
3000788 Intro to Comp Molec Biol

Week 3: Sequence alignment and phylogenetics

Fall 2024



Sira Sriswasdi, PhD

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

Part I: Sequence alignment

- Similarity in sequence implies similarity in other characteristics
- Alignment is the foundation of many bioinformatics techniques
 - How to make it fast? How to score a hit?
- Protein alignment vs nucleotide alignment
 - How to define similarity between amino acids?



Sequence homology

Evolution occurs at the sequence level

Histone H1 (residues 120-180)

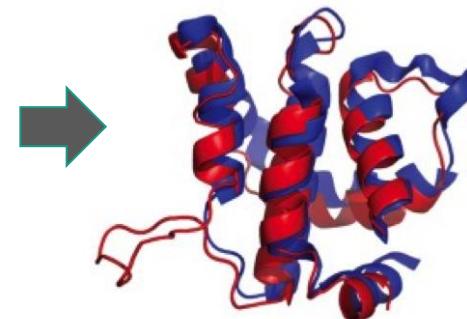
HUMAN	KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPCTVKAKPVKASKPKKAKPVK
MOUSE	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPCKVVKPVKASKPKKAKTVK
RAT	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPCKIVKVKPVKASKPKKAKPVK
COW	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKPKTVKAKPVKASKPKKTPVK
CHIMP	KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPCTVKAKPVKASKPKKAKPVK
	**** : ***** : ***** : ***** : ***** : * . ***** : * **

[https://en.wikipedia.org/wiki/Homology_\(biology\)](https://en.wikipedia.org/wiki/Homology_(biology))

- Genes / proteins originating from the same ancestor will have similar sequence
- High sequence similarity → functional similarity, structural similarity, etc.

Inference using sequence similarity

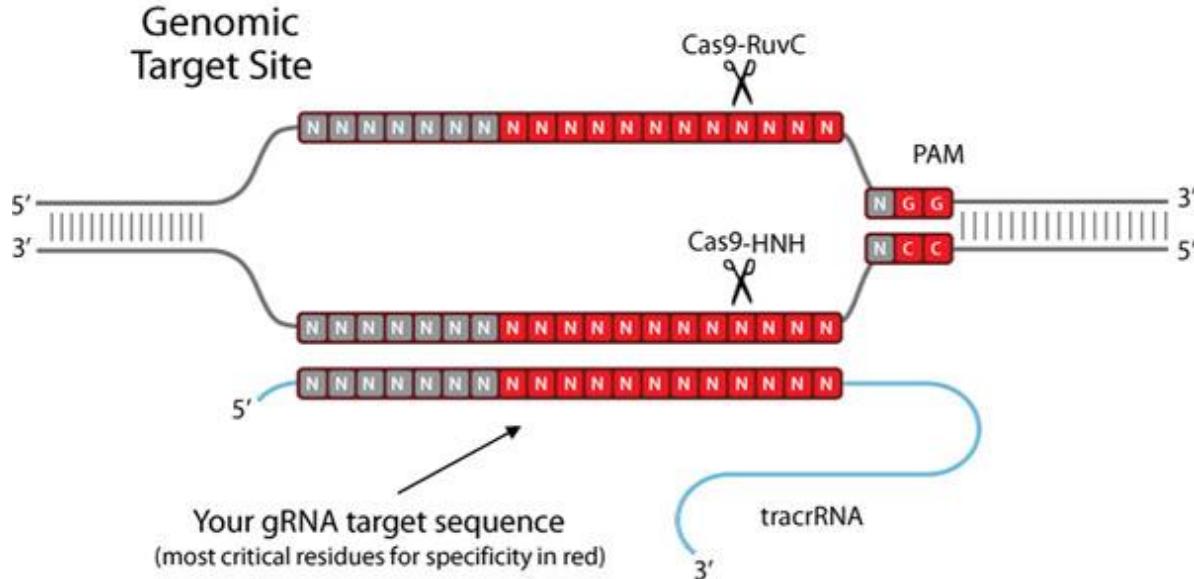
		α_1		α_2		α_3
N1	1	MRTLLIRYI	LWRND	NDQTQQNDDF	KKLM	LLDELVDDGDVCT
N2	53	IIAILNRFL	TMNKDELNNTQCHII	KEFMTYE	QMAIDHYGEYVNAILYQIR	
		α_4			α_5	
N1	51	SDGPLLDRLN	-----	QPVNNIEDAKRMI	AISAKVARDIGERSE	
N2	103	KRPNQHHT	IDLFKKIKRTPYDTFK	VDPVEFV	KKVIGFVSI	LNKYKPVSY
		α_6		α_7		
N1	90	IRWEESFTIL	FRMIETY	FDDLMIDLYG		
N2	153	VLYENVLYDEF	FKCKINY	VETKYF	---	



Ferguson et al. J General Virology, 94: 2070-2081 (2013)

- Same amino acid residue positions are involved in similar secondary structure
- Properties of amino acid side chains are important

Molecular probe design



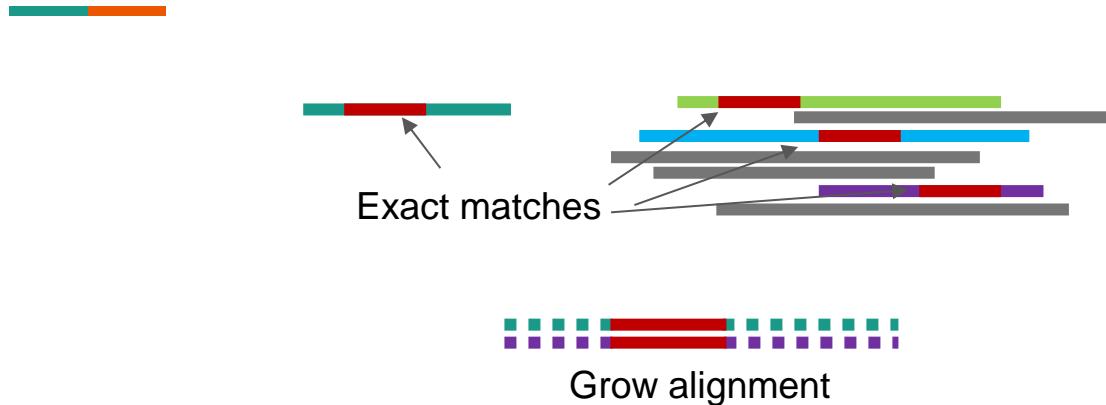
<http://www.sigmaaldrich.com/technical-documents/articles/biology/crispr-cas9-genome-editing.html>

- Sequence alignment can check the specificity of your probes



Components of sequence alignment

Starting from exact match (seed / word)



- Input sequence length = 300
- Expected similarity between input and reference = 95% (genome re-sequencing)
- Expected 15 mismatches
- If mismatches are random, there should be a run of $285/16 \sim 18$ positions with matches
 - MM...MEM...MEM.....MEM...MM
 - NCBI's **MEGABLAST** searches for a run of 28 matches

Alignment scores

The image shows two separate windows for 'Scoring Parameters'.

Top Window:

Match/Mismatch Scores	1,-2
Gap Costs	Linear: 1

Bottom Window:

Match/Mismatch Scores	2,-3
Gap Costs	Existence: 5 Extension: 2

Ref: ACCGTATCG
|| |
Query: AC— ATCG

$$\begin{aligned} \text{Score} &= +1 + 1 - 1 - 1 + 1 + 1 + 1 + 1 \\ &= +3 \end{aligned}$$

$$\begin{aligned} \text{Score} &= +2 + 2 - 5 - 2 - 2 + 2 + 2 + 2 + 2 \\ &= +1 \end{aligned}$$

- Gap cost models
 - Constant = Same penalty regardless of length
 - Linear = Penalty x Length
 - Affine = Existence + (Extension x Length)

Alignment score interpretation

- **Match / Mismatch = +1 / -2**
 - To permit a mismatch, there must be >2 matches afterward to gain score
 - Want hits with high identity
- **Match / Mismatch = +2 / -3**
 - A mismatch followed by two matches = net +1 score
 - Want hits with intermediate identity
- **Gap cost**
 - **Constant** = An insertion/deletion can be of any length
 - **Linear** = Long indel is less likely than short indel
 - **Affine** = **Existence** + (**Extension** x Length)
 - Balance between constant and linear



Global and local alignment

Global vs local alignment

Local Alignment

Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTAGAGGGCTTGCAACCA 3'
||||| ||||| ||||| ||||| |||||

Query Sequence

5' TACTCACGGATGAGGTACTTAGAGGC 3'

Global Alignment

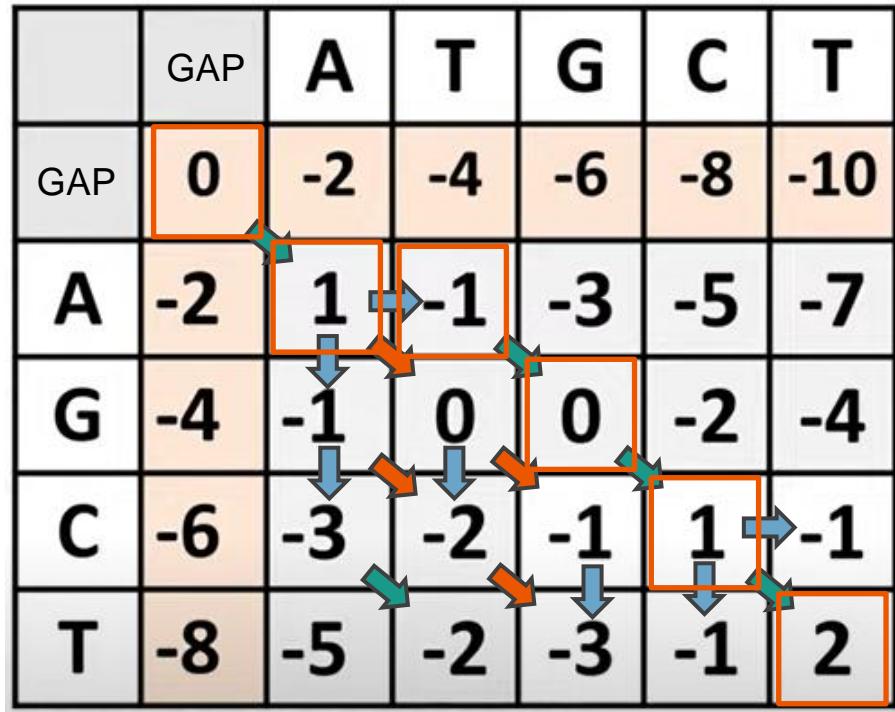
Target Sequence

5' ACTACTAGATTACTTACGGATCAGGTACTTAGAGGGCTTGCAACCA 3'
||||| ||||| ||||| ||||| ||||| |||||

5' ACTACTAGATT----ACGGATC--GTACTTAGAGGCAGCAACCA 3'

Query Sequence

Global alignment



Match : 1
Mismatch : -1
GAP : -2

Seq1 : ATGCT

| |||

Seq2 : A-GCT

Local alignment



		A	T	G	C	T
	0	0	0	0	0	0
A	0	1	0	0	0	0
G	0	0	0	1	0	0
C	0	0	0	0	2	0
T	0	0	0	0	0	3

Match : 1 
Mismatch : -1 
GAP : -2 

Seq1 : ~~ATGCT~~
 |||

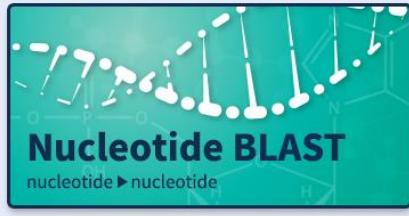
Seq2 : AGCT

- Ignore possibilities with negative score
 - Start over is better



Basic Local Alignment Search Tool

BLAST



NCBI's nucleotide BLAST interface

Enter Query Sequence BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

```
CACCATCACAAACAAAGGAACCTGGAACTGTCTATGAGGTCACTGGGTCAAGAACCCAAACAGAAAGCTGAATTGCAGGAT  
ATGATCAATGAAGTGGATGCTGATGGTAAGAGCTTAAACCATGAATGAGGCCATTGTTGTAAATTCAAGTTC  
AGACATGTTACAGGATTGCTTTCAGGCCAGAGCAAAGCAAATGTGCAAAGATCCTTCTGTGGTTGCCAG  
GGCATTGACAA
```


From

To

Or, upload file No file chosen [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
 RefSeq Representative genomes ([refseq_representative_genomes](#)) [?](#)

Organism **Optional** Exclude [+](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Nucleotide BLAST algorithms



Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.

Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.

BlastN is slow, but allows a word-size down to seven bases.

- **MEGABLAST**: word size = 28, match/mismatch score = +1/-2, linear gap
- **BLASTN**: word size = 11, match/mismatch score = +2/-3, affine gap

Interpreting BLAST result

Job title: Nucleotide Sequence (240 letters)

RID [VCC9FM9501R](#) (Expires on 09-12 14:46 pm)

Query ID [Idl/Query_58243](#)

Description None

Molecule type nucleic acid

Query Length 240

Database Name refseq_representative_genomes (GPIPE/9606/108/ref_top_level)

Description [See details](#)

Program BLASTN 2.7.0+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [MSA viewer](#)

Graphic Summary

Distribution of the top 2 Blast Hits on 2 subject sequences [?](#)
Mouse over to see the title, click to show alignments

Color key for alignment scores

<40	40-50	50-80	80-200	>=200		
■	■	■	■	■		
1	40	80	120	160	200	240

Query

Descriptions

Sequences producing significant alignments:

Select: All None Selected: 0

All Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Homo sapiens chromosome 14, GRCh38.p7 Primary Assembly	444	444	100%	1e-122	100%	NC_000014.9
<input type="checkbox"/>	Homo sapiens chromosome X, GRCh38.p7 Primary Assembly	91.6	91.6	43%	2e-16	84%	NC_000023.11

Query coverage = % of input sequence used in the alignment

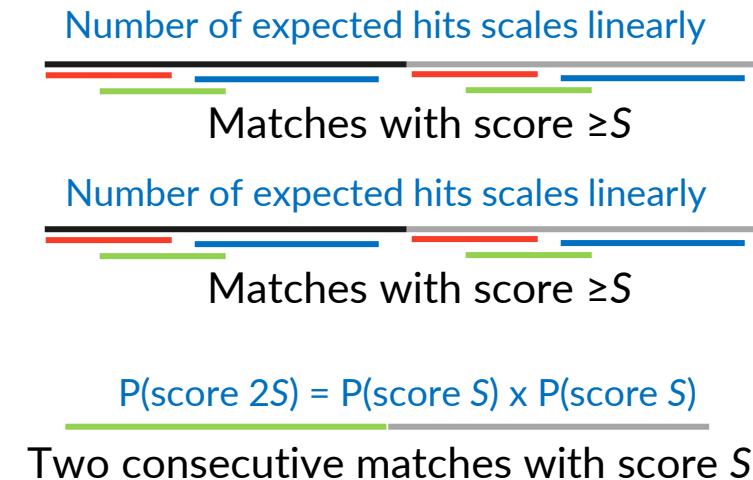
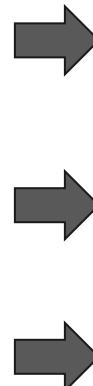
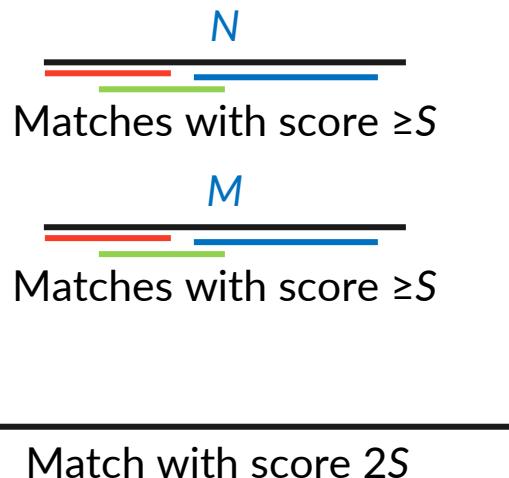
Identity = % of identity between input and matched sequences **in the aligned region**

E value = expected number of hits with the same or higher score by chance (given input length and database size)

Typical cutoff is 1e-5

Understanding E value

- Given an input sequence of length N and a reference sequence of length M
- E value for a hit with score S is proportional to $N \times M \times e^{-\lambda S}$



E value as Poisson distribution



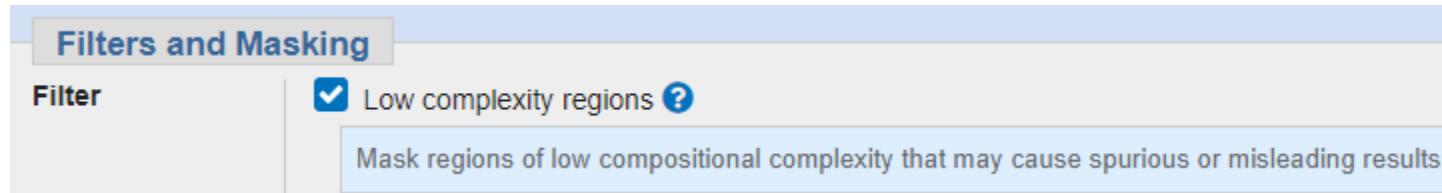
- Event of interest = hits with score $>S$ occurs on the sequence of length N
- Expected value = E value
- Probability of observing k hits with score $>S$ = $\frac{E^k e^{-E}}{k!}$

Low complexity region



CG island

CCCGCGCGCCCCGGCGCCCGATGCAACTAGC



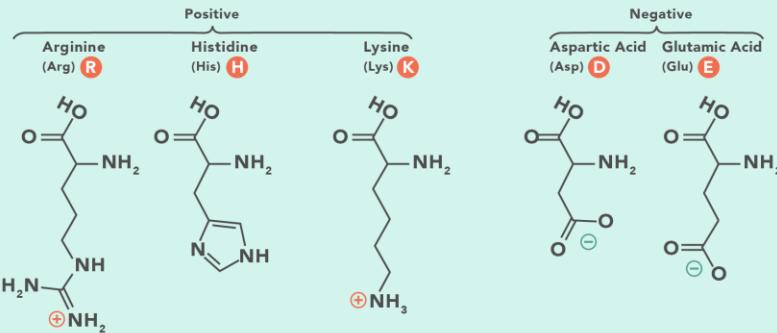
- Probability of getting a hit with score $>S$ will be high if both sequences contain only C's and G's
- BLAST withholds these regions from score calculation



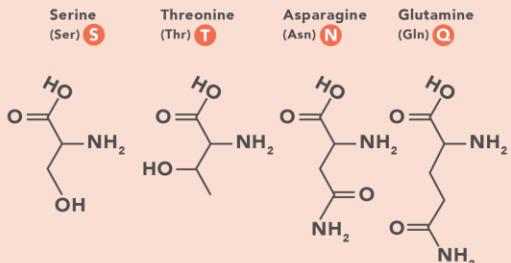
Protein sequence alignment

Amino acid side chains

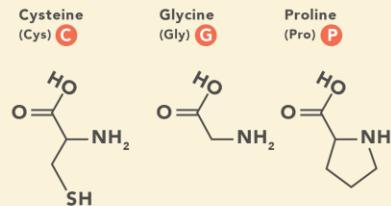
A. Amino Acids with Electrically Charged Side Chains



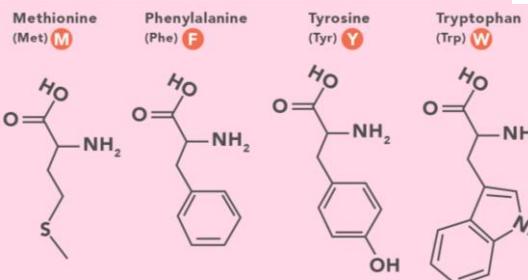
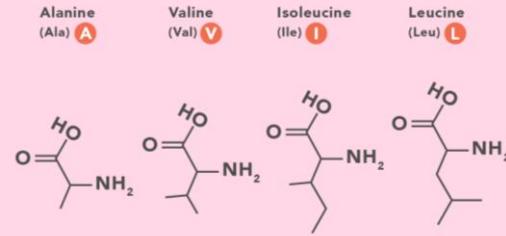
B. Amino Acids with Polar Uncharged Side Chains



C. Special Cases



D. Amino Acids with Hydrophobic Side Chains

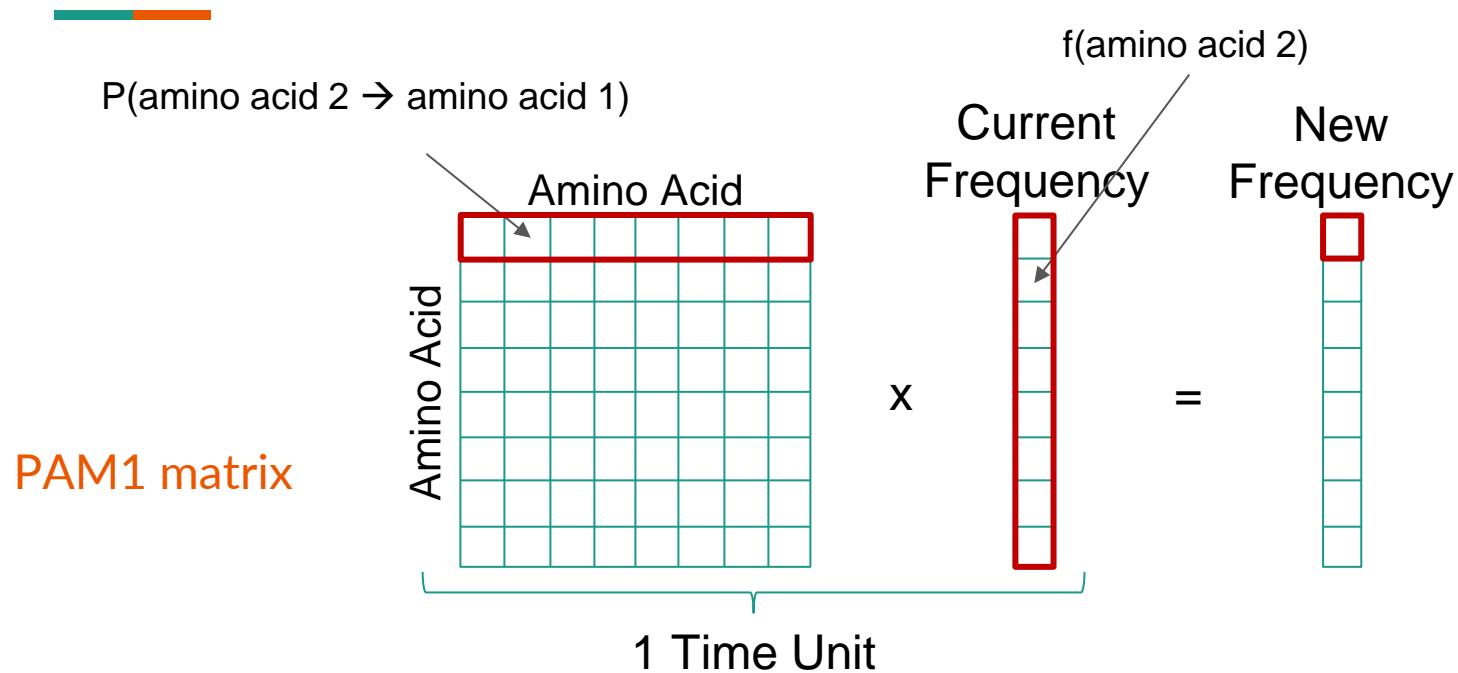


Block Substitution Matrix (BLOSUM)

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

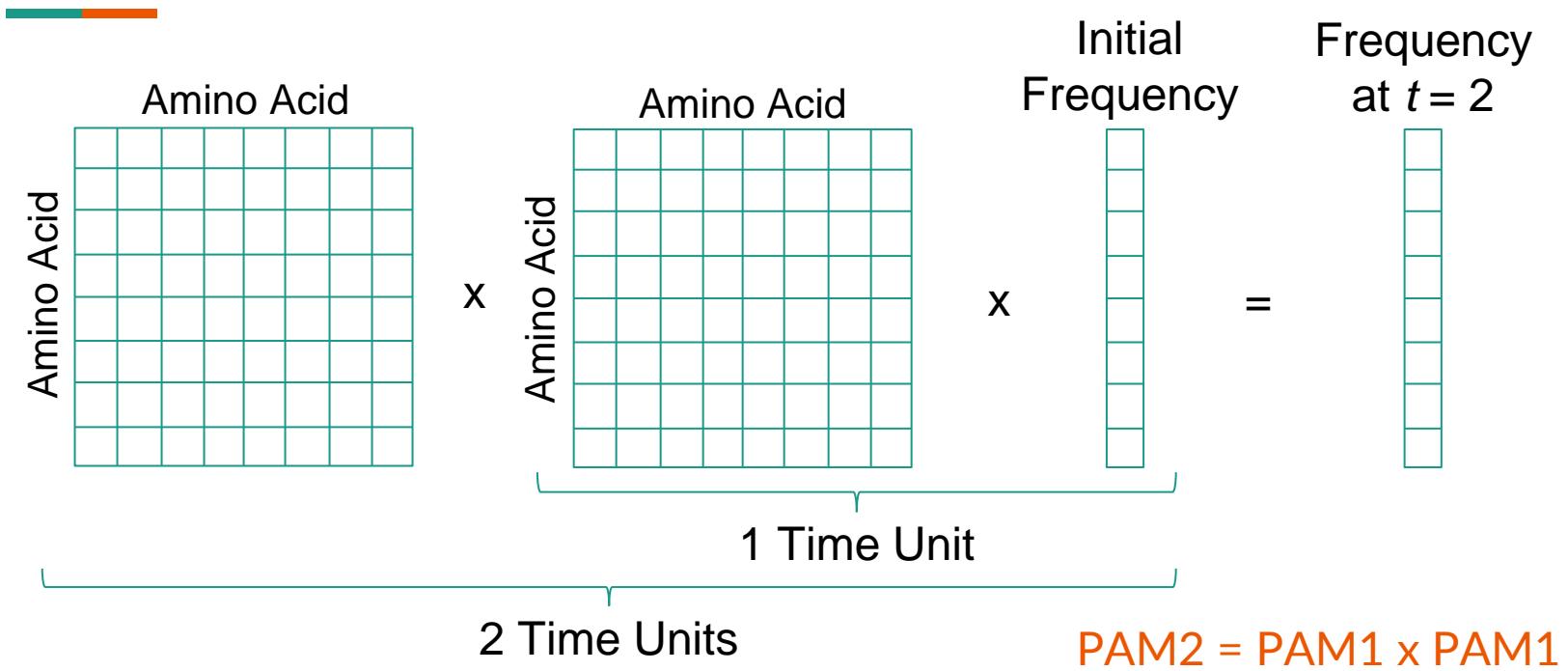
- Constructed using substitution rates from actual protein families
- BLOSUM45 was constructed using protein families with >45% conservation

Point Accepted Mutation (PAM)



- Estimate amino acid substitution rate between highly similar proteins (>85%)

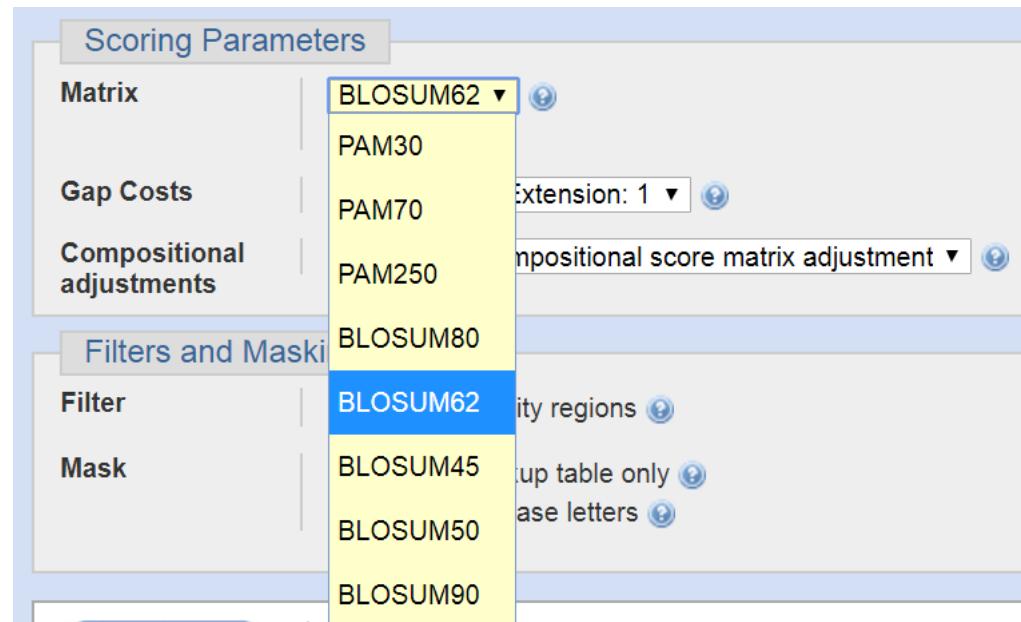
Point Accepted Mutation (PAM)



- Extrapolate substitution rates for more distant proteins

PAM vs BLOSUM

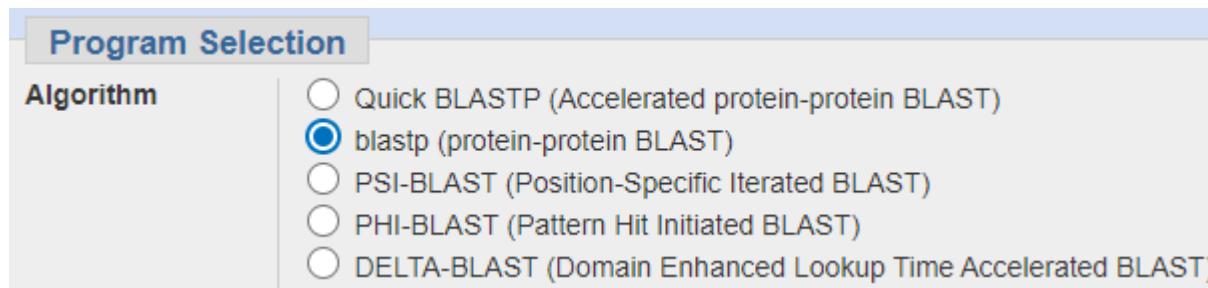
PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45



Data from <https://en.wikipedia.org/wiki/BLOSUM>

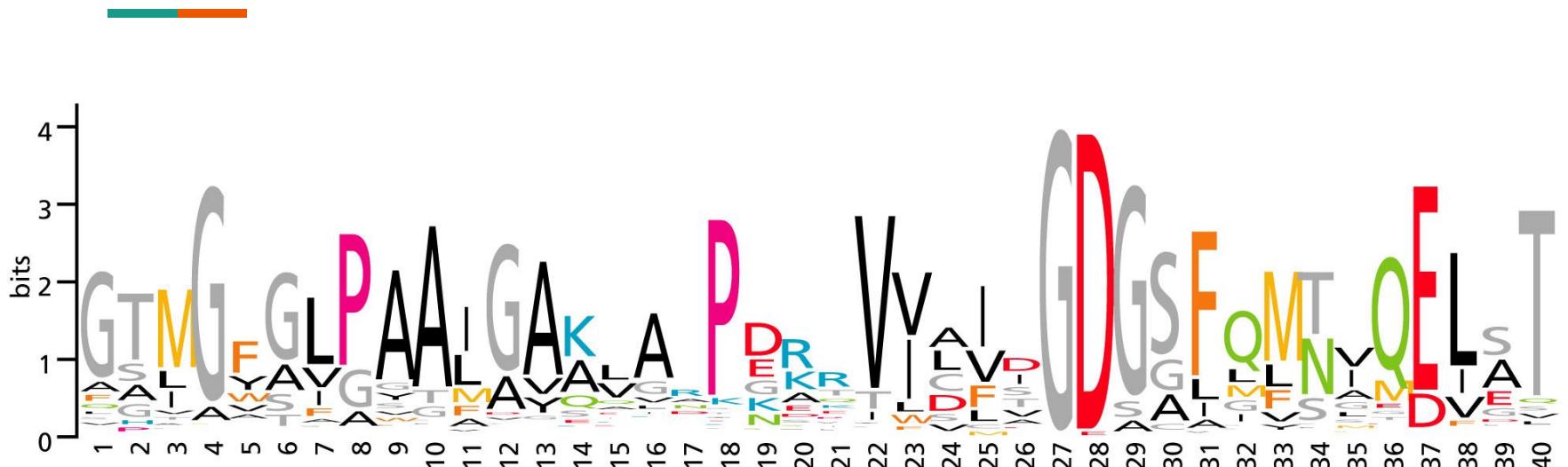
- BLOSUM for low identity, PAM for high identity

Protein BLAST algorithms



- Standard BLASTP assumes that all amino acid residue positions are the same
- But there are protein domains & motifs with specific patterns

Position-specific scoring matrix (PSSM)



www.nemates.org/uky/520/Lecture/Lect6/BIO520_2010_Lect6.pp

weblogo.berkeley.edu

- Different scoring matrix for each position in the motif
- But how do we know the position-specific amino acid profile?

Pattern hit initiated (PHI-BLAST)



x = any amino acid

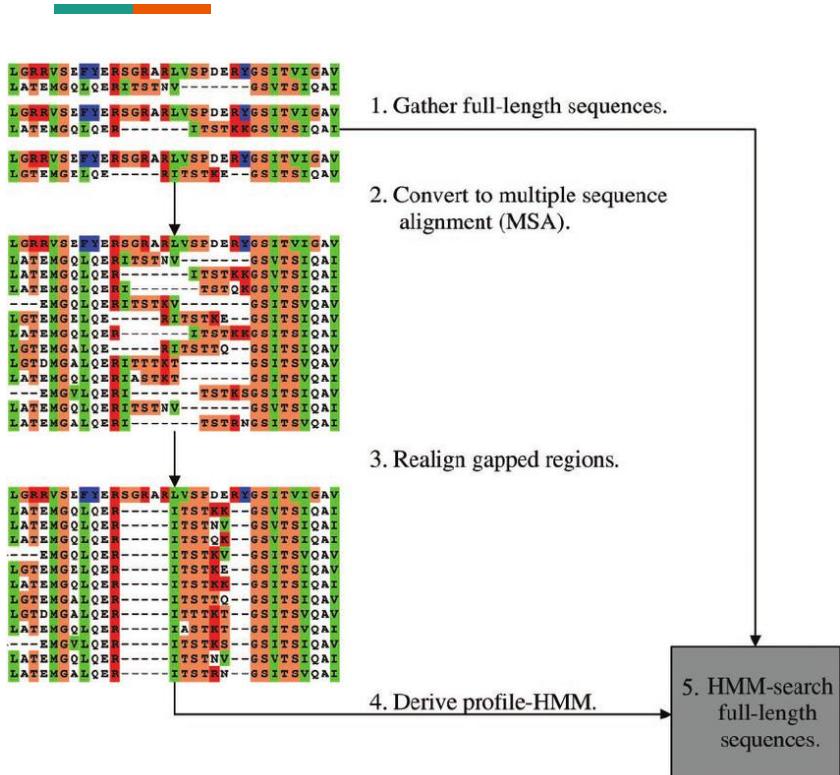
[LIVMF] -G-E-x- [GAS] - [LIVM] -x (5,11) -R- [STAQ]

L, I, V, M, or F

any sequences of 5-11 amino acids

- Combine regular BLASTP with user-specified pattern
- Hits must be similar to the input sequence AND match the pattern
- Search for known protein domain

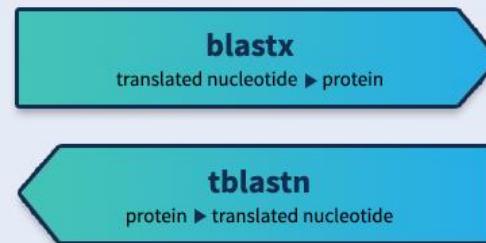
Position-specific iteratred (PSI-BLAST)





More techniques

BLASTX and TBLASTN



- For alignment of coding DNA sequence
 - Codon structure = not all nucleotide positions evolve in the same manner
 - Similarity in protein is more informative than similarity in DNA
- Align translated DNA to protein database
- Align protein to translated DNA database

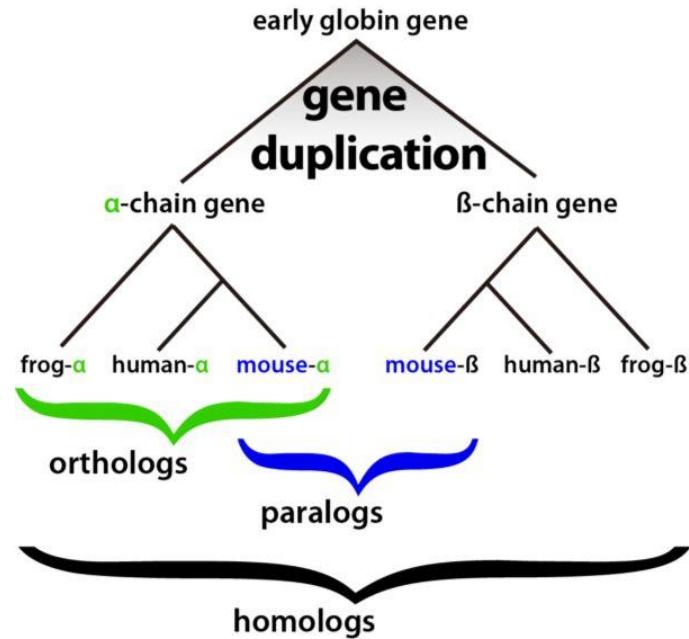
Example use cases



- BLASTX = align translated DNA to protein database
 - You perform RNA-seq
 - Unsure which open reading frame is correct
 - Check whether this RNA translated to known protein or function
- TBLASTN = align protein to translated DNA database
 - You identified novel protein
 - No evidence in protein database
 - But there might be transcriptomics studies that identified the RNA of related proteins

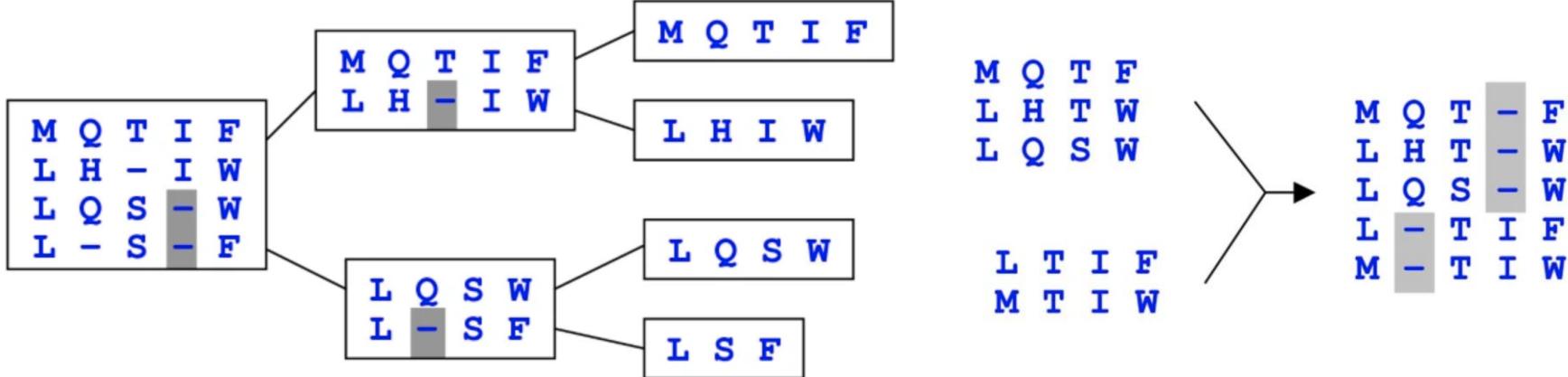
All-vs-all BLAST

- Compare genes between related species to identify genes originated from a common ancestor
 - {Mouse-a, Human-a}, {Mouse-b, Human-b}
- BLAST mouse to human
- BLAST human to mouse
- **Reciprocal best hit:**
 - Human-a should be the best hit for Mouse-a
 - Mouse-a should be the best hit for Human-a



<https://sites.google.com/site/jkim339n/part2a>

Multiple sequence alignment (MSA)



- Dynamic programming is not feasible because of too many possibilities for grouping sequences
- Rely on **heuristic** algorithm

Alignment output format

Aligned FASTA

```
>TRY2_RAT/24-239
-----IVGGYTCQENSVPYQVSLNSGY-----HFC
GGSLI-----NDQ-NV-VSAAHCYKS-----RIQVRLGE-HNINVLEGN-----
-----EQFVNAAKIIKHPNFDRKT-L-----NNDIMLKLSP
-----VKNARVATVALPS---SCA---PAGTQLCISGWGN-----TLSSGV-----
-----NEPDLLQ-CLDAP-LLPQADCEAS---YPGK-----ITONMVCVGFL-----
-EGG-KDSCQGDGGPVVCNGE-----LQGIVSWG-YGCALPDN---PGVYTKVCNY
VDWI-----
```

```
>Q16LB2_AEDAE/136-374
-----ILNGIEADLEDFPYL GALALLDNYT-----STVSYRC
GANLI-----SDR-FM-LTAHCLFG-----KQAIHVRMGTLSLTDNPDED-----
-----APVIIGVERVFFHRYTRRPIT-----RNDALEIKLN
RT-----VVEDFLIPVCLYT-----EQNDP-LPTVPLTIAGWGG-----NDSAS-----
-----LMSSSLM-KASVT-TYERDECNSL---LAKKI-----VRLSNQLCALGRSEF
NDGLRNNDTCVGDSGGPLELSIGR---RKYIVGLTSTG-IVCGNE-F---PSIYTRISQF
IDWT-----
```

PHYLIP

5 42

Turkey	AAGCTNGGGC ATTCAGGGT GAGCCCGGGC AATACAGGGT AT
Salmo	gairAAGCCTTGGC AGTCAGGGT GAGCCGTGGC CGGGCACGGT AT
H. Sapiens	ACGGTTGGC CGTTCAAGGGT ACAGGTTGGC CGTTCAAGGGT AA
Chimp	AAACCCTTGC CGTTACGCTT AAACCGAGGC CGGGACACTC AT
Gorilla	AAACCCTTGC CGGTACGCTT AAACCATTGC CGGTACGCTT AA

ClustalW

Caballeronia_arvi	MNSRIDSHVKHLIFFCGHAGTGKTTLAKRLFAPLMQAAGEPFCLLDKDCTLGAYSAAMG
Caballeronia_choica	-----MTHLVFFCGHAGTGKTTLAKRLFPRLMRATGEPFCLLDKDCTLGAYSAAMG
Caballeronia_arationis	-----MTYLIFFCGHAGTGKTTLAKRLFPRLVRATGEPFCLLDKDCTLGAYSAAMG
Caballeronia_telluris	-----MTHLIFFCGHAGTGKTTLAKRLFPRLAQASGEPFCLLDKDCTLGAYSAAMN :::*****. * .*****.*****.
Caballeronia_arvi	ALTGDPHDRSPLFIEHFRDPEYRCLVDTAAENLALGVSVVVVAPLTREVRSRLFDRAW
Caballeronia_choica	ALTGDPNDRDPLFLQHLDPEYRALIDTARENLELGVSVAVAPLSREVRDGRLFDRQW
Caballeronia_arationis	ALTGDPNDRDPLFLQHFRDPEYRALIDTARENLDLGVSVAVAPLTREVREERLFDRAW
Caballeronia_telluris	ALTGDPNDRDPLFLQHLDPEYRALIDTARENLDLGVSVAVAPLTREVREGRLFDRTW *****:*****:*****.*** *** *** *****.*****.***** *

Any question?



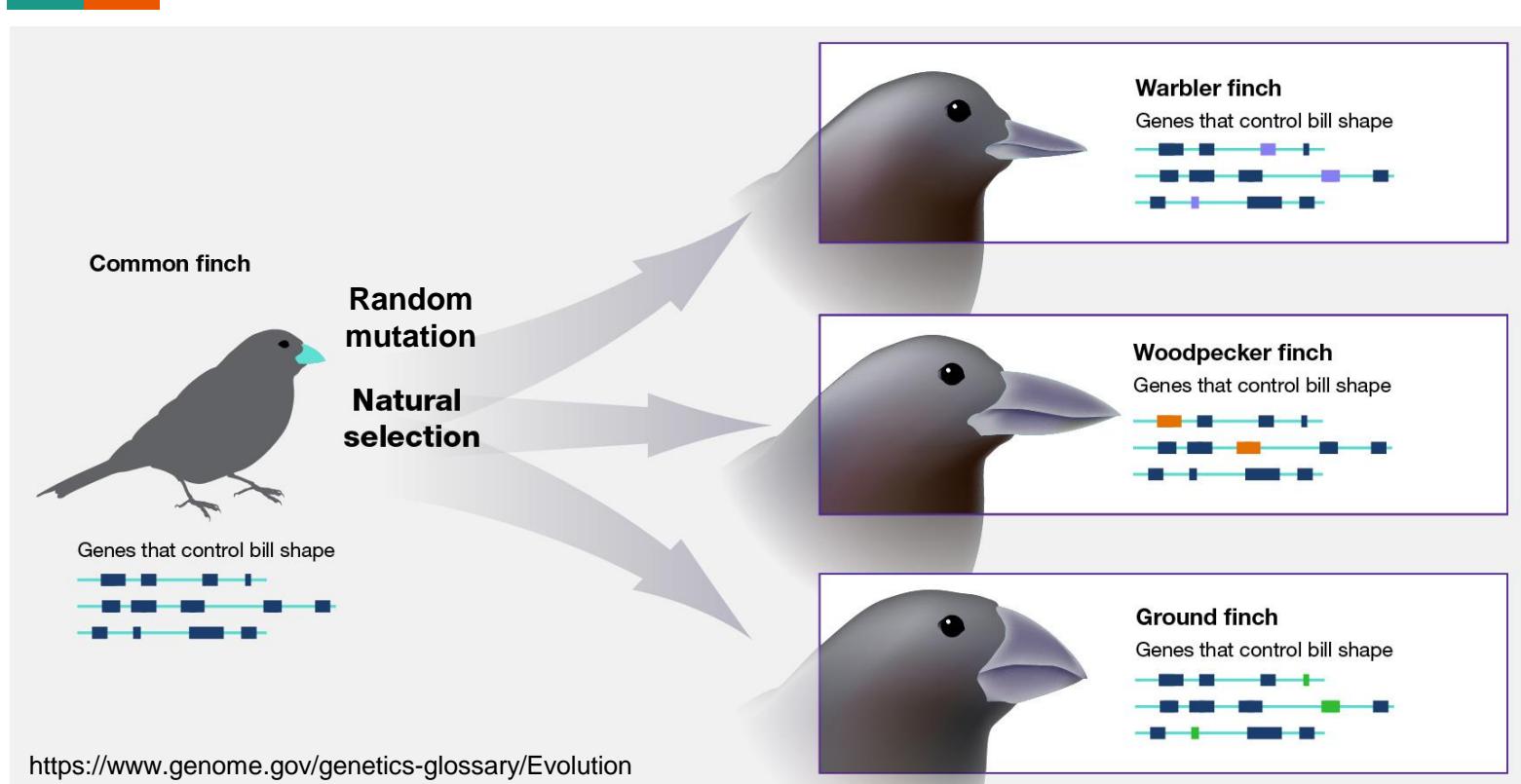
Part II: Phylogenetics

- Evolution = nature's experimentation = source of information

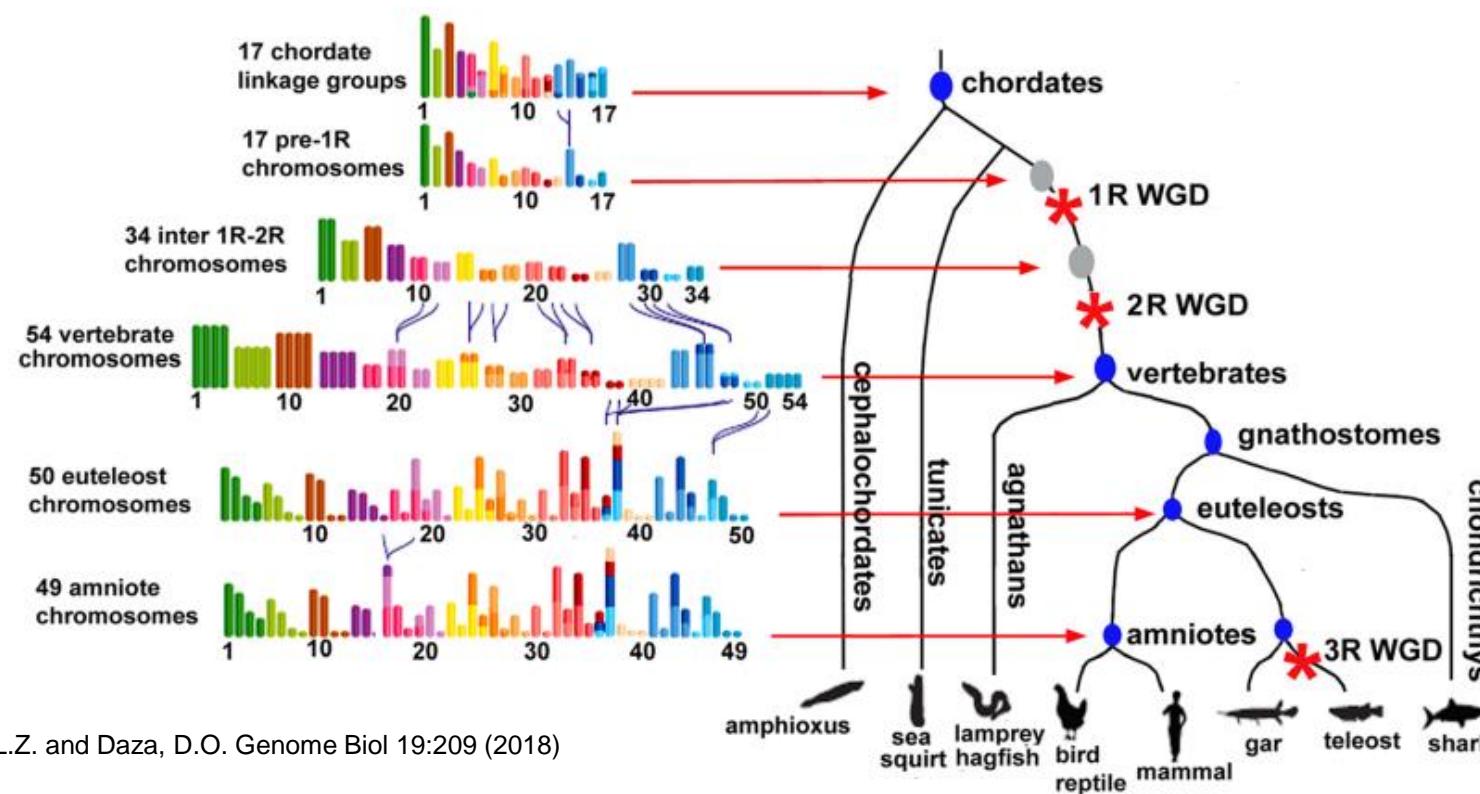


Evolution

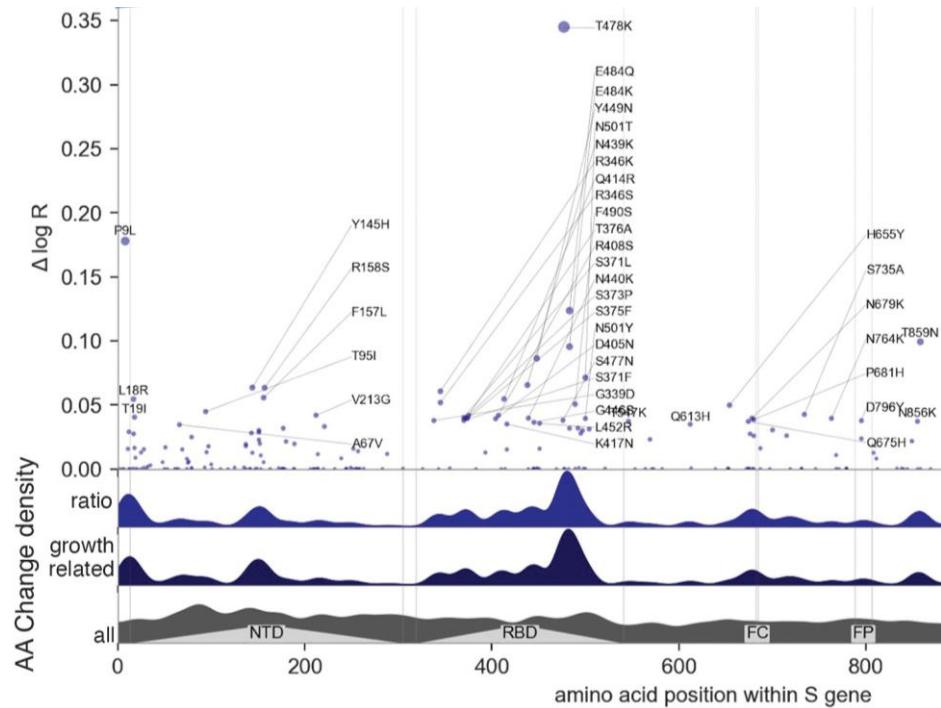
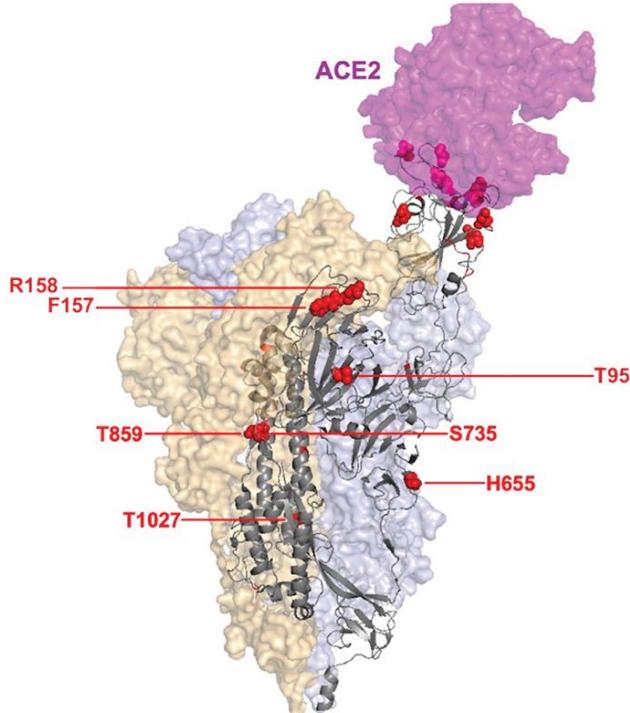
Random mutation with natural selection



History of whole-genome duplications in animal



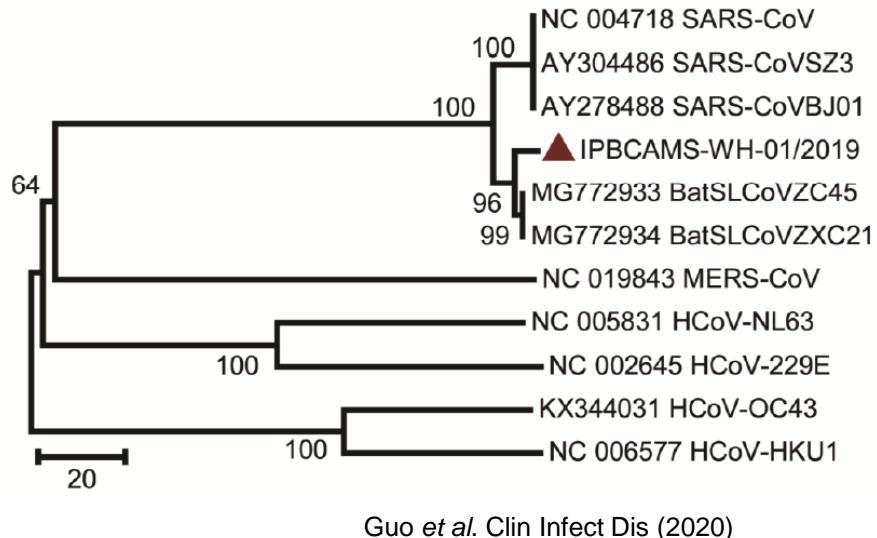
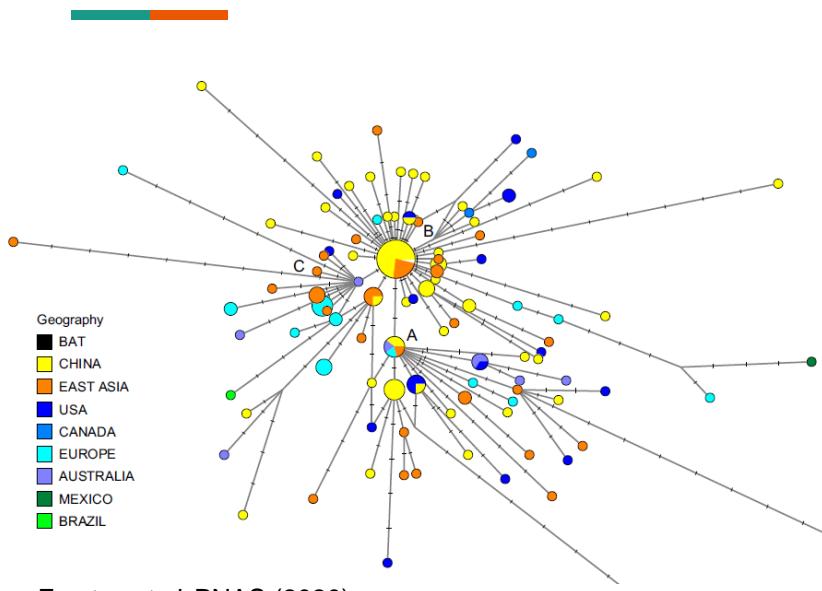
Linking evolution to function





Phylogenetics

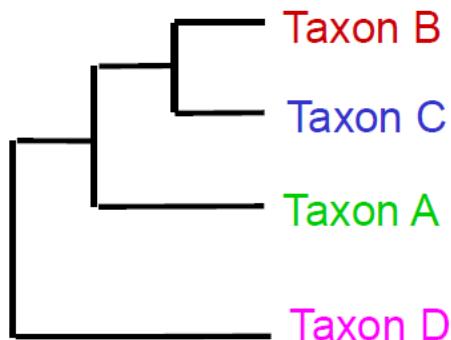
Phylogeny



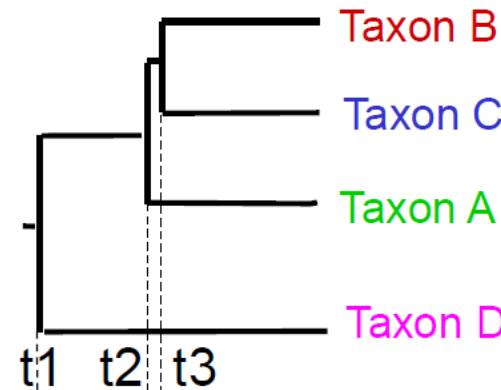
- Clustering of similar taxa with respect to evolutionary distances
- Branch length reflect **time** and **mutation rate**

Types of phylogeny

Cladogram



Chronogram



Phylogram

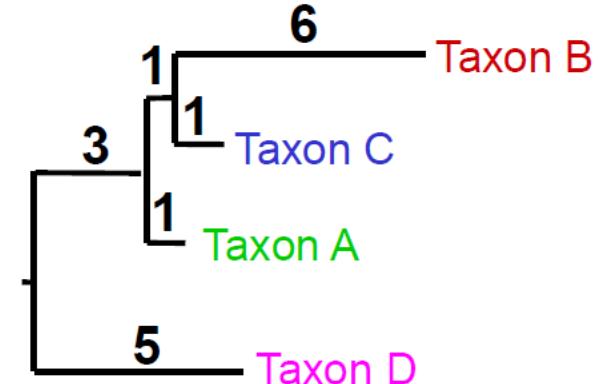


Image from Lecture 18 6.047 MIT OCW

- The choice depends on research question

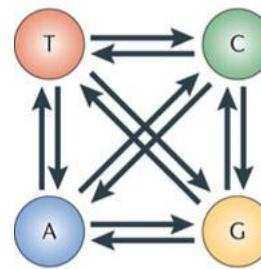
Ingredients for phylogenetic reconstruction

Multiple sequence alignment

Scarites	C T T A G A T C G T A C C A A - - A T T T T A C
Carenum	C T T A G A T C G T A C C A A - - T A C - T T T T A C
Pasimachus	A T T A G A T C G T A C C A C T - T A G T T T T A C
Pheropsophus	C T T A G A T C G T T C C C C - - - A C A T T T A C
Brachinus armiger	A T T A G A T C G T A C C A C - - - A T A T T T T C
Brachinus hirsutus	A T T A G A T C G T A C C A C - - - A T A T T T T C
Aptinus	C T T A G A T C G T A C C A C - - - A C A T T T A C
Pseudomorpha	C T T A G A T C G T A C C A C - - - A C A T T T A C

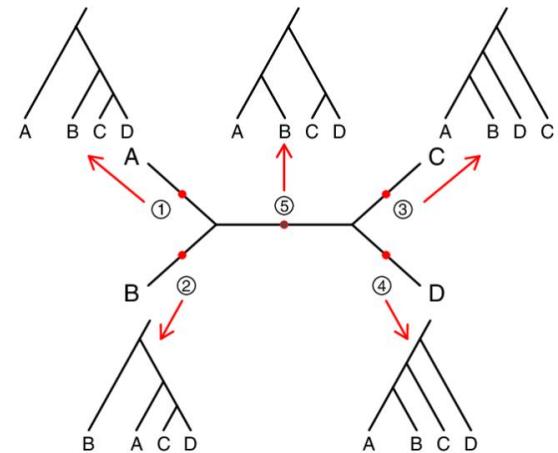
Image from www.mcqbiology.com

Evolutionary model



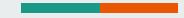
Yang & Rannala. Nat Rev Genetics (2012)

Tree building algorithm



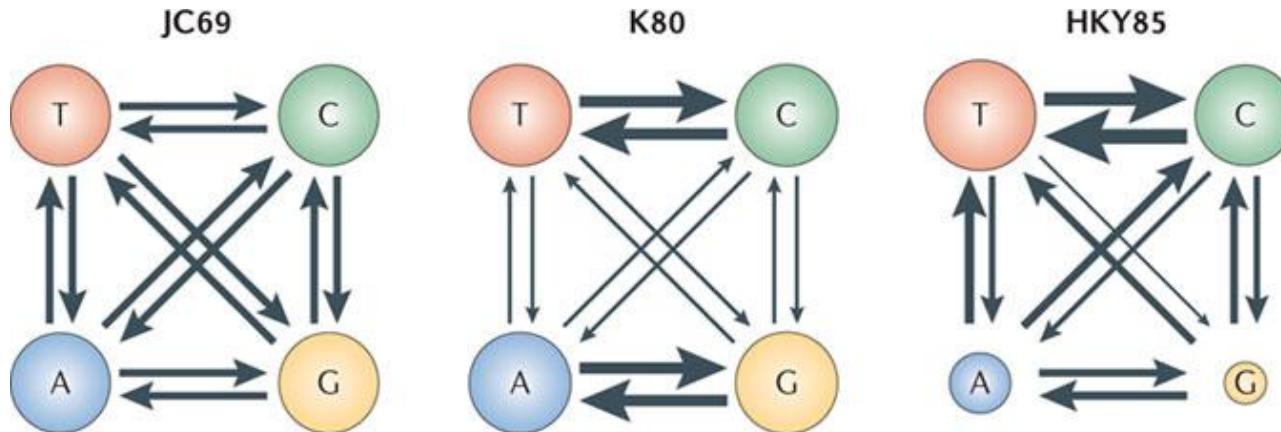
Tian, Y. and Kubatko, L.S. BMC Evol Biol 17(1) (2017)

- Sequence data + substitution model + tree topology and branch length
- Other constraints and non-molecular data can also aid the reconstruction



Evolutionary (substitution) model

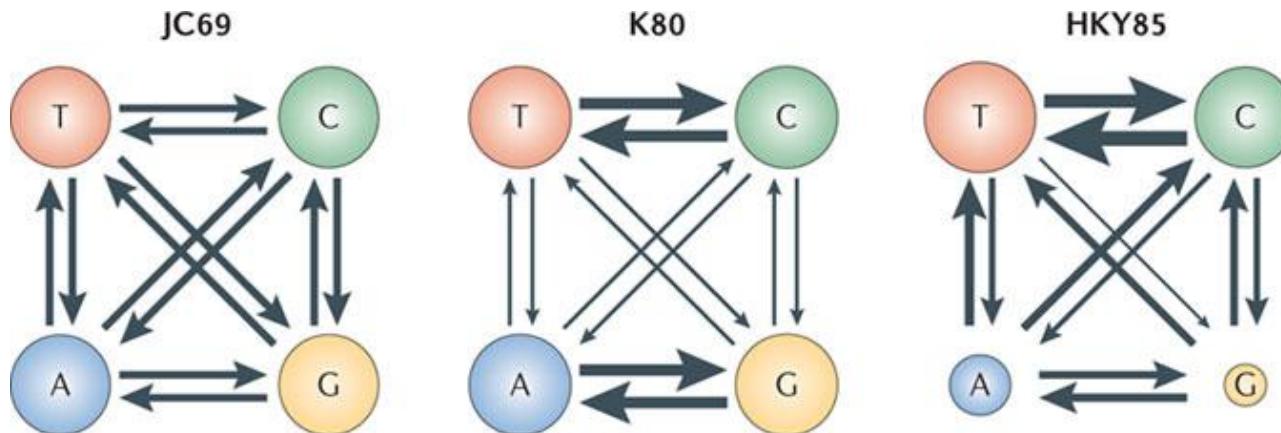
Basic nucleotide substitution models



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?



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

- 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: $P(A \rightarrow G) = P(G \rightarrow A)$
 - 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 base frequencies
 - 3 parameters
- This is the most generalized time-reversible model possible
- Often used

Basic 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	2	11	7
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	1	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Codon alignment: amino acid → nucleotide

A. DNA alignment

Q9FPK4	ATGGGTGTTTCAGCTACGAGATGAGGCCACC	TCCGTTATCCTCCGGCTA	GGCTGTTCAAGTCC	TTTGTCTTAGATGCCGAC	AACCTCATT
Q9FPK3	ATGGGTGTTTCCTGCTACGAGATGAGGCCACC	TCCGTTATCCTCCGGCTA	GGCTGTTCAAGTCC	TTTGTCTTAGATGCCGAC	AACCTCATT
Q945E7	ATGGGTGTTGTGAGTTATGAGTTTGAGGTAACC	TCCCGAAATTGCTCCAGCCA	GGCTTTCAAGGCT	TTTGTCTTGAGGCTGCC	AAGATTGG
Q6XC94	ATGGGTGTTGCCAGTTATGAGTTTGAGGTAACC	TCCCGAAATTGCTCCAGCCA	GGCTTTCAAGGCT	TTTGTCTTGAGGCTGCC	AAGATTGG
Q6Q4B5	ATGGGTGTTGTGAGTTATGACTTGAGGTAACC	CTCCCCAATTGCTCCAGCCA	TCAAGGCTTTGTTCTGAACCTGCC	AAAGGTTTGG	
Q43549	ATGGGTGTTTCATTACGAAACTGAGTTTACCC	TCCGTCATTCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AACCTCATC
Q4VPJ1	ATGGGTGTTTCACATACGAACTCTGAGTCCACC	TCCGTCATCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AACCTCATC
Q84LA7	ATGGGTGTCCTCACATACGAACTCCGAATTGACC	TCTATCATCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AACCTCATC
Q4VPI3	ATGGGTGTTTCACATACGAACTCCGAATTGACC	TCTATCATCCCCCTGCTA	GGTTGTTCAATGCC	TTTGTCTTGATGCTGAC	AACCTCATC

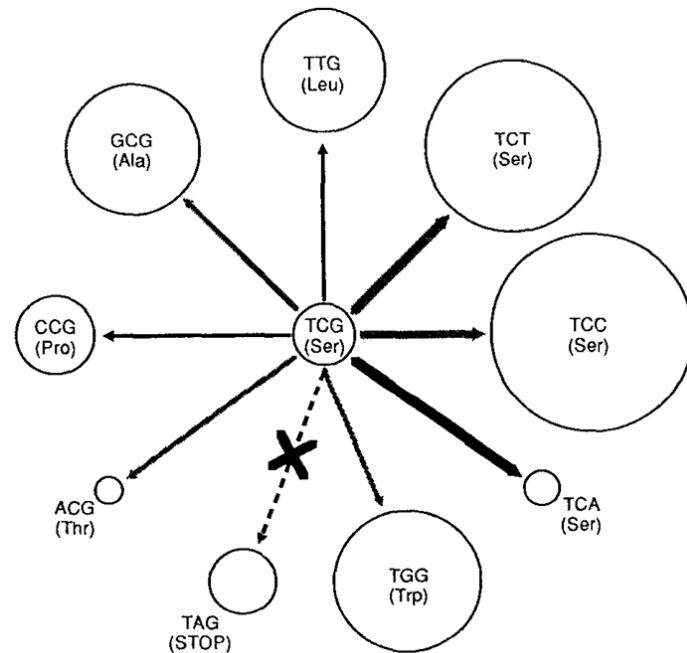
B. Back-translation from protein alignment

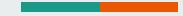
Q9FPK4	ATGGGTGTTTCAGCTACCACGATGAGGCCACC	TCCGTTATCCTCCGGCTAGGCTGTT	--AAGCTCTTGTCCTAGATGCCGACAACCTCATT	
Q9FPK3	ATGGGTGTTTCCTGCTACCACGATGAGGCCACC	TCCGTTATCCTCCGGCTAGGCTGTT	--AAGCTCTTGTCCTAGATGCCGACAACCTCATT	
Q945E7	ATGGGTGTTGTGAGTTATGAGTTTGAGGTAACC	TCCCGAACTCCCGAACATTGCTCCAGGCCAGGCTT	--AAGGCTTTGTTCTTGAGGCTGCCAAGATTGG	
Q6XC94	ATGGGTGTTGCCAGTTATGAGTTTGAGGTAACC	TCCCGAACTCCCGAACATTGCTCCAGGCCAGGCTT	--AAGGCTTTGTTCTTGAGGCTGCCAAGATTGG	
Q6Q4B5	ATGGGTGTTGTGAGTTATGACTTGAGGTAACC	CTCCCCAATTGCTCCAGGCCAGGCTT	--TCAAGGCTTTGTTCTGAACCTGCCAAGGTTTGG	
Q43549	ATGGGTGTTTCATTACGAAACTGAGTTTACCC	TCCGTCATTCCCCCTGCTAGGTT	--AATGCCCTTGTTCTGATGCTGACAAACCTCATC	
Q4VPJ1	ATGGGTGTTTCACATACGAACTCTGAGTCCACC	TCCGTCATCCCCCTGCTAGGTT	--AATGCGACTGCTCTTGATGGTGAACAAACCTCATC	
Q84LA7	ATGGGTGTCCTCACATACGAACTCCGAATTGACC	TCTATCATCCCCCTGCTAGGTT	--AATGCCCTTGTTCTGATGCTGACAAACCTCATC	
Q4VPI3	ATGGGTGTTTCACATACGAACTCCGAATTGACC	TCTATCATCCCCCTGCTAGGTT	--AATGCCCTTGTTCTGATGCTGACAAACCTCATC	



Codon substitution model

- GY94 (Goldman-Yang)
- Codon frequencies
 - Nucleotide and amino acid frequencies
- Codon neighborhood
 - Substitution rate depends on similarity between coded amino acids
- Natural selection
 - Non-synonymous / synonymous rate





Tree building algorithms

UPGMA: Unweighted pair-group method arithmetic mean

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

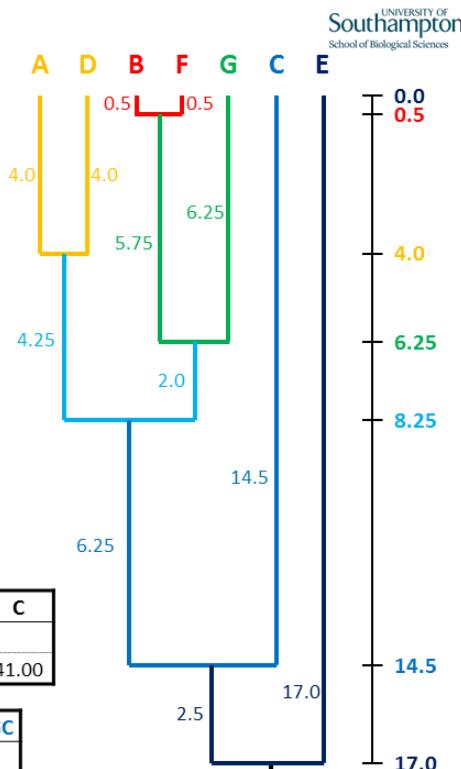
4

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

5

	ADBF	FGC
C		29.00
E	32.60	41.00

	ADBF	FGC
E		34.00

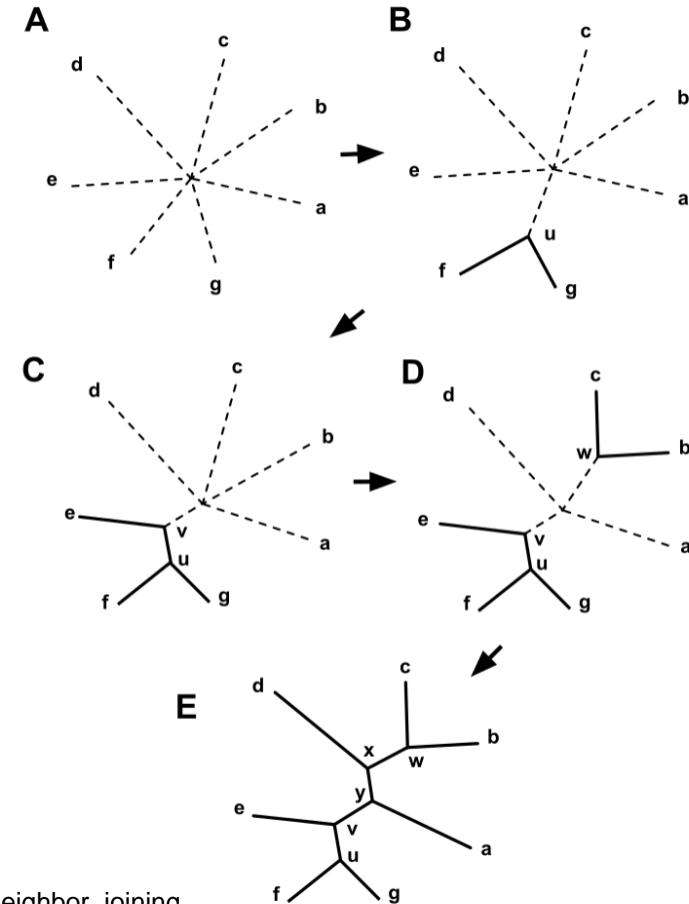


Neighbor joining

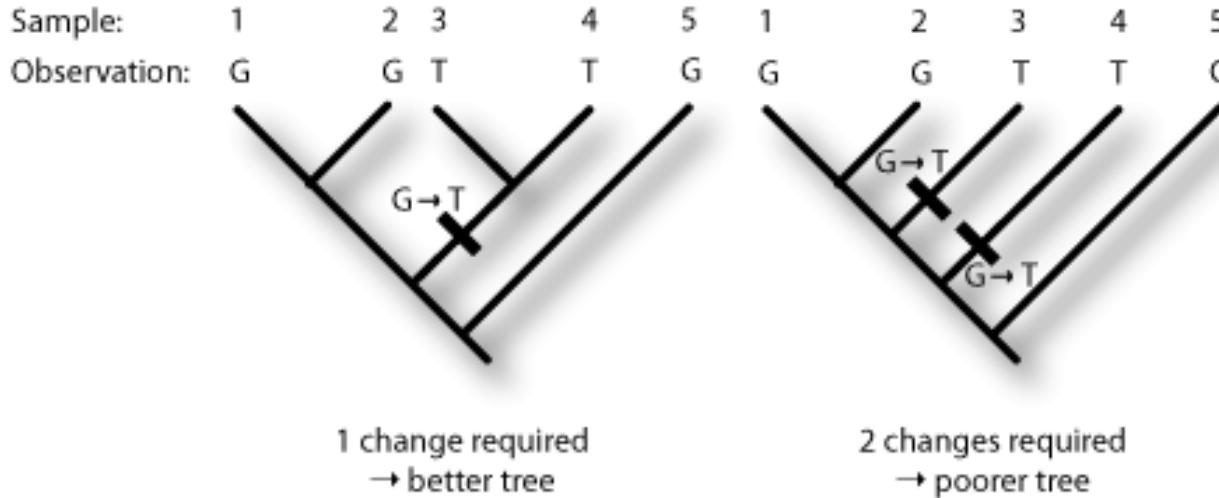
- Start with a star tree
- Join taxa i, j with smallest Q score

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

- Join taxa that are **similar to each other** but **dissimilar from other taxa**



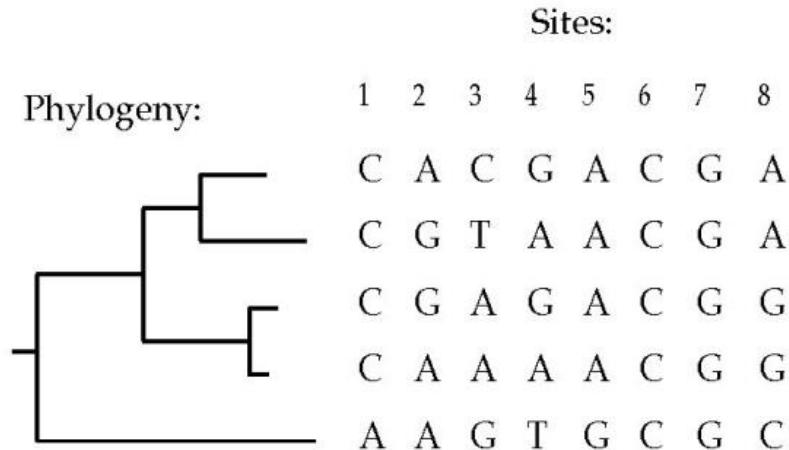
Maximum parsimony / Minimum evolution



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

- Explanation requiring minimum number of changes is the most likely
- Cannot handle distant evolution: A → T → A

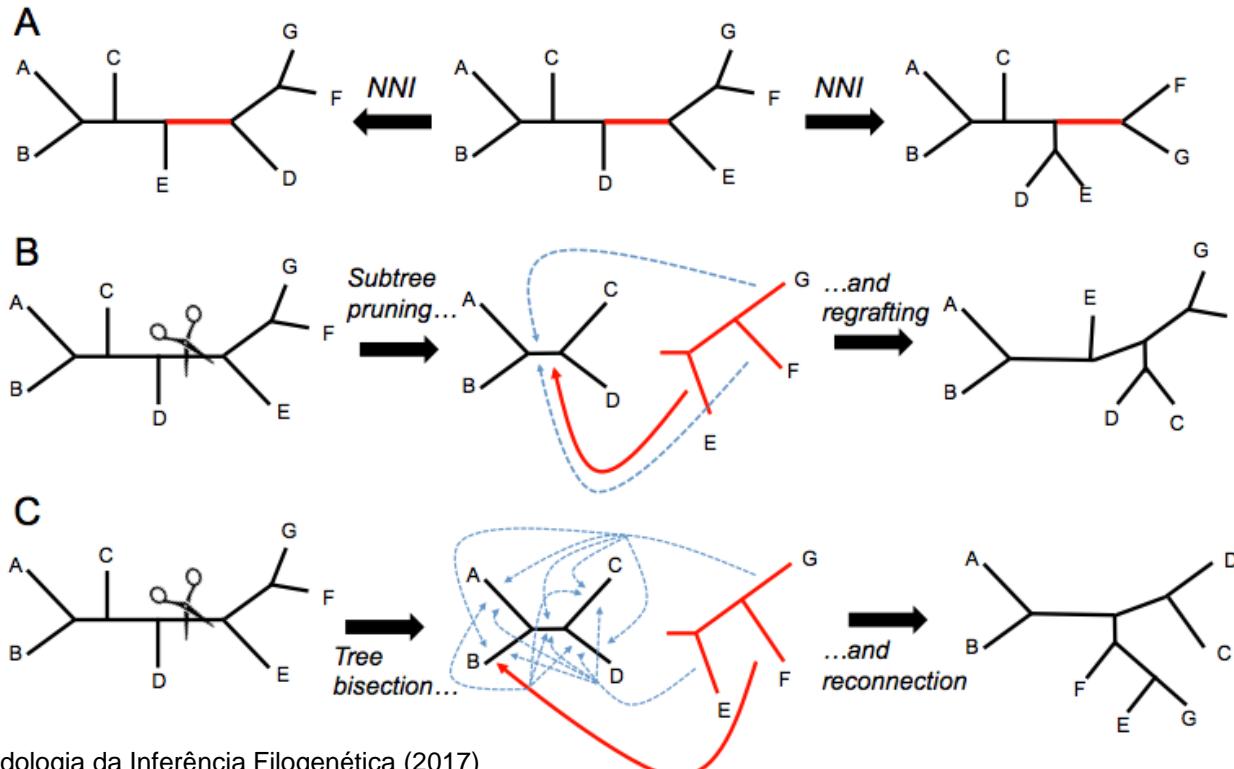
Maximum likelihood approach



https://homes.cs.washington.edu/~ruzzo/courses/gs559/09wi/lectures/8A_likelihood.pdf

- Likelihood = $P(\text{sequence data} \mid \text{substitution model, tree topology, branch lengths})$
- Given a phylogenetic tree topology, we can alter branch lengths (one branch at a time) to search for the best answer
- The problem is how to find the best tree topology

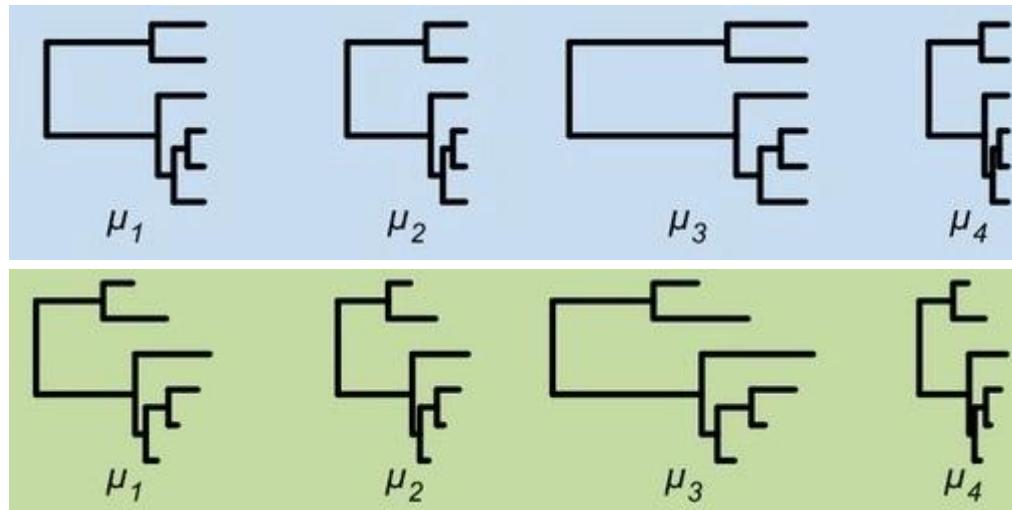
Heuristic tree search algorithms





Additional evolutionary parameters

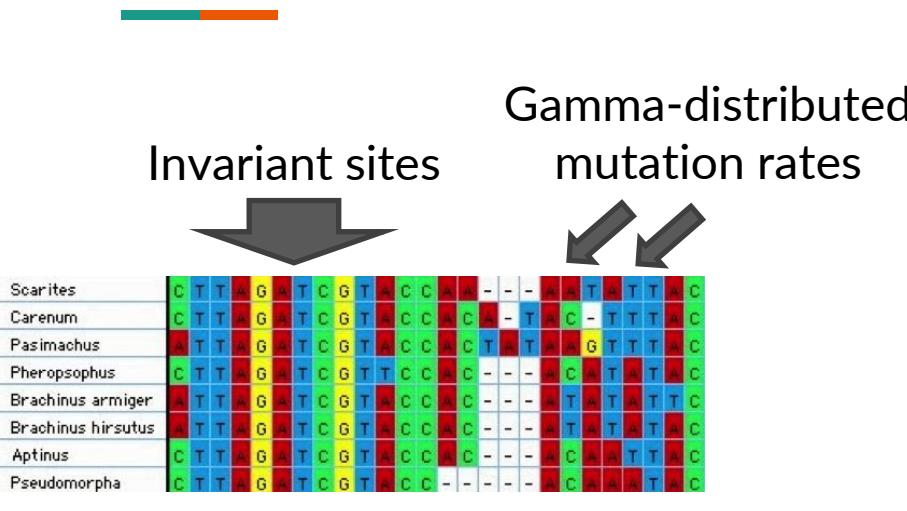
Molecular clock assumption



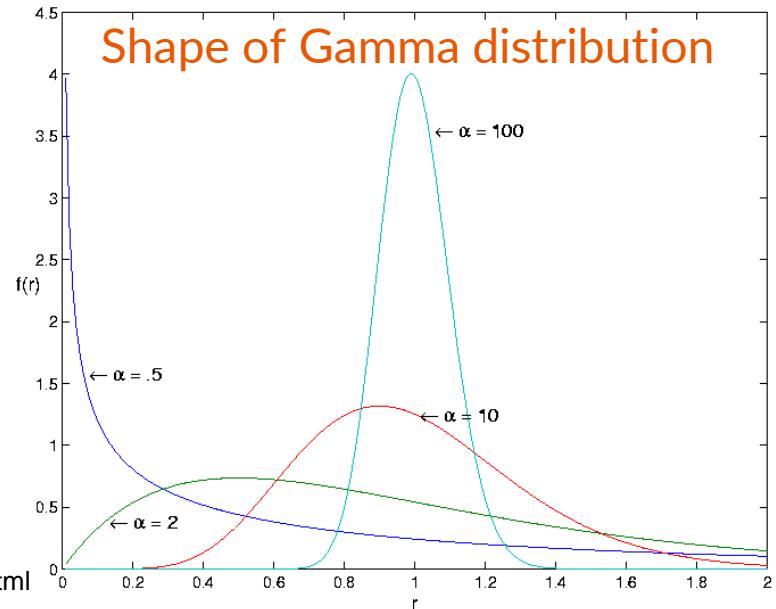
Ho, S.Y.W. and Duchene, S. Molecular Ecology 23:5947-65 (2014)

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

Site-specific evolutionary models



<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

Nonsynonymous / Synonymous substitution



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 ↓

<u>Model</u>	<u>-lnL</u>
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

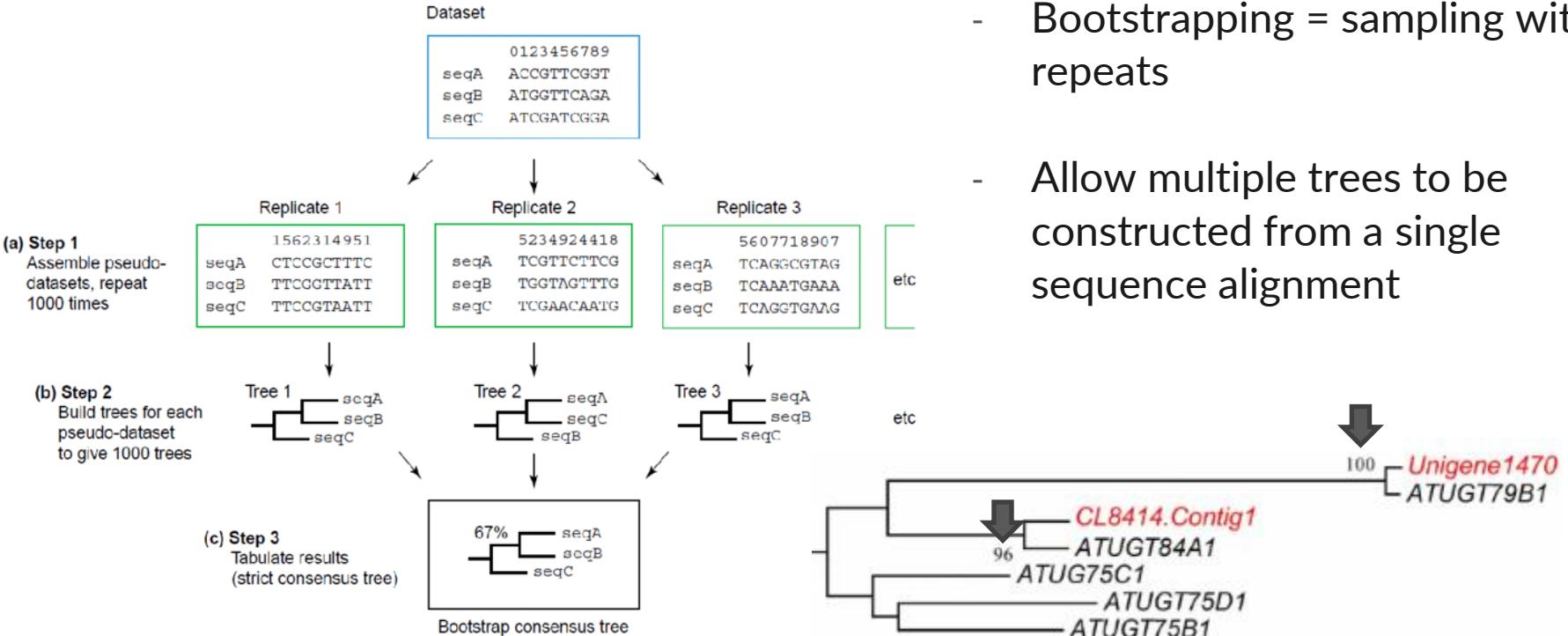
<u>models</u>	<u>diff. DF = q</u>	<u>X²</u>	<u>P</u>	
JC-F81	3 - 0 = 3	155	0	F81 model fits the data significantly better than JC
F81-HKY85	4 - 3 = 1	549.4	0	HKY85 model fits the data significantly better than F81
KHY-TrN	5 - 4 = 1	2.1	0.15	TrN model does not fit the data significantly better than HKY85

<u>Model</u>	<u>-lnL</u>
HKY85	3233.34395
HKY85 +G	3145.29031

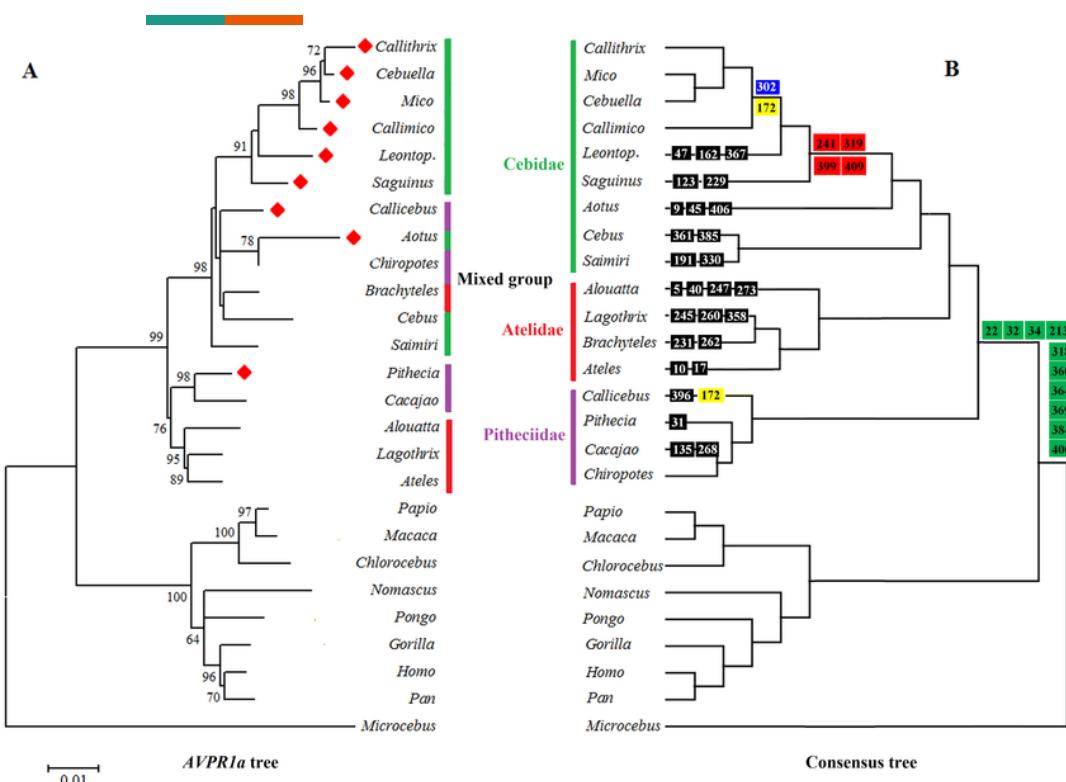
<u>models</u>	<u>diff. DF = q</u>	<u>X²</u>	<u>P</u>	
HKY85-vs. +G	1	176	0	Adding site-specific rate fits the data significantly better
HKY85+G vs. I+G	1	5.85	0.015	

Adding invariant sites does not fit the data significantly better

Bootstrap support for taxa group



Gene tree vs concatenated tree



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



Example of how to setup phylogenetic reconstruction

MEGA: A Windows GUI tool for phylogenetic analysis

Filter tutorials by topic:

- Align Sequences
- BLAST Search
- Bootstrap Tree
- Calibrate
- Distance Estimation
- Edit Sequences
- Edit Tree
- Evolutionary Probability
- Gene Duplication
- Grouping Taxa
- Installation
- Model Selection
- Pairwise Distances
- Phylogeny Construction
- Substitution Matrix
- Trace Files



TUTORIALS



Below are links to online video lectures and tutorials for multiple versions of MEGA. The first section of videos were created by members of Dr. Sudhir Kumar's lab at the Institute for Genomics and Evolutionary Medicine ([iGEM](#)) at Temple University. The rest of the videos were produced by users of MEGA. To assemble this collection of videos, the MEGA team performed a search of YouTube for instructional MEGA videos and assembled this collection of the most popular videos found. If you would like to suggest additions to this collection, please contact us by using the [feedback page](#).

KUMAR LAB VIDEOS

- Molecular Dating with MEGA
- Choosing and Acquiring Sequences Part 1
- Choosing and Acquiring Sequences Part 2

Reconstructing Ancestral
Relative Rate Framework for
Inferring Selection with MEGA

Substitution model choices

PHYLOGENY mode

MX: Analysis Preferences

Phylogeny Reconstruction |

Option	Setting
ANALYSIS	
Statistical Method	→ Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny	→ None
No. of Bootstrap Replications	→ Not Applicable
SUBSTITUTION MODEL	
Substitutions Type	→ Amino acid
Genetic Code Table	→ Standard
Model/Method	→ Jones-Taylor-Thornton (JTT) model
RATES AND PATTERNS	
Rates among Sites	→ Poisson model
No of Discrete Gamma Categories	→ JTT with Freqs. (+F)
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ WAG model

Amino acid-based models

MX: Analysis Preferences

Phylogeny Reconstruction |

Option	Setting
ANALYSIS	
Statistical Method	→ Maximum Likelihood
PHYLOGENY TEST	
Test of Phylogeny	→ None
No. of Bootstrap Replications	→ Not Applicable
SUBSTITUTION MODEL	
Substitutions Type	→ Nucleotide
Genetic Code Table	→ Not Applicable
Model/Method	→ Tamura-Nei model
RATES AND PATTERNS	
Rates among Sites	→ Jukes-Cantor model
No of Discrete Gamma Categories	→ Kimura 2-parameter model
DATA SUBSET TO USE	
Gaps/Missing Data Treatment	→ Hasegawa-Kishino-Yano model

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

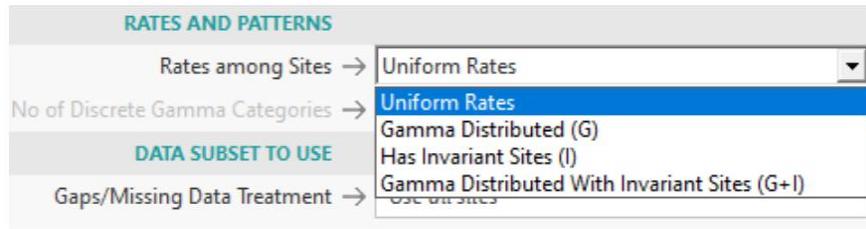
Site-specific evolutionary rate

How to handle gap (indel) and missing data

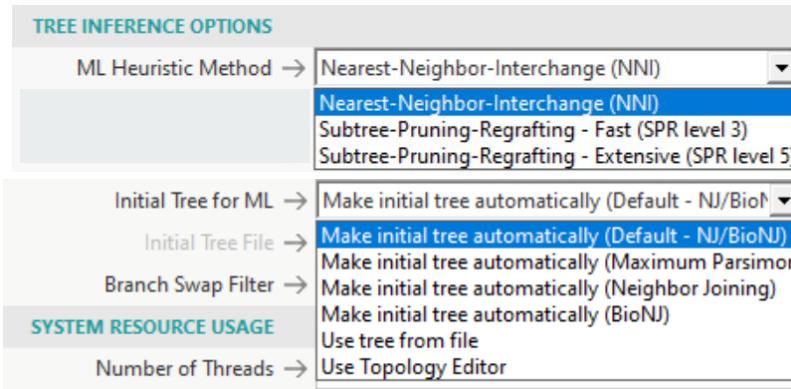
Detailed tree building method

Parallel processing

Maximum likelihood parameters



Invariant site and Gamma-distributed site-specific substitution rates



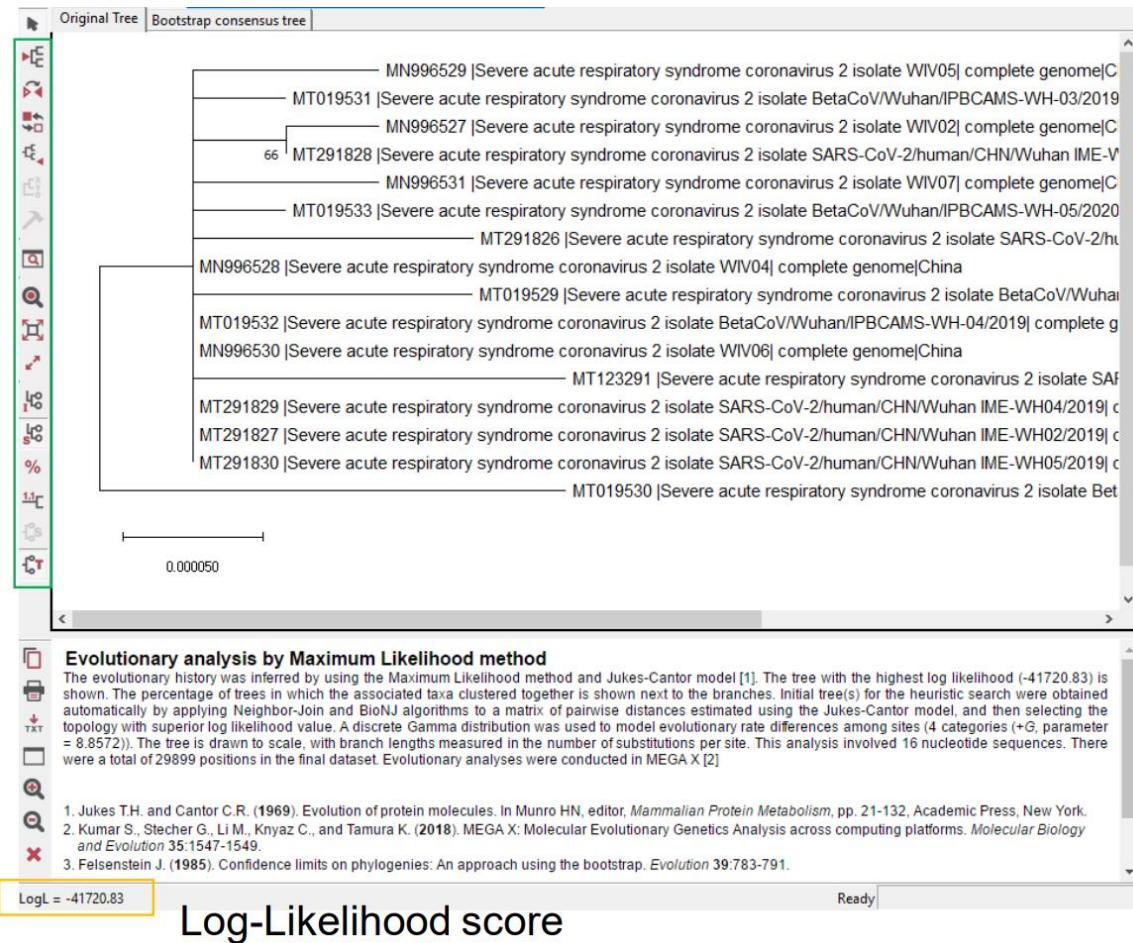
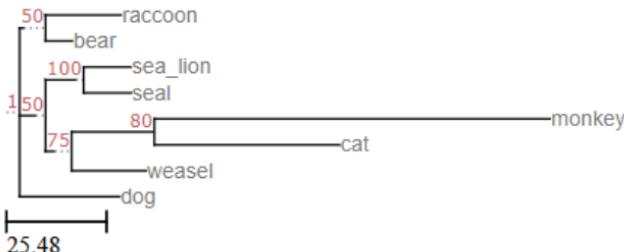
Tree search method

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

Output

NEWICK format

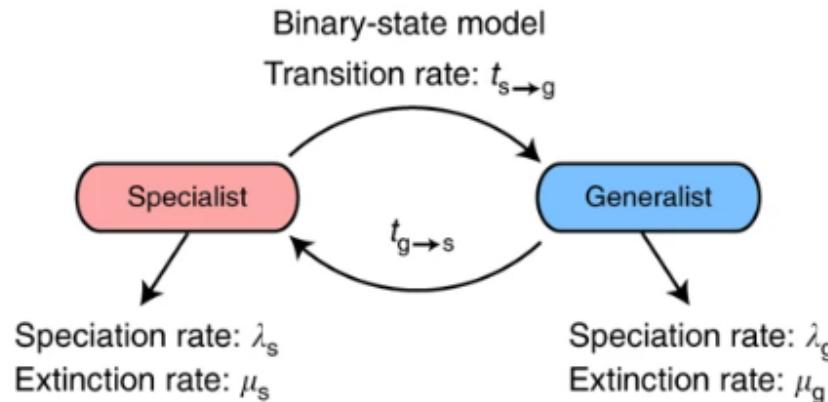
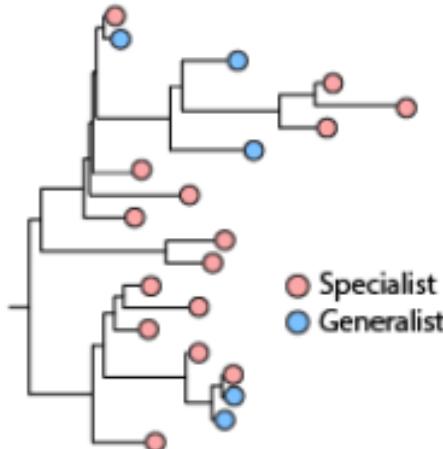
```
((raccoon:19.19959,bear:6.80041)50:0.84  
600,((sea_lion:11.99700,seal:12.00300)  
100:7.52973,((monkey:100.85930,cat:47.  
14069)80:20.59201,weasel:18.87953)75:  
2.09460)50:3.87382,dog:25.46154);
```





More molecular evolution analyses

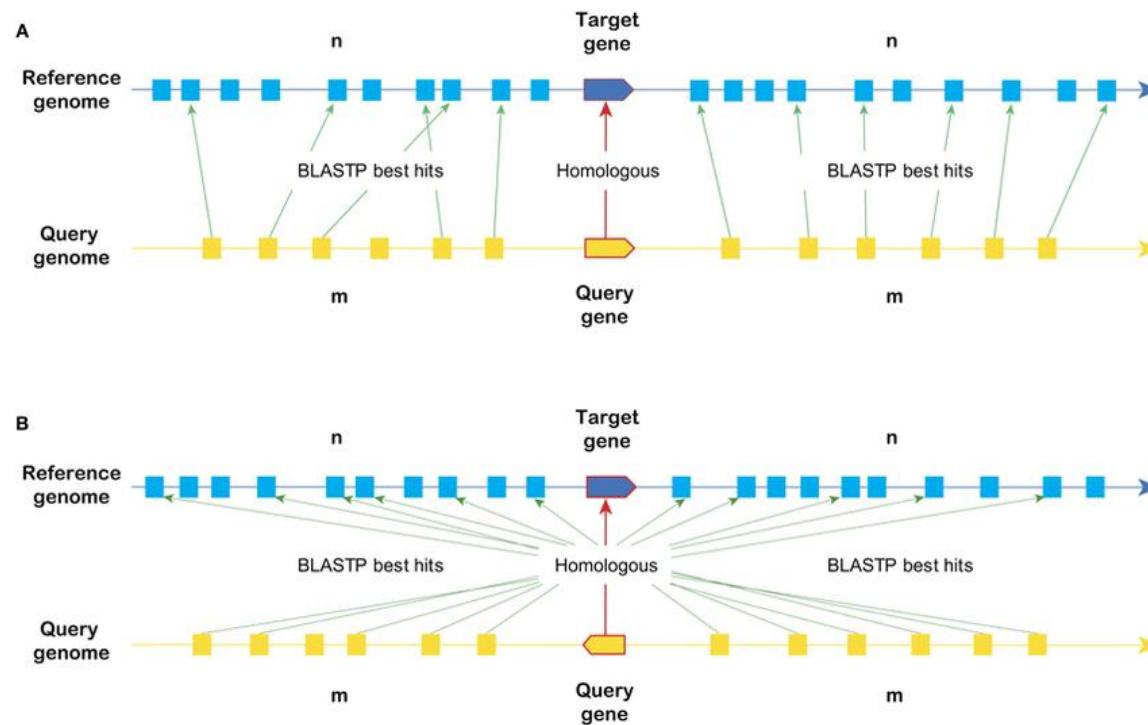
Impact of character state on evolutionary process



Sriswasdi et al. Nat Comm (2017)

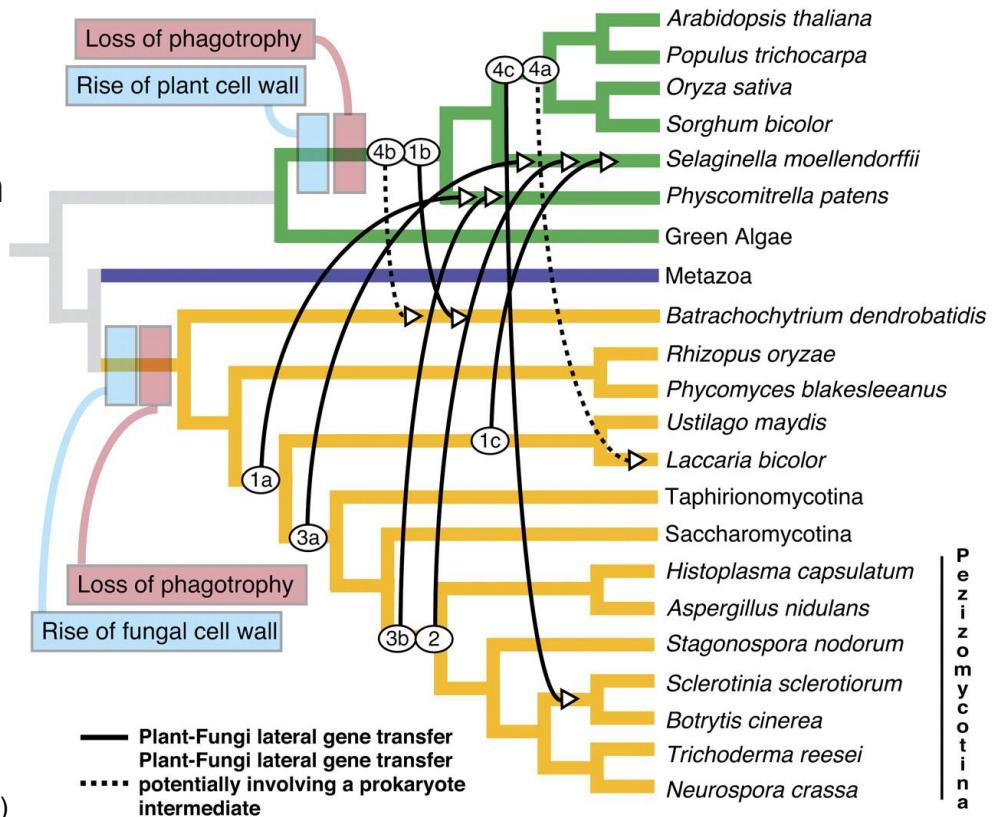
- Assume that being a generalist or a specialist would impact the speciation rate and extinction rate
- The two state can switch during evolution

Synteny: conservation of gene order on chromosome



Horizontal gene transfer (HGT)

- Evolutionary models assume vertical transmission of DNA from ancestor to offspring
- HGT creates similarity in sequence over long evolutionary distance



Any question?

