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

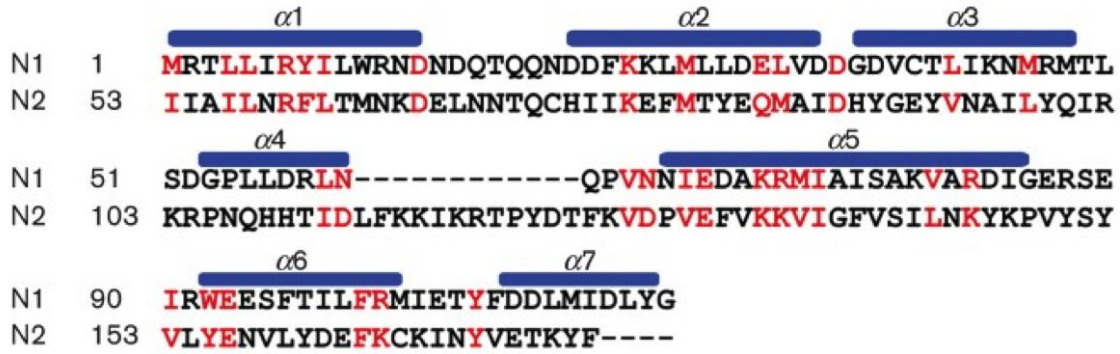
Histone H1 (residues 120-180)

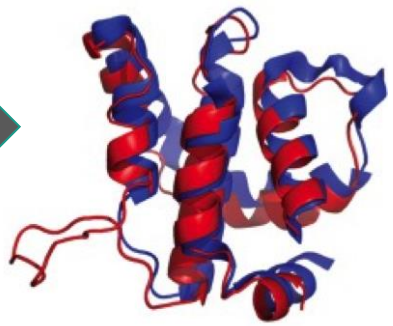
```
HUMAN  KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE  KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
RAT    KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
COW    KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
CHIMP  KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
       ***:**********:**************  ******:****  **.**********:*  **
```

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

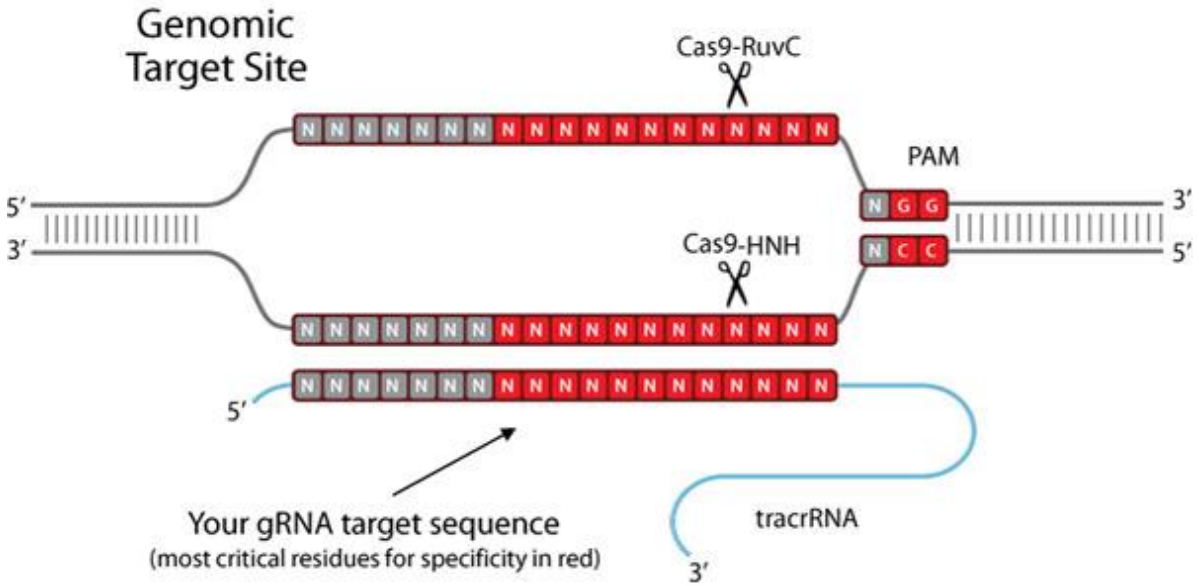




Ferguson et al. J General Virology, 94: 2070-2081 (2013)

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

Molecular probe design

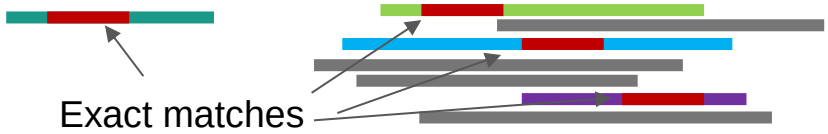


http://www.sigmaaldrich.com/technical-documents/articles/biology/crispr-cas9-genome-editing.html

- Sequence alignment can check the specificity of your probes

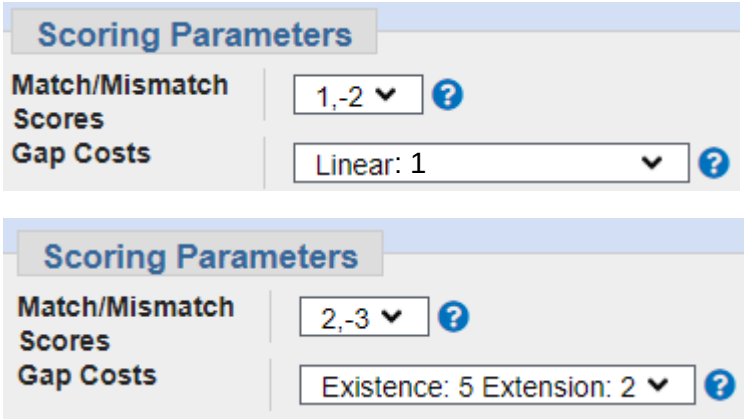
Components of sequence alignment

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 ~ 18 positions with matches
  - MM…MEM…MEM……..MEM…MM
  - NCBI's **MEGABLAST** searches for a run of 28 matches

Alignment scores




```
Ref:   ACCGTATCG
         ||   ||||
Query: AC---ATCG
```

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

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

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

Alignment score interpretation

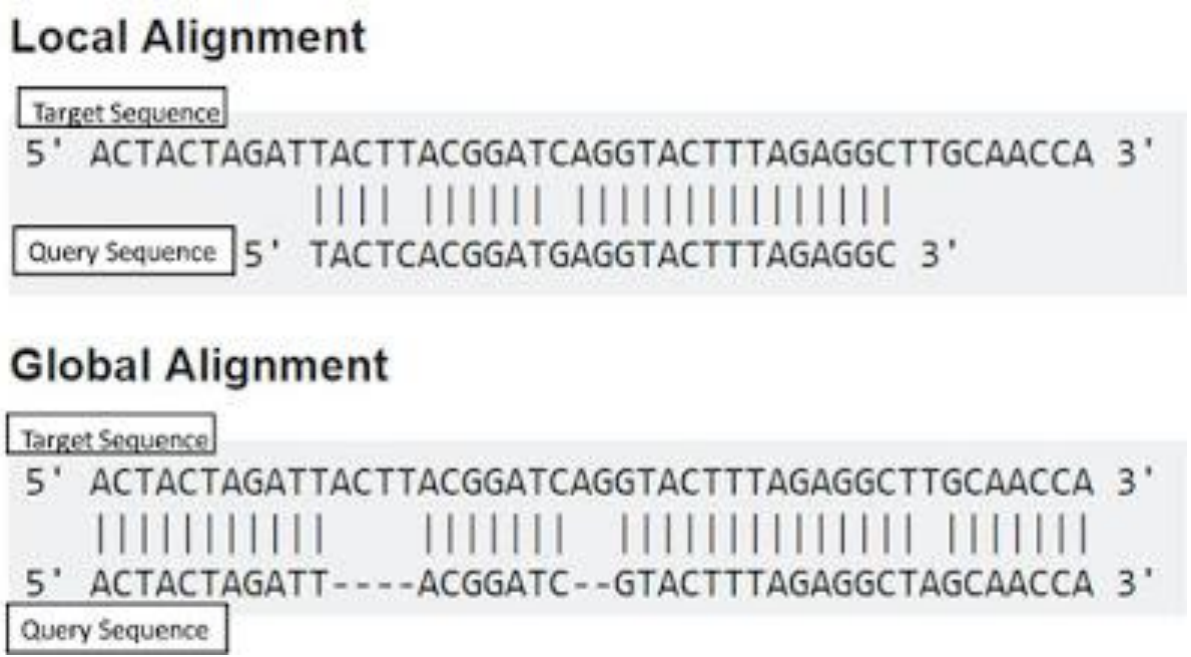
- **Match / Mismatch = +1 / −2**
  - To permit a mismatch, there must be >2 matches afterward to gain score
  - Want hits with high identity

- **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



Images generated from BABA http://baba.sourceforge.net/

Global alignment

	GAP	A	T	G	C	T
GAP	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
G	-4	-1	0	0	-2	-4
C	-6	-3	-2	-1	1	-1
T	-8	-5	-2	-3	-1	2

Match : 1
Mismatch : -1
GAP : -2

```
Seq1:  ATGCT
       |  |||
Seq2:  A-GCT
```

https://www.youtube.com/watch?v=ipp-pNRIp4g

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

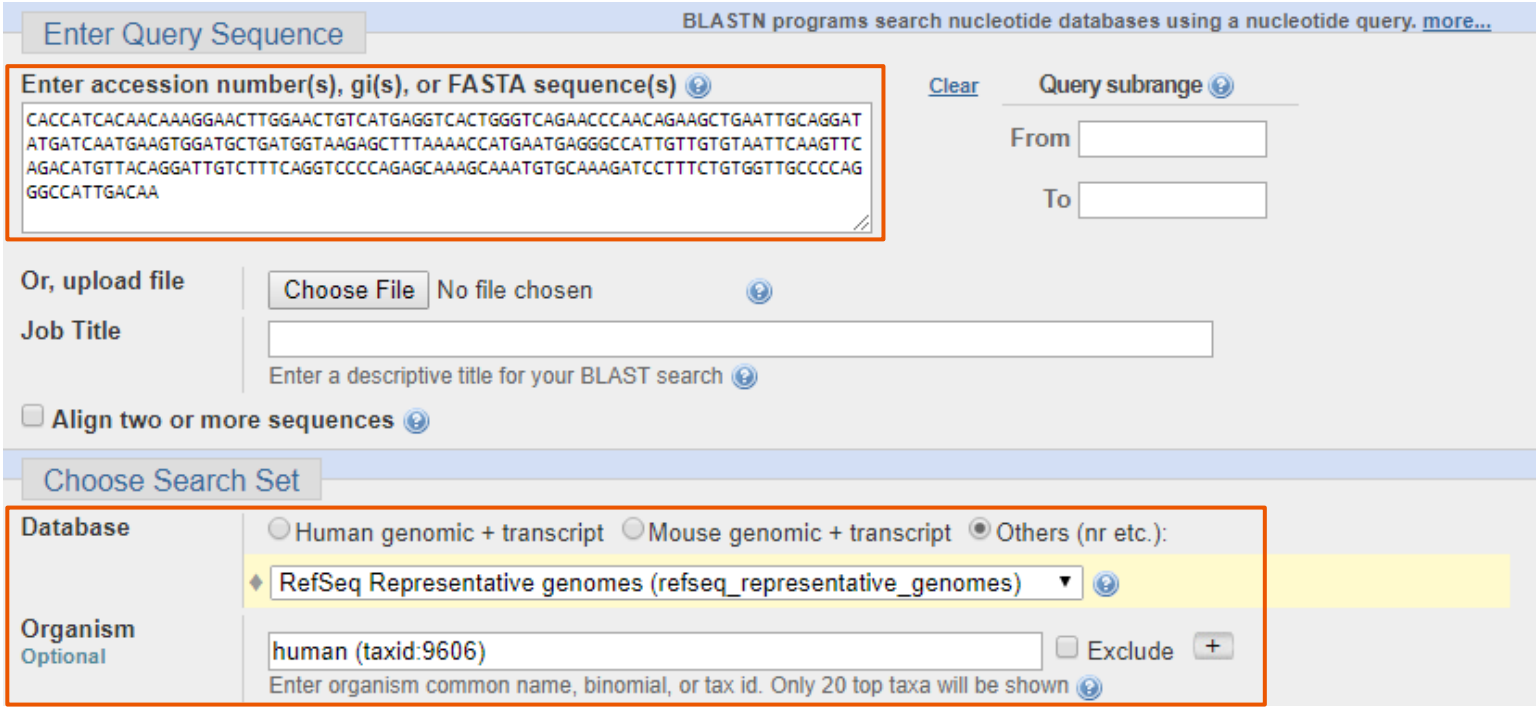
```
Seq1:  ATGCT
          |||
Seq2:   AGCT
```

- Ignore possibilities with negative score
  - Start over is better

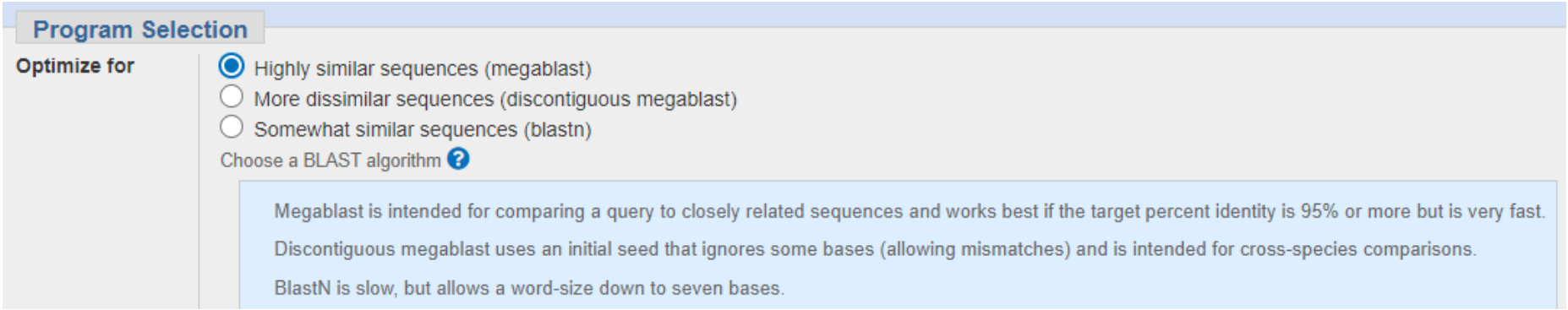
https://www.youtube.com/watch?v=lu9ScxSejSE

Basic Local Alignment Search Tool BLAST

NCBI's nucleotide BLAST interface

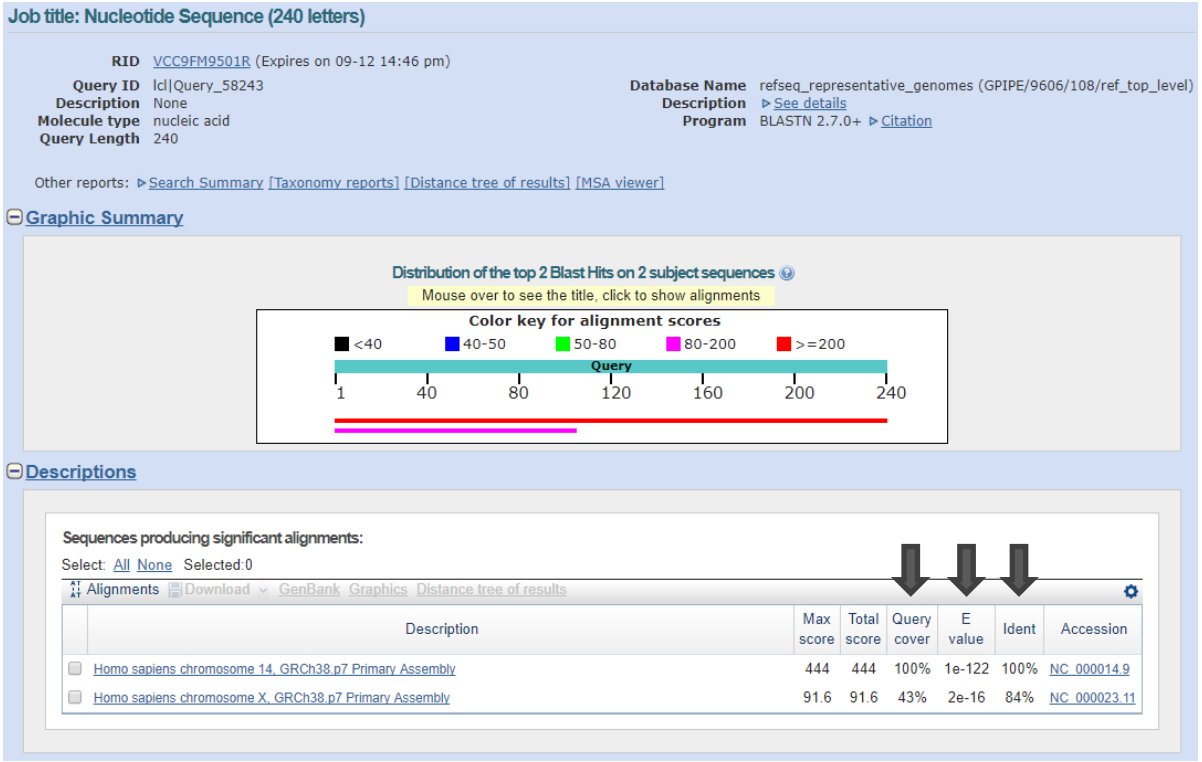


Nucleotide BLAST algorithms



- 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



Query coverage = % of input sequence used in the alignment

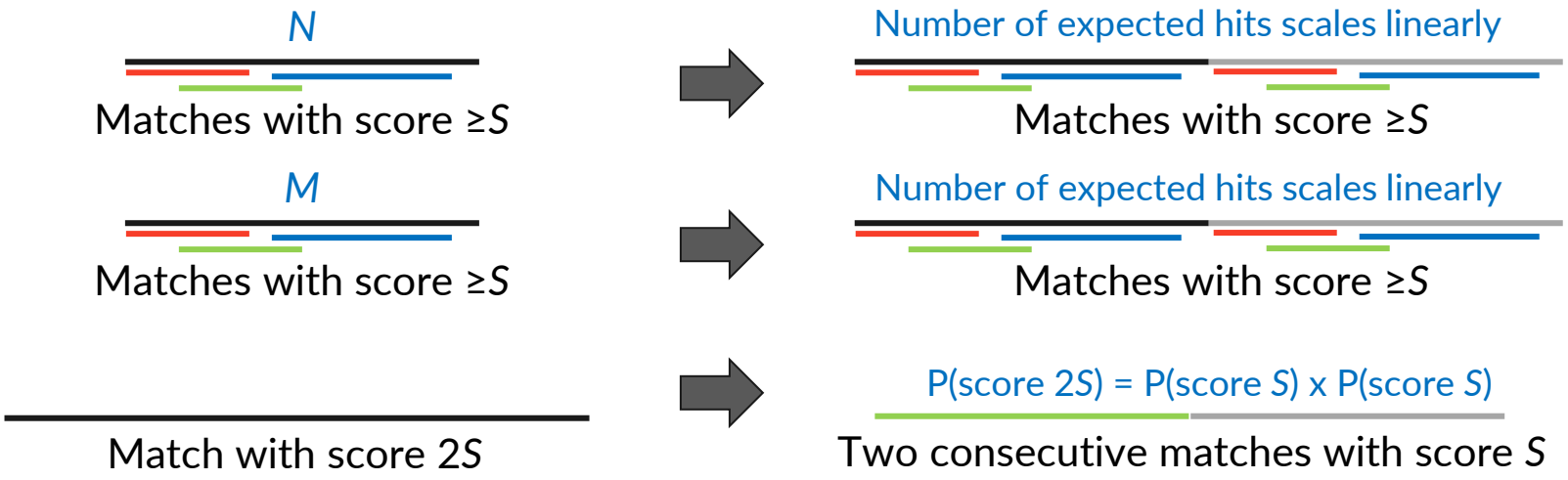
Identity = % of identity between input and matched sequences in the aligned region

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 $N \times M \times e^{-\lambda S}$



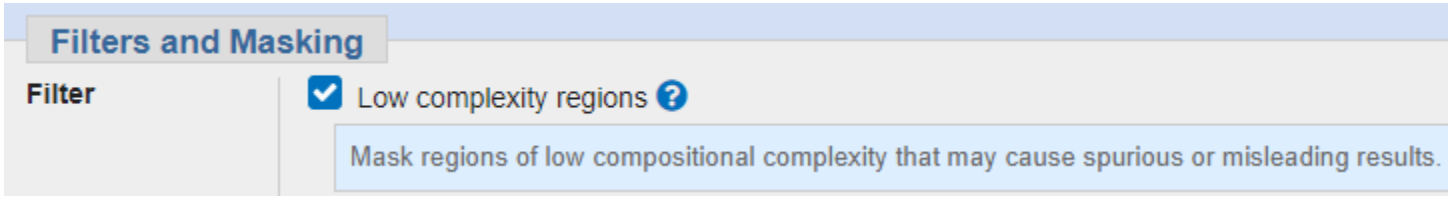
E value as Poisson distribution

- 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

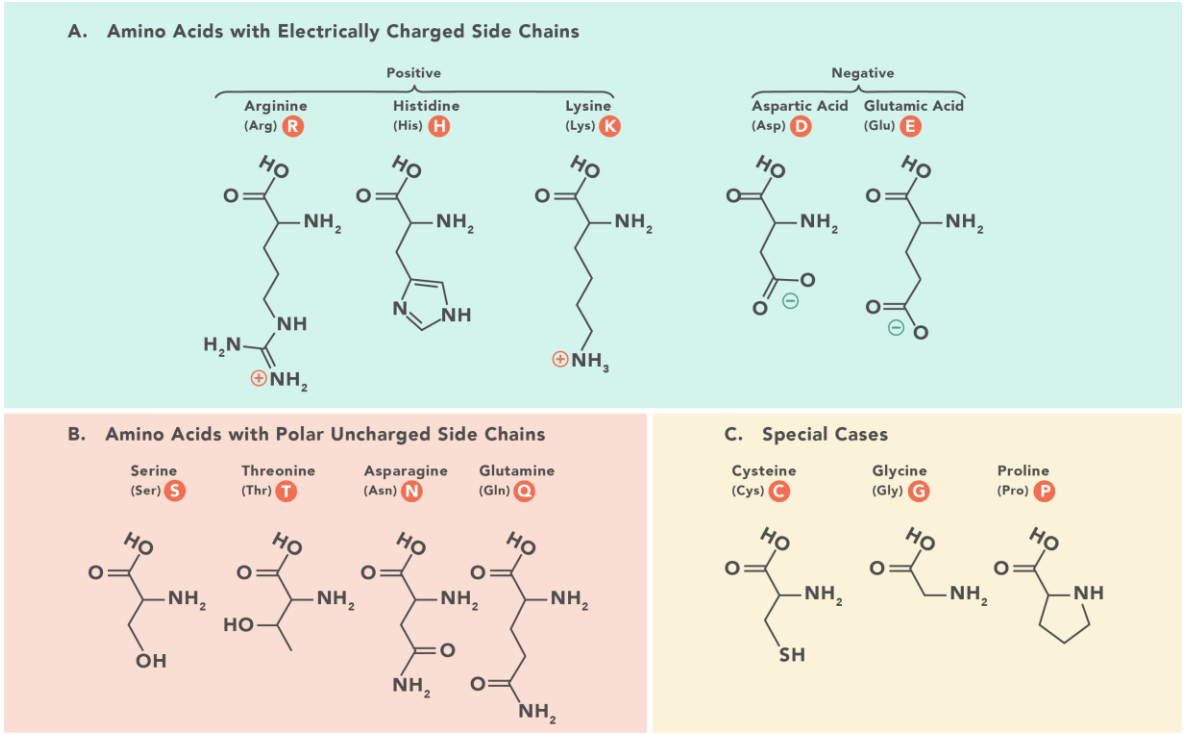
CCCGCGCGCCCCGGCGCCCGATGCAACTAGC

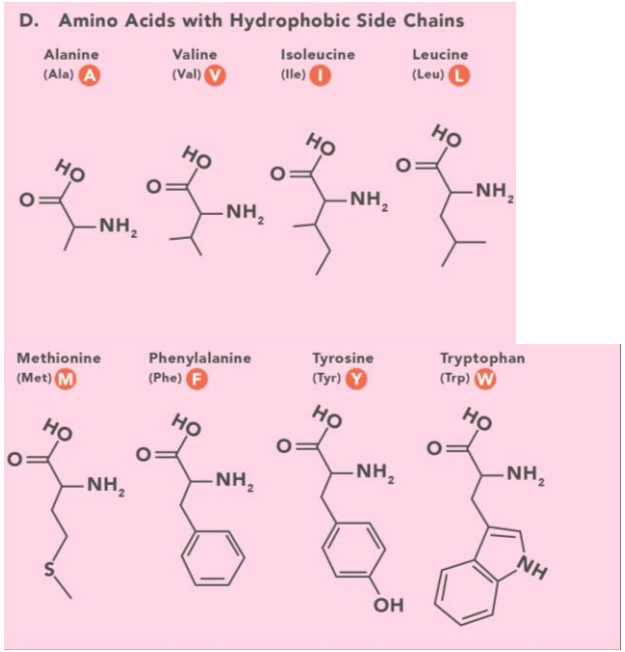


- Probability of getting a hit with score >*S* will be high if both sequences contain only C's and G's
- BLAST withholds these regions from score calculation

Protein sequence alignment

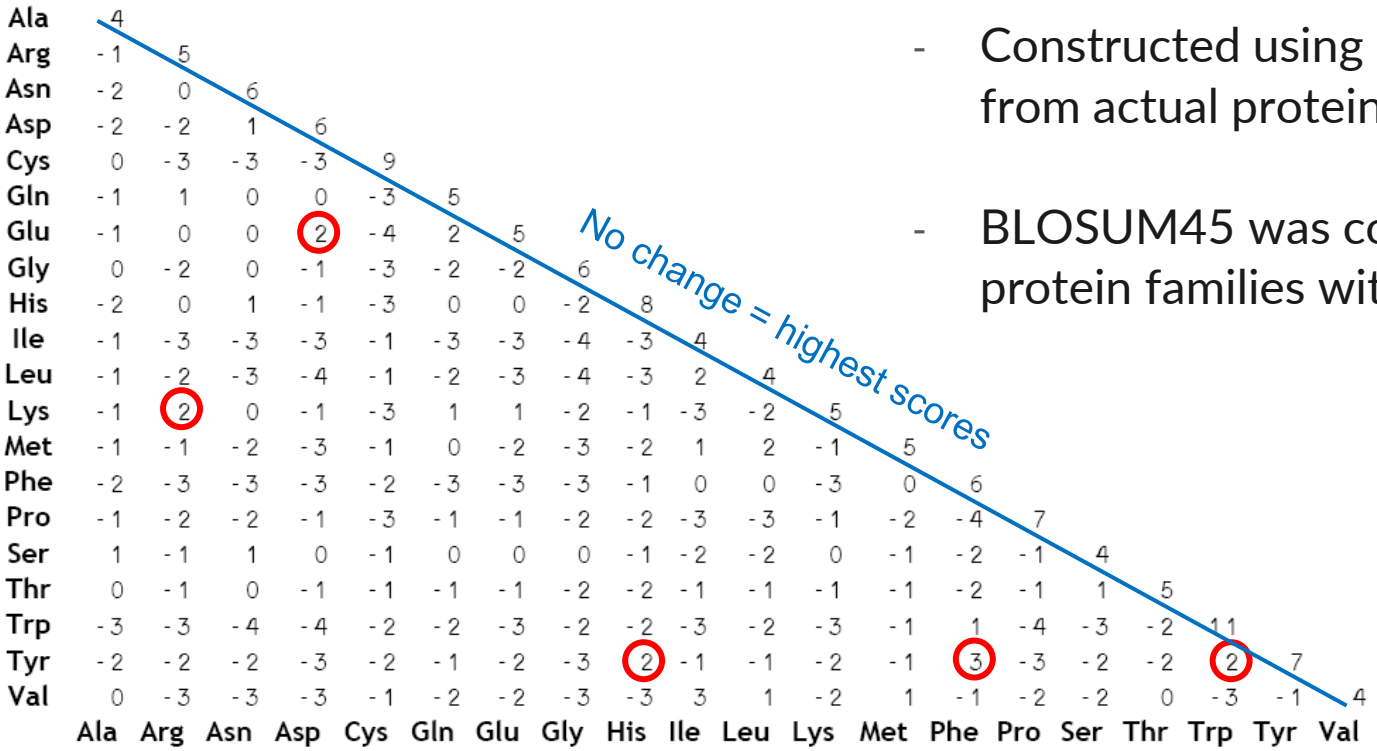
Amino acid side chains





https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357

Block Substitution Matrix (BLOSUM)

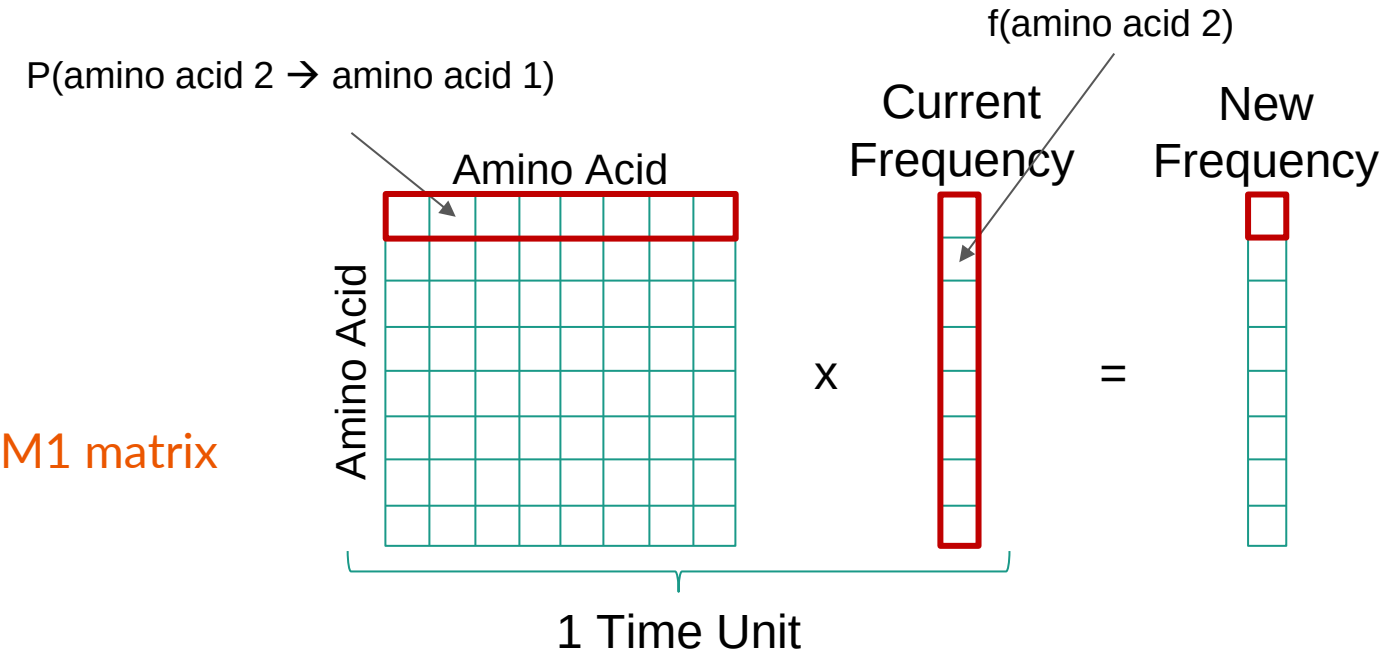


	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1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

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

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

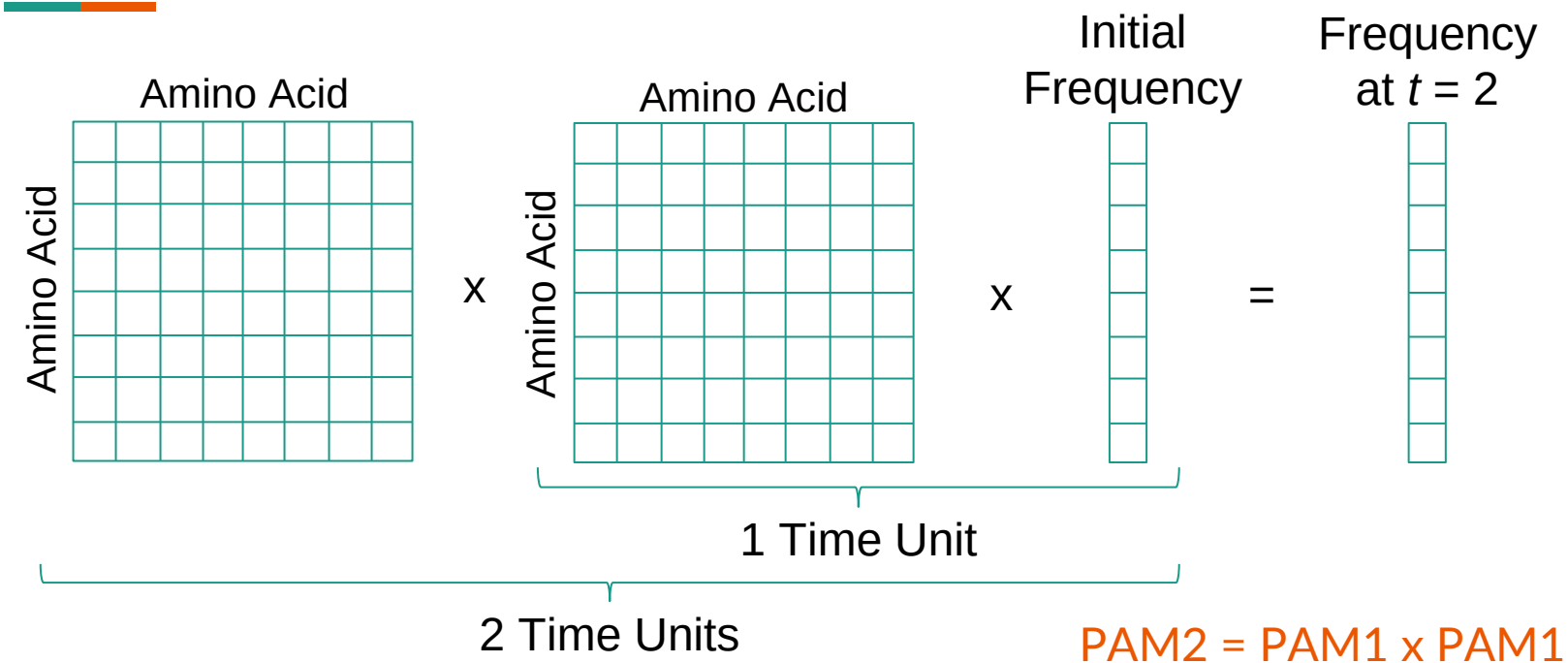
Point Accepted Mutation (PAM)



PAM1 matrix

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

Point Accepted Mutation (PAM)



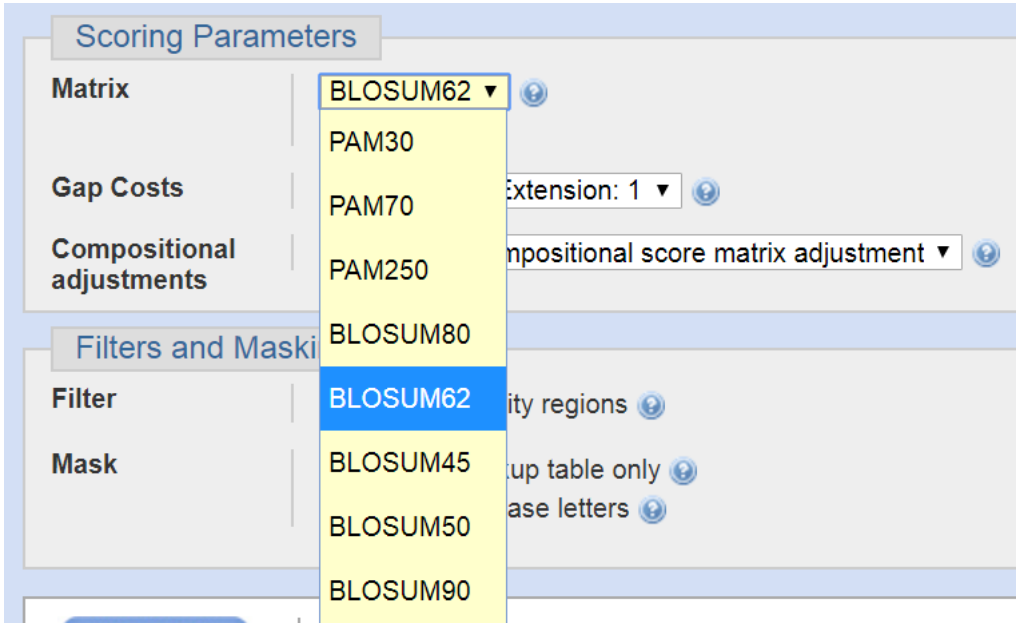
PAM2 = PAM1 x PAM1

- Extrapolate substitution rates for more distant proteins

PAM vs BLOSUM

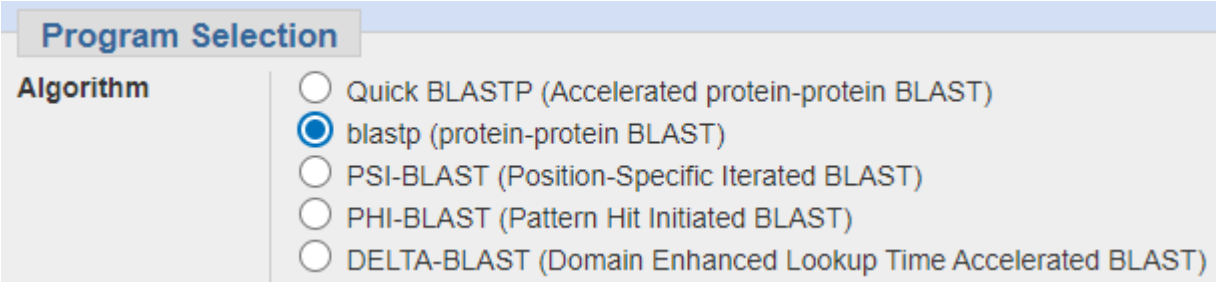
PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

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



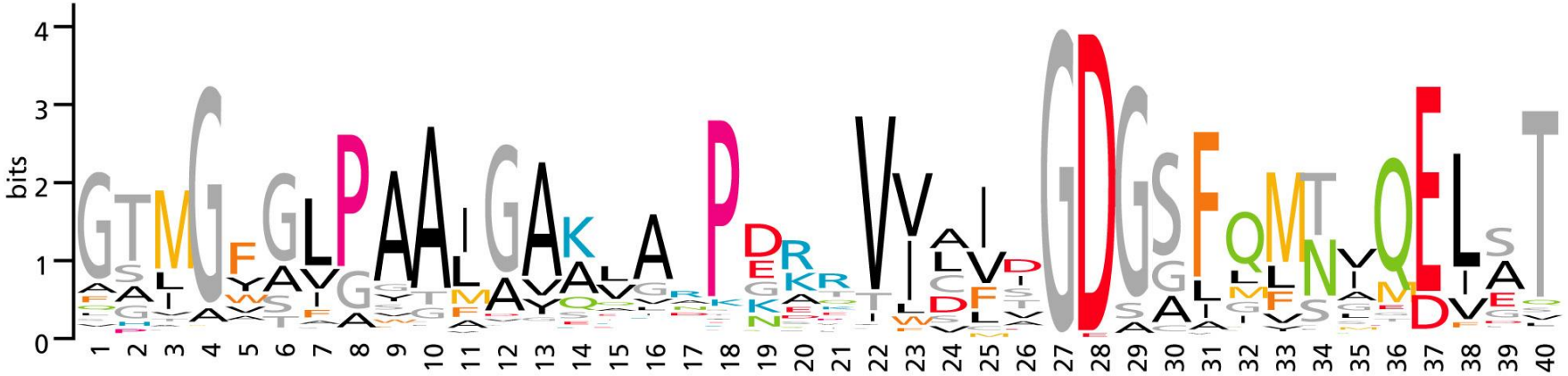
- BLOSUM for low identity, PAM for high identity

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

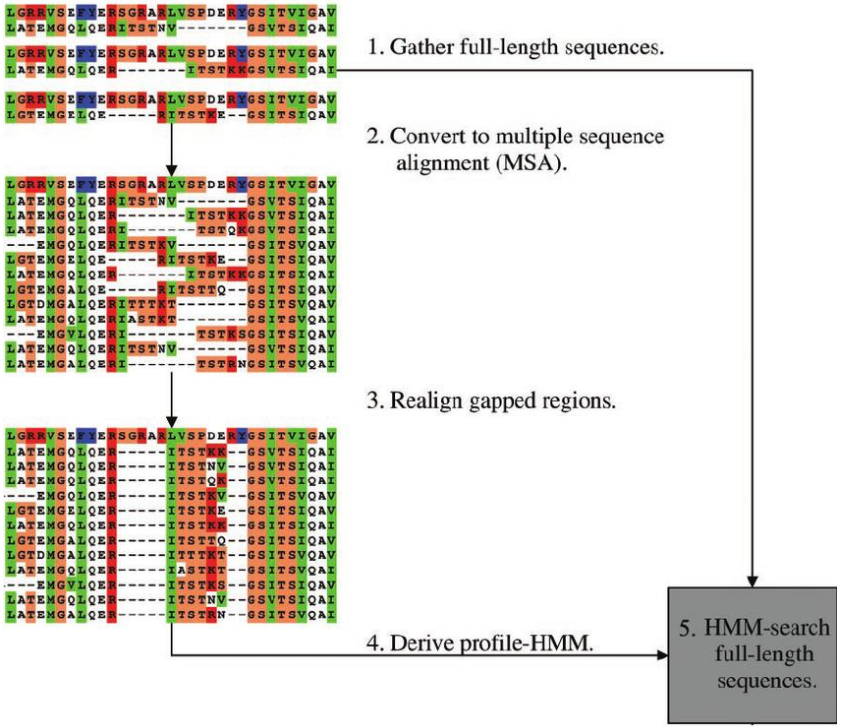
[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]

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)

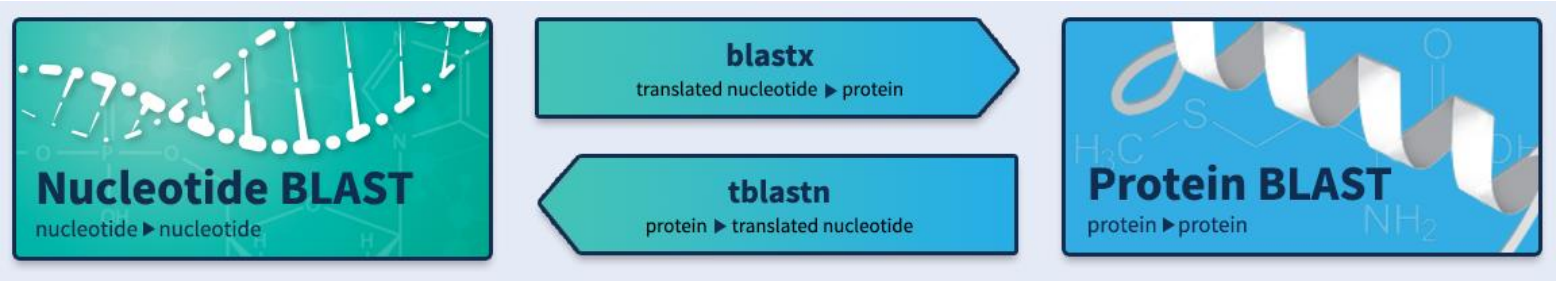


- Start from user inputs
- First round of BLASTP
- Construct PSSM from hits
- Re-search using the PSSM
- Repeat

Frickey, T. and Lupas, A. NAR 32:5231-8 (2004)

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

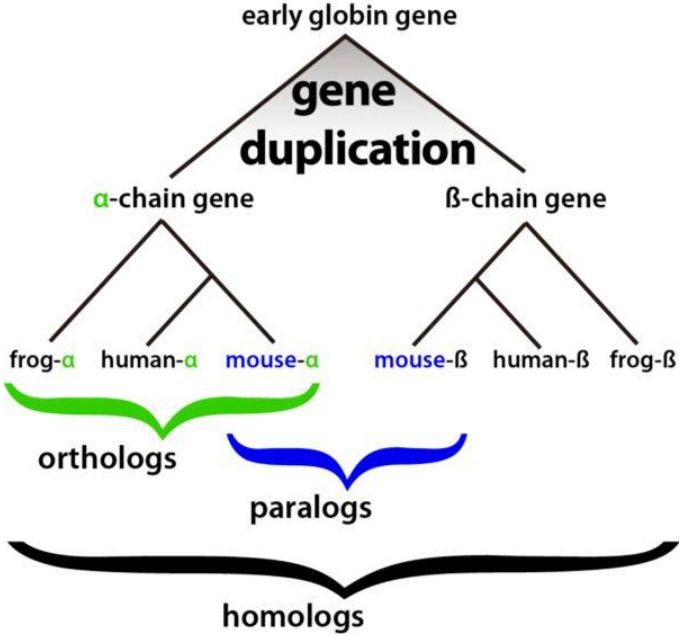
- TBLASTN = align protein to translated DNA database
  - You identified novel protein
  - No evidence in protein database
  - But there might be transcriptomics studies that identified the RNA of related proteins

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}

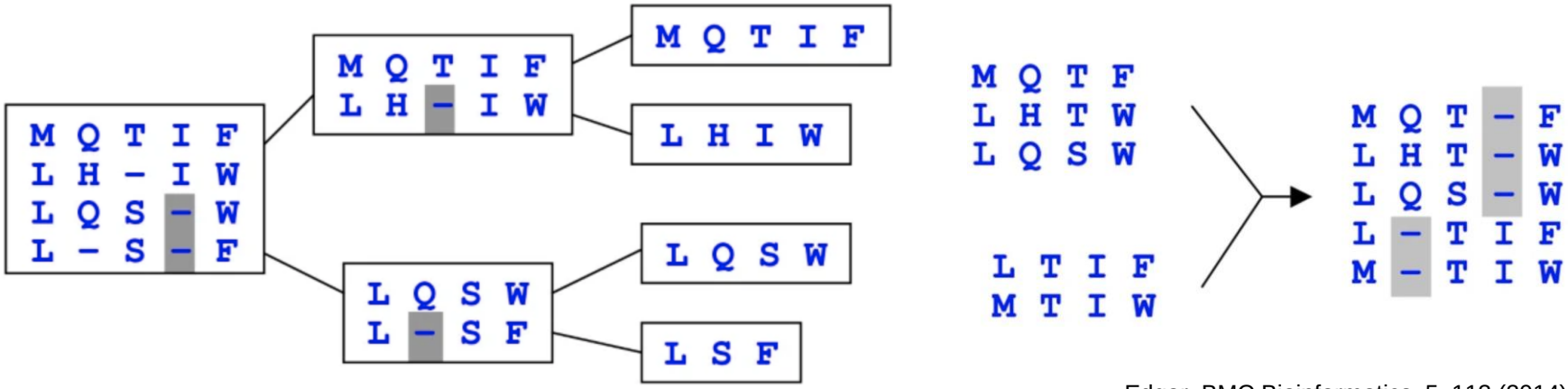
- BLAST mouse to human
- BLAST human to mouse

- 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)



Edgar, BMC Bioinformatics, 5, 113 (2014)

- 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-WV-VSAAHCYKS----------RIQVRLGE-HNINVLEGN-----
-----EQFVNAAKIIKHPNFDRKT-L------------------------NNDIMLIKLS
SP--VKLNARVATVALPS---SCA---PAGTQCLISGWGN----------TLSSGV----
-------NEPDLLQ-CLDAP-LLPQADCEAS---YPGK--------ITDNMVCVGFL---
-EGG-KDSCQGDSGGPVVCNGE--------LQGIVSWG-YGCALPDN---PGVYTKVCNY
VDWI---------------

>Q16LB2_AEDAE/136-374
-----------------------ILNGIEADLEDFPYLGALALLDNYT-----STVSYRC
GANLI------SDR-FM-LTAAHCLFG--------KQAIHVRMGTLSLTDNPDED-----
----APVIIGVERVFFHRNYTRRPIT-----------------------RNDIALIKLN
RT---VVEDFLIPVCLYT---EQNDP-LPTVPLTIAGWGG----------NDSAS-----
-------LMSSSLM-KASVT-TYERDECNSL---LAKKI-----VRLSNDQLCALGRSEF
NDGLRNDTCVGDSGGPLELSIGR----RKYIVGLTSTG-IVCGNE-F---PSIYTRISQF
IDWI---------------
```

PHYLIP

```
      5    42
Turkey     AAGCTNGGGC ATTTCAGGGT GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT ACAGGTTGGC CGTTCAGGGT AA
Chimp      AAACCCTTGC CGTTACGCTT AAACCGAGGC CGGGACACTC AT
Gorilla    AAACCCTTGC CGGTACGCTT AAACCATTGC CGGTACGCTT AA
```

ClustalW

```
Caballeronia_arvi          MNSRIDSHVKHLIFFCGHAGTGKTTLAKRLFAPLMQAAGEPFCLLDKDTLYGAYSAAAIG
Caballeronia_choica        --------MTHLVFFCGHAGTGKTTLAKRLFPRLMRATGEPFCLLDKDTLYGGYSAAAMG
Caballeronia_arationis     --------MTYLIFFCGHAGTGKTTLAKRLFPRLVRATGEPFCLLDKDTLYGAYSAAAMG
Caballeronia_telluris      --------MTHLIFFCGHAGTGKTTLAKRLFPRLAQASGEPFCLLDKDTLYGAYSAAAMN
                                  :.:*:******************. * .*:***************.*****:.

Caballeronia_arvi          ALTGDPHDRDSPLFIEHFRDPEYRCLVDTAAENLALGVSVVVVAPLTREVRSSRLFDRAW
Caballeronia_choica        ALTGDPNDRDSPLFLQHLRDPEYRALIDTARENLELGVSVAVVAPLSREVRDGRLFDRQW
Caballeronia_arationis     ALTGDPNDRDSPLFLQHFRDPEYRALIDTARENLDLGVSVAVVAPLTREVREERLFDRAW
Caballeronia_telluris      ALTGDPNDRDSPLFLQHLRDPEYRALIDTARENLDLGVSVAVVAPLTREVREGRLFDRTW
                           ******:*******::*:*******.*:*** *** *****.*****:****. ***** *
```

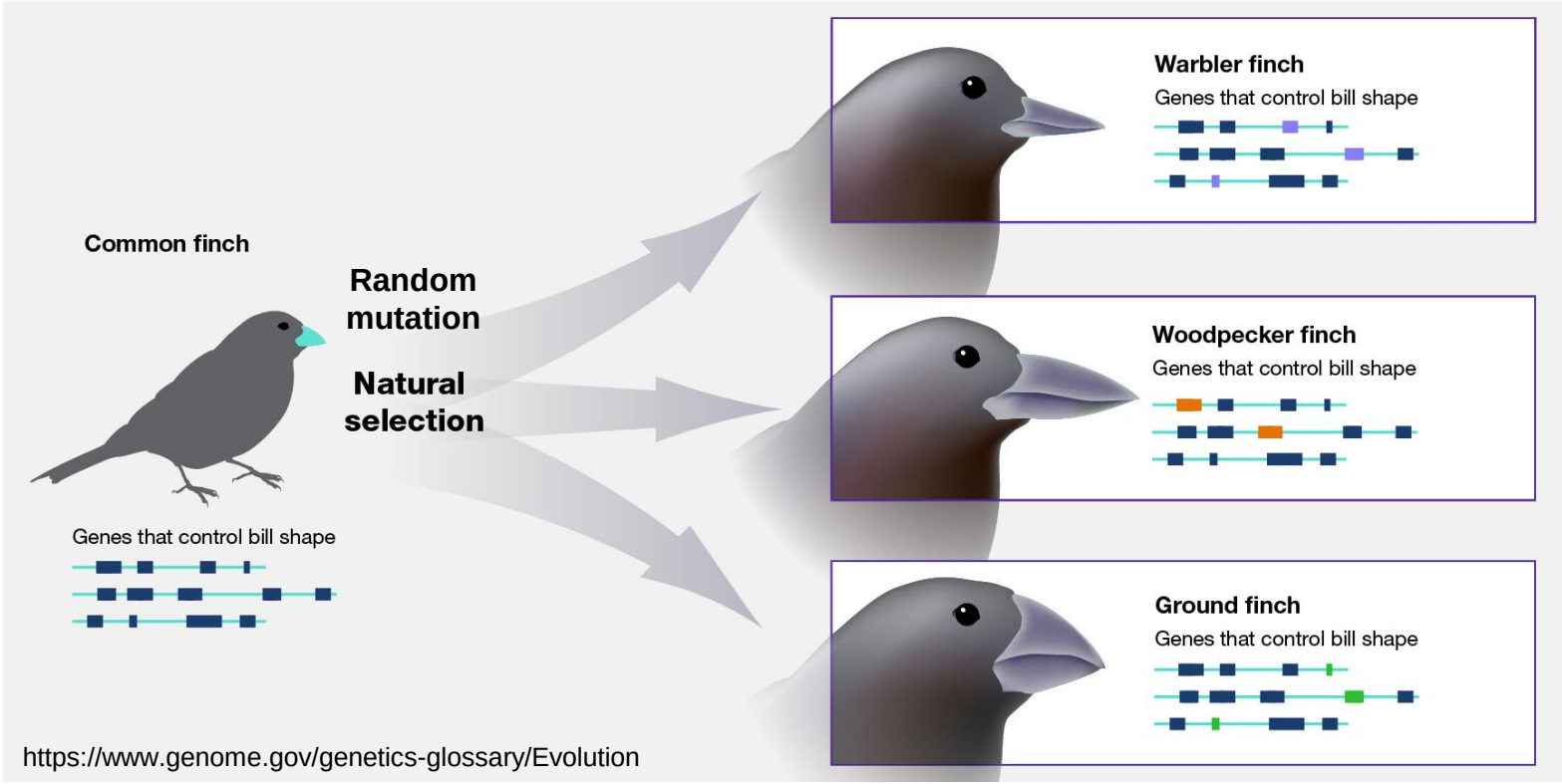
Any question?

Part II: Phylogenetics

- Evolution = nature’s experimentation = source of information

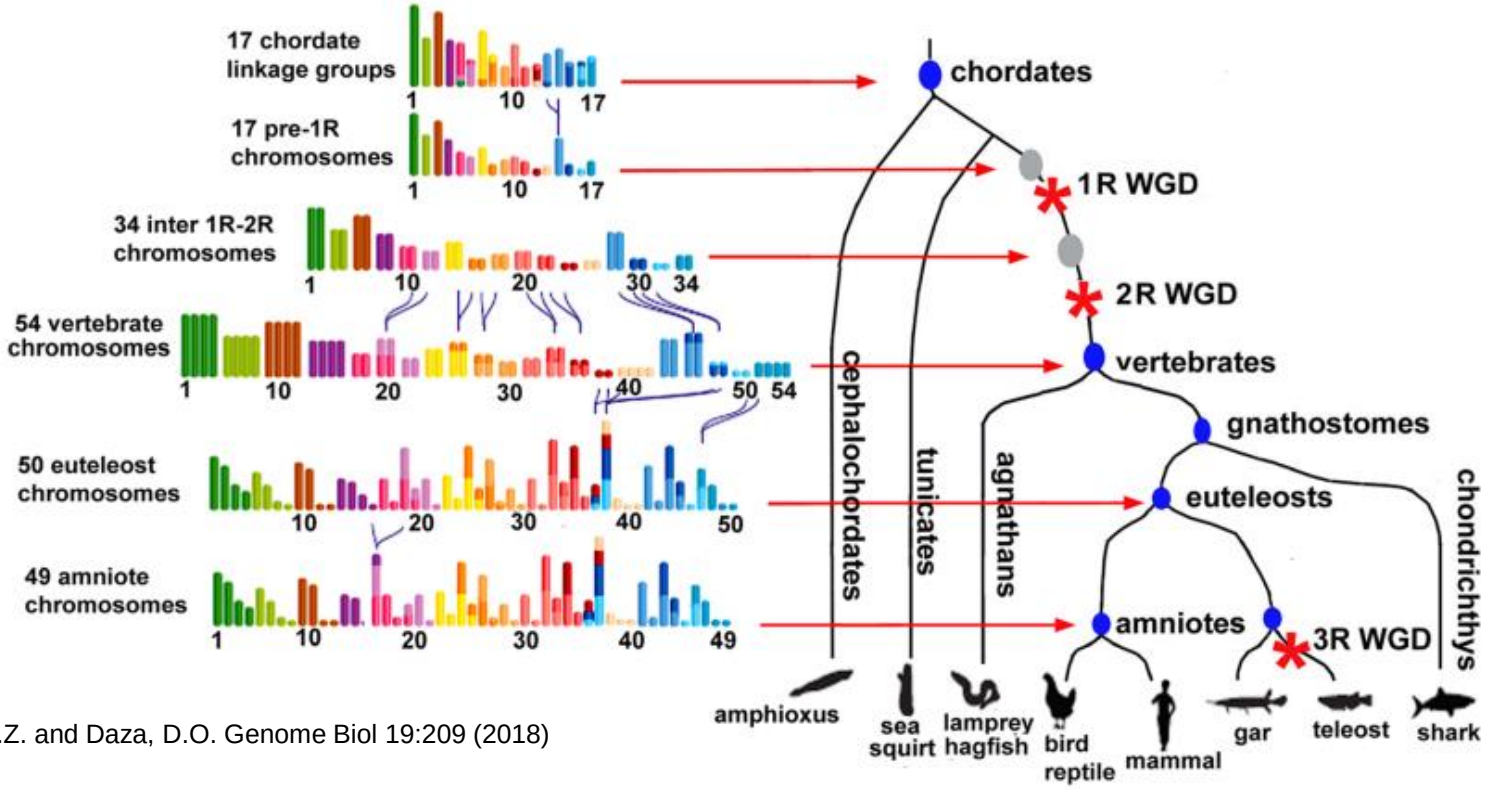
Evolution

Random mutation with natural selection



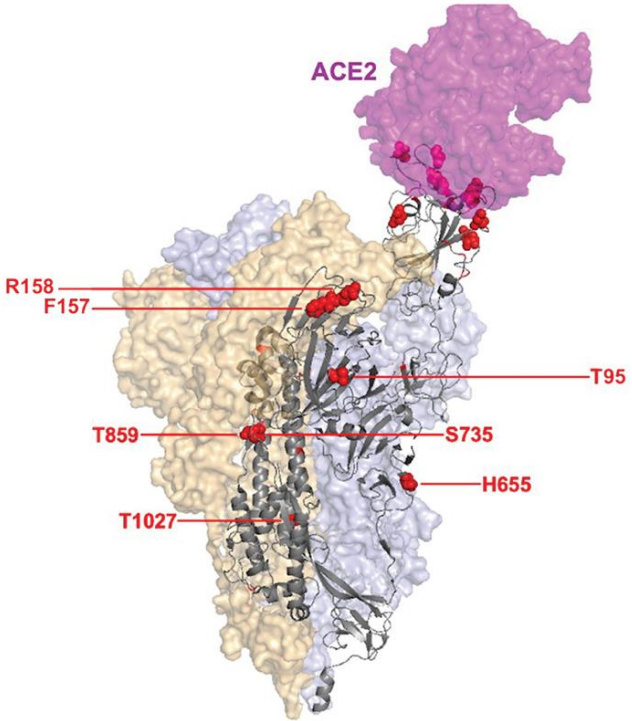
https://www.genome.gov/genetics-glossary/Evolution

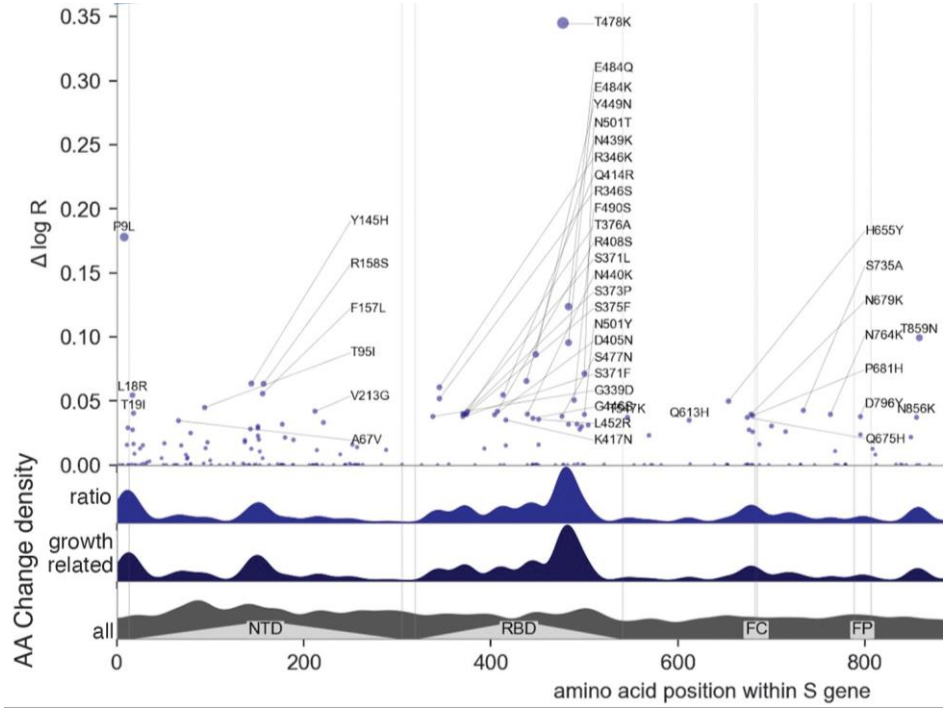
History of whole-genome duplications in animal



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

Linking evolution to function

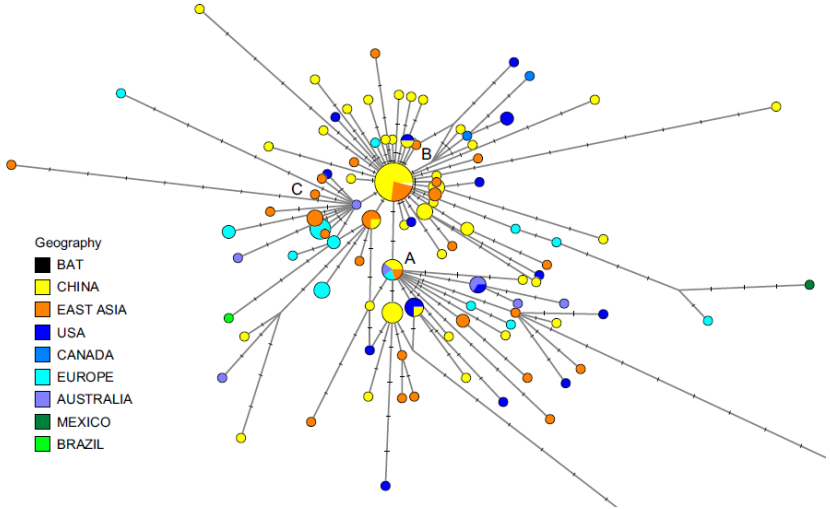




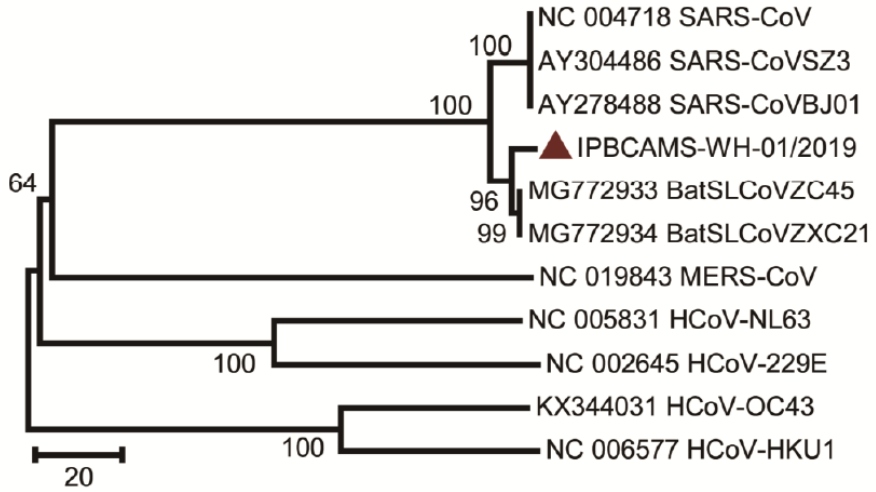
Obermeyer, F. et al. Science 376:1327-1332 (2022)

Phylogenetics

Phylogeny



Forster *et al*. PNAS (2020)



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

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

Types of phylogeny

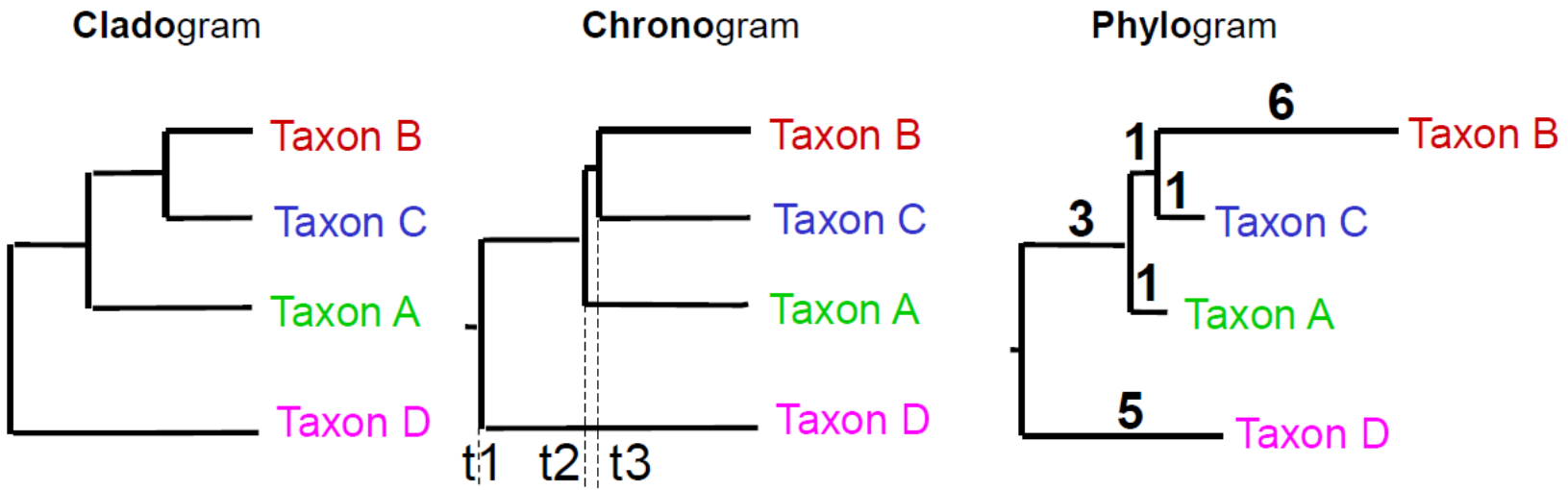


Image from Lecture 18 6.047 MIT OCW

- The choice depends on research question

Ingredients for phylogenetic reconstruction

Multiple sequence alignment

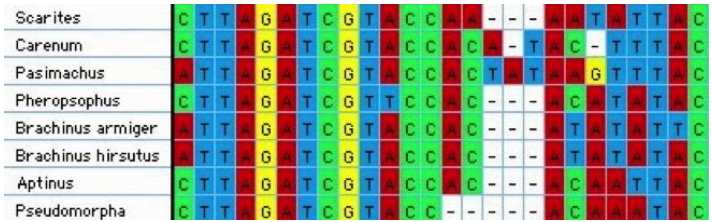
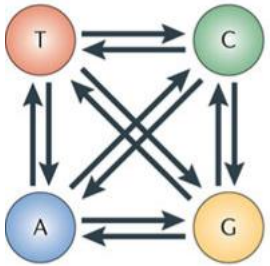


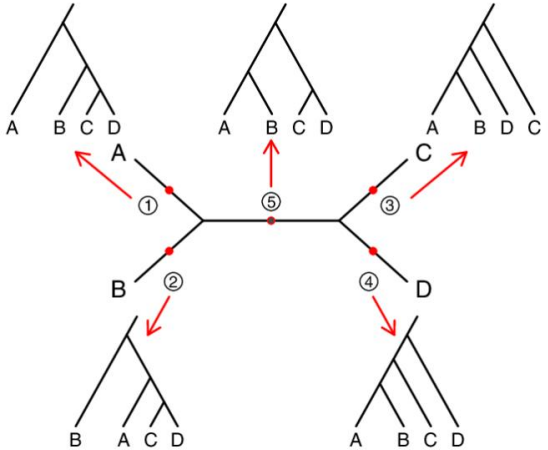
Image from www.mcqbiology.com

Evolutionary model



Yang & Rannala. Nat Rev Genetics (2012)

Tree building algorithm

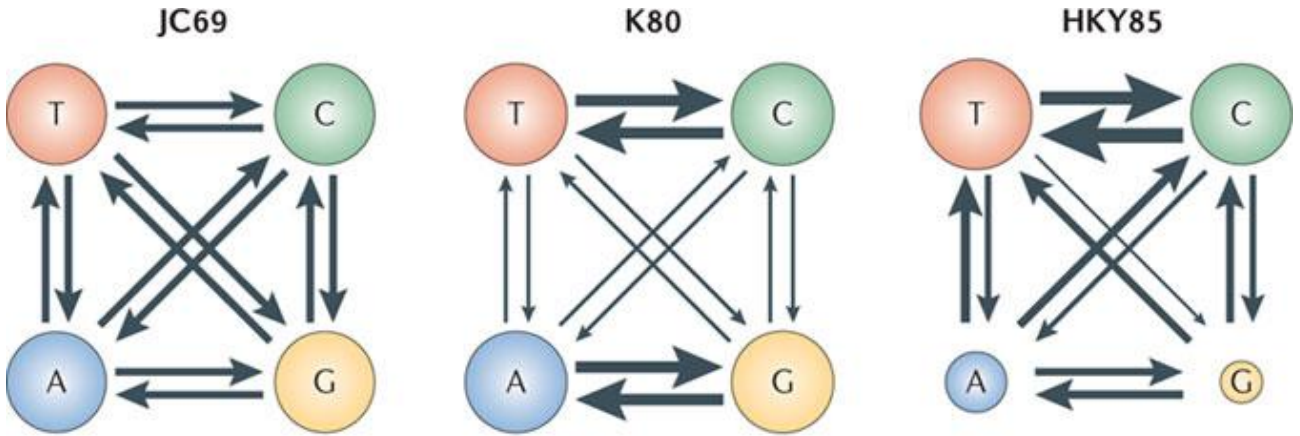


Tian, Y. and Kubatko, L.S. BMC Evol Biol 17(1) (2017)

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

Evolutionary (substitution) model

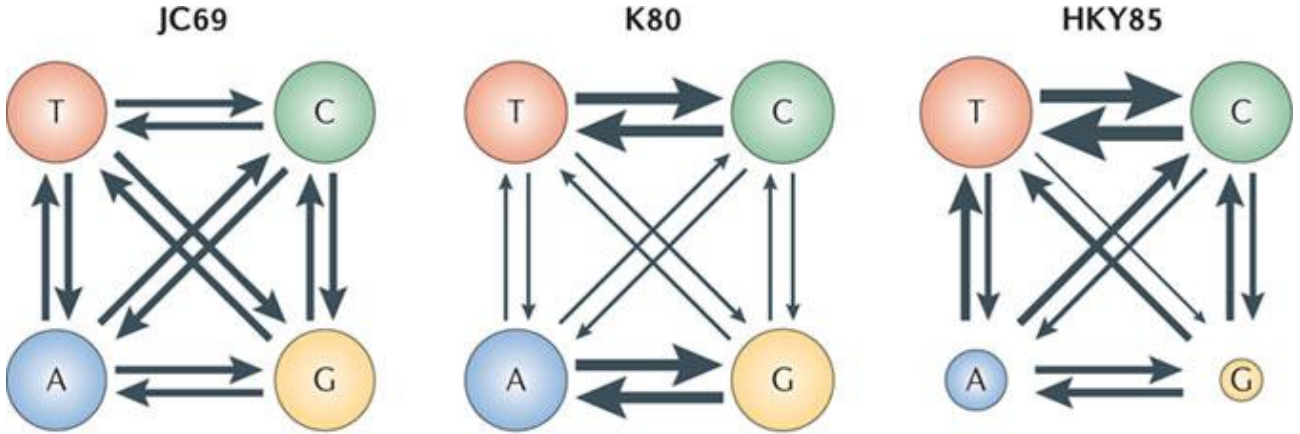
Basic nucleotide substitution models



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)

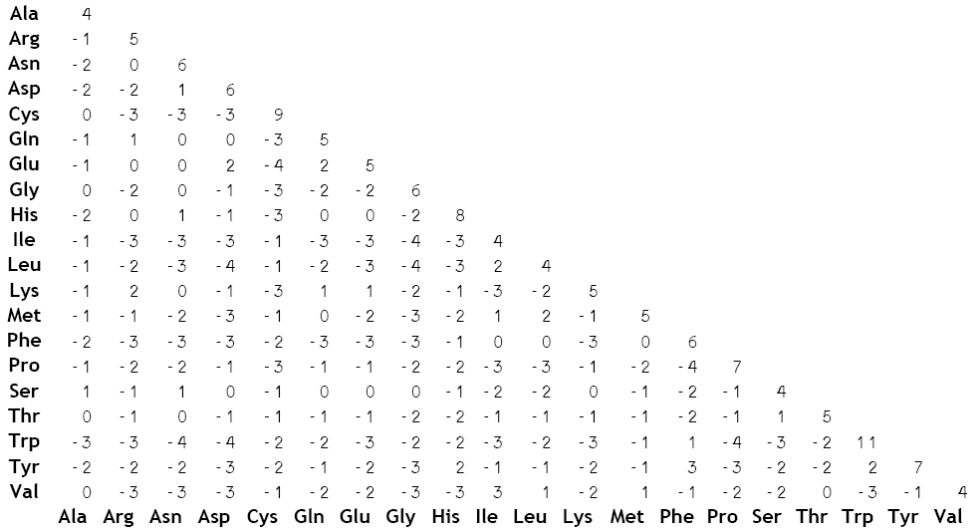
- Symmetric substitution rates: P(A→G) = P(G→A)
  - 6 parameters (4 nucleotides choose 2)
  - Time-reversible: switching ancestral and descendant taxa does not change the calculation
  - Useful in practice because we often don't know which taxon came first

- Different base frequencies
  - 3 parameters

- This is the most generalized time-reversible model possible
- Often used

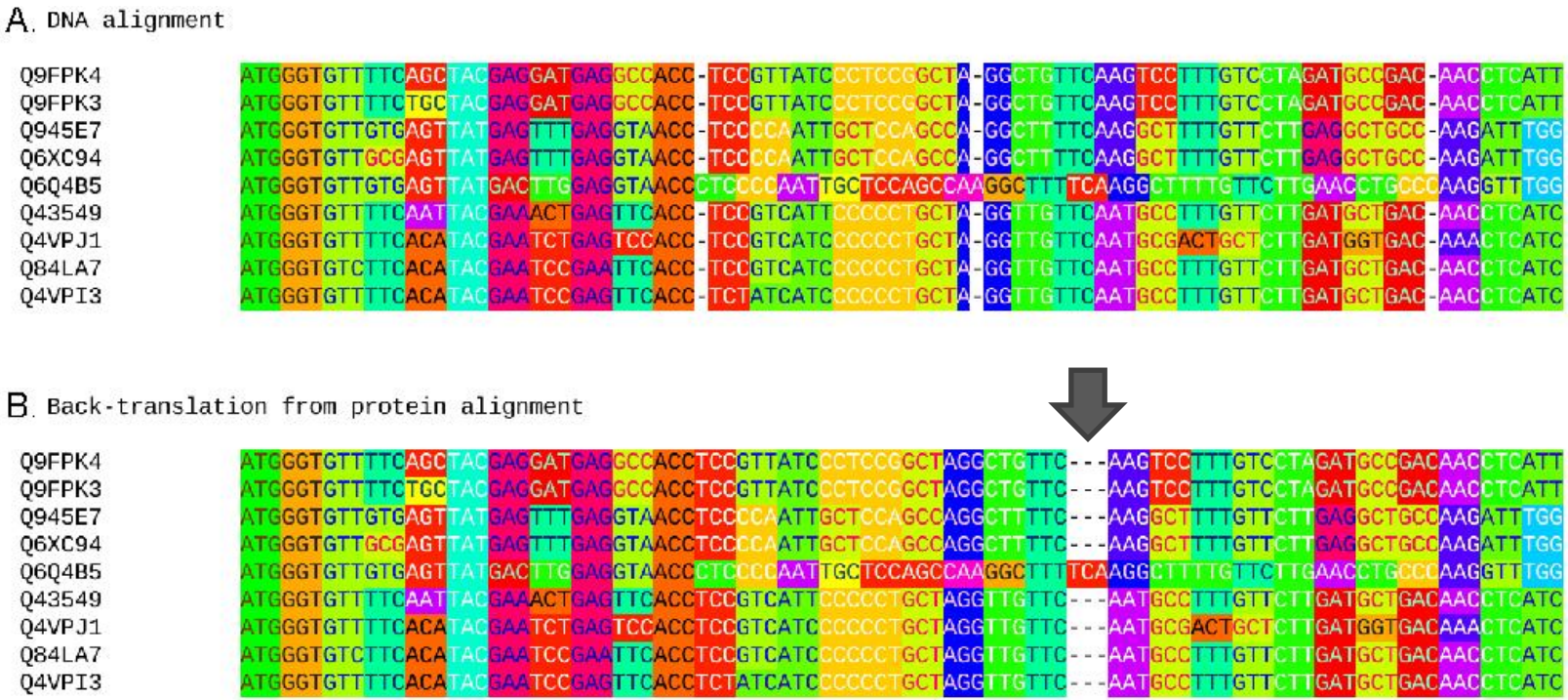
Basic amino acid substitution models

- 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	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

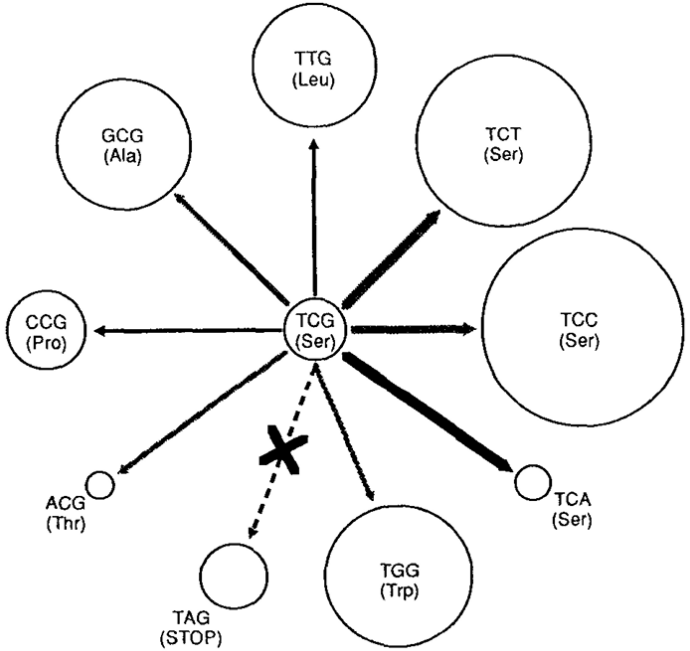
Codon alignment: amino acid → nucleotide



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

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



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

Tree building algorithms

UPGMA: Unweighted pair-group method arithmatic 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

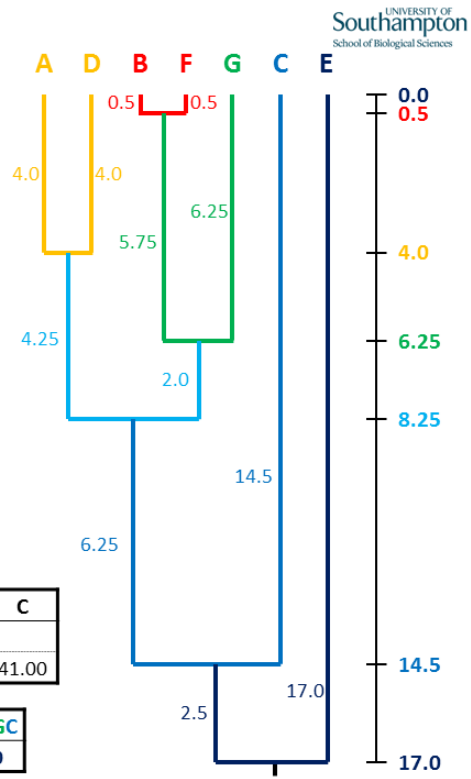
4

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

5

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**



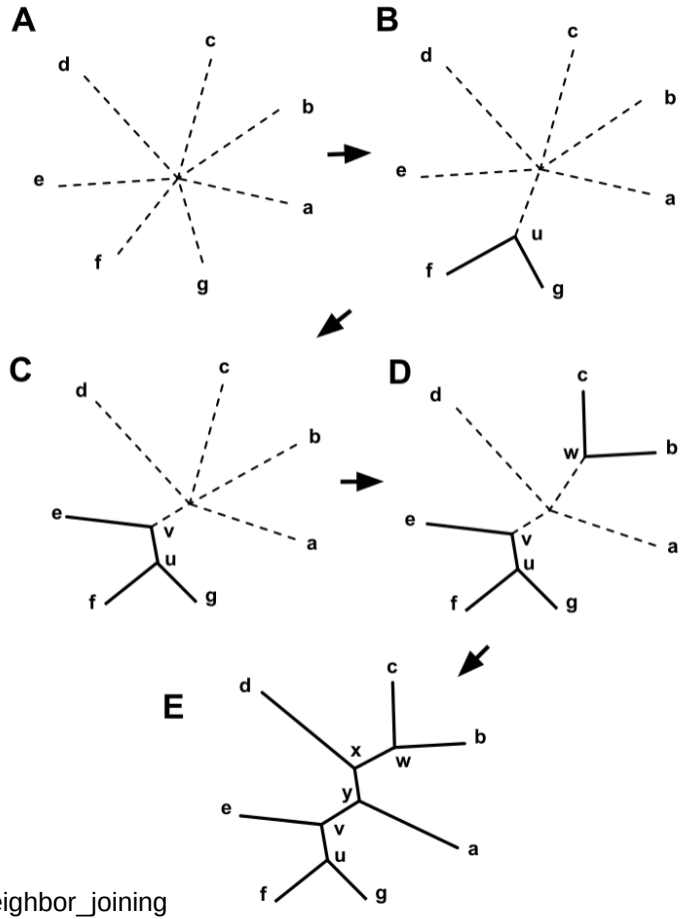
http://www.slimsuite.unsw.edu.au/teaching/upgma/

Neighbor joining

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

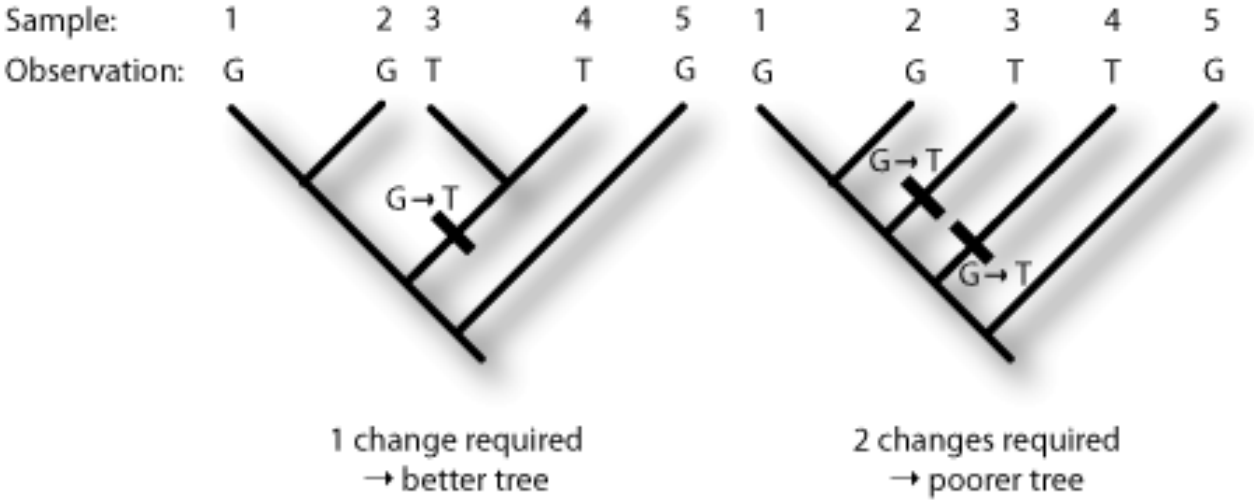
$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n}[d(i,k) + d(j,k)]$$

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



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

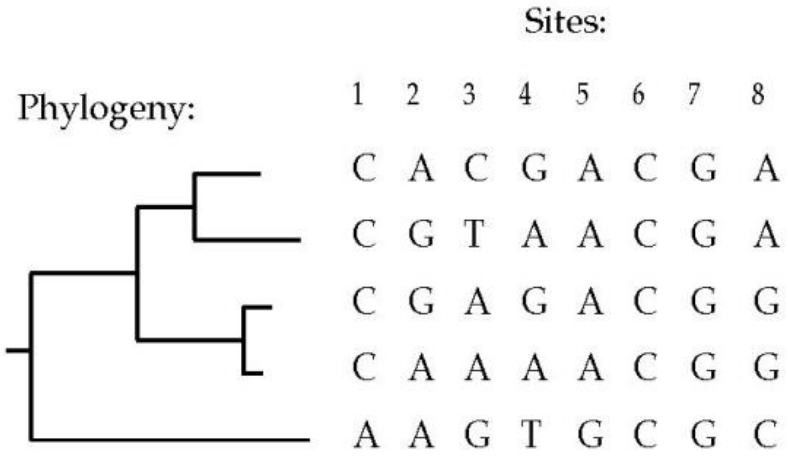
Maximum parsimony / Minimum evolution



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

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

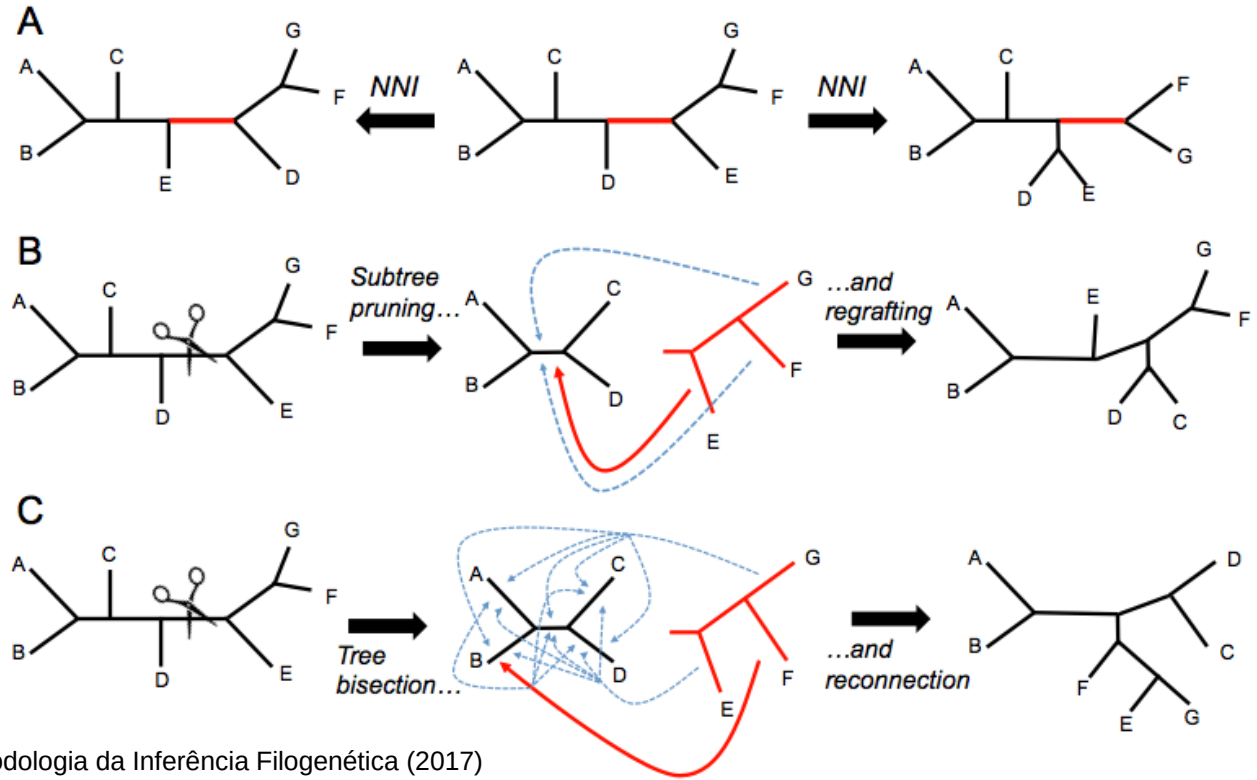
Maximum likelihood approach



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

- Likelihood = P(sequence data | 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

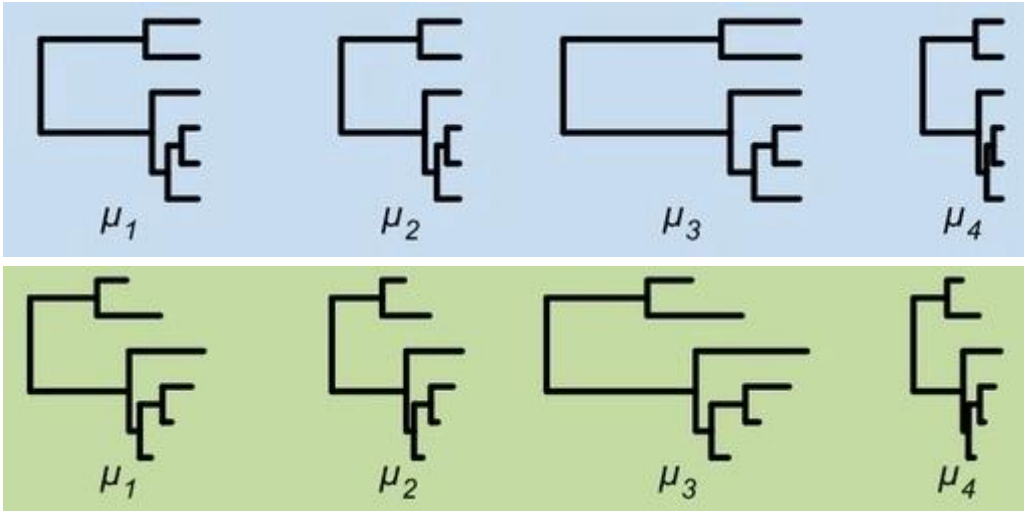
Heuristic tree search algorithms



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

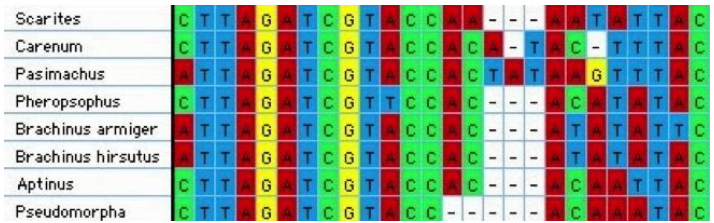
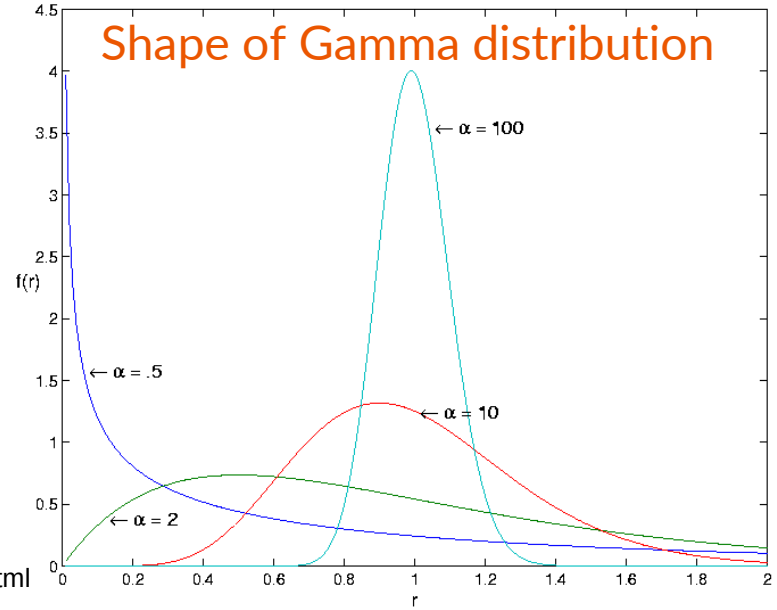


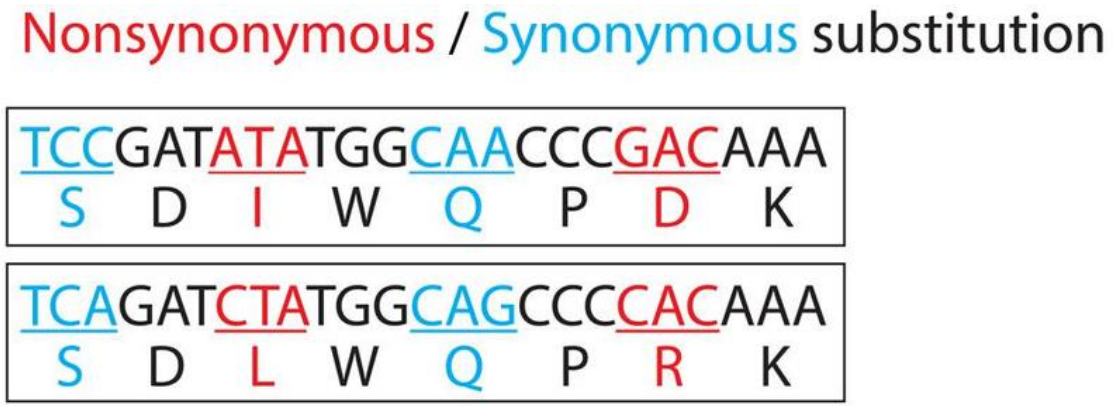
Image from www.mcqbiology.com



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

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

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	-lnL
JC69	3585.54820
F81	3508.04085
HKY85	3233.34395
TrN93	3232.29439

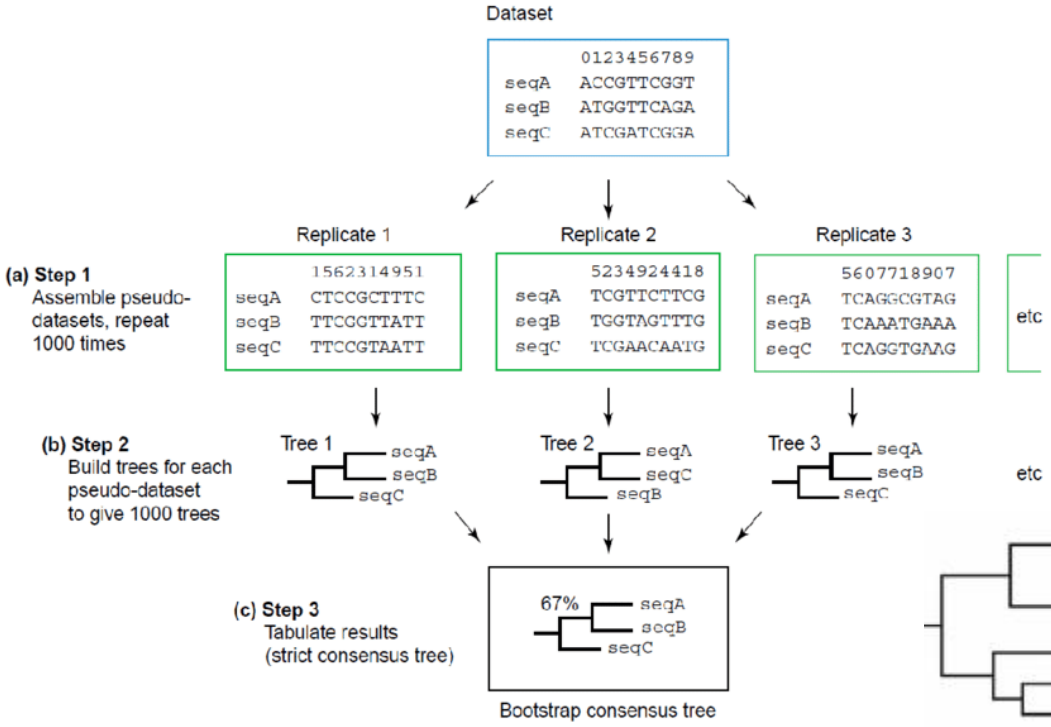
models	diff. DF = q	X^2	P	
JC-F81	3 - 0 = 3	155	0	F81 model fits the data significantly better than JC
F81-HKY85	4 – 3 = 1	549.4	0	HKY85 model fits the data significantly better than F81
KHY-TrN	5 – 4 = 1	2.1	0.15	TrN model does not fit the data significantly better than HKY85

Model	-lnL
HKY85	3233.34395
HKY85 +G	3145.29031
HKY85 +I+G	3142.36439

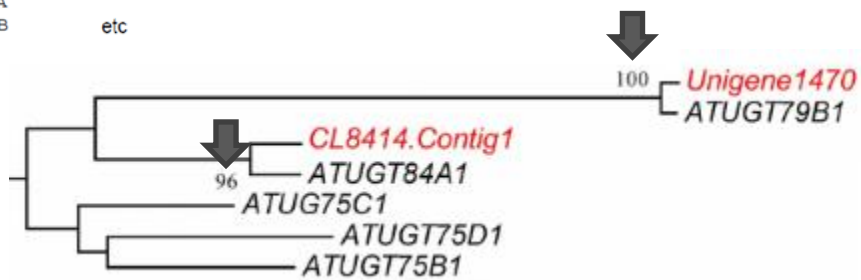
models	diff. DF = q	X^2	P	
HKY85-vs. +G	1	176	0	Adding site-specific rate fits the data significantly better
HKY85+G vs. I+G	1	5.85	0.015	Adding invariant sites does not fit the data significantly better

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

Bootstrap support for taxa group



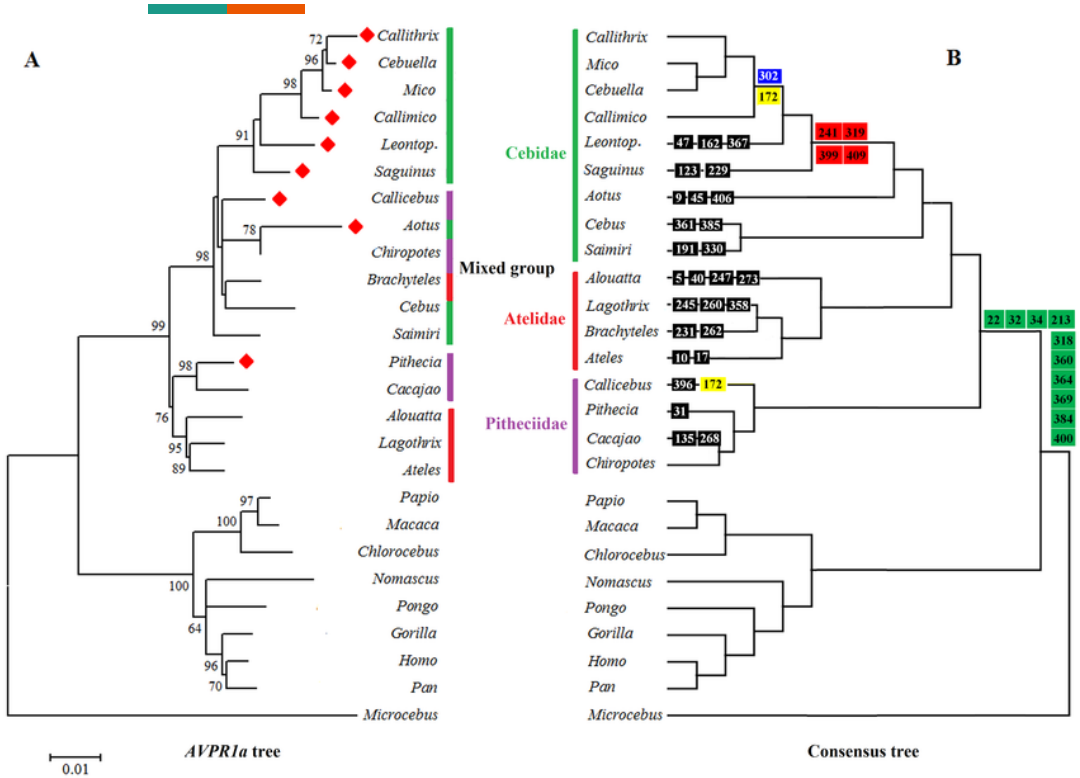
- Bootstrapping = sampling with repeats
- Allow multiple trees to be constructed from a single sequence alignment



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

Wei et al. PLoS ONE 2013

Gene tree vs concatenated tree

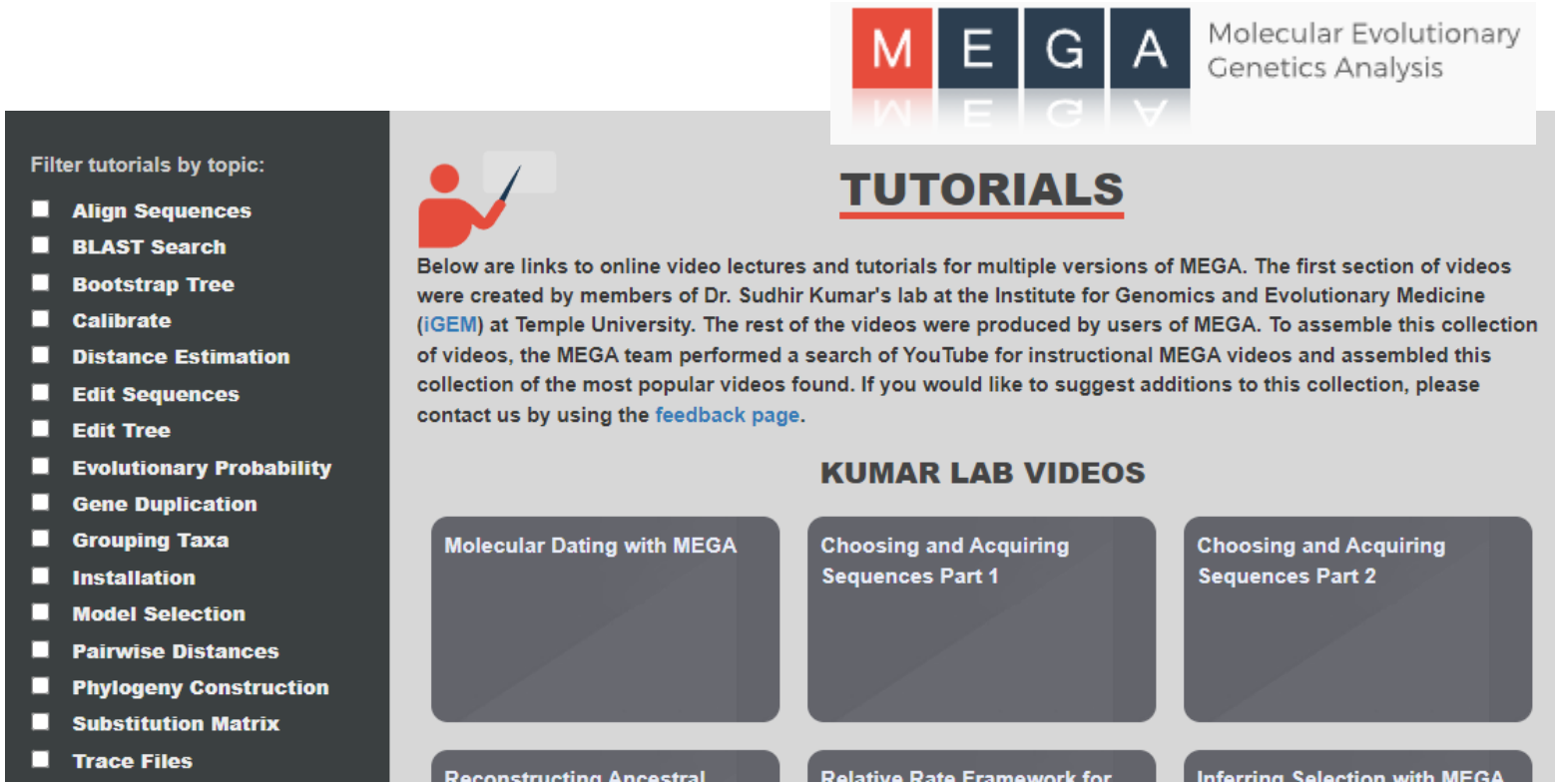


Ren, D. et al. PLoS ONE 9:e222638 (2014)

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

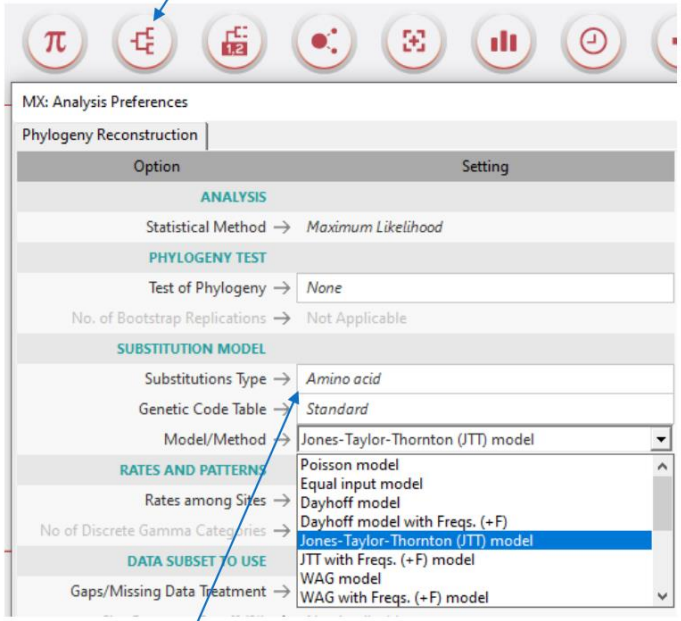
Example of how to setup phylogenetic reconstruction

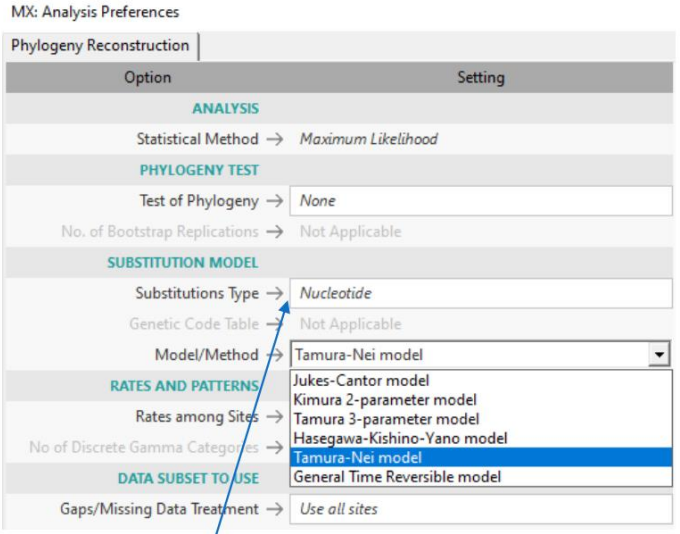
MEGA: A Windows GUI tool for phylogenetic analysis



Substitution model choices

PHYLOGENY mode

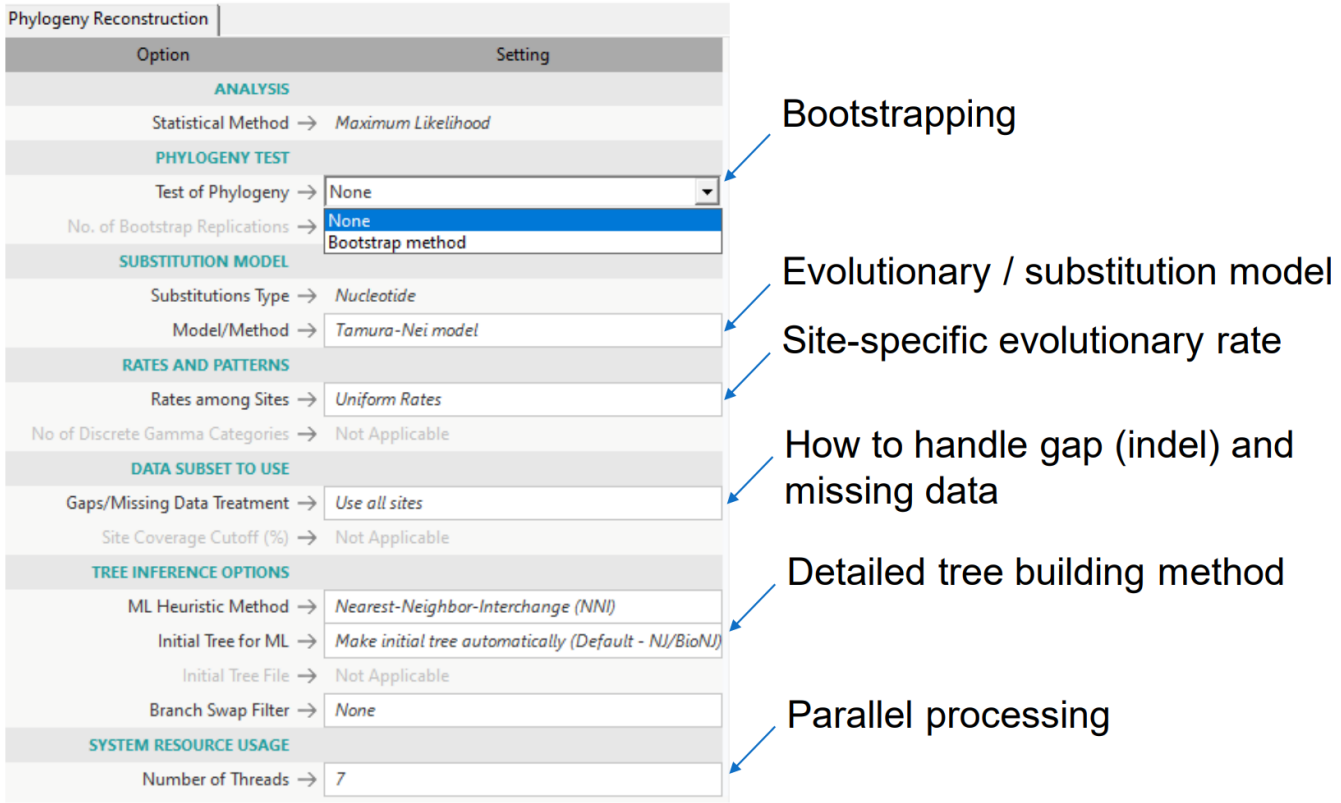




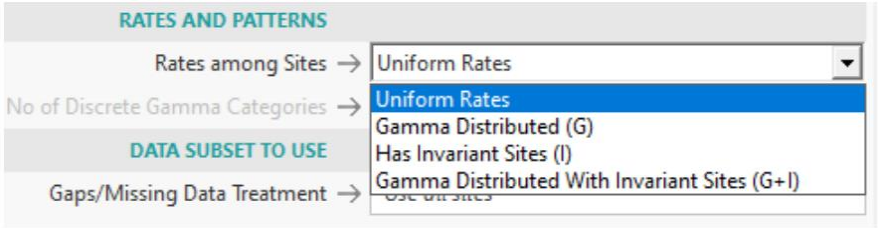
Amino acid-based models

Nucleotide-based models

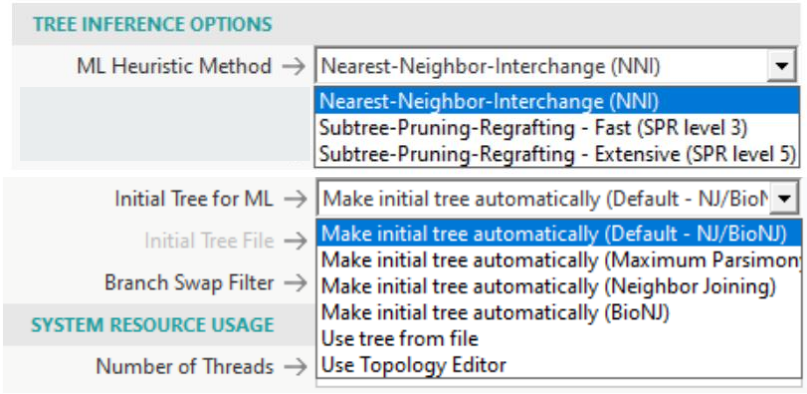
Maximum likelihood parameters



Maximum likelihood parameters



Invariant site and Gamma-distributed site-specific substitution rates



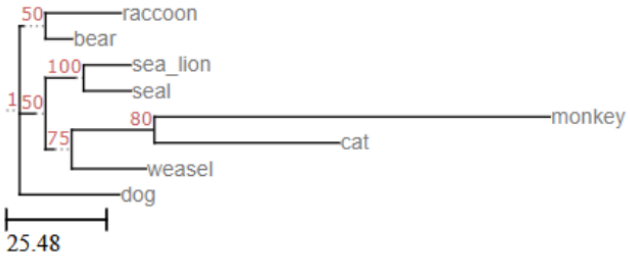
Tree search method

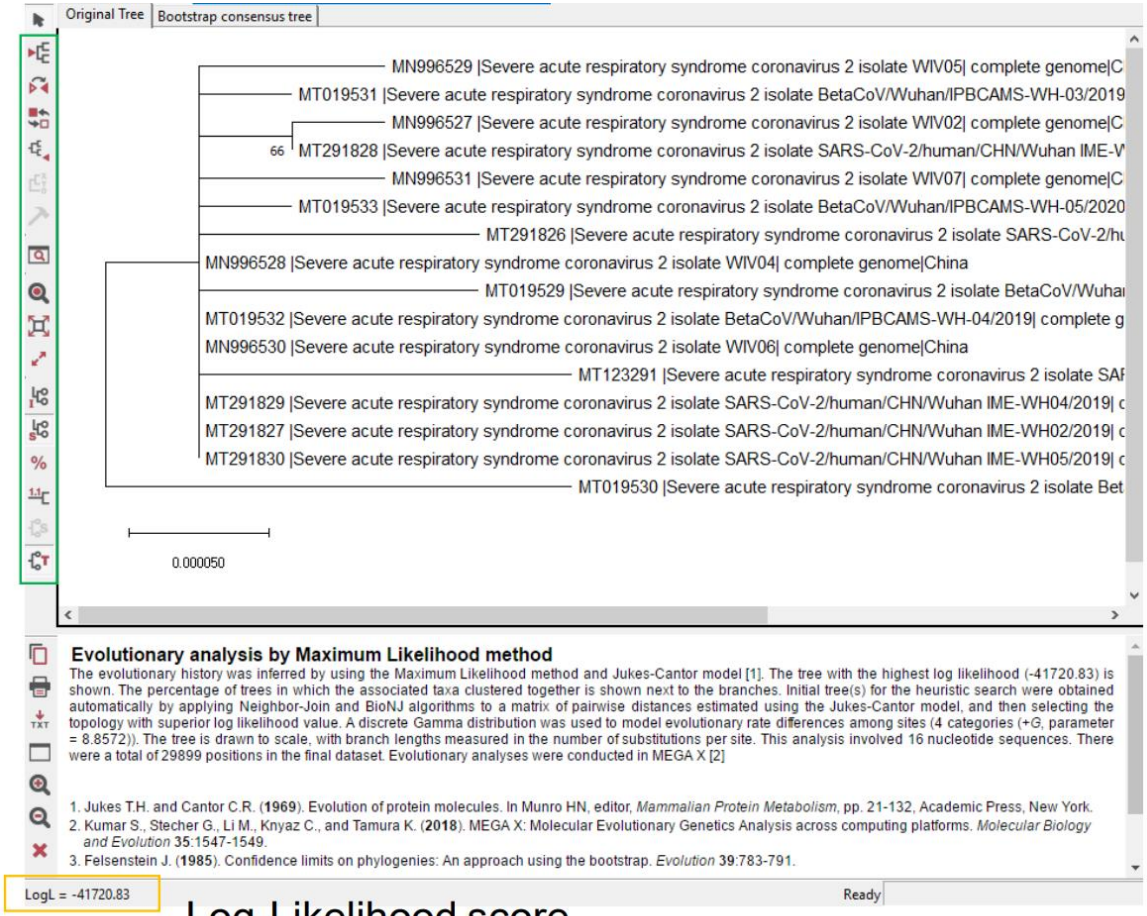
Initial tree can be built using quick and simple method like neighbor joining

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);

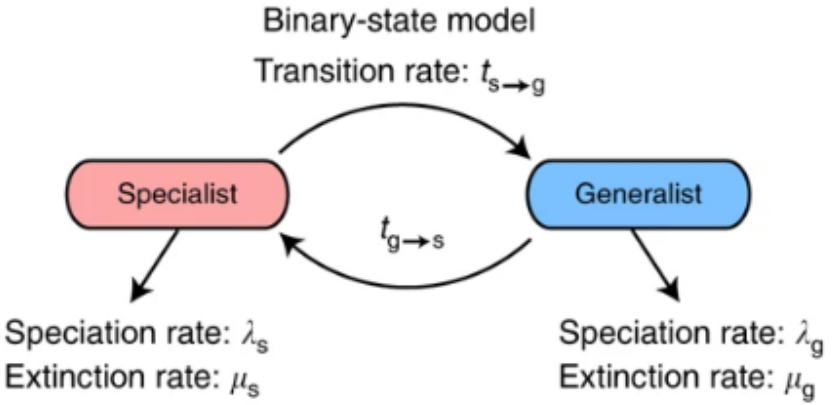




Log-Likelihood score

More molecular evolution analyses

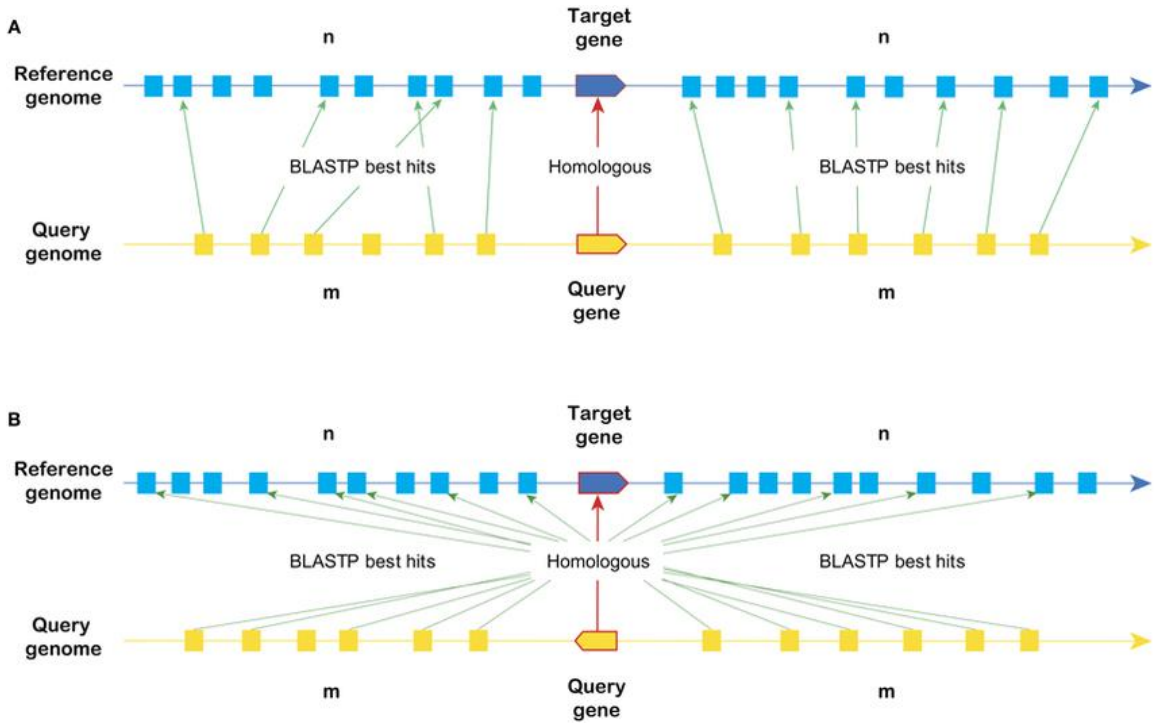
Impact of character state on evolutionary process



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

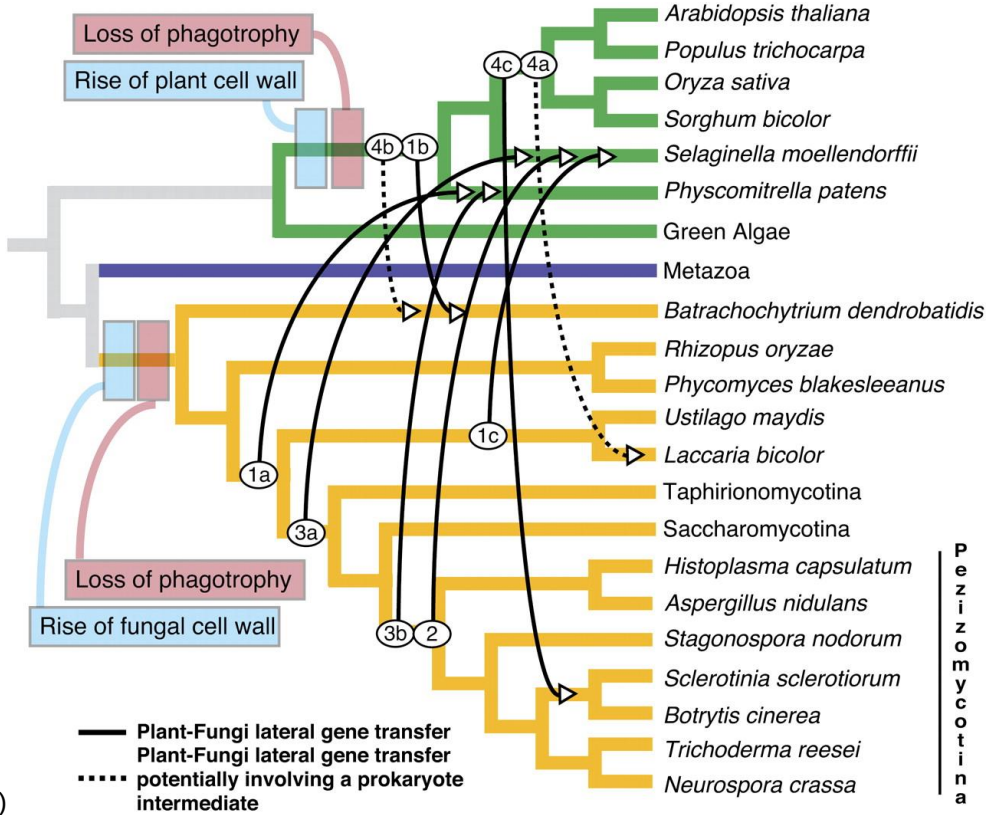
Synteny: conservation of gene order on chromosome



Cheng, F. et al. Frontiers in Plant Science 3:198 (2012)

Horizontal gene transfer (HGT)

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



Richards *et al*. The Plant Cell (2009)

Any question?