3000788 Intro to Comp Molec Biol

Lecture 7: Phylogenetics and evolutionary models

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

- Evolutionary perspective
- Phylogenetics

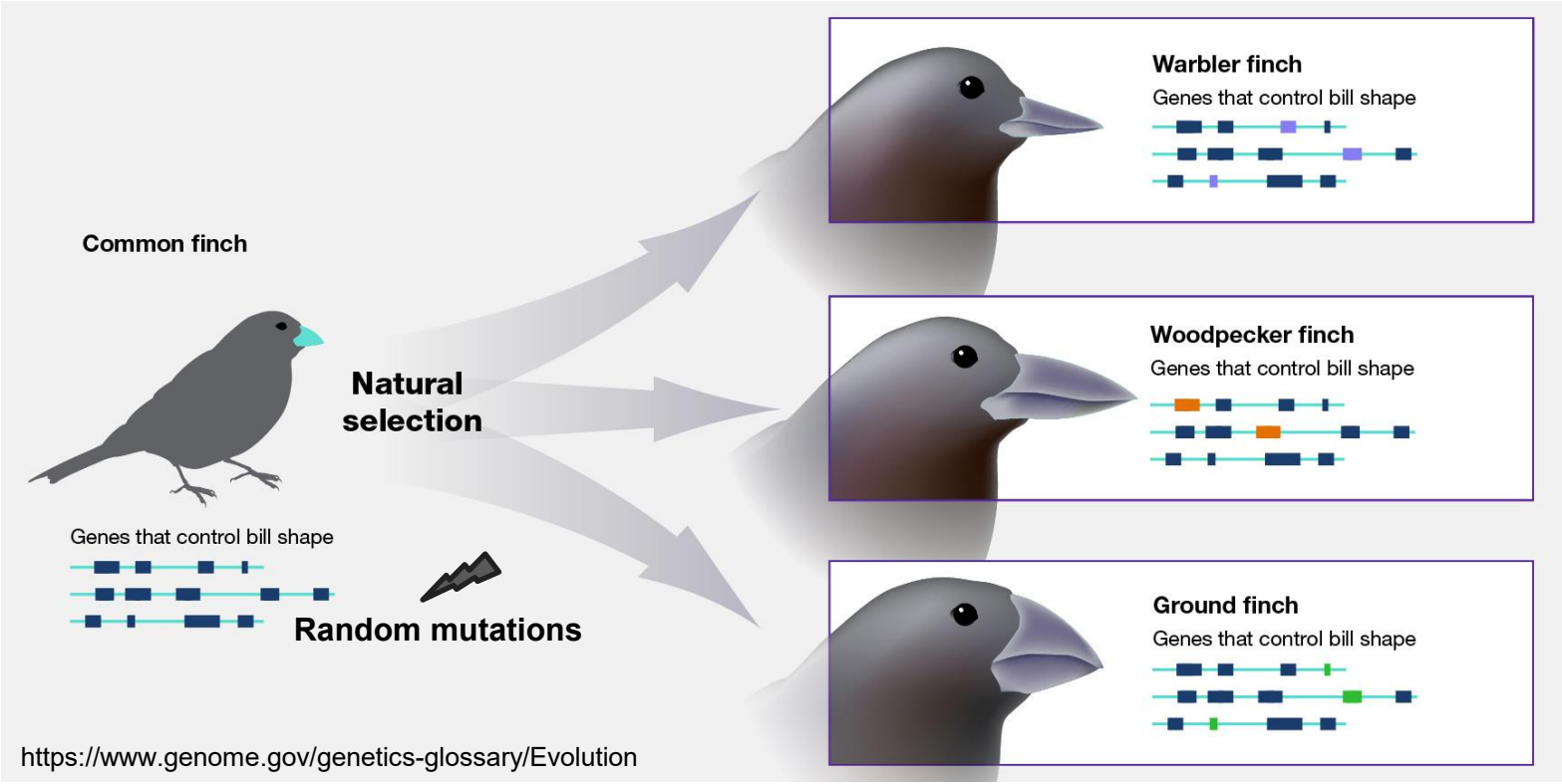
Evolutionary perspective

Biology developed through and follows evolution

shutterstock.com · 1652460871

- Mutations occurred randomly without knowing their consequences
- Evolution is the process of competition and selection that weeds out unfit mutations and retains beneficial ones
- Nature is one large experiment

Natural selection



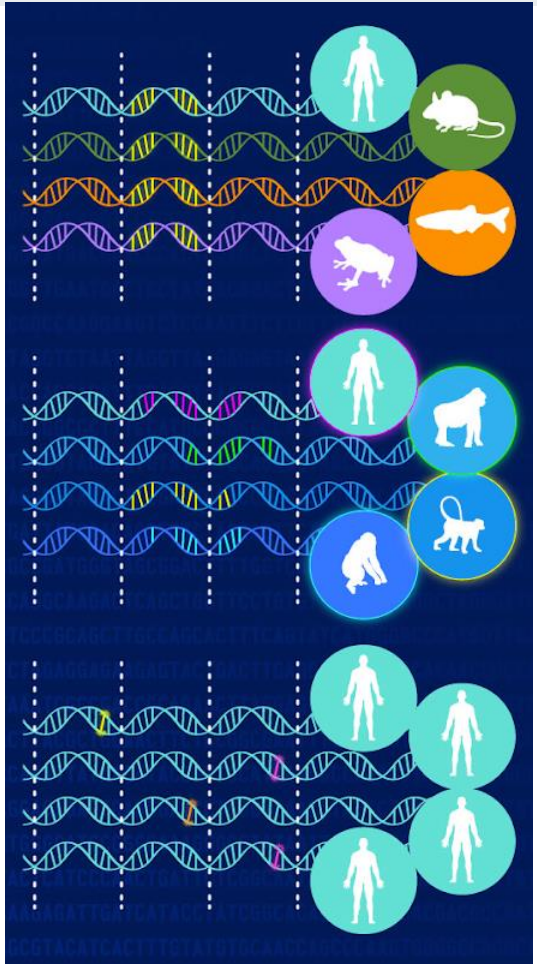
https://www.genome.gov/genetics-glossary/Evolution

Comparative genomics analysis

- Compare genomes of individuals/organisms with and without a phenotype

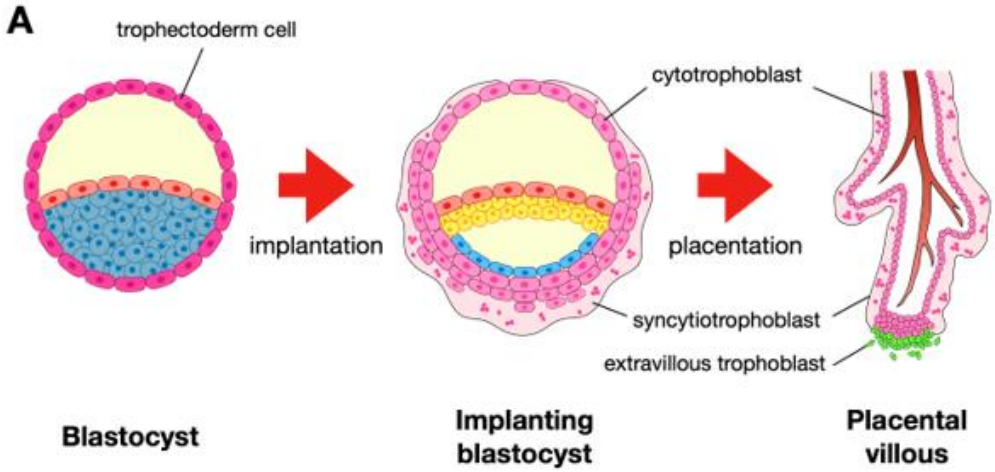
- Pinpoint causative genetic factors of the phenotype

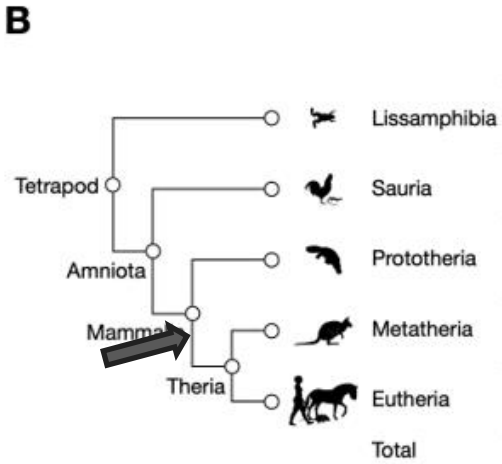
- Many scopes
  - **Genetic disease**: Human populations
  - **Brain development**: Primates
  - **Novel enzymes**: Microorganisms



https://neurosnap.ai/blog/post/comparative-genomics-analysis-of-evolutionary-relationships-among-species/64e14ceb033cf680aecc43c9

The rise of placenta



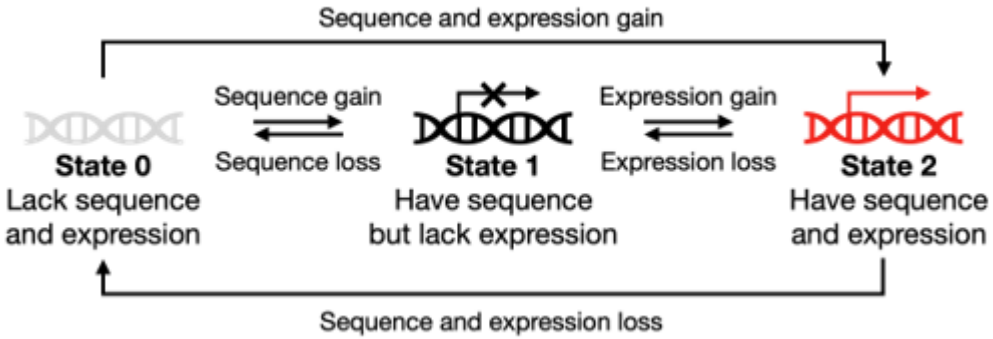


	Genome (spp.)	Transcriptome (spp.)
Lissamphibia	6	0
Sauria	106	1 (chorioallantoic membrane)
Prototheria	2	0
Metatheria	6	2 (placenta)
Eutheria	147	11 (placenta)
Total	267	14

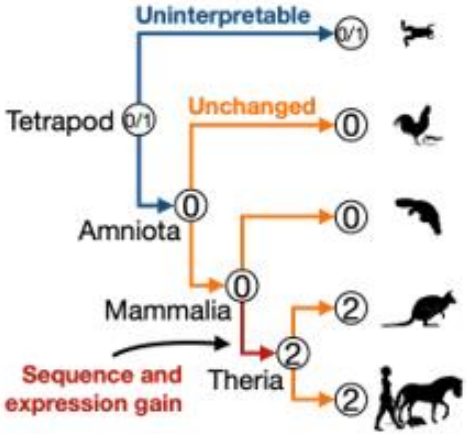
Plianchaisuk, A. et al. Mol Biol Evol msac176 (2022)

- Placenta emerged as a new organ in mammalian ancestor
- Genes essential for placenta development should be acquired during evolution **around the same period when placenta emerged**

The search for genes involved in placenta development



Plianchaisuk, A. et al. Mol Biol Evol msac176 (2022)

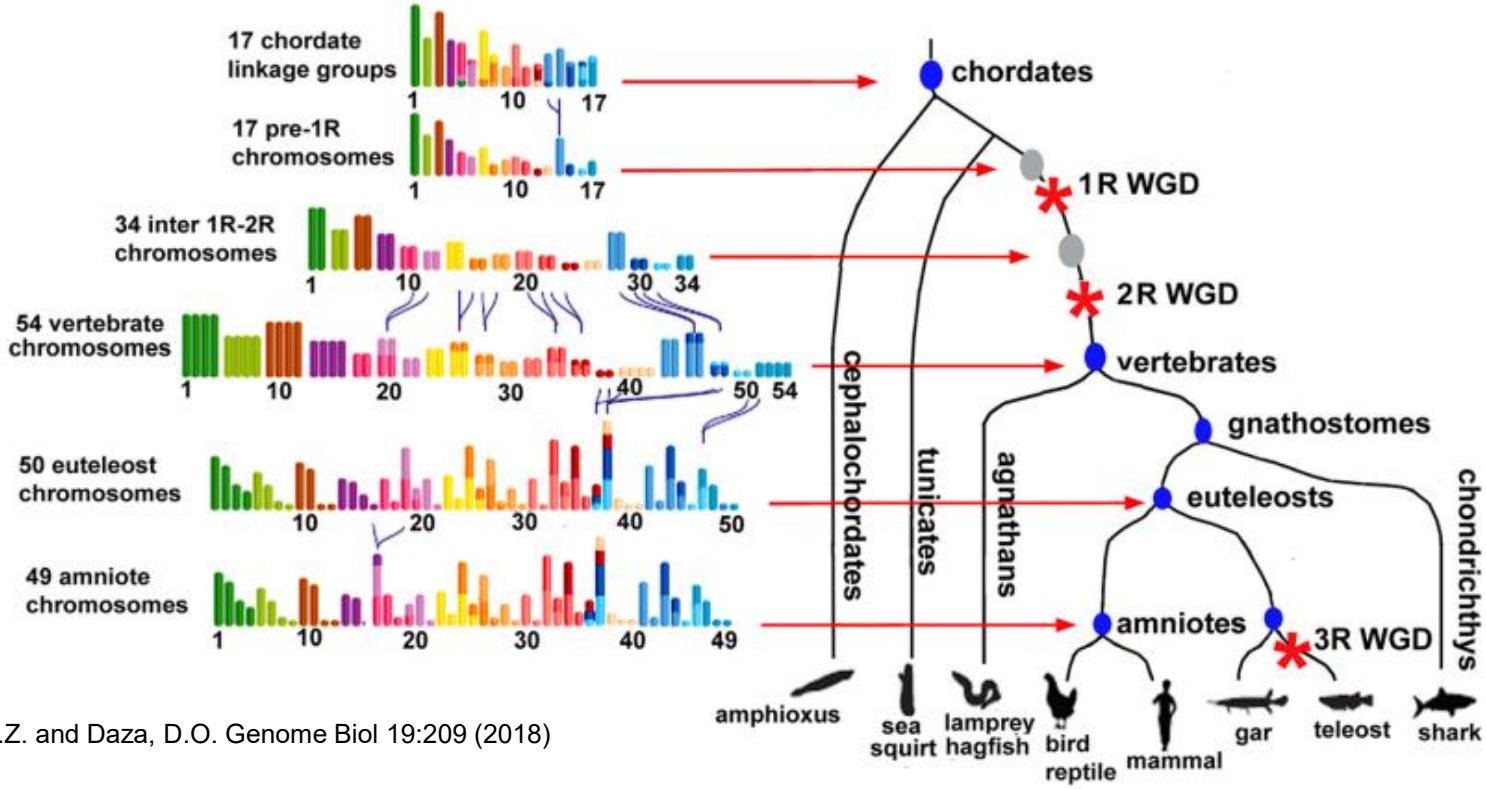


Interpreting character state changes from ancestors to descendants to transition events

Without evolutionary perspective, we have to develop an animal model with defects in placenta development

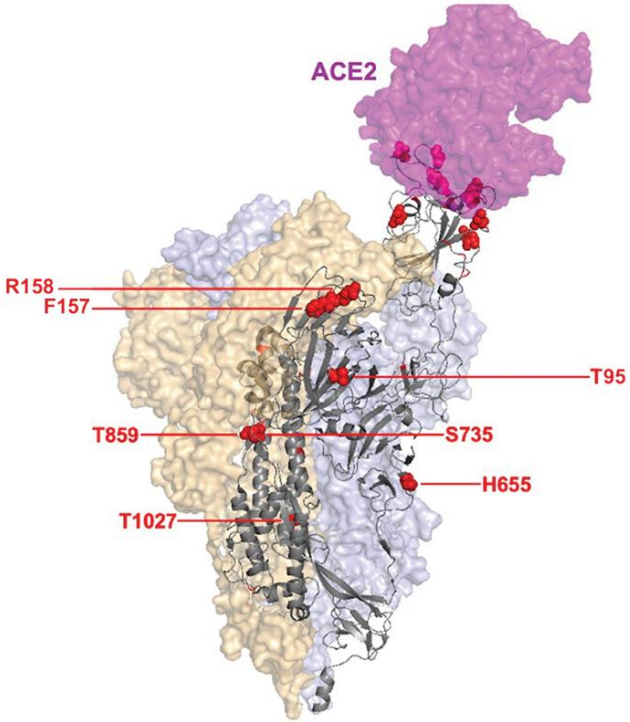
- Model function acquisition as two stages:
  - Gain of gene sequence (viral integration, gene duplication, etc.)
  - Activation of gene expression

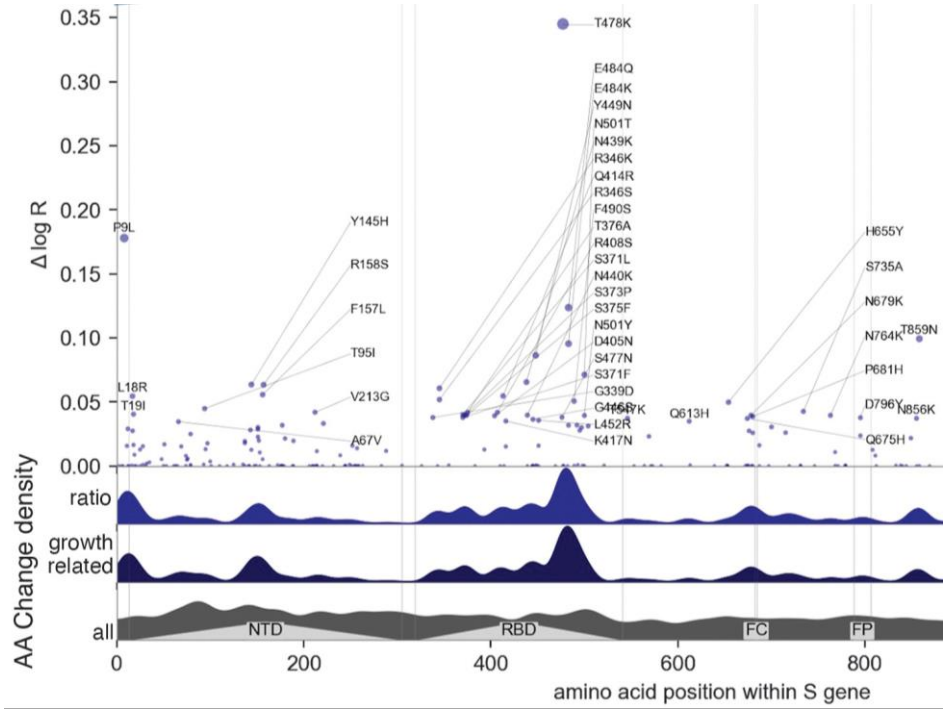
Genome expansions that gave rise to us



Holland, L.Z. and Daza, D.O. Genome Biol 19:209 (2018)

Amino acid evolution in Coronavirus's Spike protein





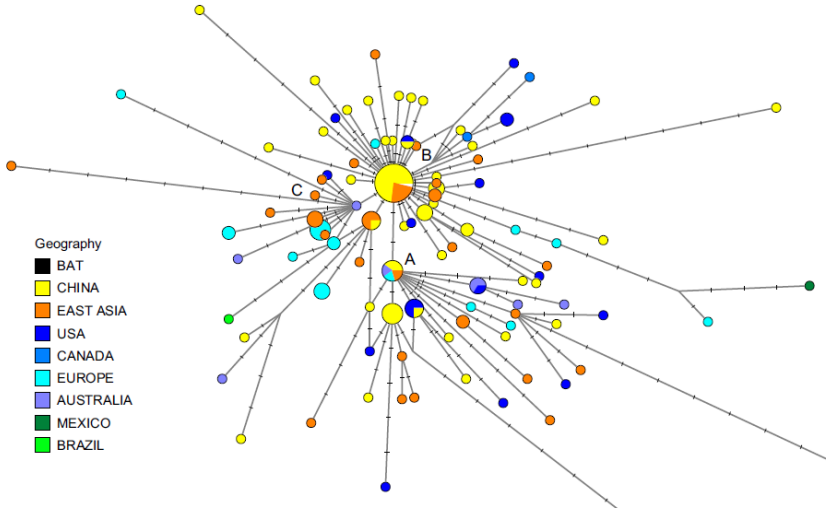
Obermeyer, F. et al. Science 376:1327-1332 (2022)

Applications of evolutionary perspective

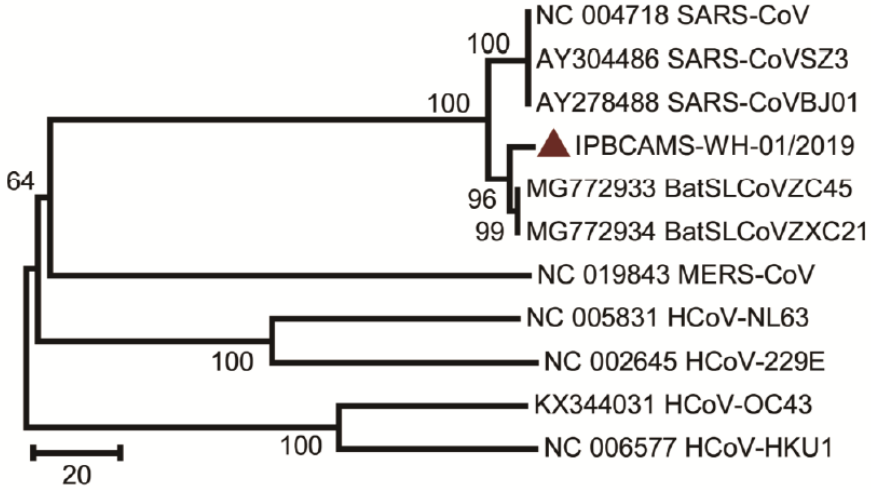
- Can be utilized in broad biology, not just evolutionary research
- Mutations in **conserved regions** probably have more chance to cause some phenotypes than those in **variable regions**
- Where to modify your enzyme to achieve new specificity?
- Where on a viral protein to target with antibody?
- Which group of genes are necessary for certain functions?

Phylogenetics

Phylogeny



Forster *et al*. PNAS (2020)



Guo *et al*. Clin Infect Dis (2020)

- Clustering of similar taxa with respect to evolutionary similarity
- Branch length reflect time and mutation rate

Types of phylogeny

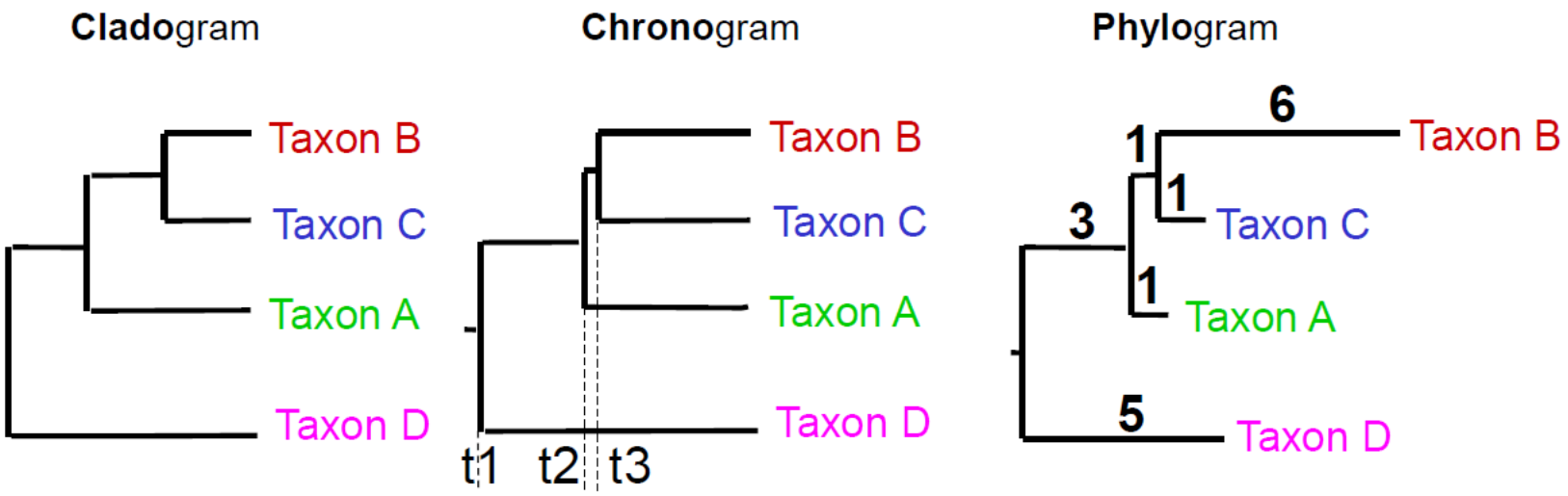


Image from Lecture 18 6.047 MIT OCW

- Depends research question: Is grouping or time more important?

Ingredients for phylogenetic reconstruction

Multiple sequence alignment

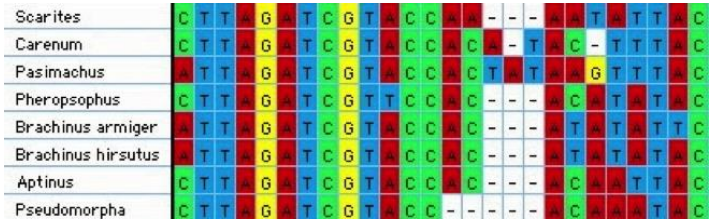
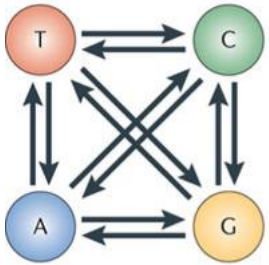


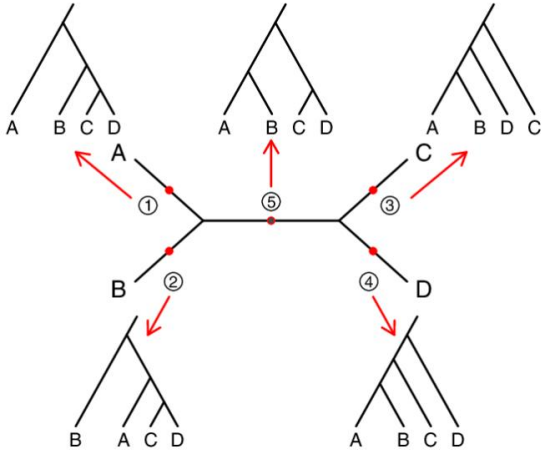
Image from www.mcqbiology.com

Evolutionary model



Yang & Rannala. Nat Rev Genetics (2012)

Tree building algorithm



Tian, Y. and Kubatko, L.S. BMC Evol Biol 17(1) (2017)

- Sequence data + **evolutionary model** + tree construction
- Other constraints and non-molecular data can also aid the reconstruction
  - Sampling date, phenotype, lifestyle, geography, etc.

Extra features for phylogenetic reconstruction

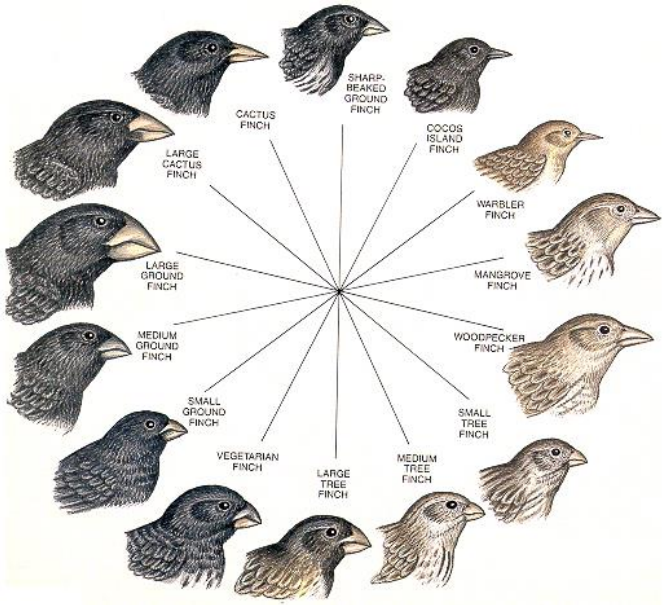
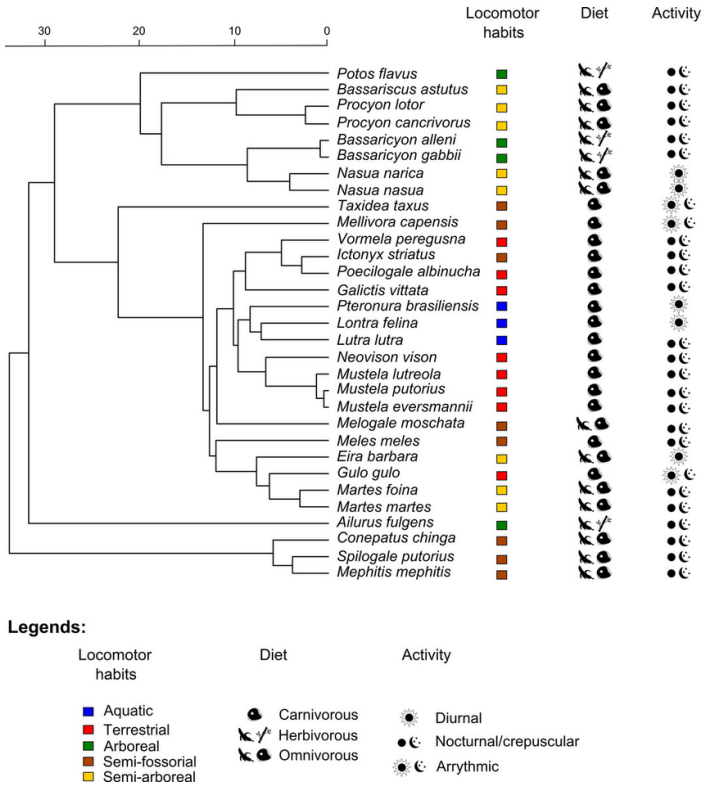


Image from pinterest



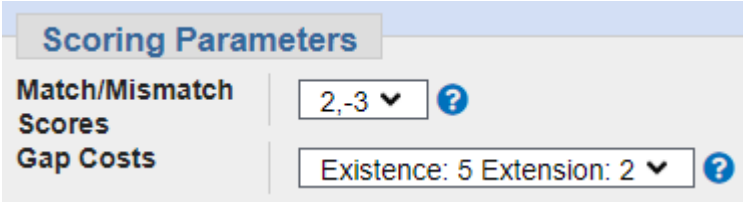
Dumont, M. et al. Biol J Linnean Soc (2015)

Evolutionary models:

Nucleotide/amino acid substitution models

Recall nucleotide alignment scores

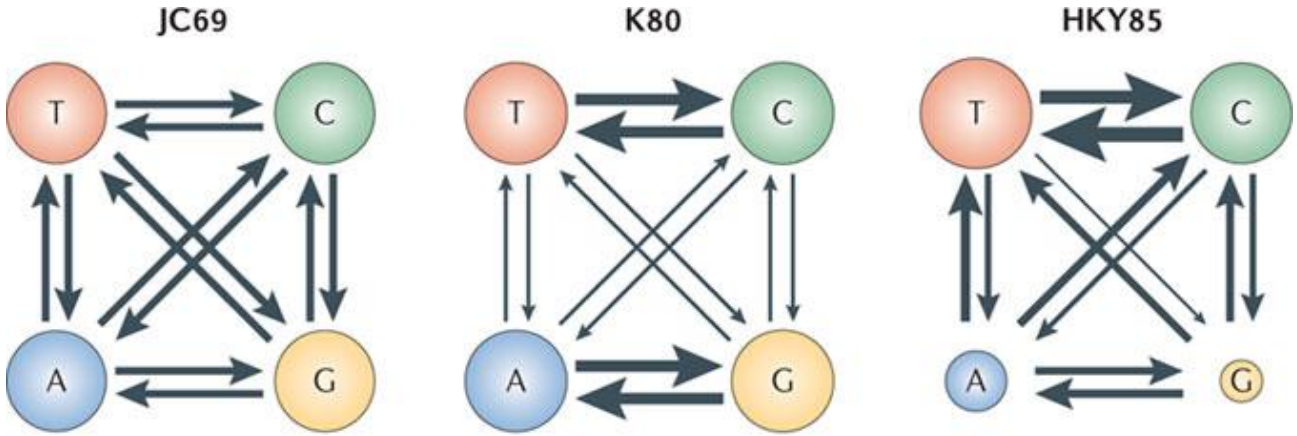
Score = +1+1-1-1-1+1+1+1+1 = +3



Score = +2+2-5-2-2-2+2+2+2+2 = +1

- A special case of substitution model that considers only match or mismatch, each position independently
- That's **not really how mutations occurred in nature**

Theoretical nucleotide substitution models



Yang & Rannala. Nat Rev Genetics, 13: 303-314 (2012)

- Juke-Cantor assumes equal base frequencies and equal substitution rates
- Kimura adds transition rate and transversion rate
- Hasegawa-Kishino-Yano adds different base frequencies

How many parameters do these models used?

JC69	A	C	G	T
A	$1-3x_1$	x_1	x_1	x_1
C	x_1	$1-3x_1$	x_1	x_1
G	x_1	x_1	$1-3x_1$	x_1
T	x_1	x_1	x_1	$1-3x_1$

K80	A	C	G	T
A	$1-x_1-x_2$	x_2	x_1	x_2
C	x_2	$1-x_1-x_2$	x_2	x_1
G	x_1	x_2	$1-x_1-x_2$	x_2
T	x_2	x_1	x_2	$1-x_1-x_2$

- Juke-Cantor = 1 (substitution rate)
- Kimura = 2 (transition rate, transversion rate)
- Hasegawa-Kishino-Yano = 5 (transition rate, transversion rate, 3 base frequencies)

General time-reversible model (GTR)

- **Symmetric** substitution rates: P(A→G) = P(G→A)
  - 6 parameters (4 nucleotides choose 2)
  - Time-reversible: switching ancestral and descendant taxa does not change the calculation
  - Useful in practice because we often don't know which taxon came first

- Different nucleotide frequencies: 3 parameters

- This is the most generalized time-reversible model possible

Amino acid substitution models

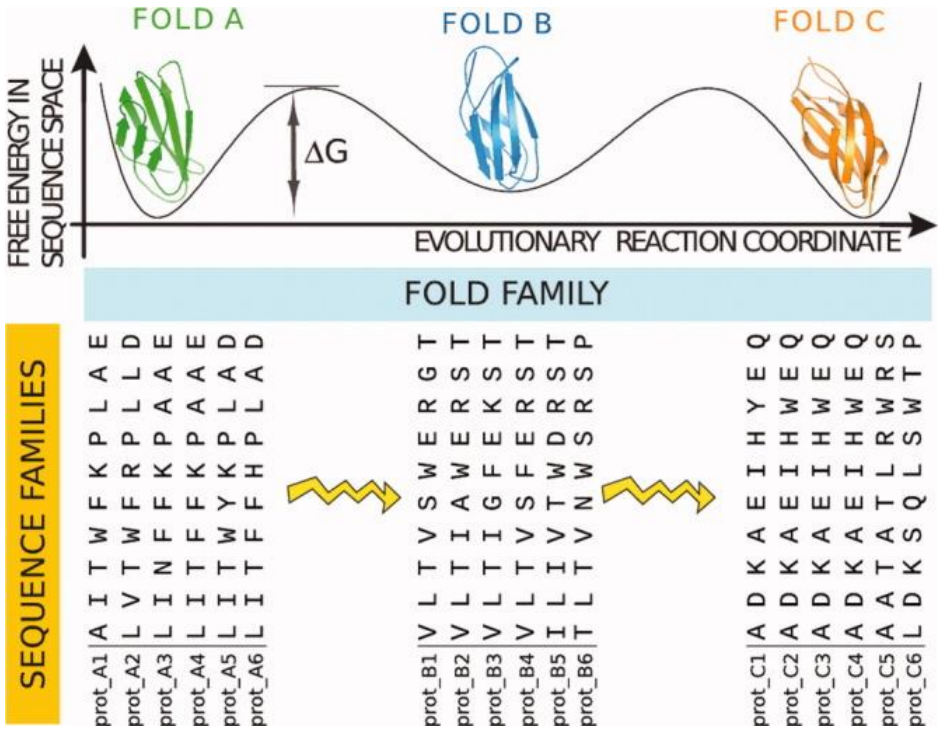
- BLOSUM & PAM

- Dayhoff 1978
- JTT (Jones-Taylor-Thornton, 1992)
  - Essentially PAM250

- WAG (Whelan-Goldman, 2001)
  - More protein families included

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Structural context for substitution model

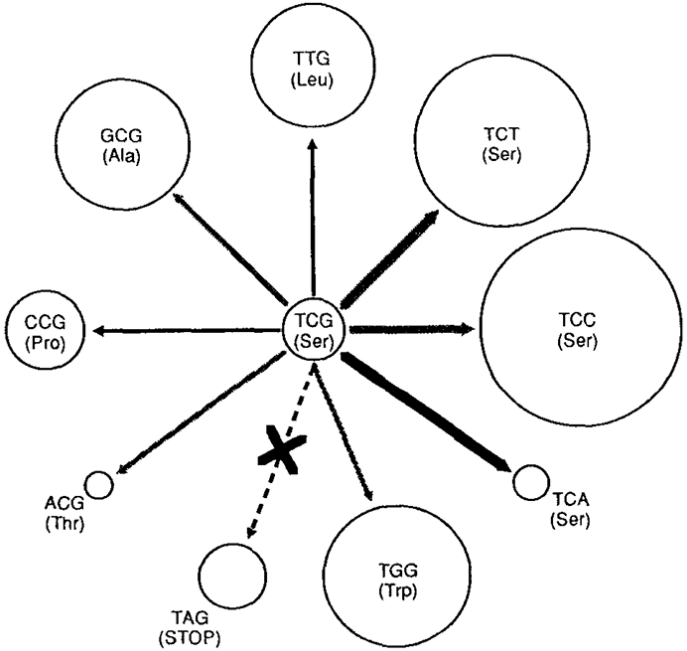


Liberles, D.A. Protein Sci. 21: 769-785 (2012)

- Different positions on different protein families have different evolutionary constraints
- Use energy from interactions between residues in 3D to develop substitution matrix
- Similar idea to **Position-Specific Substitution Matrix** (PSSM) used by PSI-BLAST

Codon substitution model

- GY94 (Goldman-Yang)
- Learn base codon frequencies from nucleotide and amino acid data
- Codon neighborhood
  - Substitution rate depends on similarity between coded amino acids
- Non-synonymous / synonymous rate
- Similar idea to **BLASTX** and **tBLASTN**



Goldman, N. and Yang, Z. Mol Biol Evol 11:725-736 (1994)

Codon alignment: amino acid → nucleotide

A. DNA alignment

DNA alignment can disrupt codon frames

```
Q9FPK4   ATGGGTGTTTTCAGCTACGAGGATGAGGCCACC-TCCGTTATCCCTCCGGCTA-GGCTGTTCAAGTCCTTTGTCCTAGATGCCGAC-AACCTCATT
Q9FPK3   ATGGGTGTTTTCTGCTACGAGGATGAGGCCACC-TCCGTTATCCCTCCGGCTA-GGCTGTTCAAGTCCTTTGTCCTAGATGCCGAC-AACCTCATT
Q945E7   ATGGGTGTTGTGAGTTATGAGTTTGAGGTAACC-TCCCCAATTGCTCCAGCCA-GGCTTTTCAAGGCTTTTGTTCTTGAGGCTGCC-AAGATTTGG
Q6XC94   ATGGGTGTTGCGAGTTATGAGTTTGAGGTAACC-TCCCCAATTGCTCCAGCCA-GGCTTTTCAAGGCTTTTGTTCTTGAGGCTGCC-AAGATTTGG
Q6Q4B5   ATGGGTGTTGTGAGTTATGACTTGGAGGTAACCCTCCCCAATTGCTCCAGCCAAGGCTTTTCAAGGCTTTTGTTCTTGAACCTGCCCAAGGTTTGG
Q43549   ATGGGTGTTTTCAATTACGAAACTGAGTTCACC-TCCGTCATTCCCCCTGCTA-GGTTGTTCAATGCCTTTGTTCTTGATGCTGAC-AACCTCATC
Q4VPJ1   ATGGGTGTTTTCACATACGAATCTGAGTCCACC-TCCGTCATCCCCCCTGCTA-GGTTGTTCAATGCGACTGCTCTTGATGGTGAC-AAACTCATC
Q84LA7   ATGGGTGTCTTCACATACGAATCCGAATTCACC-TCCGTCATCCCCCCTGCTA-GGTTGTTCAATGCCTTTGTTCTTGATGCTGAC-AACCTCATC
Q4VPI3   ATGGGTGTTTTCACATACGAATCCGAGTTCACC-TCTATCATCCCCCCTGCTA-GGTTGTTCAATGCCTTTGTTCTTGATGCTGAC-AACCTCATC
```

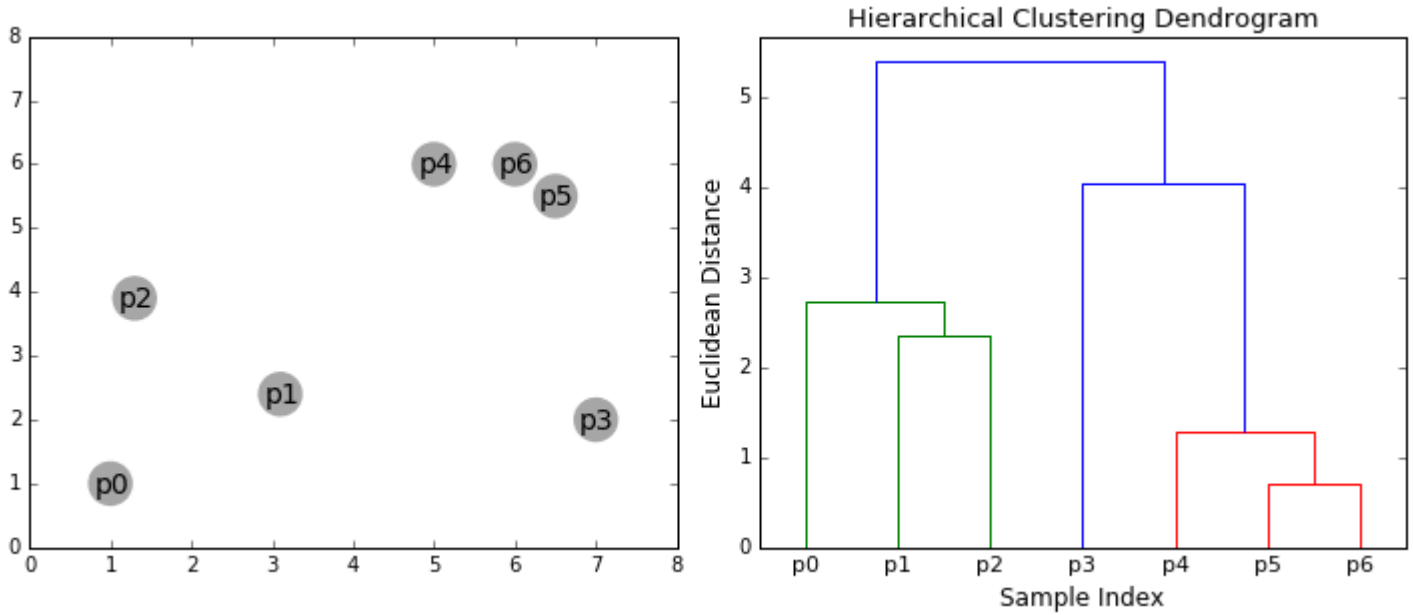
B. Back-translation from protein alignment

```
Q9FPK4   ATGGGTGTTTTCAGCTACGAGGATGAGGCCACCTCCGTTATCCCTCCGGCTAGGCTGTTC---AAGTCCTTTGTCCTAGATGCCGACAACCTCATT
Q9FPK3   ATGGGTGTTTTCTGCTACGAGGATGAGGCCACCTCCGTTATCCCTCCGGCTAGGCTGTTC---AAGTCCTTTGTCCTAGATGCCGACAACCTCATT
Q945E7   ATGGGTGTTGTGAGTTATGAGTTTGAGGTAACCTCCCCAATTGCTCCAGCCAGGCTTTTC---AAGGCTTTTGTTCTTGAGGCTGCCAAGATTTGG
Q6XC94   ATGGGTGTTGCGAGTTATGAGTTTGAGGTAACCTCCCCAATTGCTCCAGCCAGGCTTTTC---AAGGCTTTTGTTCTTGAGGCTGCCAAGATTTGG
Q6Q4B5   ATGGGTGTTGTGAGTTATGACTTGGAGGTAACCCTCCCCAATTGCTCCAGCCAAGGCTTTTCAAGGCTTTTGTTCTTGAACCTGCCCAAGGTTTGG
Q43549   ATGGGTGTTTTCAATTACGAAACTGAGTTCACCTCCGTCATTCCCCCTGCTAGGTTGTTC---AATGCCTTTGTTCTTGATGCTGACAACCTCATC
Q4VPJ1   ATGGGTGTTTTCACATACGAATCTGAGTCCACCTCCGTCATCCCCCCTGCTAGGTTGTTC---AATGCGACTGCTCTTGATGGTGACAAACTCATC
Q84LA7   ATGGGTGTCTTCACATACGAATCCGAATTCACCTCCGTCATCCCCCCTGCTAGGTTGTTC---AATGCCTTTGTTCTTGATGCTGACAACCTCATC
Q4VPI3   ATGGGTGTTTTCACATACGAATCCGAGTTCACCTCTATCATCCCCCCTGCTAGGTTGTTC---AATGCCTTTGTTCTTGATGCTGACAACCTCATC
```

Scherer, N. and Basso, D.M. Genet. Mol. Res. 7:853-860 (2008)

Phylogenetic tree construction

Hierarchical/agglomerative clustering



https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

- **Iterative** grouping of the **most similar** samples/groups until all are connected

UPGMA: A hierarchical clustering approach

1

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

2

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	8.00	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

3

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	12.50	29.00	28.00

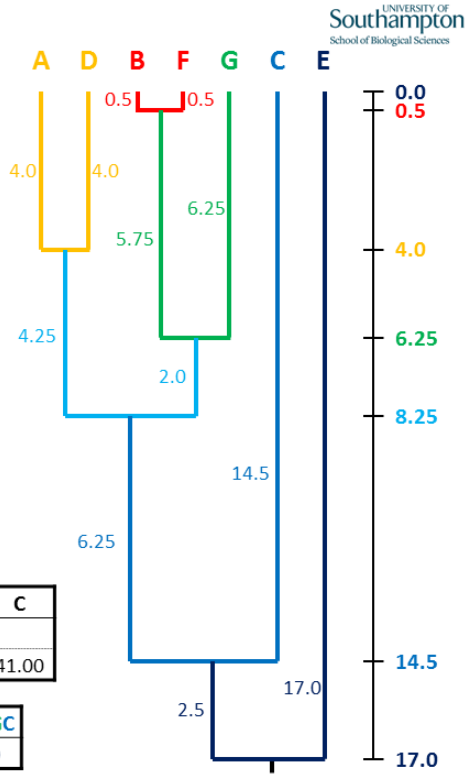
4

	AD	BFG	C
BFG	16.50		
C	26.50	30.67	
E	32.00	33.00	41.00

5

	ADBFG	C
C	29.00	
E	32.60	41.00

	ADBFGC
E	34.00



- Distance between groups of taxa to average distance between all pairs
- Calculated based on a selected **substitution model**

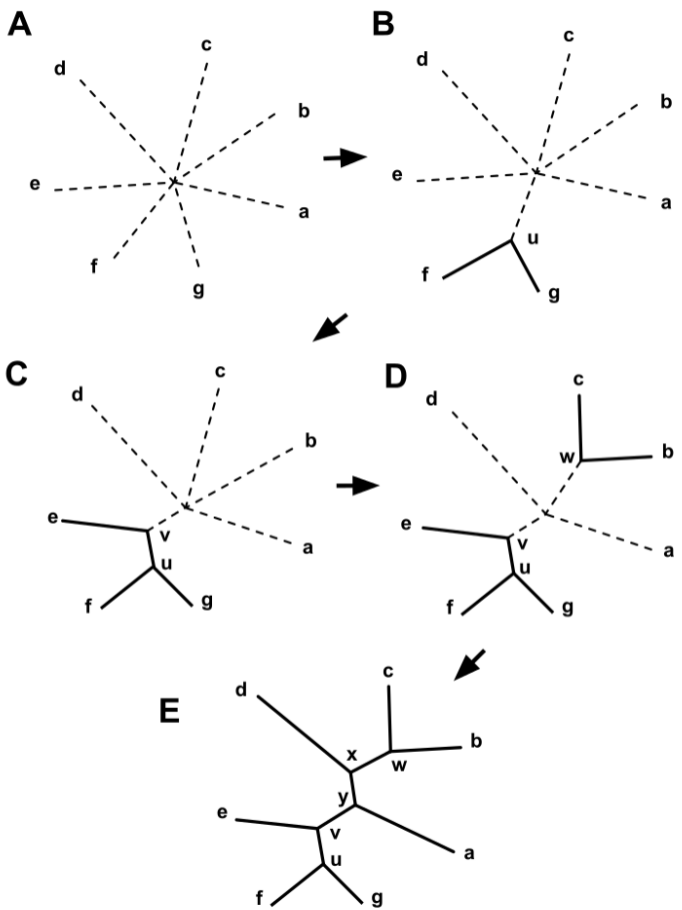
http://www.slimsuite.unsw.edu.au/teaching/upgma/

NJ: Neighbor joining

- Start with a star tree
- Join taxa *i*, *j* with smallest Q score

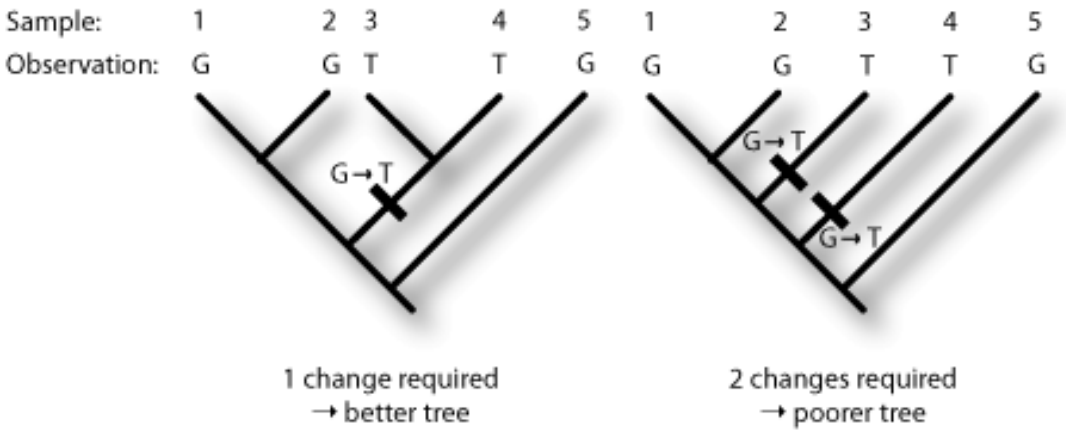
$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n}[d(i,k) + d(j,k)]$$

- Join taxa that are similar to each other but dissimilar from other taxa
- *d(i, j)* is based on a substitution model



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

Maximum parsimony / Minimum evolution



https://biology.stackexchange.com/questions/60161/which-nucleotide-as-start-point-for-maximum-parsimony

- Explanation requiring **minimum number of changes** is the most likely
- Fail under **long evolutionary time**: A → T → A
- Cannot distinguish **convergent evolution**

Maximum likelihood principle

Let's review some terminology

- P(A) = probability that A occurs

Joint probability: P(A, B) = probability that A and B occurs

Conditional probability:
- P(A | B) = probability that A occurs given that B already occurred
- P(A | B) = P(A, B) / P(B)
- P(A, B) = P(A | B) P(B) = P(B | A) P(A)

- P(Shopping | Sunny) = ? How about P(Sunny | Shopping)?
  - **Which one is more sensible?**

Bayes' rule

$$P(A \mid B) / P(A) = P(A, B) = P(B \mid A) / P(B)$$

- A = **Evolutionary data** is observed
- B = **Proposed phylogenetics** is true

- We want to identify phylogenetics that maximize P(B | A)

- Using **Bayes' rule**:
  - P(B | A) = P(A | B) x P(B) / P(A)
  - P(A | B) is much easier (and more sensible) to compute!
  - Prior belief, P(B), can be integrated with likelihood to get a better estimate

Maximum likelihood principle

- **Likelihood score** = P(data | model), or P(data | hypothesis)

- Find the model or hypothesis that maximize the likelihood score

- Become an **optimization problem**!
  - How to identify such model or hypothesis if there are many of them?
  - **Calculus**
  - **Brute-force search**

Maximum likelihood answer is often intuitive

- Gene A has two alleles, *A* and *a*
- A study of 1,000 Thai individuals found 700 with genotype *AA*, 200 with genotype *Aa*, and 100 with genotype *aa*
- What is the estimated allele frequency of *A*?

- **Total allele counts**
  - ***A***: 2 x 700 + 200 = 1,600
  - ***a***: 2 x 100 + 200 = 400
  - Frequency of *A* = 1,600 / 2,000 = 0.8

Maximum likelihood answer is often intuitive

- In a study of 5 pancreatic cancer patients, they passed away after 1, 5, 3, 4, and 5 years, respectively
- What is the estimated yearly survival rate?

- **Observation**: SD, SSSSSD, SSSD, SSSSD, SSSSSD
  - 18 *S*'s and 5 *D*'s
  - 23 years total
  - Probability of *S* = 18/23

MLE: Maximum likelihood estimator

- Gene A has two alleles, *A* and *a*
- A study of 1,000 Thai individuals found 700 with genotype *AA*, 200 with genotype *Aa*, and 100 with genotype *aa*
- What is the estimated allele frequency of *A*?

  - Parametrize the allele frequencies: $f_A = p$ and $f_a = 1 - p$
  - $P(AA) = p^2$, $P(Aa) = 2p(1 - p)$, and $P(aa) = (1 - p)^2$
  - **Likelihood** $= P(AA)^{700}\, P(Aa)^{200}\, P(aa)^{100} = p^{1400} 2^{200} p^{200} (1 - p)^{200} (1 - p)^{200}$
    $= 2^{200} p^{1600} (1 - p)^{400}$
  - Which p maximize the likelihood?
    - Solve the equation $\frac{dLikelihood}{dp} = 0 \rightarrow p_{\text{MLE}} = 0.8$

MLE: Maximum likelihood estimator

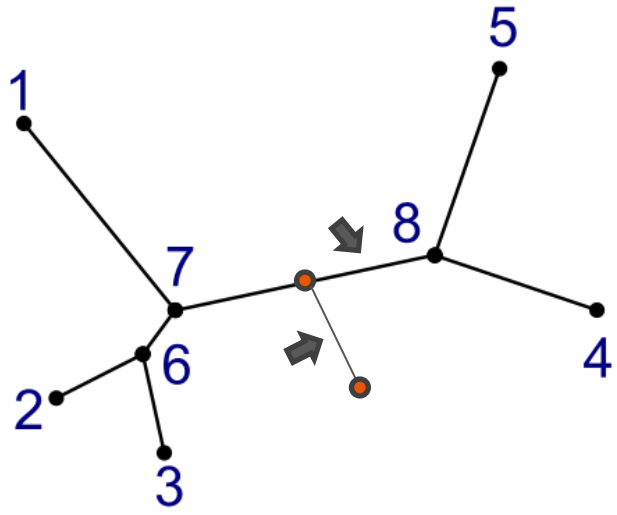
- In a study of 5 pancreatic cancer patients, they passed away after 1, 5, 3, 4, and 5 years, respectively
- What is the estimated yearly survival rate?

  - Let's set the yearly survival rate = r
  - P(survive exactly k years) = $r^k(1 - r)$
  - P(data | r) = $r^1(1 - r)\, r^5(1 - r)\, r^3(1 - r)\, r^4(1 - r)\, r^5(1 - r) = r^{18}(1 - r)^5$
  - Which r maximize the likelihood?
    - Solve the equation $\frac{dLikelihood}{dr} = 0 \rightarrow r_{\text{MLE}} = 18/23$

MLE for phylogenetics

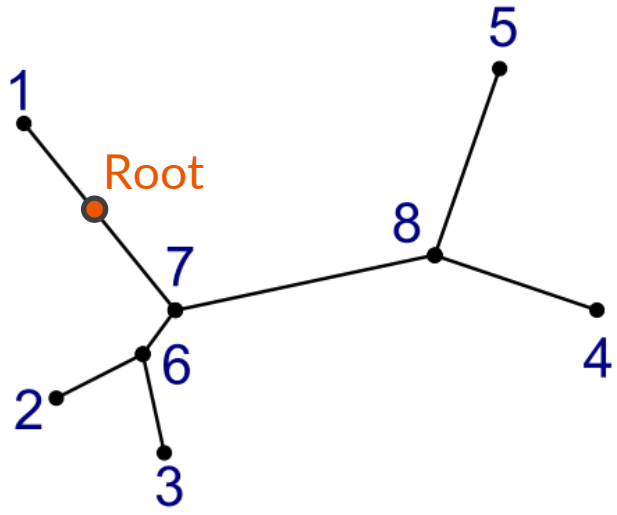
- **Likelihood** = P(sequence data | substitution model, phylogenetics tree)
- We have a limited number of substitution models
- Given a fixed phylogenetic tree topology, we can alter the lengths of the branches to find the best answer:
- The problem is how to find the best tree topology

Number of possible trees with distinct leaves



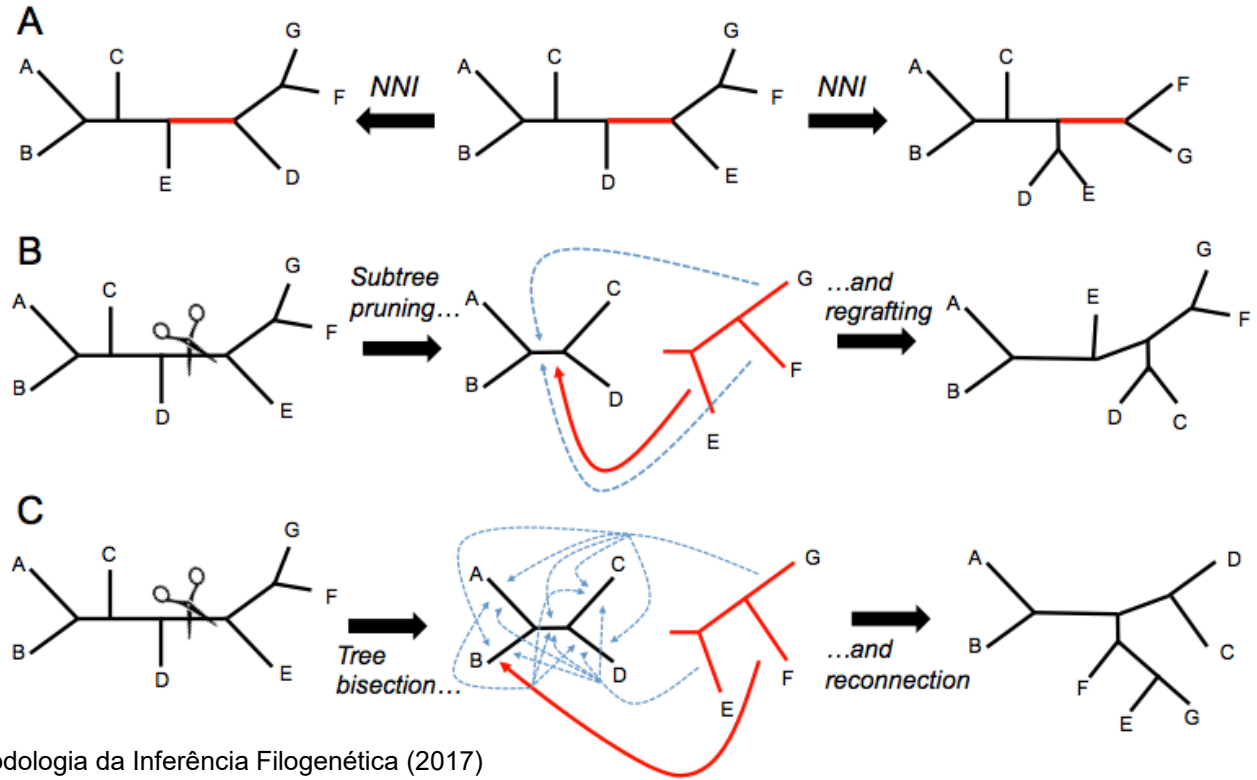
- Unrooted tree with *n* leaves has 2*n*−2 nodes
  - Adding a leaf will also create one internal node
- There are 2*n*−3 branches
  - Adding a leaf will create two branches
  - One new branch and split an existing branch
- From each unrooted tree with n leaves, there are 2*n*−3 locations to attach a new leaf
- Number of unrooted trees with *n* leaves, for n>2, is (2*n*−5)x(2*n*−7)x(2*n*−9)x … x1

Number of possible trees with distinct leaves



- For each unrooted tree with *n* leaves, there are 2*n*−3 locations to designate as the root (common ancestor)
- Number of rooted trees with *n* leaves, for n>2, is (2*n*−3)x(2*n*−5)x(2*n*−7)x(2*n*−9)x … x1
- Number of rooted trees with 10 leaves = 17x15x13x11x9x7x5x3 = 34,459,425

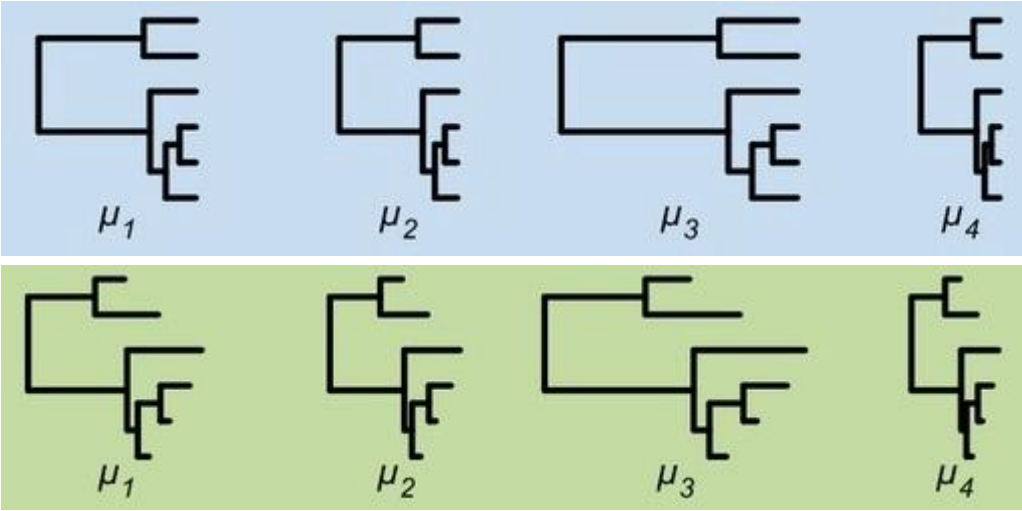
Heuristic tree search algorithms



Gainett *et al*. Metodologia da Inferência Filogenética (2017)

Additional evolutionary parameters

Molecular clock assumption



Ho, S.Y.W. and Duchene, S. Molecular Ecology 23:5947-65 (2014)

- Molecular clock assumes constant evolutionary rate throughout the tree
- Same root-to-tip distance → allow dating of evolutionary events

Site-specific evolutionary models

Invariant sites

Gamma-distributed mutation rates

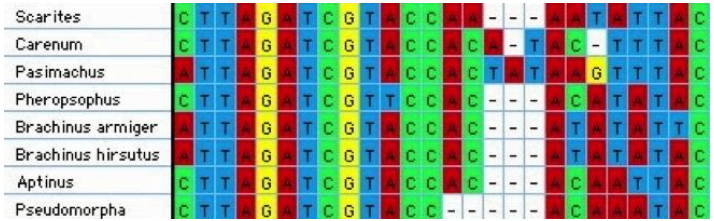
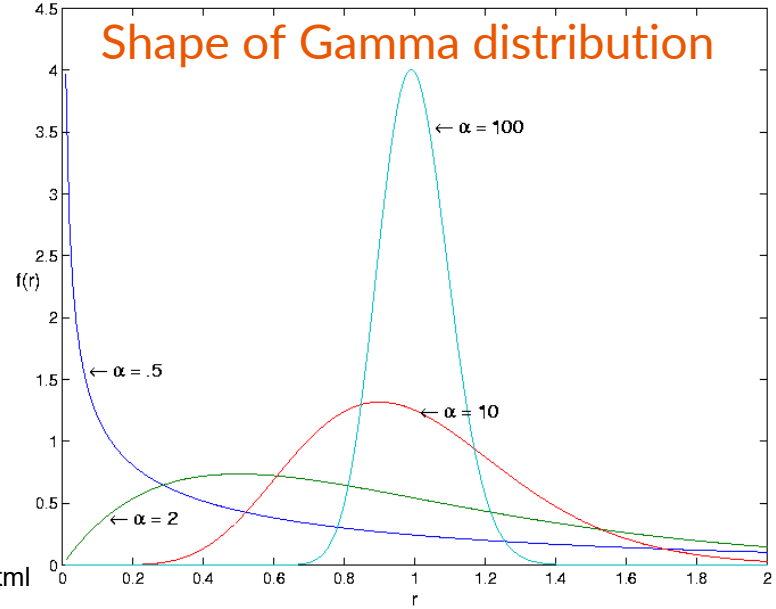


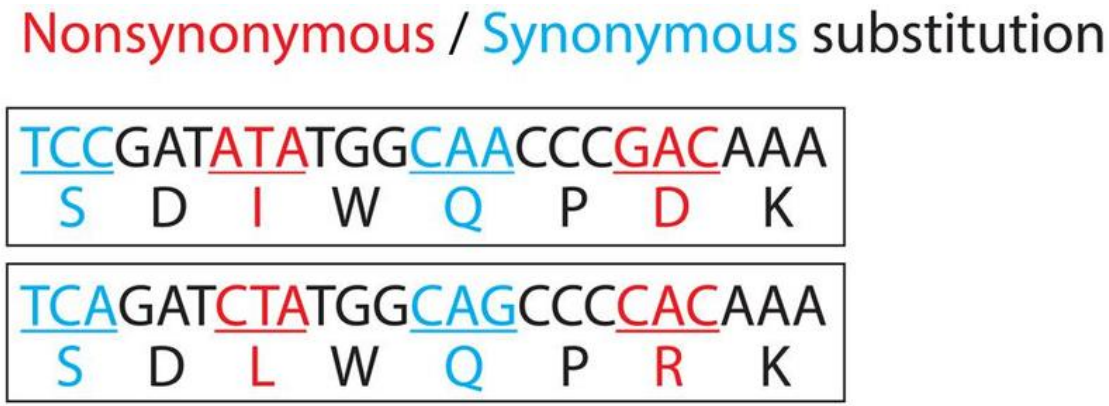
Image from www.mcqbiology.com



http://www.bioinf.man.ac.uk/resources/phase/manual/node81.html

- Gamma distribution $\Gamma(\alpha, \alpha)$ can mimic diverse shapes with mean = 1
- High α results in bell shape while low α yields higher variance = $1/\alpha$

Natural selection on codon model



Luo, H. Frontiers in Microbiology 6:191 (2015)

- **Null hypothesis**: synonymous and non-synonymous occurs proportional to the number of corresponding codon positions (dN/dS = 1)
- **Alternative hypothesis**: dN/dS can differ from 1

Nested model testing (chi-squared & likelihood)

Model complexity

Model	-ln$L$
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

models	diff. DF = q	X^2	P
JC-F81	3 - 0 = 3	155	0
F81-HKY85	4 – 3 = 1	549.4	0
KHY-TrN	5 – 4 = 1	2.1	0.15

F81 model fits the data significantly better than JC

HKY85 model fits the data significantly better than F81

TrN model does not fit the data significantly better than HKY85

Model	-ln$L$
HKY85	3233.34395
HKY85 +G	3145.29031
HKY85 +I+G	3142.36439

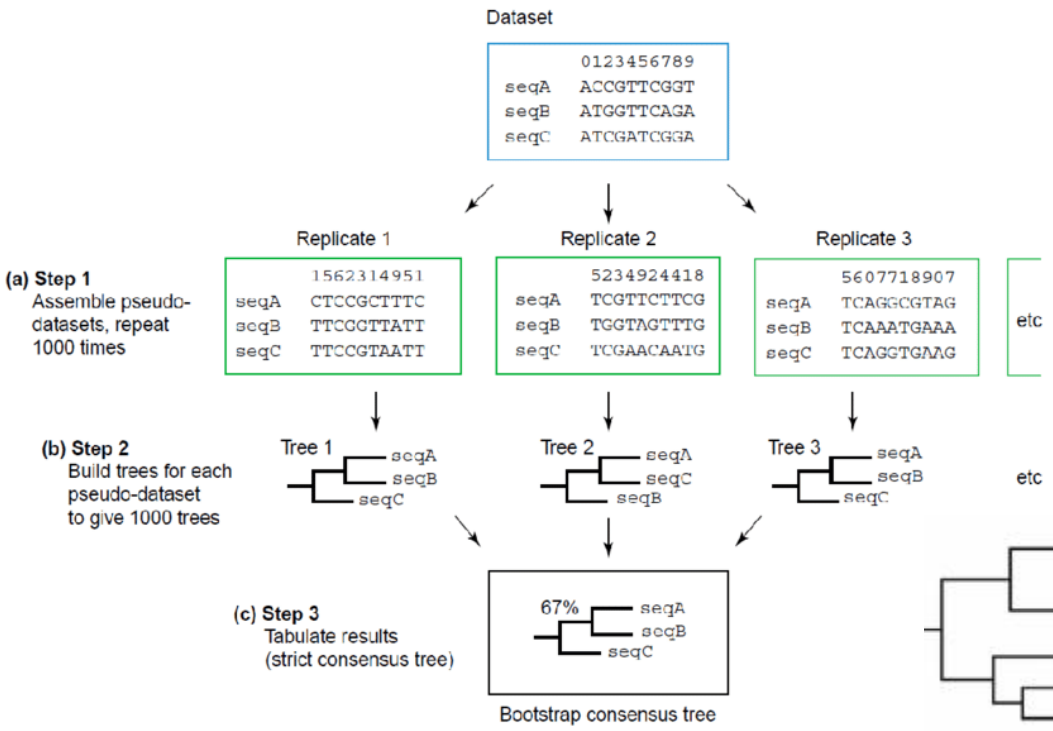
models	diff. DF = q	X^2	P
HKY85-vs. +G	1	176	0
HKY85+G vs. I+G	1	5.85	0.015

Adding site-specific rate fits the data significantly better

Adding invariant sites does not fit the data significantly better

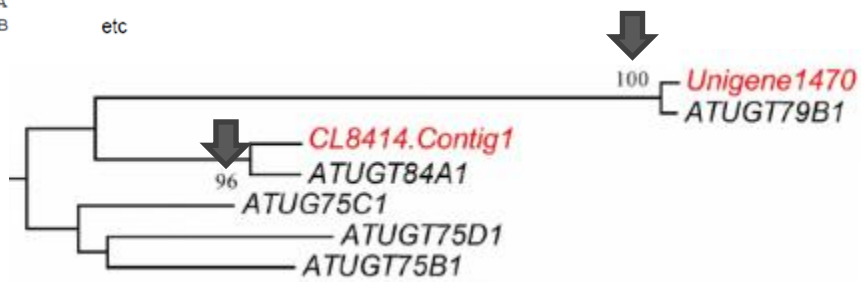
https://www.ccg.unam.mx/~vinuesa/Model_fitting_in_phylogenetics.html

Bootstrap support for taxa group



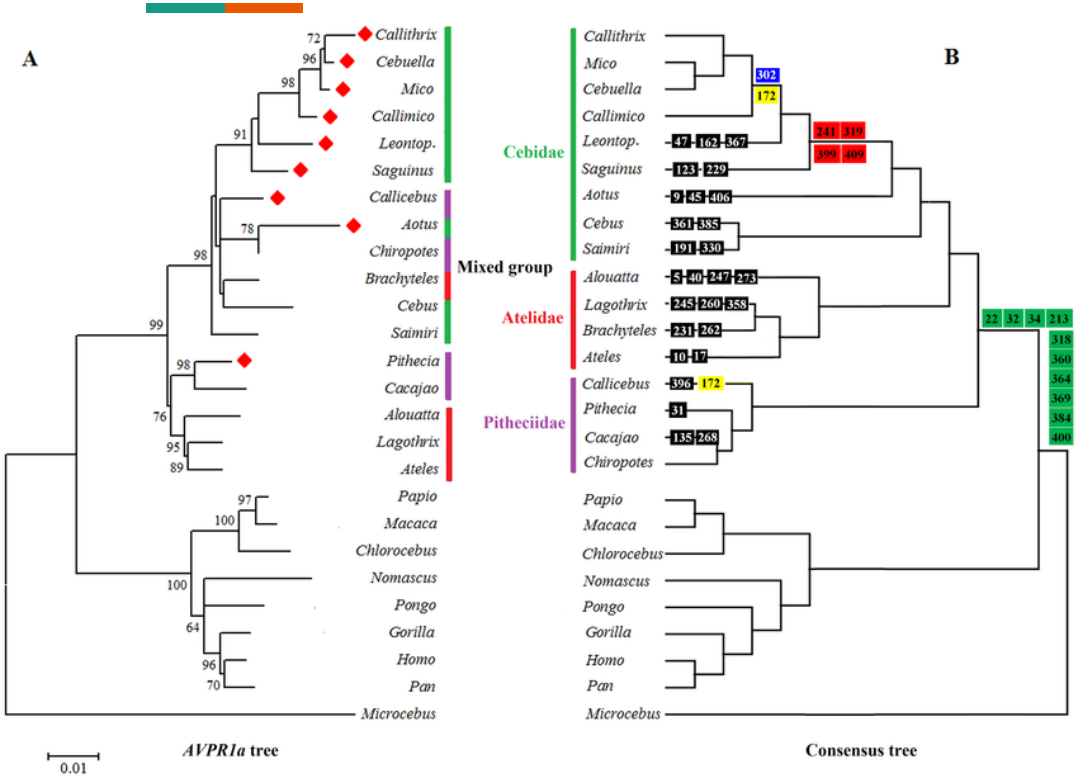
Wyant *et al*. Archives of Virology (2011)

- Bootstrapping = sampling with repeats
- Allow multiple trees to be constructed from a single sequence alignment



Wei et al. PLoS ONE 2013

Gene tree vs concatenated tree

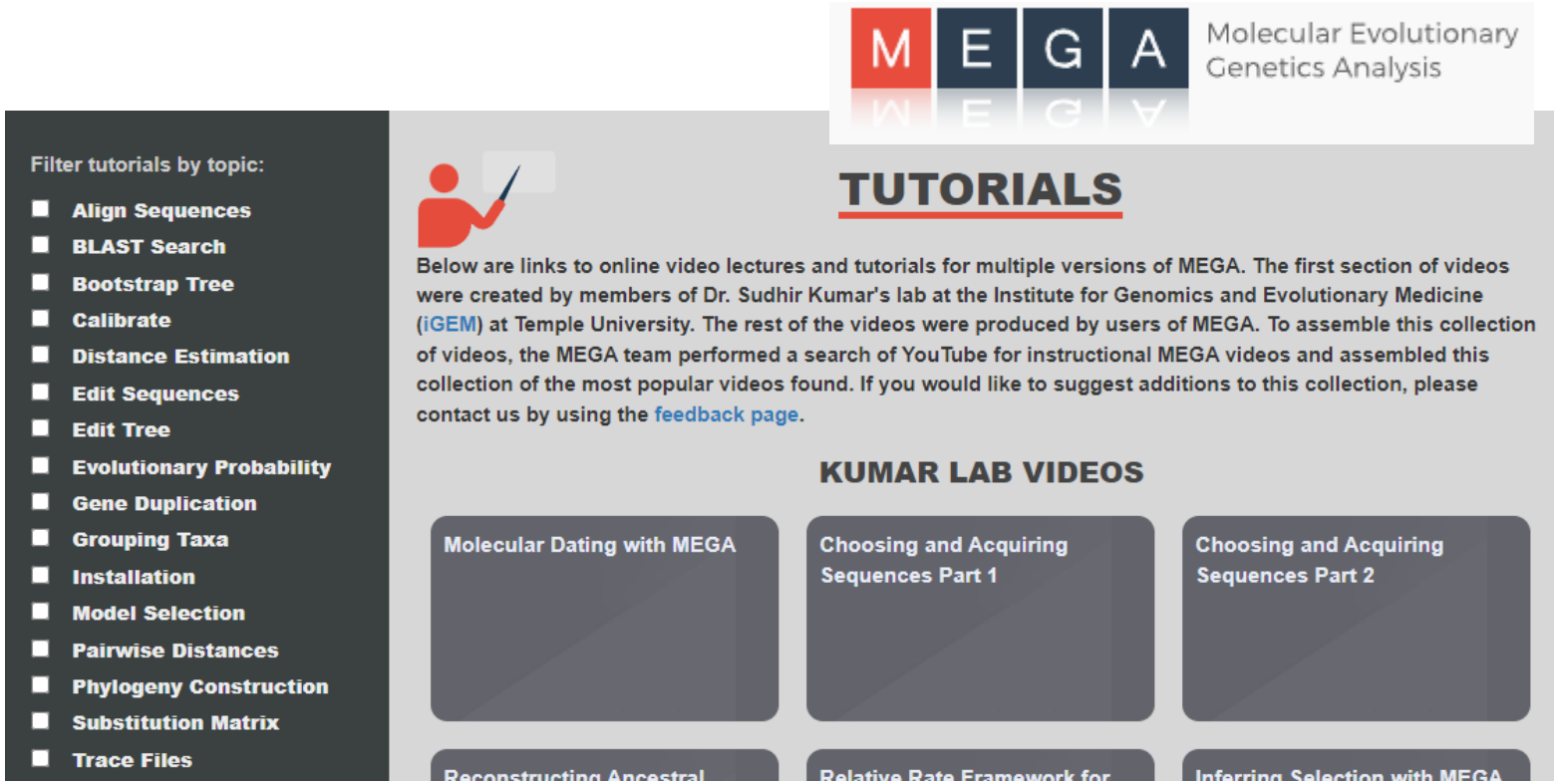


Ren, D. et al. PLoS ONE 9:e222638 (2014)

- Different gene may have different evolutionary history
- Housekeeping genes vs taxa-specific genes
- Can provide insights into the evolutionary process

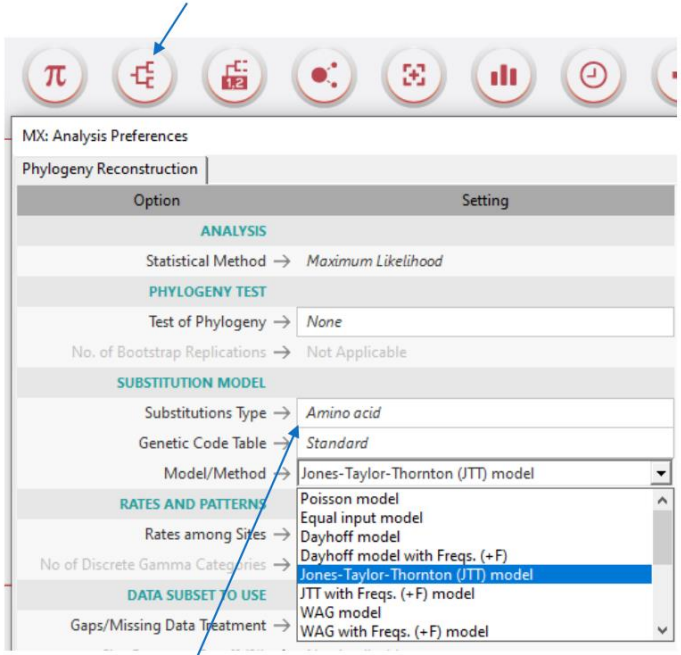
Example of how to setup phylogenetic reconstruction

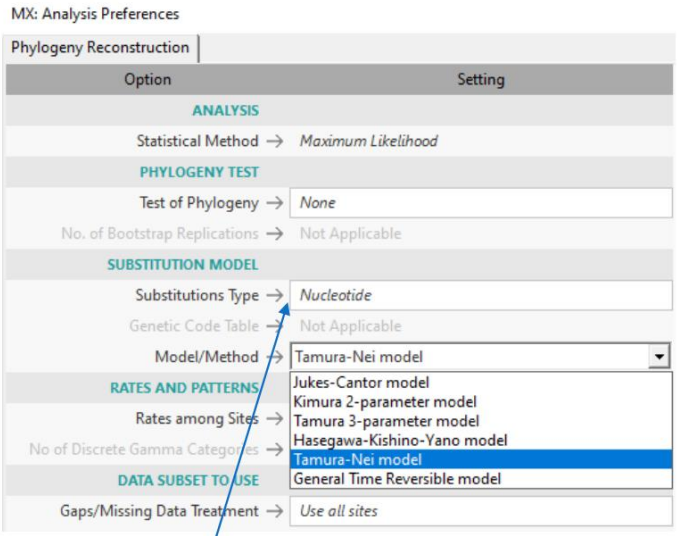
MEGA: A GUI tool for phylogenetic analysis



Substitution model choices

PHYLOGENY mode





Amino acid-based models

Nucleotide-based models

Maximum likelihood parameters

Phylogeny Reconstruction

Option	Setting
ANALYSIS	
Statistical Method →	*Maximum Likelihood*
PHYLOGENY TEST	
Test of Phylogeny →	None
No. of Bootstrap Replications →	None / Bootstrap method
SUBSTITUTION MODEL	
Substitutions Type →	*Nucleotide*
Model/Method →	*Tamura-Nei model*
RATES AND PATTERNS	
Rates among Sites →	*Uniform Rates*
No of Discrete Gamma Categories →	Not Applicable
DATA SUBSET TO USE	
Gaps/Missing Data Treatment →	*Use all sites*
Site Coverage Cutoff (%) →	Not Applicable
TREE INFERENCE OPTIONS	
ML Heuristic Method →	*Nearest-Neighbor-Interchange (NNI)*
Initial Tree for ML →	*Make initial tree automatically (Default - NJ/BioNJ)*
Initial Tree File →	Not Applicable
Branch Swap Filter →	*None*
SYSTEM RESOURCE USAGE	
Number of Threads →	*7*

Bootstrapping

Evolutionary / substitution model

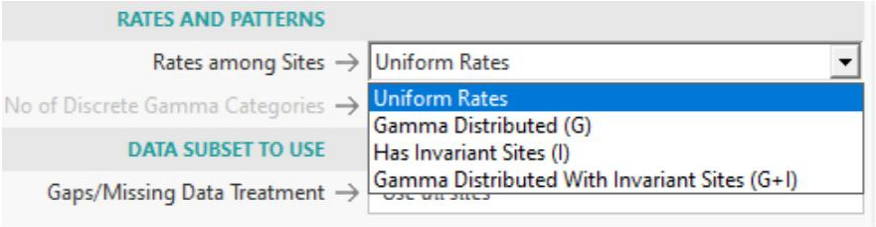
Site-specific evolutionary rate

How to handle gap (indel) and missing data

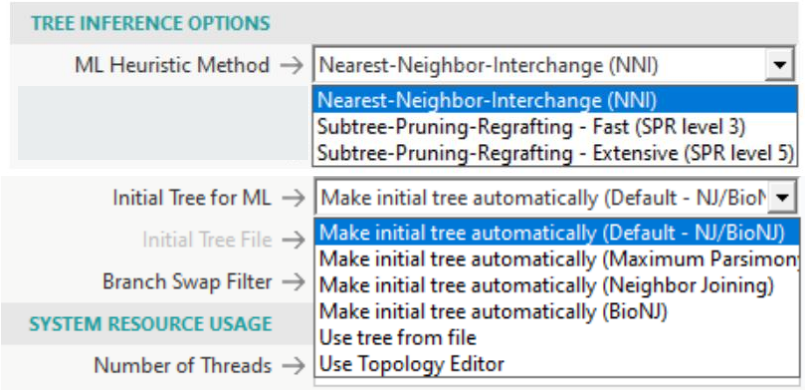
Detailed tree building method

Parallel processing

Maximum likelihood parameters



Invariant site and Gamma-distributed site-specific substitution rates



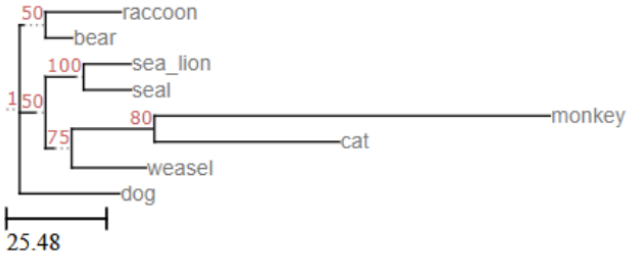
Tree search method

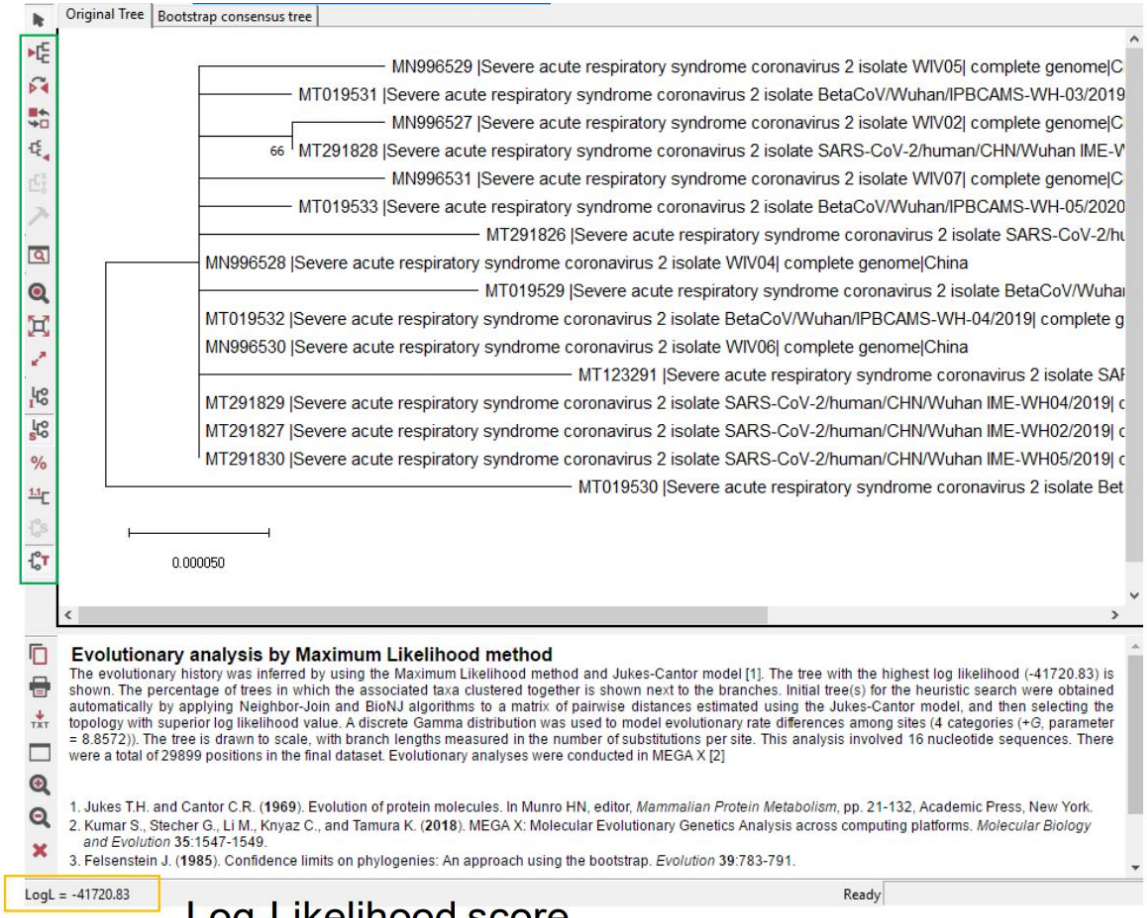
Initial tree can be built using quick and simple method like neighbor joining

Output

NEWICK format

((raccoon:19.19959,bear:6.80041)50:0.84600,((sea_lion:11.99700,seal:12.00300)100:7.52973,((monkey:100.85930,cat:47.14069)80:20.59201,weasel:18.87953)75:2.09460)50:3.87382,dog:25.46154);





Log-Likelihood score

Other useful tools

- RAxML, PAUP are standard phylogenetic reconstruction tools
- PAML specializes in natural selection analysis (dN/dS)
- FastTree specializes in speed for large dataset
- MrBayes, BEAST use Bayesian approach
  - P(tree topology, branch lengths | substitution model, sequence data)
- More: https://evolution.genetics.washington.edu/phylip/software.html#methods

Any question?

- See you next time