances (PhyloAnalysis.meg)

	1	2	3	4	5	6	7	8
1. Brachythele								
2. Daphnia pulex	0.386							
3. Hebestatis	0.068	0.391						
4. Deinopis	0.214	0.393	0.215					
5. Hypochilus	0.187	0.382	0.188	0.195				
6. Philoponella	0.221	0.393	0.221	0.140	0.203			
7. Cyclocosmia	0.075	0.391	0.053	0.215	0.181	0.218		
8. Theridion	0.238	0.396	0.236	0.166	0.221	0.174	0.235	
9. Aptscolberti	0.071	0.371	0.060	0.207	0.174	0.213	0.076	0.229
10. Habronattus	0.220	0.392	0.220	0.141	0.208	0.157	0.219	0.173
11. Tetragnatha	0.243	0.405	0.244	0.171	0.224	0.178	0.237	0.171
12. Nephilia	0.231	0.397	0.231	0.156	0.216	0.166	0.228	0.161
13. Scytodes	0.226	0.395	0.225	0.233	0.196	0.234	0.221	0.244
14. Ixodes	0.326	0.409	0.323	0.335	0.325	0.339	0.322	0.348

[1,1] (Brachythele-Brachythele) / Amino: Poisson correction

MEGA Caption Expert: Find Best-Fit Substitution Model (ML)

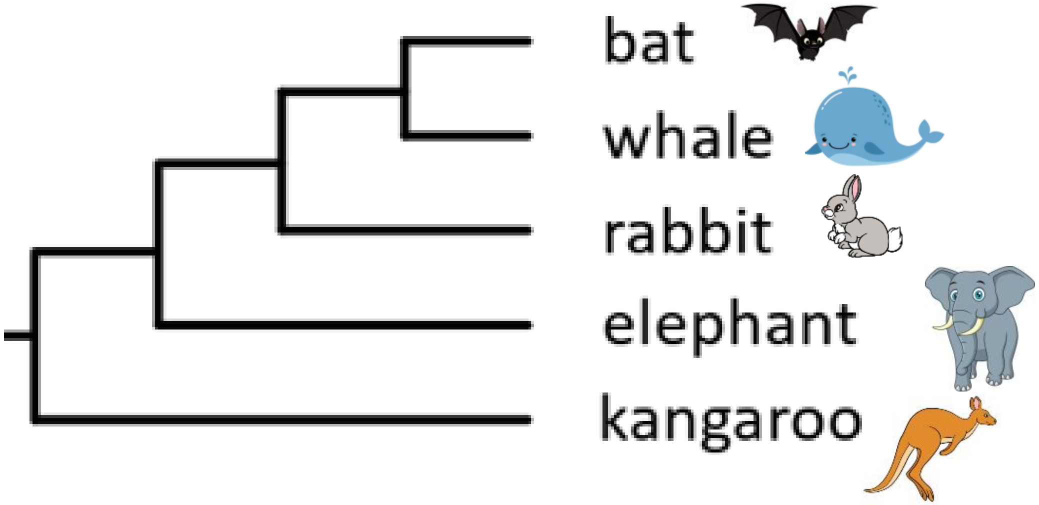
Table. Maximum Likelihood fits of 56 different amino acid substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*f*(A)
		50497.595	50211.719	-25071.823	n/a	0.59	0.077
		50507.035	50212.752	-25071.338	0.06	0.67	0.077
		50508.000	50222.124	-25077.026	n/a	0.57	0.079
		50517.169	50222.886	-25076.405	0.06	0.66	0.079
		50702.100	50256.530	-25075.179	n/a	0.59	0.092
		50709.630	50423.754	-25177.841	n/a	0.61	0.087
		50711.485	50257.511	-25074.666	0.06	0.67	0.092
		50719.663	50425.380	-25177.652	0.04	0.67	0.087
		50782.446	50336.876	-25115.352	n/a	0.58	0.092

Stecher et al., 2020

Additional topics: evaluation

- You got a ML tree, but how confident are you in it?



Evaluation: bootstrap

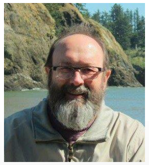
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

Randomly sample sites (with replacement)

bat	T
whale	T
rabbit	A
elephant	T
kangaroo	T

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

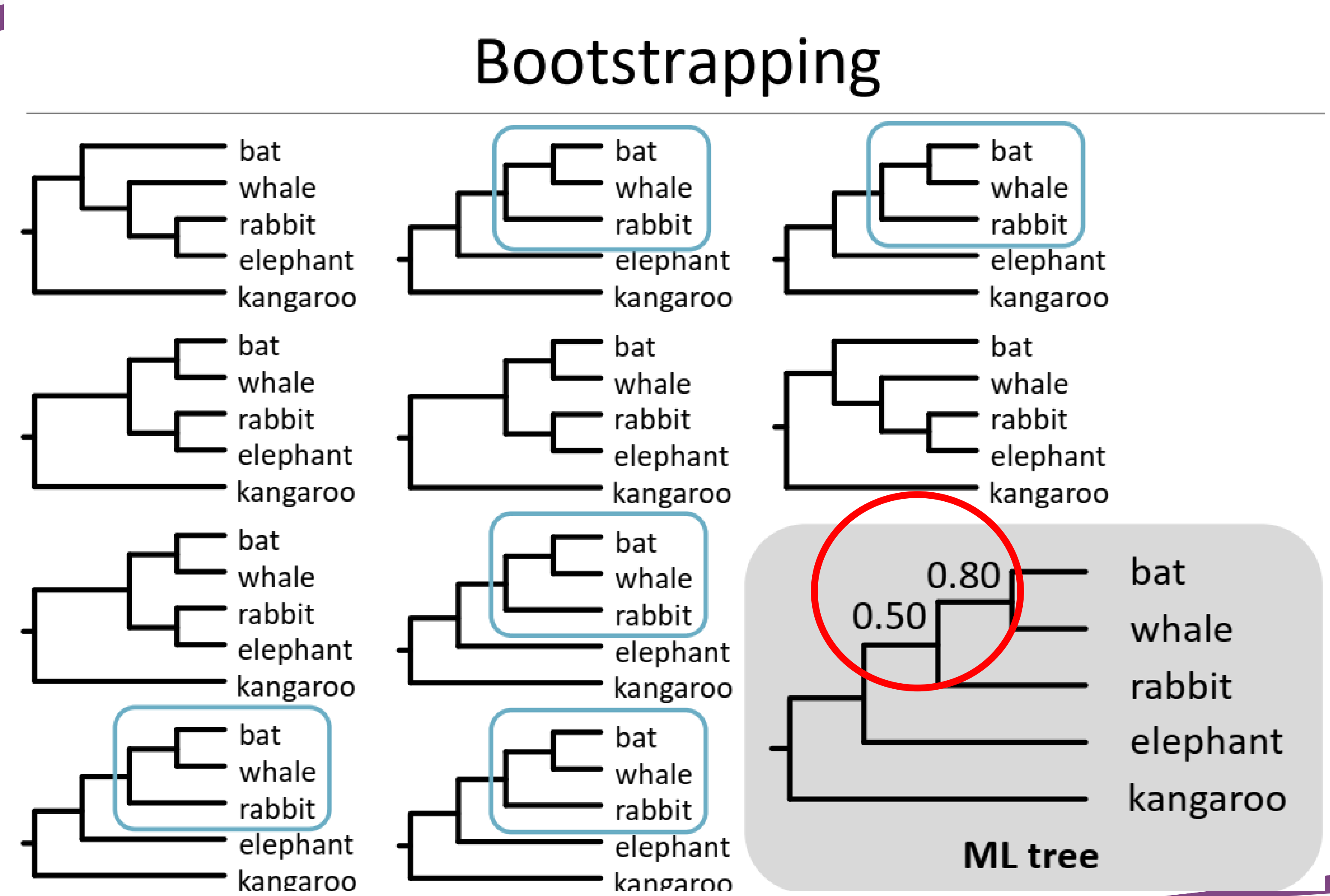
bat	TG
whale	TG
rabbit	AG
elephant	TG
kangaroo	TG



Felsenstein, 1985

Taken from Ho's slides

Evaluation: bootstrap

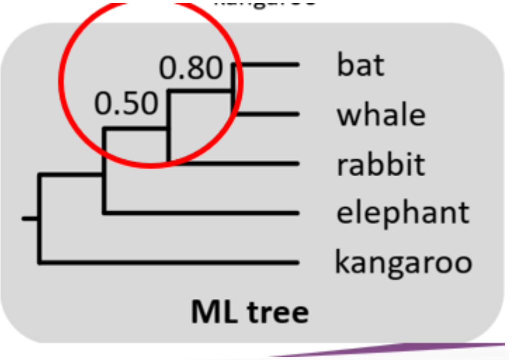


Taken from Ho's slides

Evaluation: thresholds

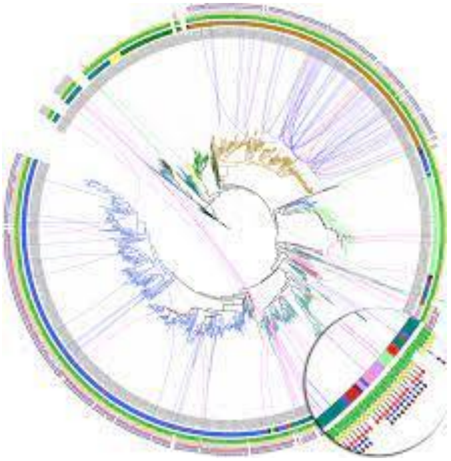
- Maximum likelihood trees, the values on the nodes usually represent bootstrap so >= 80 are considered “reliable”
  - But IQ-Tree may use ultrafast bootstrap with 95 considered as “reliable”

- Bayesian trees, those values represent posterior probabilities, >= 95% considered as reliable.



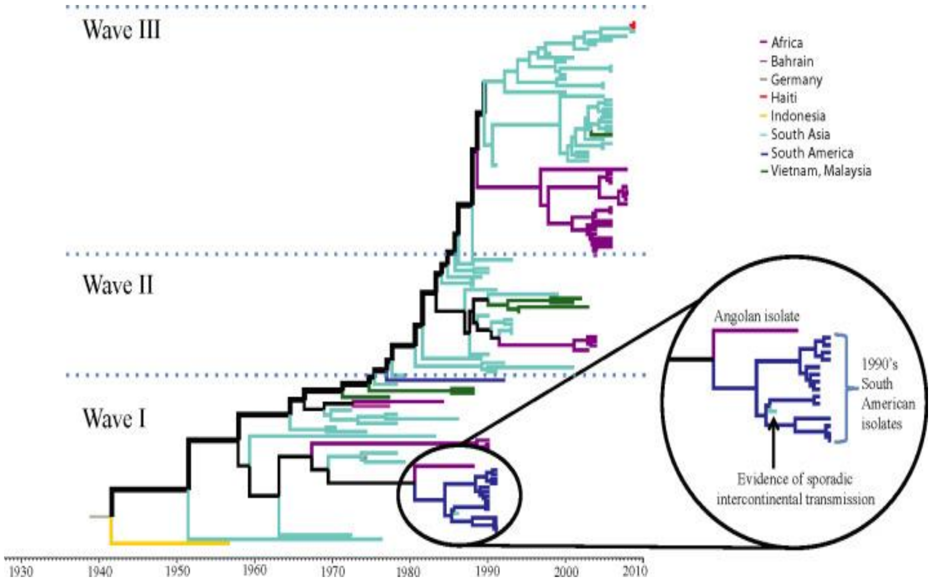
Visualization

- ITOL: https://itol.embl.de/
- Figtree: https://github.com/rambaut/figtree
- R package ggtree: https://github.com/YuLab-SMU/ggtree



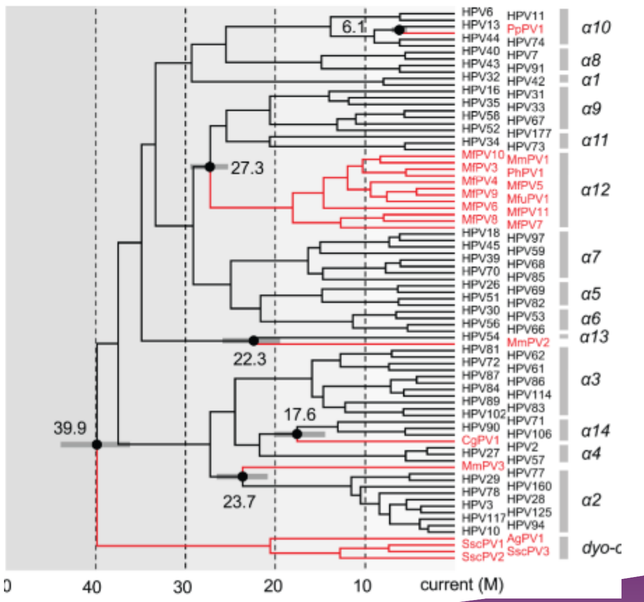
Timetree

- If you have some info about time or rates, you can convert your phylogeny into a timetree.
- Note here the branch length means substitutions/site/unit time.



Vibrio cholerae 7th pandemic lineage

HPV



Chen Zigui, et al., 2018, plos patho

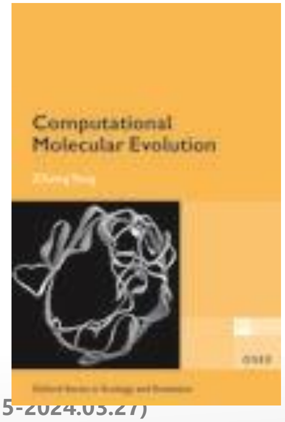
References

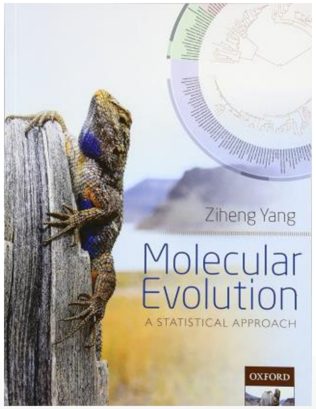
- Rokas, A. Evolution of the human pathogenic lifestyle in fungi. *Nat Microbiol* **7**, 607–619 (2022).
- Wang, Zhang, and Martin Wu. "Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite." PLoS One 9.10 (2014): e110685.
- Ravenhall M, Škunca N, Lassalle F, Dessimoz C (2015) Inferring Horizontal Gene Transfer. *PLoS Comput Biol 11(5): e1004095.*
- Felsenstein, Joseph. "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of molecular evolution 17 (1981): 368-376.
- Yang Z, Rannala B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution, 14(7):717-724.
- Felsenstein, Joseph. "Confidence limits on phylogenies: an approach using the bootstrap." evolution 39.4 (1985): 783-791.
- Zhou, X, et al. "Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets." *Molecular biology and evolution* 35.2 (2018): 486-503.
- Moore, S., et al. "Widespread epidemic cholera caused by a restricted subset of Vibrio cholerae clones." Clinical Microbiology and Infection 20.5 (2014): 373-379.
- Chen Z, DeSalle R, Schiffman M, Herrero R, Wood CE, et al. (2018) Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. PLOS Pathogens 14(11): e1007352.
- Glen Stecher, Koichiro Tamura, Sudhir Kumar, Molecular Evolutionary Genetics Analysis (MEGA) for macOS, Molecular Biology and Evolution, Volume 37, Issue 4, April 2020, Pages 1237–1239

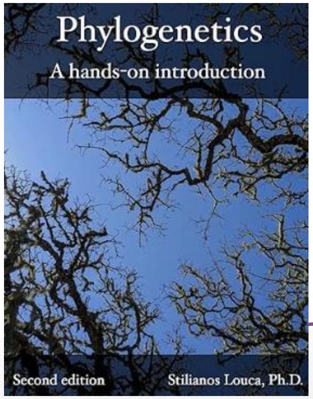
Resources

- Yang, Ziheng. Computational molecular evolution. OUP Oxford, 2006.
- Yang, Ziheng. Molecular evolution: a statistical approach. Oxford University Press, 2014.
  - solutions: https://github.com/sishuowang/
- Louca, Stilianos. Phylogenetics, a **hands-on** introduction. 2023.

- Sydney phylogenetics workshop. https://github.com/simon-ho/SydneyPhyloWorkshop/ (organized by **Simon Ho**)







lottery

https://numbergenerator.org/randomnumbergenerator/1-40

Sishuo WANG

Thanks

For Your Listening