Phylogenetic methods for tree inference

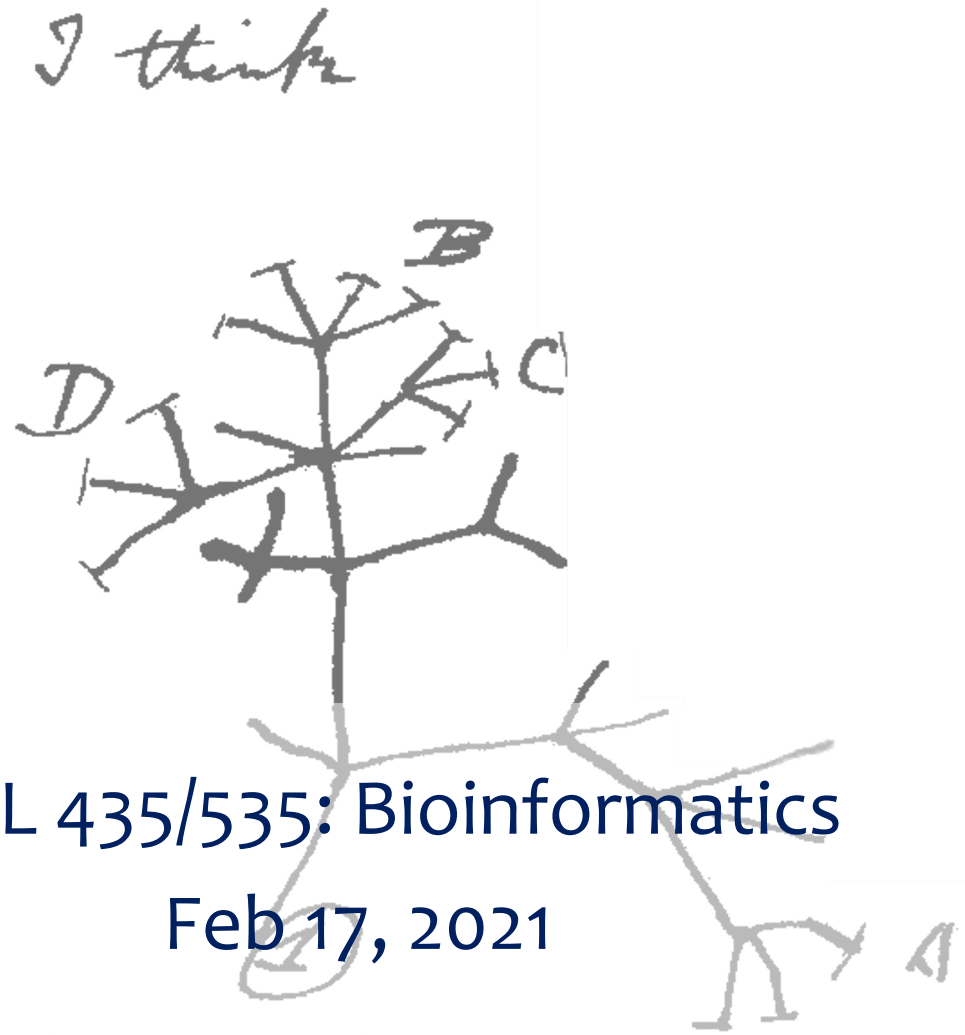
Distance-based methods

- NJ

Parsimony

Maximum Likelihood

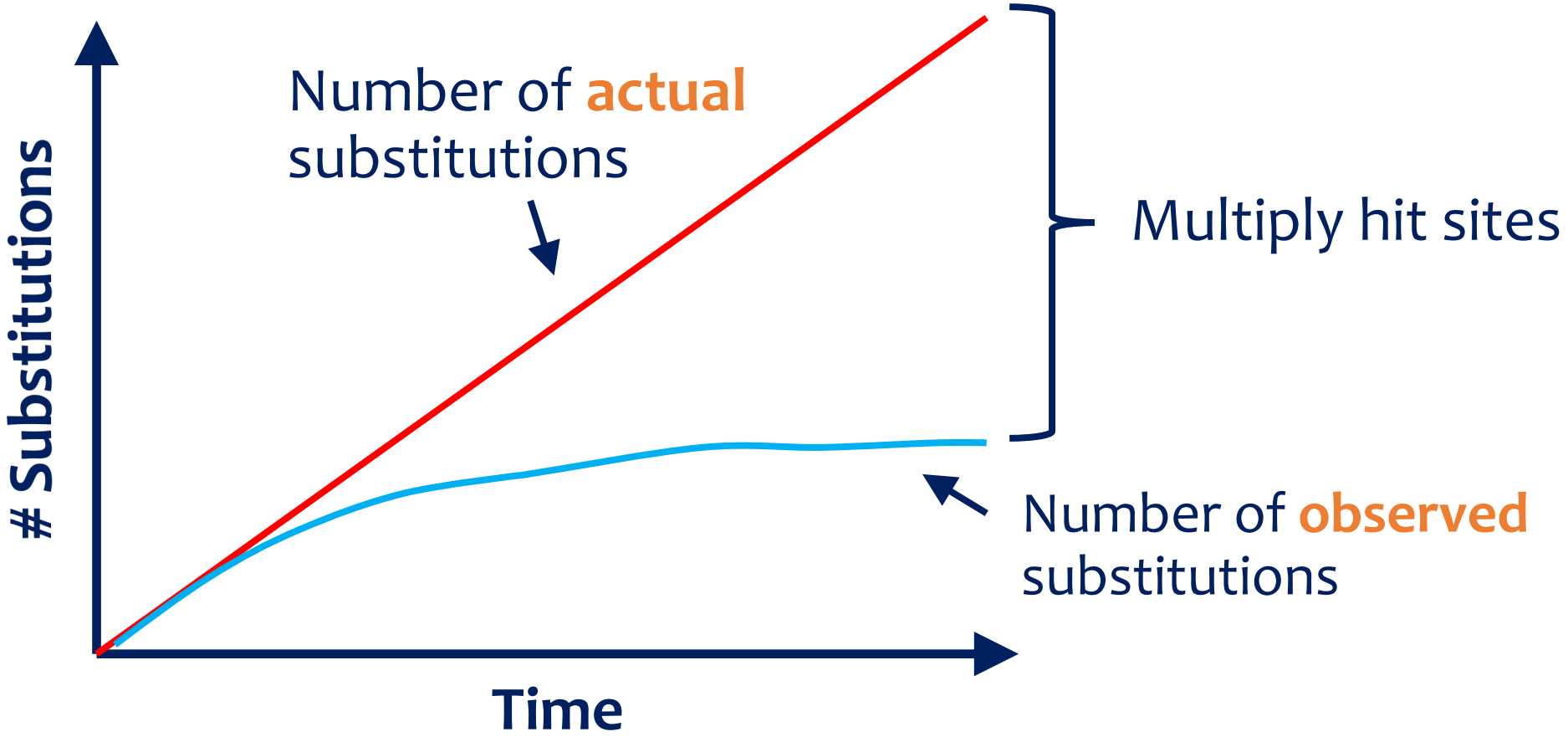
Bayesian



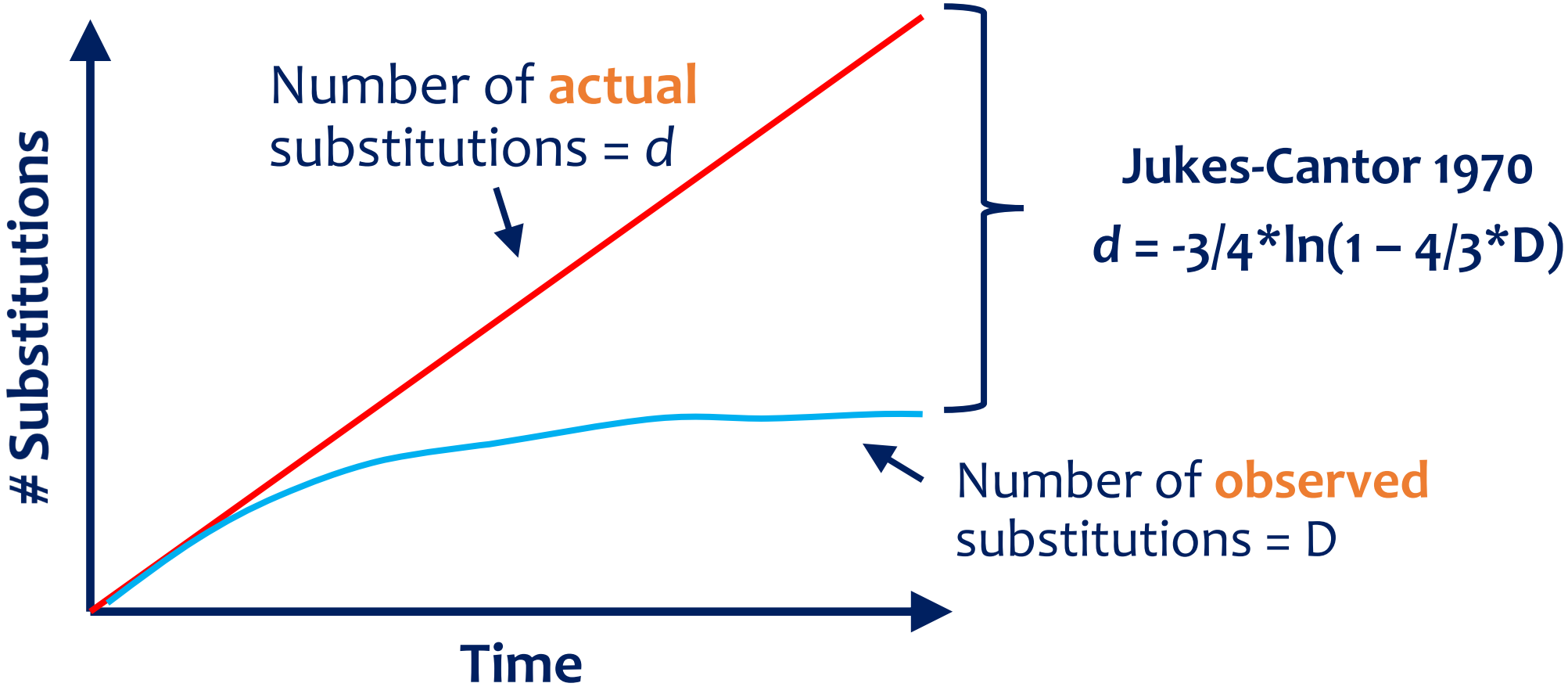
BIOL 435/535: Bioinformatics

Feb 17, 2021

Models of Molecular Evolution



Models of Molecular Evolution

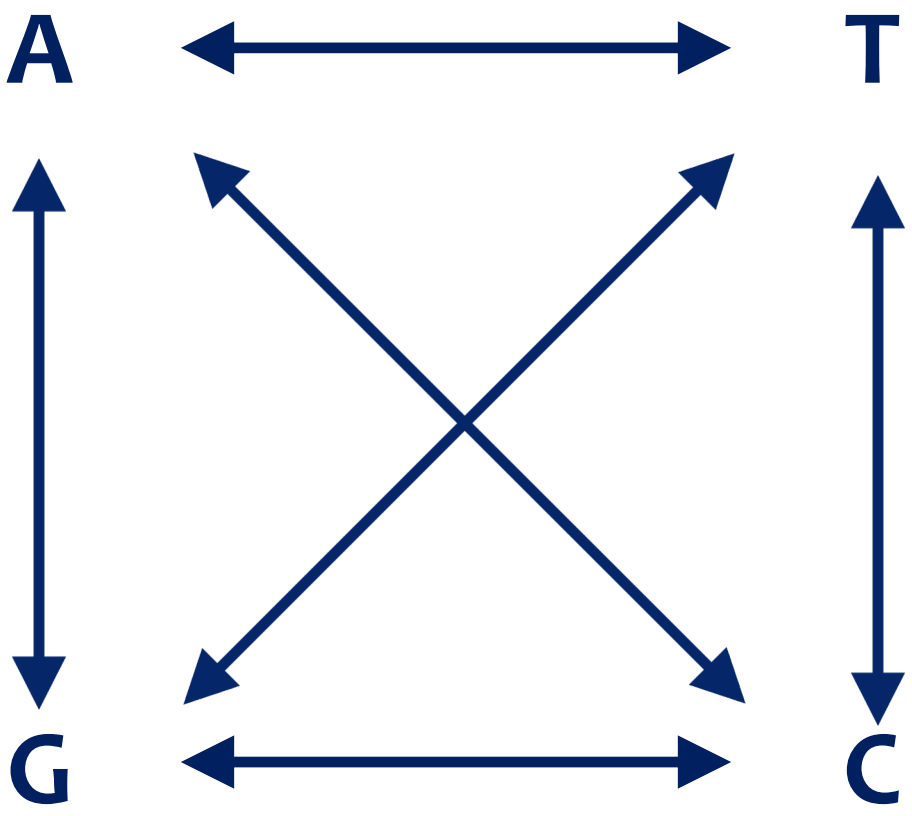


Models of Molecular Evolution

Jukes Cantor 1970

- **Equal probabilities** of different probabilities

- Useful when species are <20% divergent

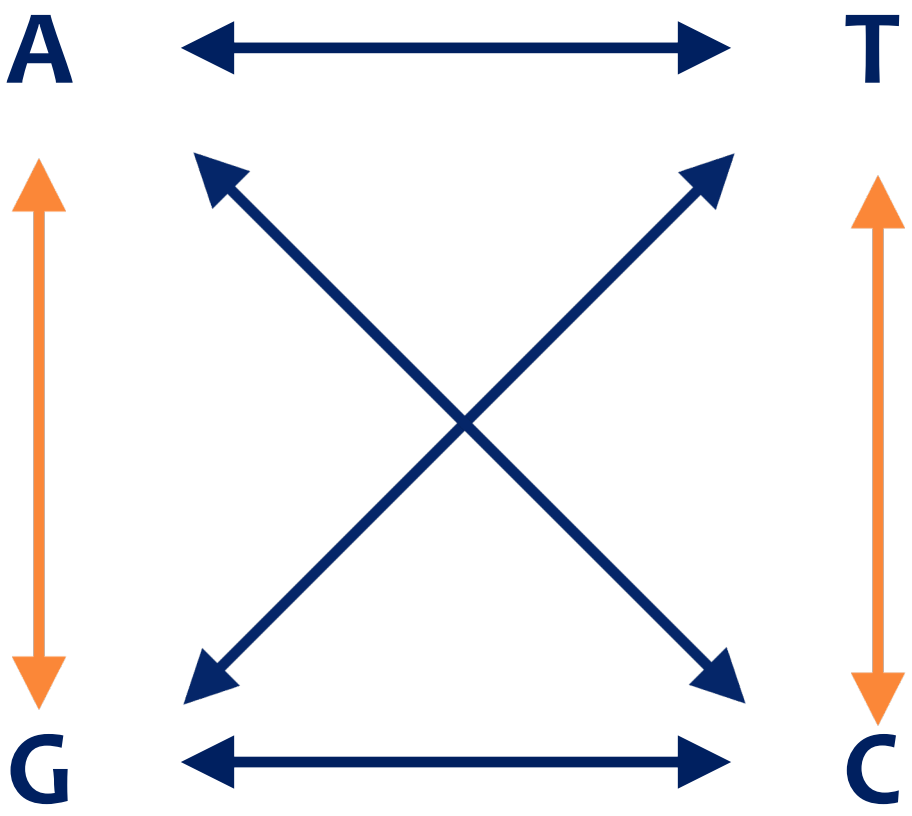


$$d = -3/4*\ln(1 - 4/3*D)$$

Models of Molecular Evolution

Kimura 2 parameter

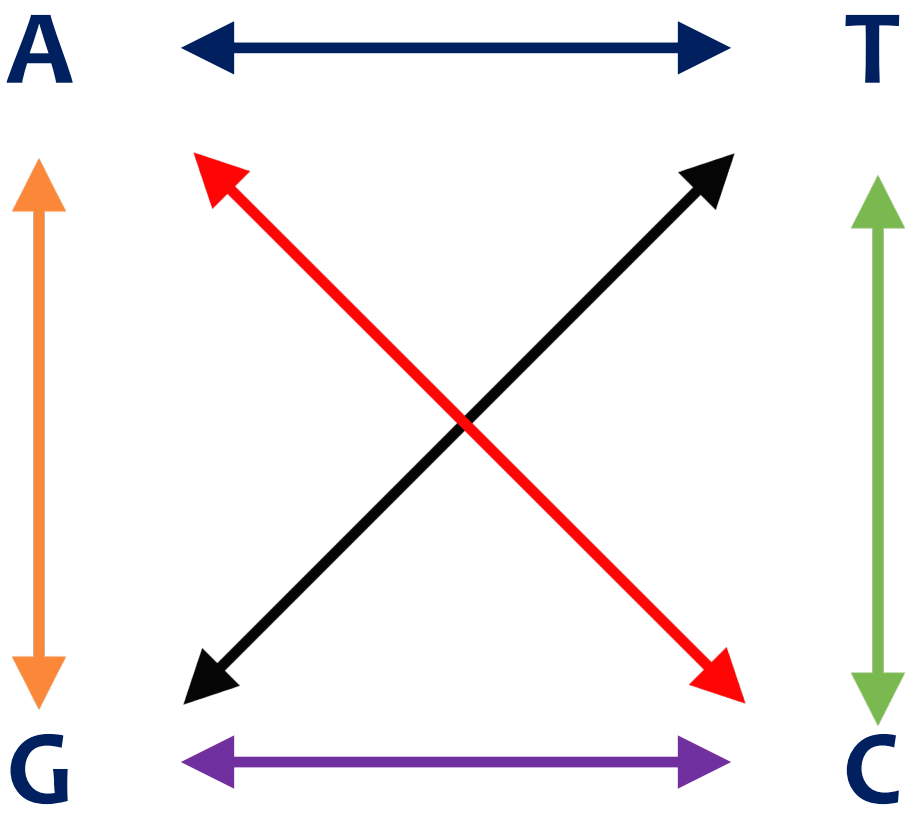
- **Transitions and transversions** have different probabilities



Models of Molecular Evolution

General Time Reversible (GTR)

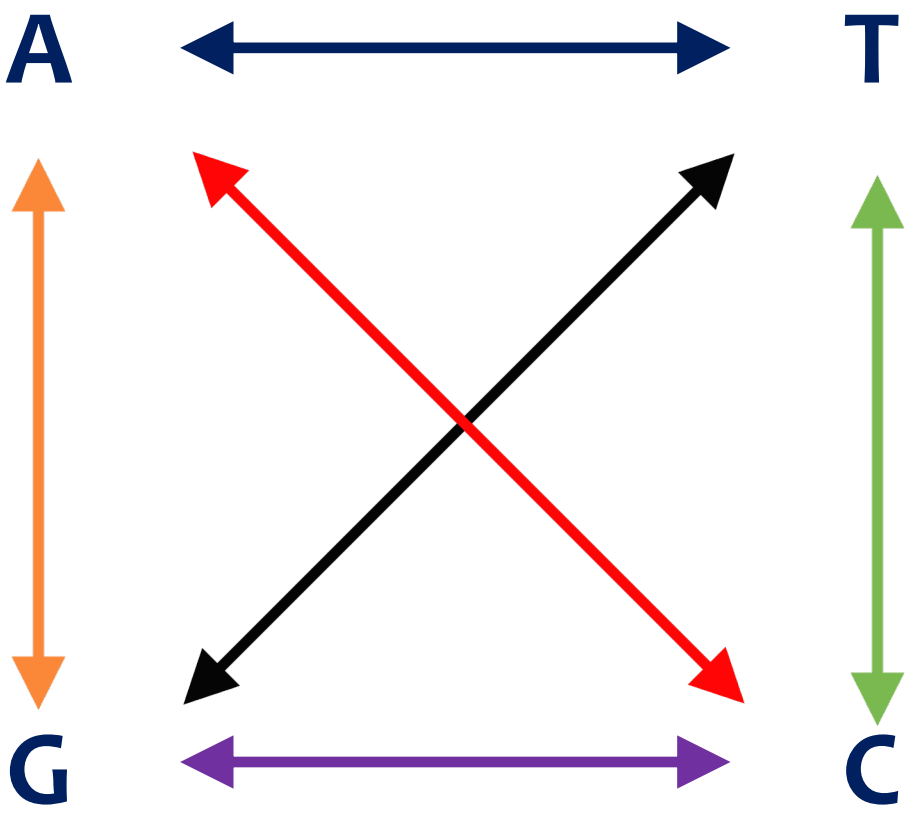
- **All mutations** have different probabilities
- Only appropriate for very divergent taxa (parameter rich)



Model selection tools

jModelTest2

ProtTest



Phylogenetic methods for tree inference

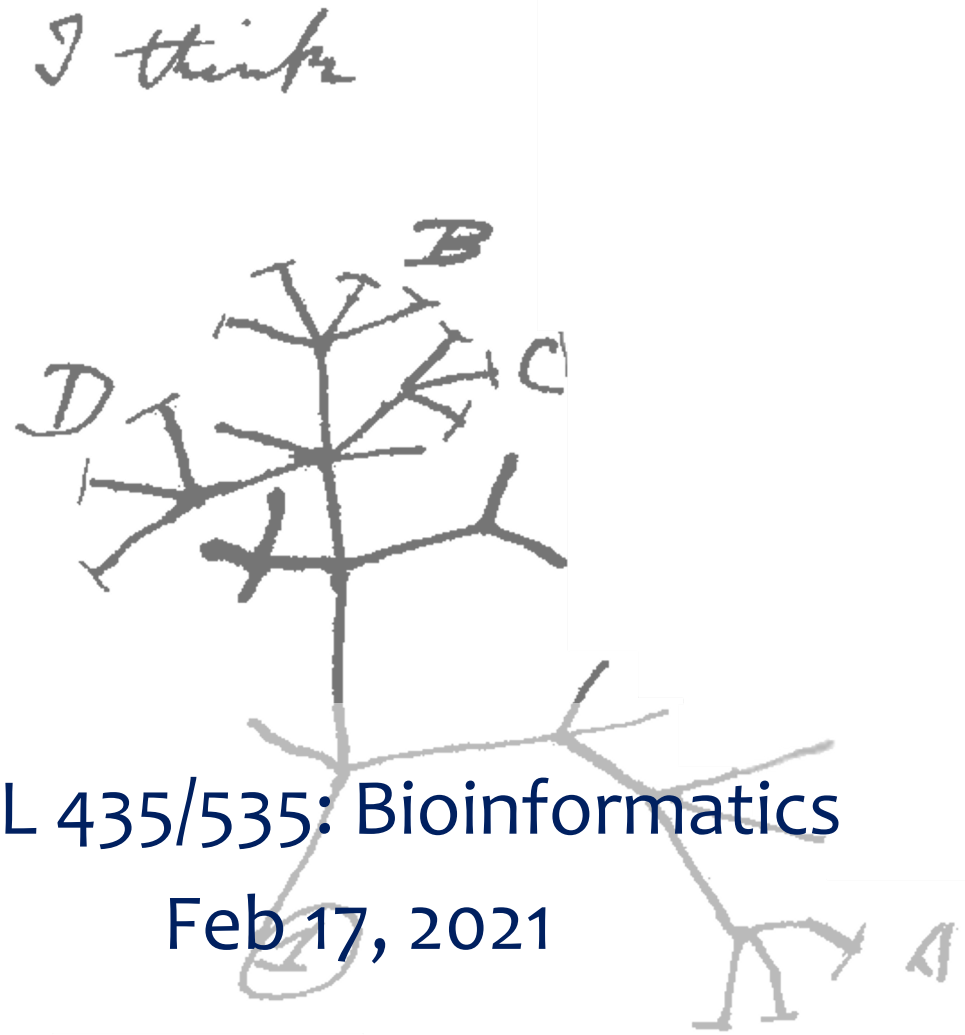
Distance-based methods

- NJ

Parsimony

Maximum Likelihood

Bayesian



BIOL 435/535: Bioinformatics

Feb 17, 2021

Distance Methods – Neighbor Joining

Pros:

- Easy, fast, computationally tractable
- Gives you a single tree

Cons:

- Distance != evolutionary history
- Gives you a single tree
- **Homoplasy – Shared character from different ancestral origins**

Works reasonably well in 95% of cases – but very poorly in the remaining 5%

Distance Methods – Neighbor Joining

1. Calculate pairwise differences (or % differences) in multiple sequence alignment, put in matrix

2. Identify pair with smallest difference, join

3. Re-calculate pairwise distance matrix

4. Repeat until all taxa are placed in the tree

Distance Methods – Neighbor Joining

MSA.NJ.fasta

Distance Methods – Neighbor Joining

MSA.NJ.fasta

RAW

H. vulgare	-	4.2%	4.5%	5.1%
T. urartu	24	-	1.1%	1.4%
A. speltoides	26	6	-	2.1%
A. tauschii	29	8	12	-
	H. vulgare	***T. urartu***	***A. speltoides***	***A. tauschii***

Distance Methods – Neighbor Joining

RAW

H. vulgare	-	4.2%	4.5%	5.1%
T. urartu	24	-	1.1%	1.4%
A. speltoides	26	6	-	2.1%
A. tauschii	29	8	12	-
	H. vulgare	***T. urartu***	***A. speltoides***	***A. tauschii***

JC-corrected

H. vulgare	-	4.3%	4.6%	5.3%
T. urartu	24.77	-	1.1%	1.4%
A. speltoides	26.59	6.35	-	2.1%
A. tauschii	30.26	8.10	12.21	-
	H. vulgare	***T. urartu***	***A. speltoides***	***A. tauschii***

Distance Methods – Neighbor Joining

T. urartu *A. speltoides*

H. vulgare	-			
T. urartu	24	-		
A. speltoides	26	**6**	-	
A. tauschii	29	8	12	-
	H. vulgare	***T. urartu***	***A. speltoides***	***A. tauschii***

Distance Methods – Neighbor Joining

T. urartu *A. speltoides*

H. vulgare	-			
T. urartu + A. speltoides	24	-		
	26			
A. tauschii	29	8	12	-
	H. vulgare	***T. urartu + A. speltoides***		***A. tauschii***

Distance Methods – Neighbor Joining

T. urartu *A. speltoides*

H. vulgare	-			
T. urartu + A. speltoides	**25.0**	-		
A. tauschii	29	8	12	-
	H. vulgare	***T. urartu + A. speltoides***		***A. tauschii***

Distance Methods – Neighbor Joining

T. urartu *A. speltoides*

H. vulgare	-		
T. urartu + A. speltoides	25.0	-	
A. tauschii	29	**10.0**	-
	H. vulgare	***T. urartu + A. speltoides***	***A. tauschii***

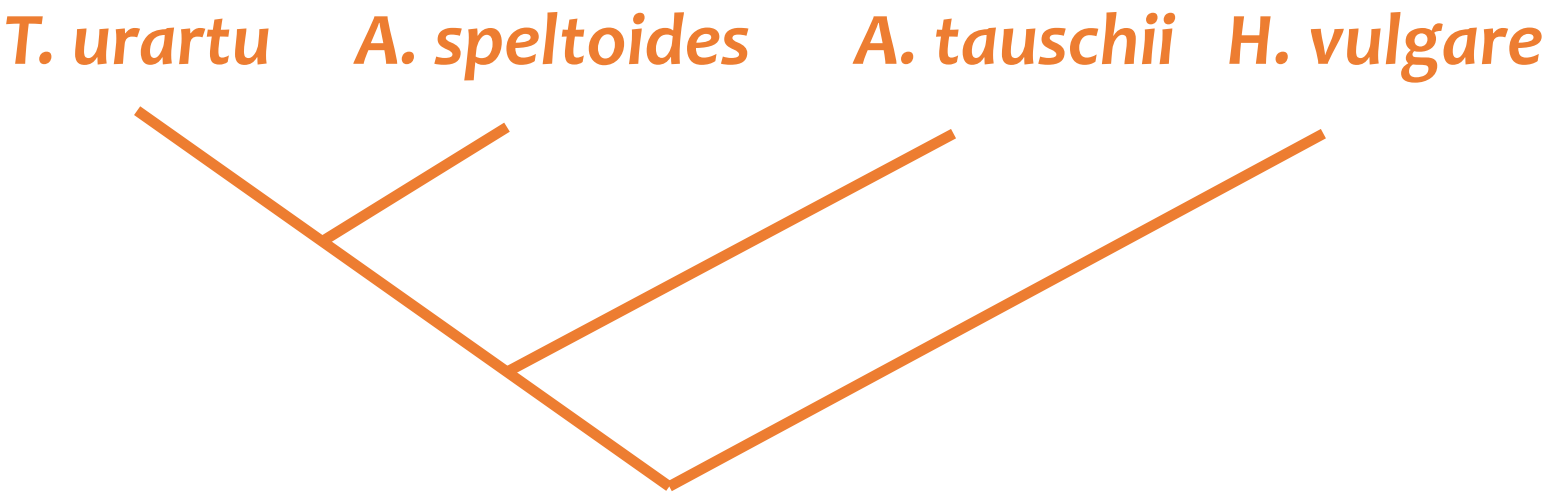
Distance Methods – Neighbor Joining

H. vulgare	-		
T. urartu + A. speltoides	25.0	-	
A. tauschii	29	**10.0**	-
	H. vulgare	***T. urartu + A. speltoides***	***A. tauschii***

Distance Methods – Neighbor Joining

H. vulgare	-		
T. urartu + A. speltoides + A. tauschii	**27.0**	-	
			-
	H. vulgare	*T. urartu + A. speltoides + A. tauschii*	

Distance Methods – Neighbor Joining



H. vulgare	-		
T. urartu + A. speltoides + A. tauschii	**27.0**	-	
		-	
	H. vulgare	*T. urartu + A. speltoides + A. tauschii*	

Distance tools

MAFFT

MEGA

PAUP*

PAML

Parsimony – Fewest number of steps

Pros:

- Searches the entire treespace (all the possible trees)
- Good for morphological data

Cons:

- Searches the entire treespace (all the possible trees)
- Bad for nucleotide sequence data, large number of taxa
- **Homoplasy – Shared character from different ancestral origins**

Parsimony – Fewest number of steps

1. Identify variable sites

2. For all possible trees, calculate the number of steps required to explain the tree given the data

3. Identify the tree with the fewest number of steps

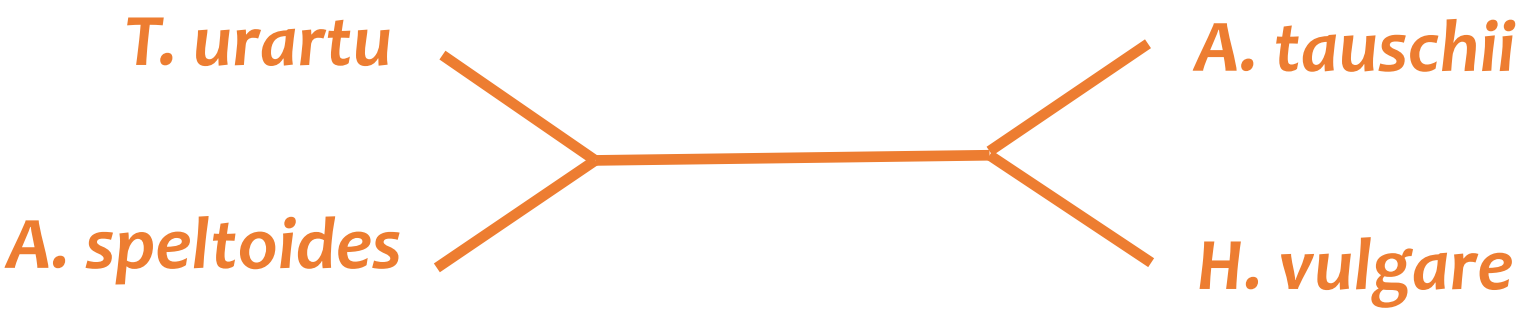
Parsimony – Fewest number of steps

1. Identify variable sites

   MSA.parsimony.fasta

Parsimony – Fewest number of steps

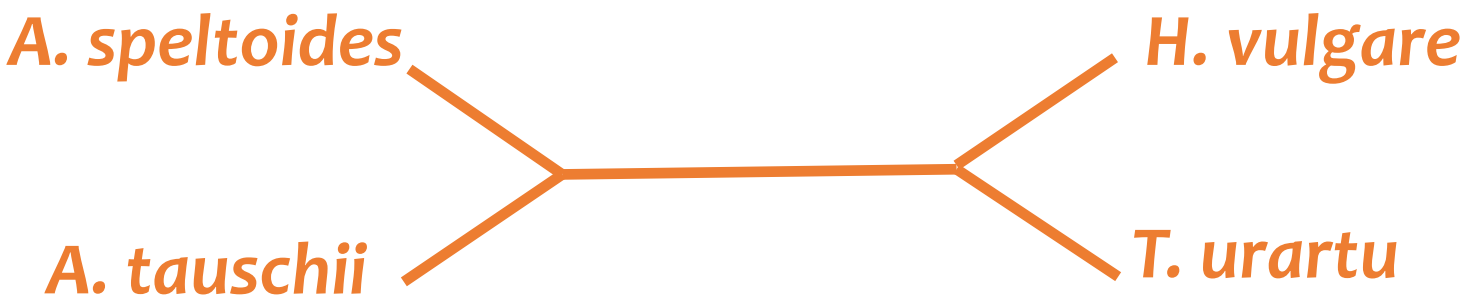
Identify all possible trees



Tree #1

Parsimony – Fewest number of steps

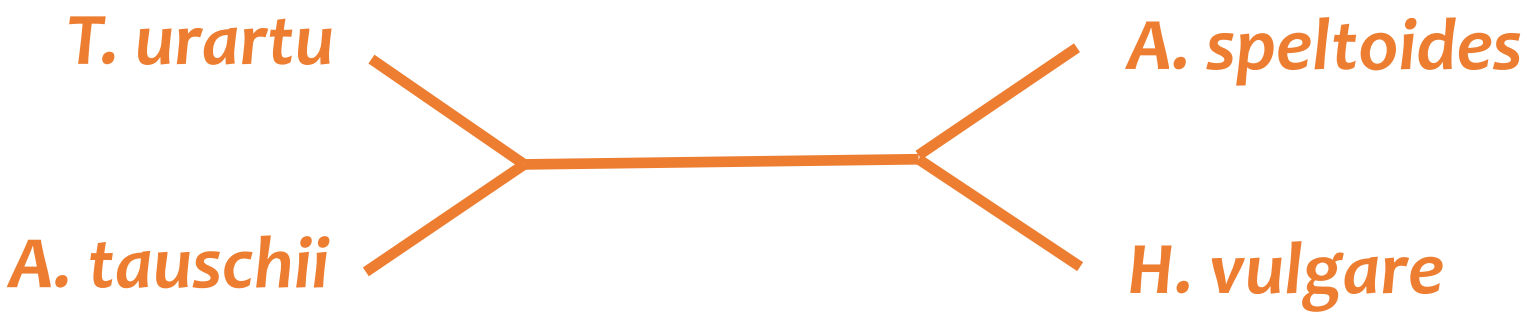
Identify all possible trees



Tree #2

Parsimony – Fewest number of steps

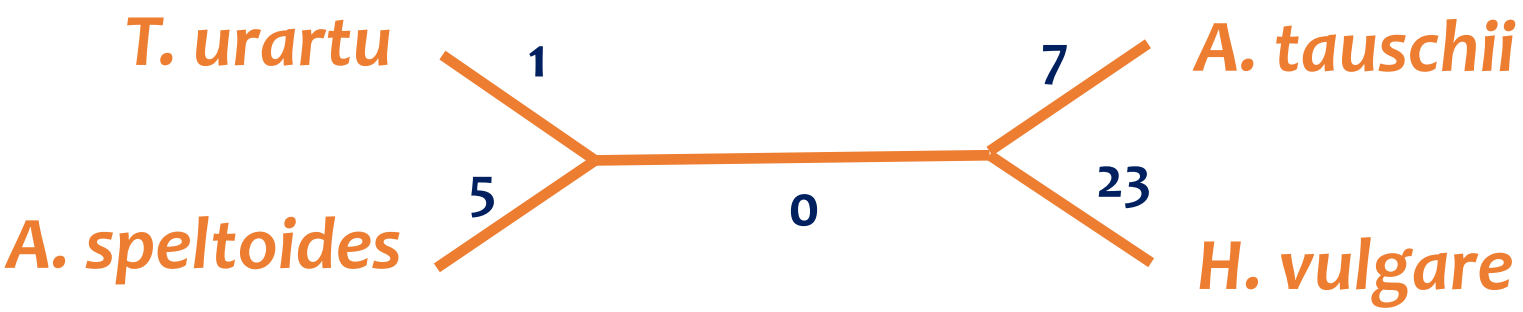
Identify all possible trees



Tree #3

Parsimony – Fewest number of steps

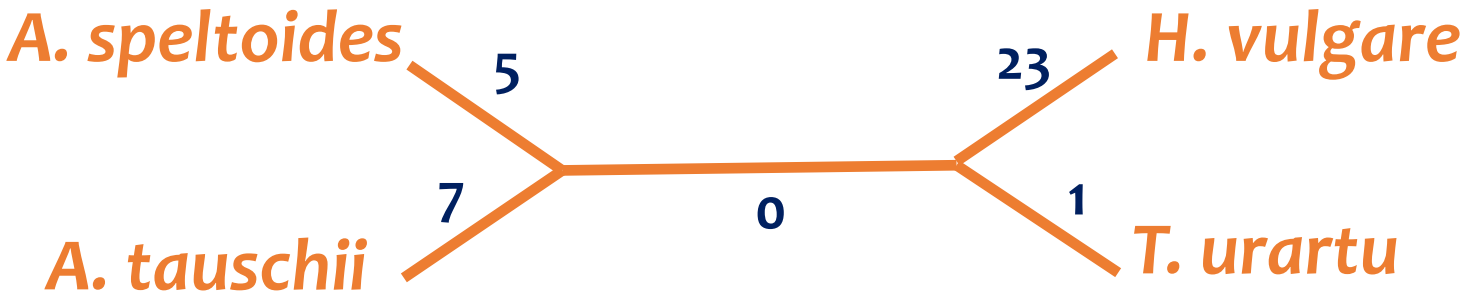
Count number of changes necessary to explain tree



Tree #1 – 36 total changes

Parsimony – Fewest number of steps

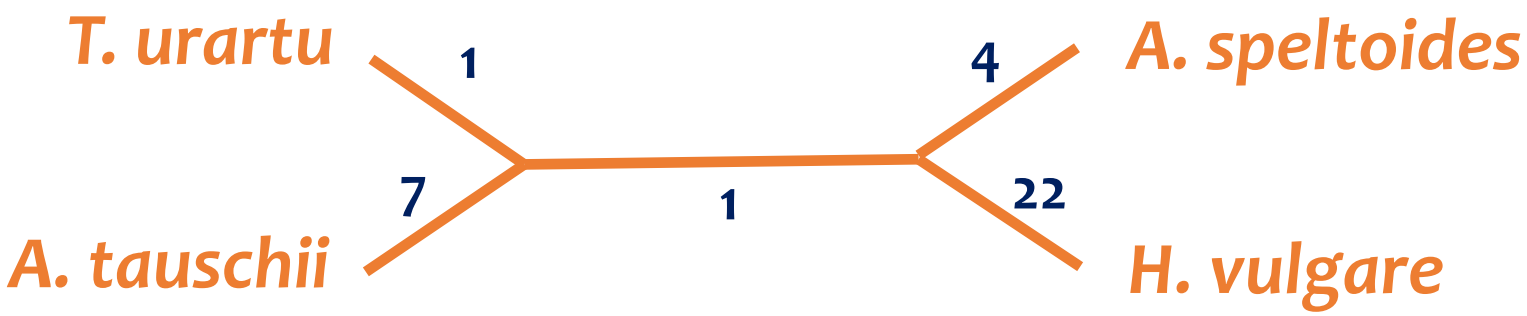
Count number of changes necessary to explain tree



Tree #2 – 36 total changes

Parsimony – Fewest number of steps

Count number of changes necessary to explain tree



Tree #3 – 35 total changes

Most parsimonious tree

Parsimony tools

MEGA

PAUP*

Maximum likelihood methods

Pros:

- Heuristically searches the tree space
- Easier to incorporate different models of molecular evolution
- Lends itself to statistical inference (e.g., bootstrapping, likelihood ratio tests)

Cons:

- Computationally intensive
- Long-branch attraction

Maximum Likelihood methods

$$P\,(data \mid tree{*}model)$$

Maximum likelihood tools

IQTree

FastTree

MEGA

PAUP*

PhyML

RAxML

PAML

Bayesian methods

Pros:

- Heuristically searches the tree space
- Directly models the underlying mechanisms of evolution
- Lends itself to statistical inference

Cons:

- Long-branch attraction
- Overconfidence

Bayesian methods

$$P\,(tree \mid data * prior * model)$$

Bayesian tools

MrBayes

BEAST

Statistical methods for tree inference

Bootstrapping:

- Randomly sample sites with replacement
- Re-infer tree
- Replicate 100-1000x count number of times a given split occurs in each of those replicates
- **Splits with bootstrap values >60 are statistically supported splits**

Jackknifing (leave-one-out analysis)

- Randomly sample sites without replacement so length of alignment = # sites - 1
- Number of replicates = Number of sites

Likelihood ratio tests

Posterior probability (Bayes)

- Probability of the split, given the data, the model, and the priors
- **PP values >80 are usually ok, better to go with 90**

Next up: Practicum in Phylogenetics