A PHYLOGENETICS SURVEY

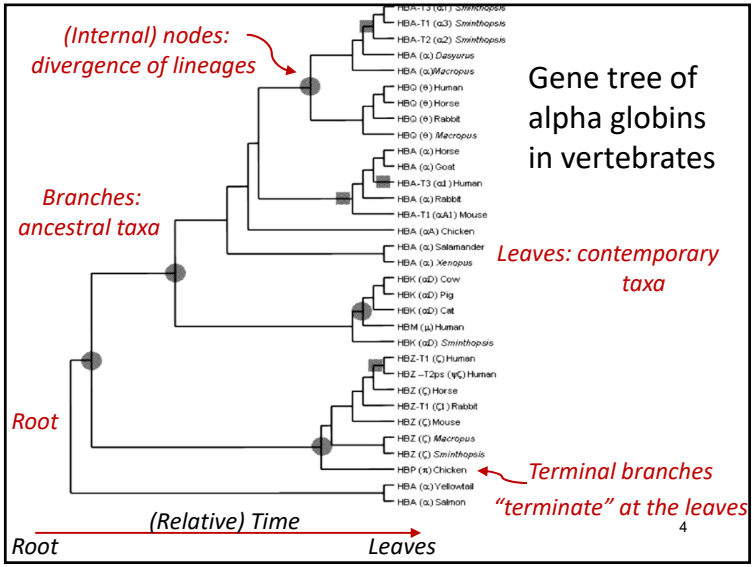
Dr. Maureen Stolzer
October 2nd, 2025

TREE TERMINOLOGY

Information encoded in trees

- Tree terminology
- Branch lengths
- Similarity, common ancestry and relatedness
- Binary and Non-binary trees
- Rooted and unrooted trees

3



4

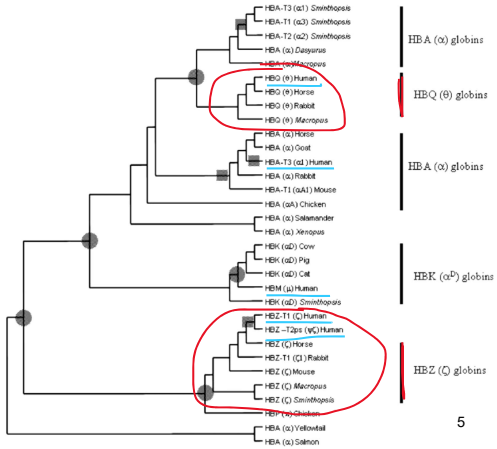
Clade: A group of taxa consisting of
a common ancestor and all of its descendants.

Clades:

- *HBQ (θ) globins*
- *HBZ (ζ) globins*
- *The tree*

Not a clade:

- *Human globins*



5

Information encoded in trees

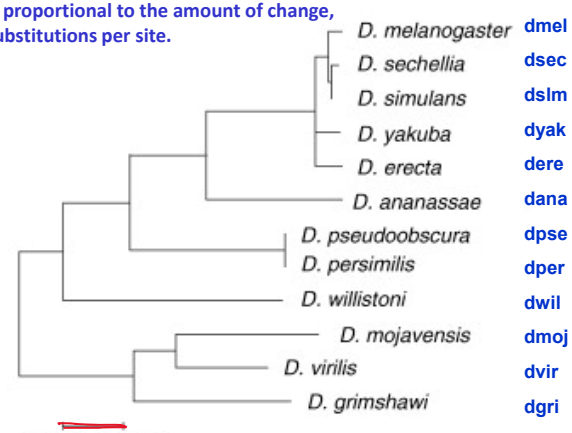
- Tree terminology
- ➢ Branch lengths vs topology
- Similarity, common ancestry and relatedness
- Binary and Non-binary trees
- Rooted and unrooted trees

6

12 Drosophila species

In this tree...

branches are proportional to the amount of change, typically in substitutions per site.



7

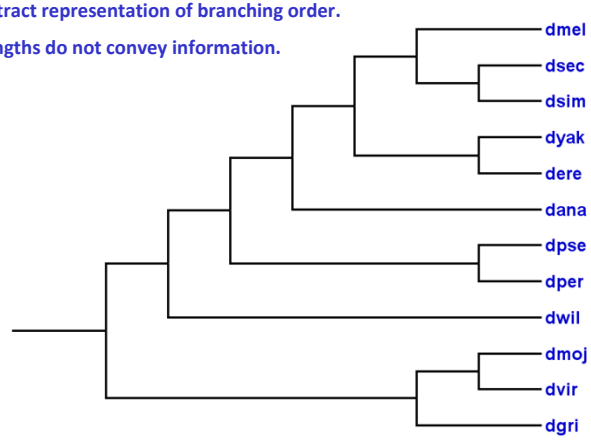
Larracuente *et al.*, Trends in Genetics, 2008

12 Drosophila species

This tree is...

an abstract representation of branching order.

Branch lengths do not convey information.



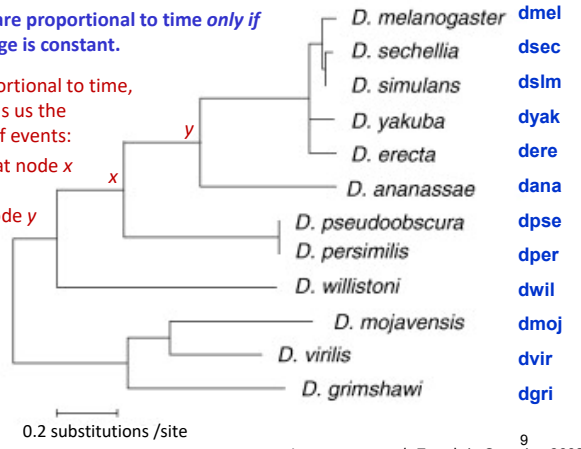
8

Branches are proportional to the amount of change, typically in substitutions per site.

Branch lengths are proportional to time *only if* the rate of change is constant.

Even if not proportional to time, *topology* still tells us the *relative* timing of events:

The divergence at node *x* came before the divergence at node *y*

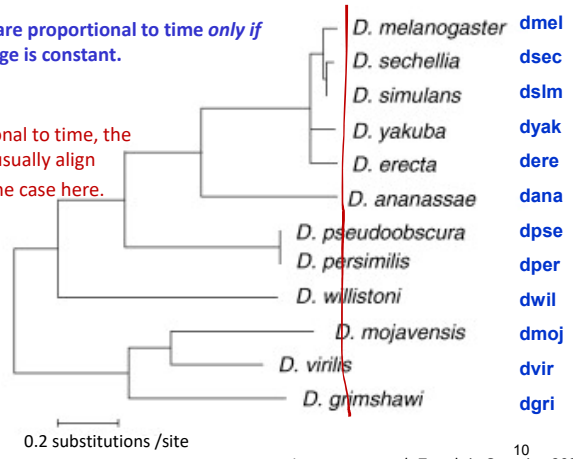


Larracuente *et al.*, Trends in Genetics, 2008

Branches are proportional to the amount of change, typically in substitutions per site.

Branch lengths are proportional to time *only if* the rate of change is constant.

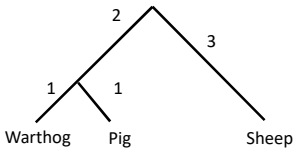
When proportional to time, the leaf nodes will usually align which is not the case here.



Larracuente *et al.*, Trends in Genetics, 2008

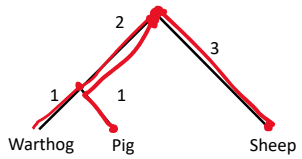
Ultrametric trees

- The rate of change is the same in all lineages
- Branch lengths are proportional to time
- The distance from the root to leaf is the same for all leaves
- The root is at the midpoint between the two most distant taxa.



Ultrametric trees

- The rate of change is the same in all lineages
- Branch lengths are proportional to time
- The distance from the root to leaf is the same for all leaves
- The root is at the midpoint between the two most distant taxa.

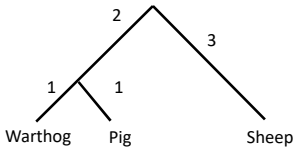


Root to:

Warthog	3
Pig	3
Sheep	3

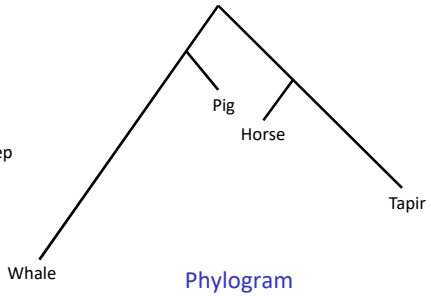
An *ultrametric* tree has branch lengths that are proportional to time.

Ultrametric



Chronogram

Not ultrametric



Phylogram

13

Information encoded in trees

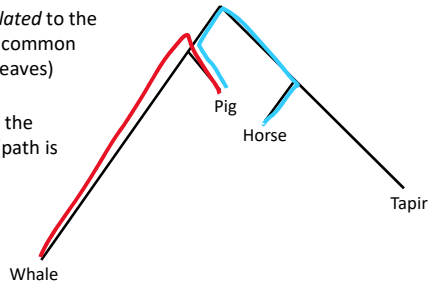
- Tree terminology
- Branch lengths
- ➢ Similarity, common ancestry and relatedness
- Binary and Non-binary trees
- Rooted and unrooted trees

14

Similarity versus common ancestry

More closely related taxa are not guaranteed to be more similar except in ultrametric trees, where the rate of change is proportional to time.

- The pig is *more closely related* to the whale than to the horse (common ancestor is closer to the leaves)
- The pig is *more similar* to the horse than to the whale (path is shorter)



15

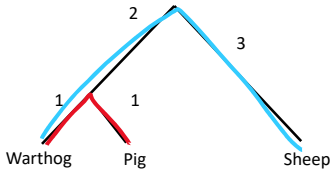
Similarity versus common ancestry

More closely related taxa are not guaranteed to be more similar except in ultrametric trees, where the rate of change is proportional to time.

.

an ultrametric tree

- The warthog is more closely related to the pig than to the sheep
- The warthog is more similar to the pig than to the sheep



16

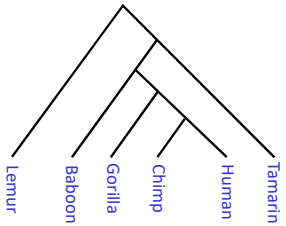
Information encoded in trees

- Tree terminology
- Branch lengths
- Similarity, common ancestry and relatedness
- ➢ Binary and Non-binary trees
- Rooted and unrooted trees

17

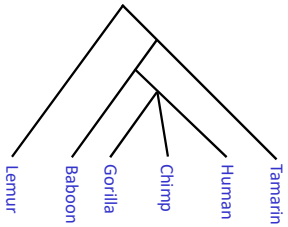
Binary versus non-binary trees

In a rooted tree, every *binary* node has 2 descendants



A tree is binary if every internal node is binary

In a rooted tree, a *non-binary* node or *polytomy* has 3 or more descendants



A tree is *non-binary* if one or more internal nodes are *non-binary*

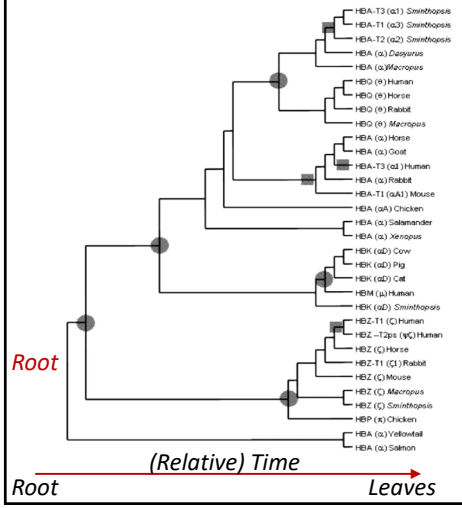
18

Information encoded in trees

- Tree terminology
- Branch lengths
- Similarity, common ancestry and relatedness
- Binary and Non-binary trees
- ➢ Rooted and unrooted trees

19

Unrooted trees versus rooted trees



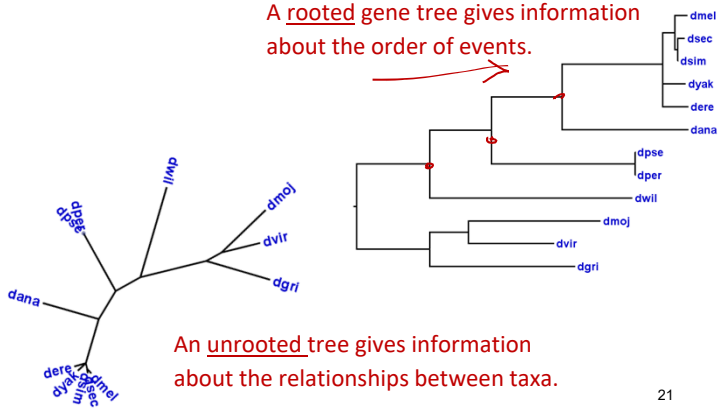
Previously, I stated that going from the root to the leaves gives us relative timing…

This is only for *rooted* trees

20

Unrooted trees versus rooted trees

A rooted gene tree gives information about the order of events.

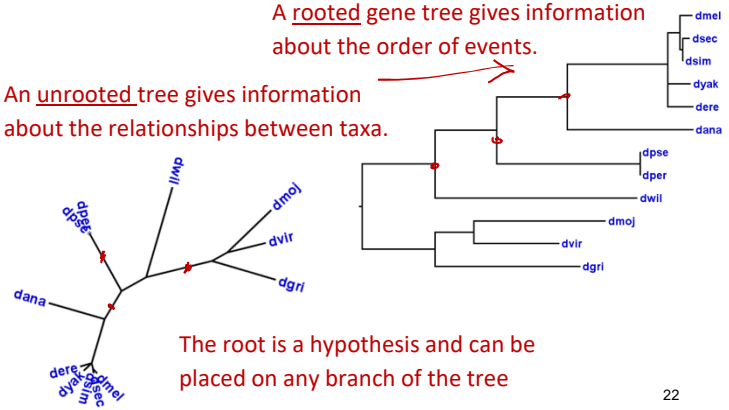


An unrooted tree gives information about the relationships between taxa.

Unrooted trees versus rooted trees

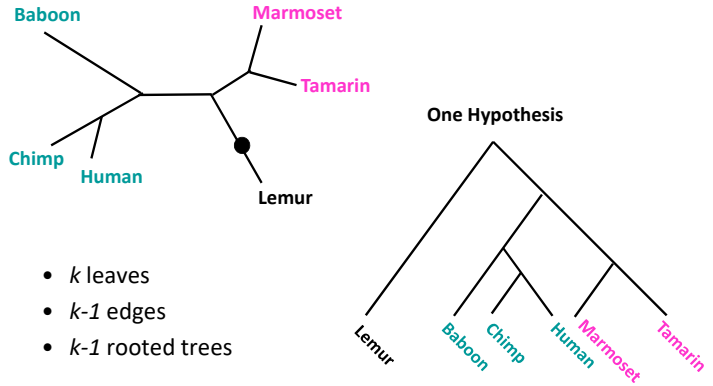
A rooted gene tree gives information about the order of events.

An unrooted tree gives information about the relationships between taxa.



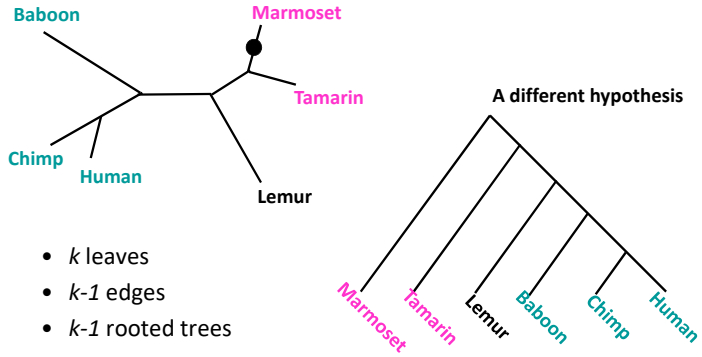
The root is a hypothesis and can be placed on any branch of the tree

- There are as many possible roots as there are edges in the tree.
- Each of these is a different hypothesis about the order of events.



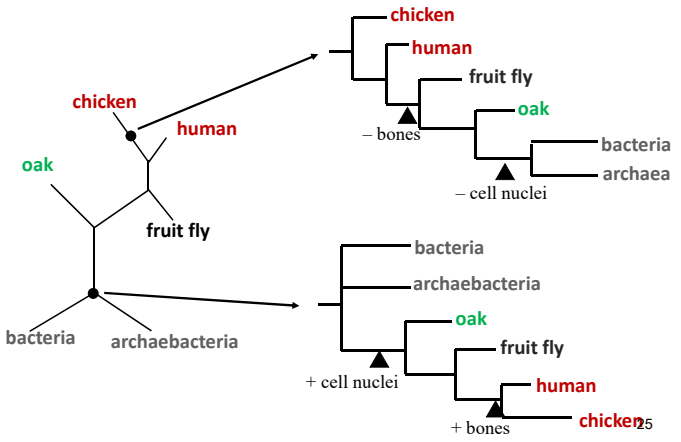
- *k* leaves
- *k-1* edges
- *k-1* rooted trees

- There are as many possible roots as there are edges in the tree.
- Each of these is a different hypothesis about the order of events.



- *k* leaves
- *k-1* edges
- *k-1* rooted trees

The position of the root of the tree can influence the interpretation dramatically

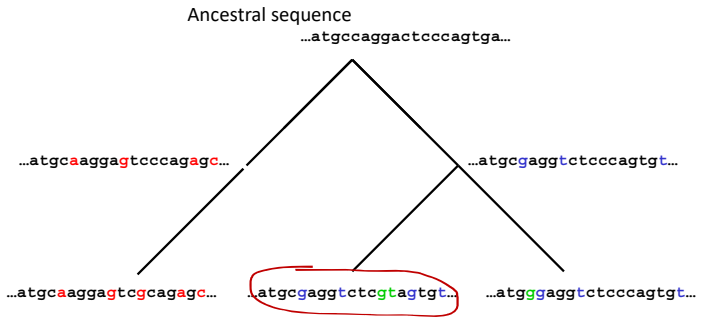


25

PHYLOGENY RECONSTRUCTION

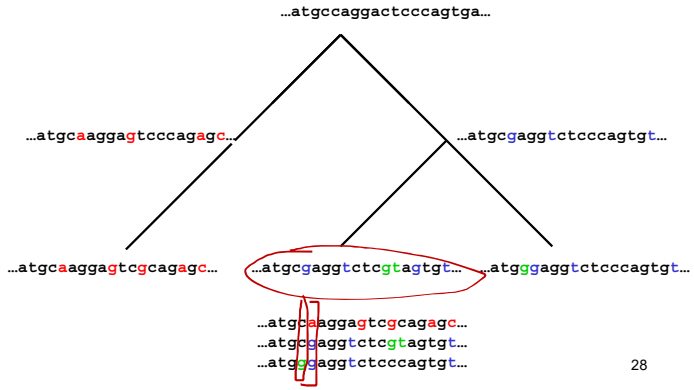
26

Gene family evolution



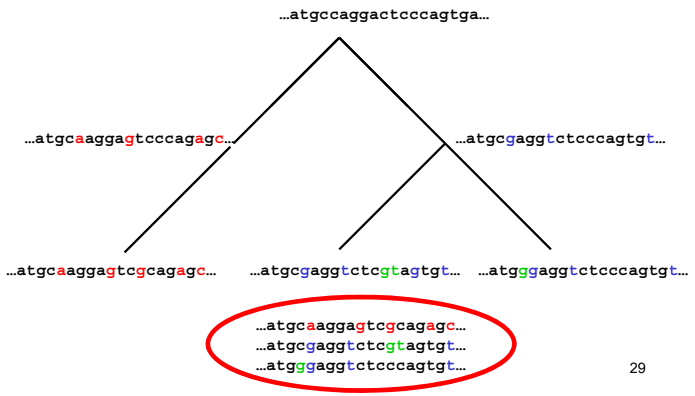
27

Reconstructing Evolutionary History



28

Reconstructing Evolutionary History

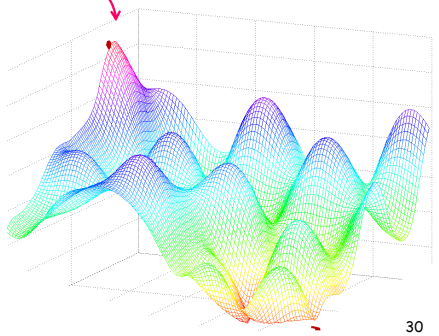


29

Space of all possible trees for a given set of taxa

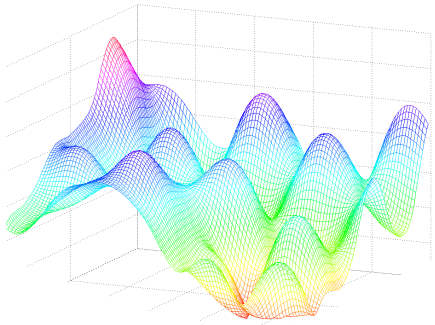
We can seek to maximize a measure of how good our solution is…

Optimality criteria



30

Space of all possible trees for a given set of taxa



Optimality criteria

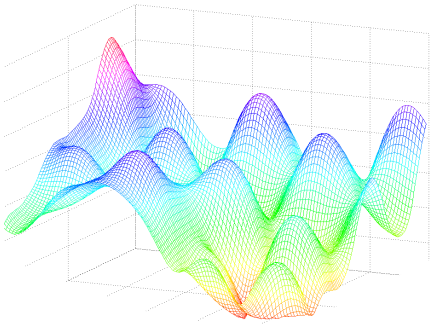
… or minimize a cost.

31

TREE SPACE

32

Space of all possible trees for a given set of taxa



Goal: Find the most parsimonious tree

33

Tree reconstruction

1. Construct a tree that fits the data

   This is only possible when the data satisfies some very restrictive conditions.

2. Score each possible tree with *k* leaves and select the best one

34

Tree reconstruction

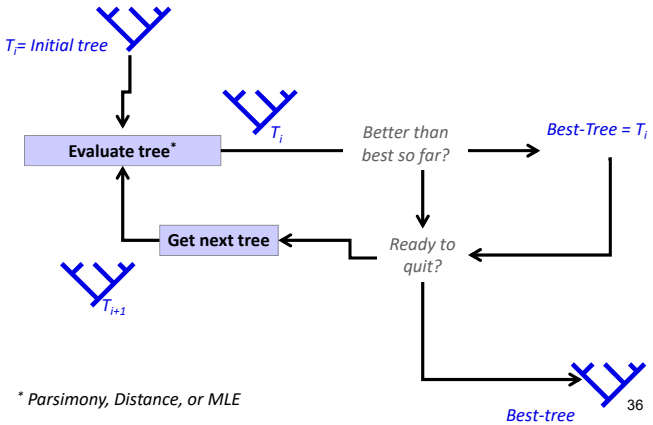
1. Construct a tree that fits the data

   This is only possible when the data satisfies some very restrictive conditions

2. Score each possible tree with *k* leaves and select the best one

35

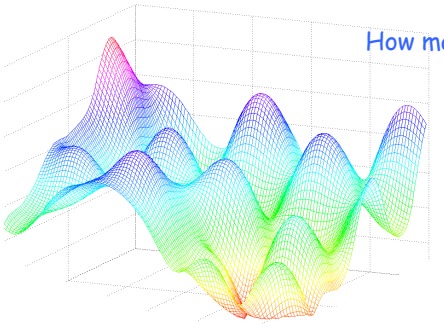
Phylogeny reconstruction



* Parsimony, Distance, or MLE

36

Space of all possible trees for a given set of taxa



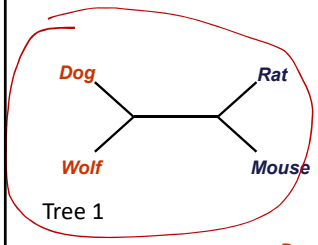
How many trees are there?

Goal: Find the most parsimonious tree

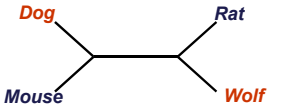
37

Given 4 taxa, there are three possible unrooted trees (hypotheses).

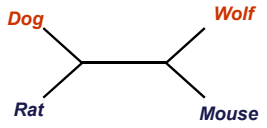
Given 5 taxa, how many unrooted trees (hypotheses) are there?



Tree 1



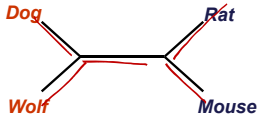
Tree 3

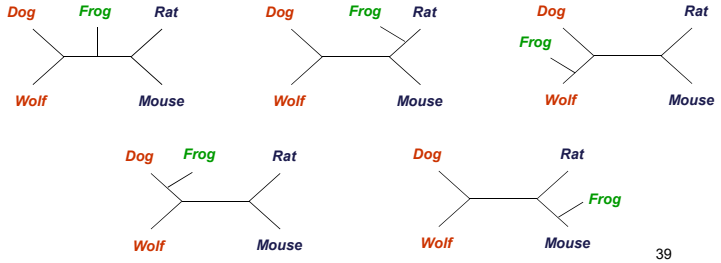


Tree 2

38

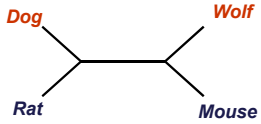
Tree 1

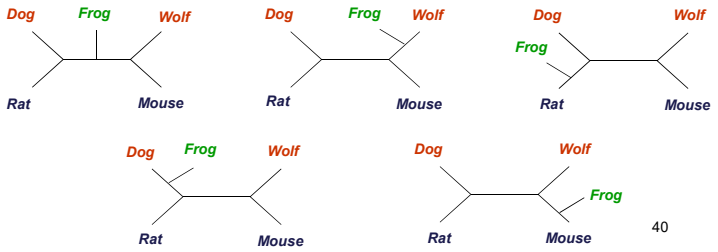




39

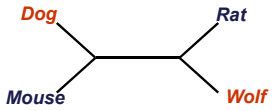
Tree 2

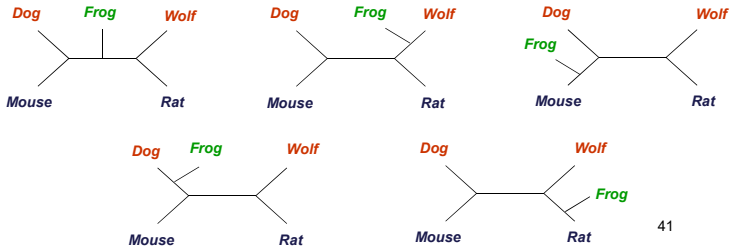




40

Tree 3





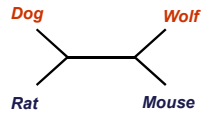
41

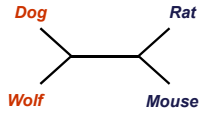
Given 5 taxa, how many unrooted trees (hypotheses) are there?

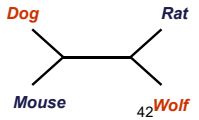
There are 3 unrooted trees with 4 leaves

Each tree has 5 edges; i.e., 5 places where a new taxon can be attached.

3 trees x 5 edges → 15 trees with 5 leaves







42

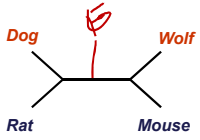
Given 5 taxa, how many unrooted trees (hypotheses) are there?

There are 3 unrooted trees with 4 leaves

Each tree has 5 edges; i.e., 5 places where a new taxon can be attached.

3 trees x 5 edges → 15 trees with 5 leaves

15t x 7e



43

In general, how many hypotheses must we consider for *k* leaf taxa?

- A tree with *k* leaves, has *2k - 3* edges
- For a given tree with *k* leaves, we can add a new taxon to any edge to get a different tree with *k + 1* leaves.
- That gives us *2k - 3* new trees for *each* tree with *k* leaves

Ergo, the number of hypotheses we need to consider gets big very, very quickly, as the number of taxa increases

44

The number of trees gets big fast

Number of taxa	Number of unrooted binary trees
3	1
4	3
5	15
6	105
10	2,027,025
20	2.2×10^{20}
50	2.8×10^{74}
500	1×10^{1074}

45

How big is that?

Number of taxa	Number of unrooted binary trees
…	…
20	2.2×10^{20}
50	2.8×10^{74}
500	1×10^{1074}

Age of the universe (seconds):	4.42×10^{17}
Diameter of the universe (cm):	2.6×10^{28}
Number of stars in the universe:	1×10^{22}

46

In general, how many hypotheses must we consider for *k* leaf taxa?

The number of hypotheses we need to consider gets big very, very quickly, as the number of taxa increases.

- If *k* is small (e.g., $k \leq 10$), we can consider all hypotheses.

➢ Otherwise ($k > 10$), sample a subset of all possible hypotheses.

"Heuristic search"

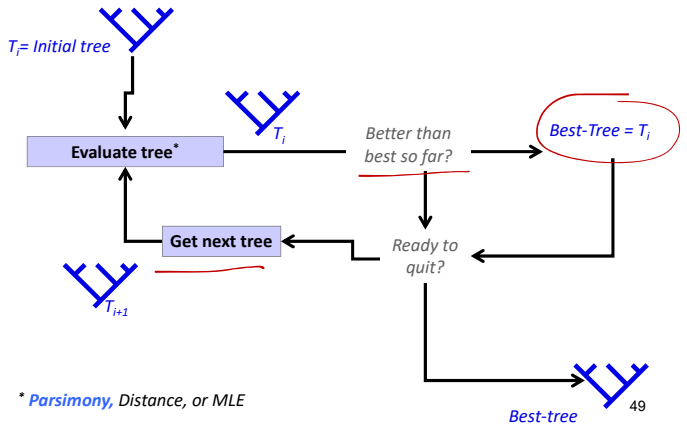
47

SEARCHING TREE SPACE

48

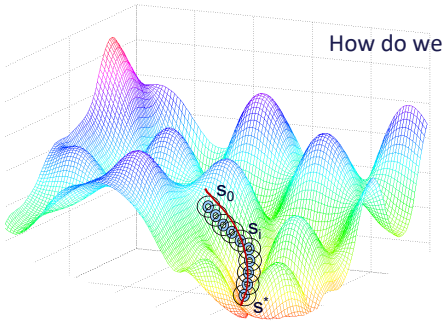
Phylogeny reconstruction

Heuristic search



* *Parsimony, Distance, or MLE*

Space of all possible trees for a given set of taxa

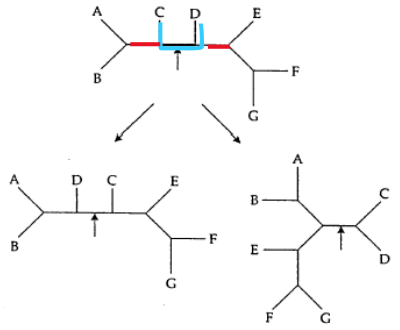


How do we sample the space of trees?

Goal: Find the most parsimonious tree

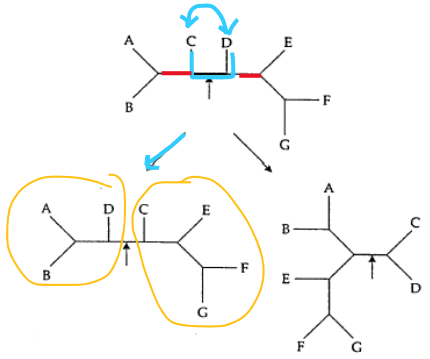
Branch swapping

Nearest-neighbor interchange (NNI)



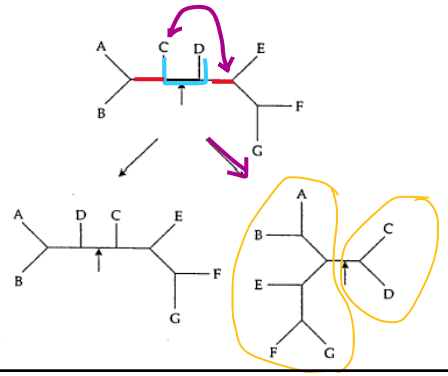
Branch swapping

Nearest-neighbor interchange (NNI)



Branch swapping

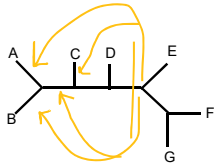
Nearest-neighbor interchange (NNI)



53

Branch swapping

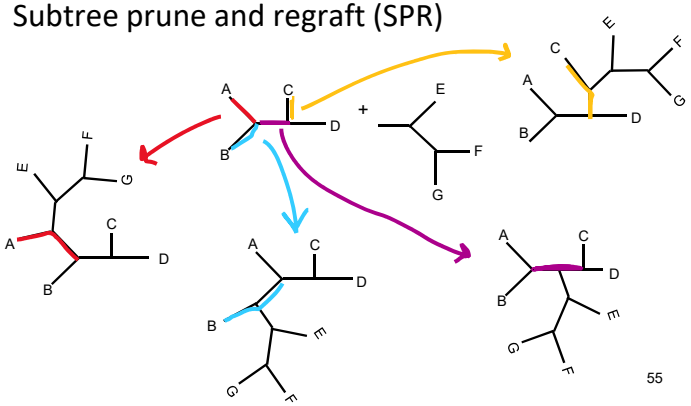
Subtree prune and regraft (SPR)



54

Branch swapping

Subtree prune and regraft (SPR)



55

Phylogeny reconstruction

- Design a mathematical function for scoring evolutionary trees such that
  - Good trees get good scores
  - Bad trees get bad scores

Optimality criterion

- Find the *optimal* tree; the tree with the best score
- Potential problems:
  - The optimal tree can be hard to find – NP Hard
  - A good numerical score does not always translate to a good evolutionary hypothesis.

56

Optimality criteria

Maximum parsimony

Tree that requires the fewest changes to explain the data is the best tree.

Maximum likelihood estimation

Tree that maximizes the likelihood of the data

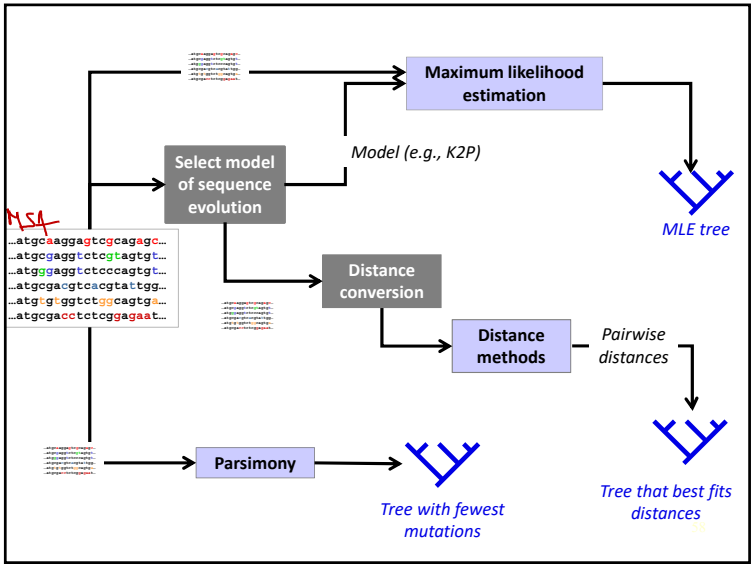
Character data

Distance-based evaluation

Tree that best fits observed distances between taxa

Distance data

57



PARSIMONY

59

The *parsimony* criterion

- In regular speech, “parsimonious” means excessively thrifty or sparing, stingy, or frugal
- Phylogenetics: the most parsimonious tree is the tree that requires the fewest substitutions to explain the data.

Evolutionary change on a tree

- Given
  - a tree
  - a set of characters that are variable for these taxa,
  - a character state matrix for the leaf taxa
- infer
  - the character states of each ancestral node and
  - the state changes along each branch
- such the *number of changes required is minimal*

Parsimony

61

An example:

Dog:	ACC
Wolf:	ATC
Rat:	CCC
Mouse:	CTG

Given
- 4 taxa
- data describing each taxon

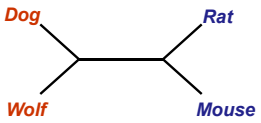
find

the most parsimonious tree with *4* leaves.

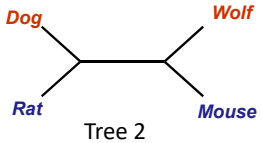
62

Which tree is most parsimonious?

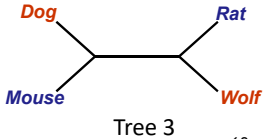
Dog:	ACC
Wolf:	ATC
Rat:	CCC
Mouse:	CTG



Tree 1



Tree 2



Tree 3

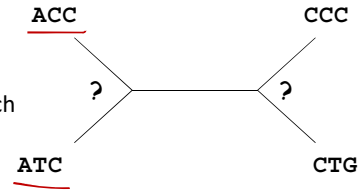
63

Evaluation using the parsimony criterion

Given a leaf labeled tree

infer the internal states (ancestral sequences) that minimize the number of changes along the branches

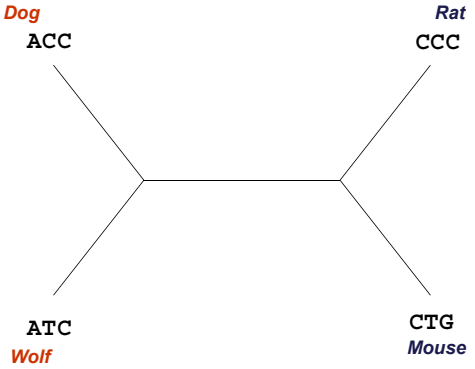
Parsimony score of this tree:
Sum of changes on each branch



64

Tree 1

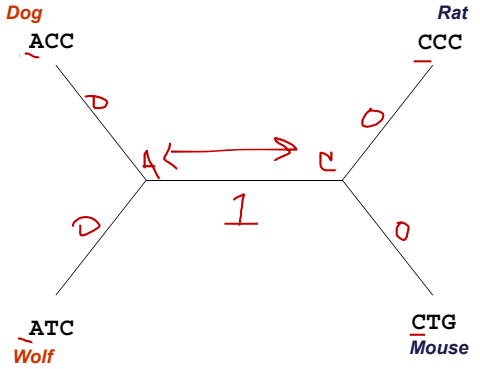
Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG



65

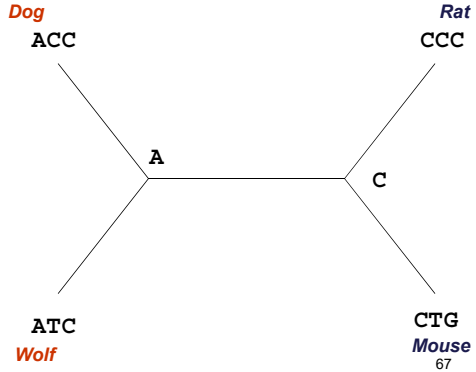
Tree 1

Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG



66

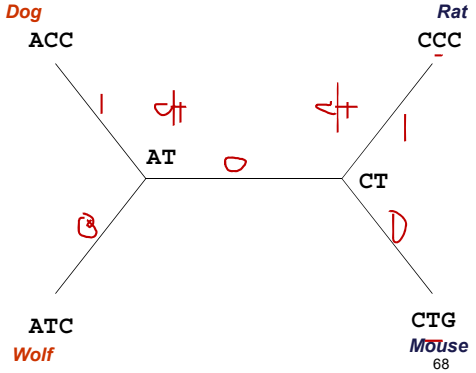
Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG



67

Inferring ancestral sequences and computing the parsimony score

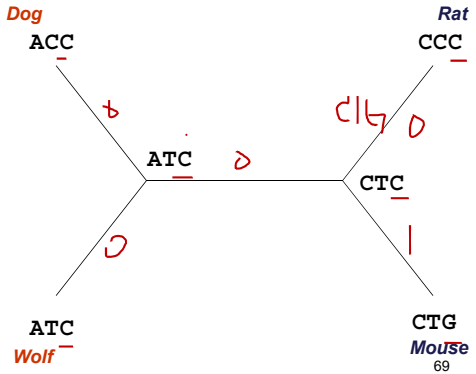
Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG



68

Inferring ancestral sequences and computing the parsimony score

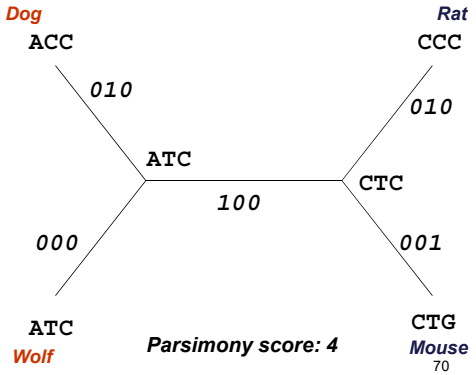
Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG



69

Inferring ancestral sequences and computing the parsimony score

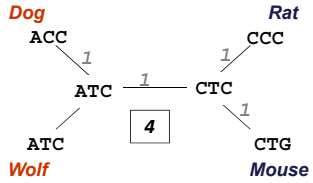
Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG

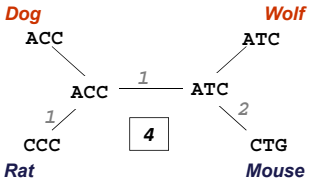


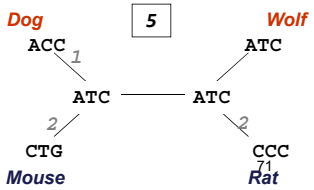
70

Parsimony scores of all trees with four leaves

Dog: ACC
Wolf: ATC
Rat: CCC
Mouse: CTG



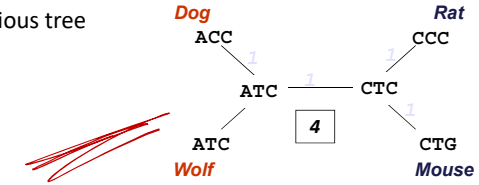


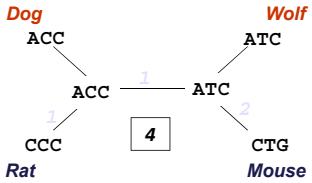


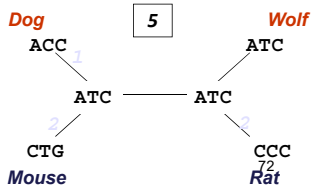
71

Note:

There can be more than one most parsimonious tree







72

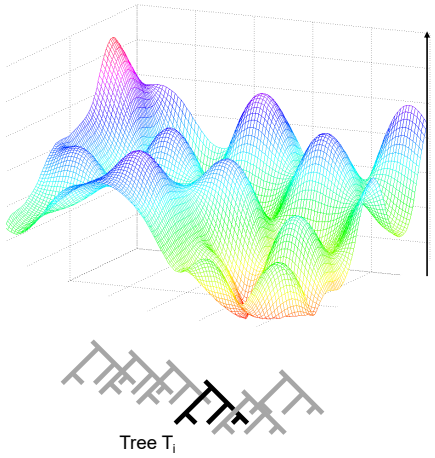
Phylogeny reconstruction

Potential problems with Maximum Parsimony:

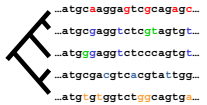
- There can be more than one most parsimonious tree.
- Not all sites are informative
- The assumption that mutations are rare may be wrong.

73

Tree reconstruction with parsimony



Measure of tree quality: The minimum number of substitutions required to explain the MSA with tree T_i



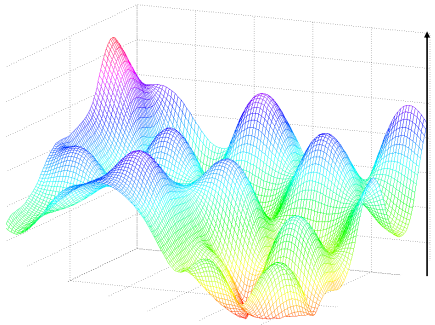
Tree T_i

74

MAXIMUM LIKELIHOOD

75

Tree reconstruction with likelihood maximization



Measure of tree quality: The probability of observing the MSA if T_i is the true tree

Tree T_i

76

Tree reconstruction

Given

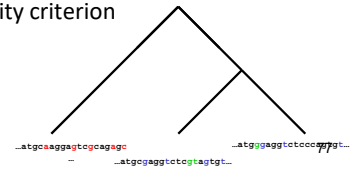
- contemporary taxa: *k* genes, strains, species…
- data describing each taxon

…atgcaaggagtcgcagagc…
…atgcgaggtctcgtagtgt…
…atgggaggtctcccagtgt…

find

the tree with *k* leaves that explains the data.

➢ Find the best tree with respect to an optimality criterion

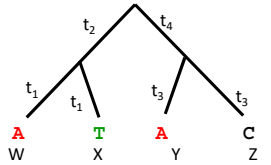


Estimating the probability of an MSA for a given tree, T_i

How to calculate the probability of the MSA given *Ti*:

- Calculate the probability of each column given *Ti*
- Multiply the probabilities over all columns.

Tree T_i : Topology and branch lengths



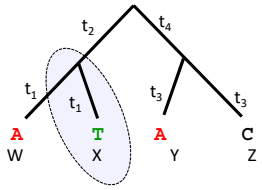

```
Sequence W: A C G C A T T G G G
Sequence X: A C G C T T T G G G
Sequence Y: A C G C A A T G A A
Sequence Z: A C A C C G G G A A
```


Estimating the probability of <u>one column</u> for a given tree, T_i

Working up from the leaves

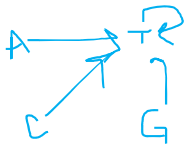
- Estimate contribution of each branch to the probability of the column, given the tree.
- Multiply the probabilities over all branches.

Tree T_i : Topology and branch lengths

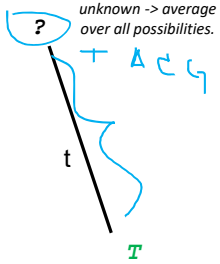



```
Sequence W:  . . A . . .
Sequence X:  . . T . . .
Sequence Y:  . . A . . .
Sequence Z:  . . C . . .
```


Estimate contribution of one branch to the probability of the column



The ancestral base is unknown -> average over all possibilities.



*P(ancestral T) * P(T after time t has elapsed) +*

*P(ancestral A) * P(changed from A→T in time t) +*

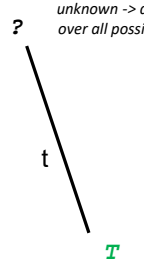
*P(ancestral C) * P(changed from C→T in time t) +*

*P(ancestral G) * P(changed from G→T in time t)*

```
Sequence W:  . . A . . .
Sequence X:  . . T . . .
Sequence Y:  . . A . . .
Sequence Z:  . . C . . .
```

Estimate contribution of one branch to the probability of the column

The ancestral base is unknown -> average over all possibilities.



Stated formally

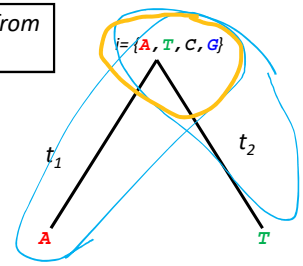
$P(A)P(t)_{AT} + P(T)P(t)_{TT} + P(C)P(t)_{CT} + P(G)P(t)_{GT}$

*P(ancestral A) * P(changed from A→T in time t)*

```
Sequence W:  ...A...
Sequence X:  ...T...
Sequence Y:  ...A...
Sequence Z:  ...C...
```

82

Extending the likelihood calculation from one branch to two branches



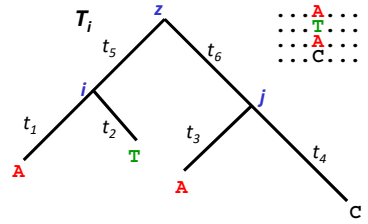
$P(i=A)P(t_1)_{AA}P(t_2)_{AT} + P(i=T)P(t_1)_{TA}P(t_2)_{TT}$

$+ P(i=C)P(t_1)_{CA}P(t_2)_{CT} + P(i=G)P(t_1)_{GA}P(t_2)_{GT}$

This calculation accounts for the fact that the two branches are not independent: They share an ancestral nucleotide

83

Extending the calculation obtain the probability of the column given the entire tree



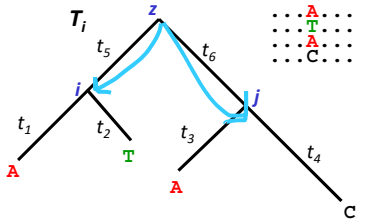
The probability of this column given T_i is

$$\sum_{i\in\{ACGT\}} \sum_{j\in\{ACGT\}} \sum_{z\in\{ACGT\}} p(z)p_{zi}(t_5)p_{zj}(t_6) \cdot p(i)p_{i\mathbf{A}}(t_1)p_{i\mathbf{T}}(t_2) \cdot p(j)p_{j\mathbf{A}}(t_3)p_{j\mathbf{C}}(t_4)$$

root | left subtree | right subtree

84

Extending the calculation obtain the probability of the column given the entire tree



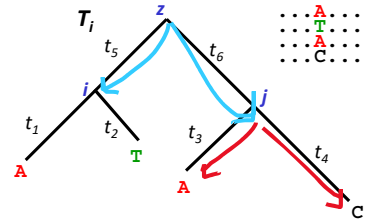
The probability of this column given T_i is

$$\sum_{i\in\{ACGT\}} \sum_{j\in\{ACGT\}} \sum_{z\in\{ACGT\}} p(z)p_{zi}(t_5)p_{zj}(t_6) \cdot p(i)p_{i\mathbf{A}}(t_1)p_{i\mathbf{T}}(t_2) \cdot p(j)p_{j\mathbf{A}}(t_3)p_{j\mathbf{C}}(t_4)$$

root | left subtree | right subtree

85

Extending the calculation obtain the probability of the column given the entire tree



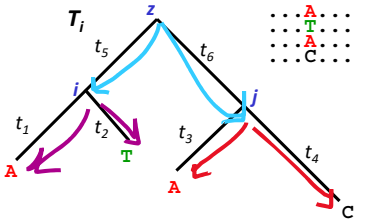
The probability of this column given T_i is

$$\sum_{i\in\{ACGT\}}\sum_{j\in\{ACGT\}}\sum_{z\in\{ACGT\}} p(z)p_{zi}(t_5)p_{zj}(t_6)\cdot p(i)p_{i\mathtt{A}}(t_1)p_{i\mathtt{T}}(t_2)\cdot p(j)p_{j\mathtt{A}}(t_3)p_{j\mathtt{C}}(t_4)$$

root | left subtree | right subtree

Same j

Extending the calculation obtain the probability of the column given the entire tree

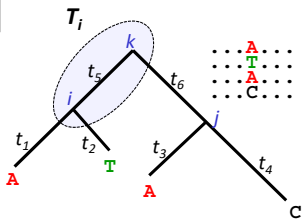


The probability of this column given T_i is

$$\sum_{i\in\{ACGT\}}\sum_{j\in\{ACGT\}}\sum_{z\in\{ACGT\}} p(z)p_{zi}(t_5)p_{zj}(t_6)\cdot p(i)p_{i\mathtt{A}}(t_1)p_{i\mathtt{T}}(t_2)\cdot p(j)p_{j\mathtt{A}}(t_3)p_{j\mathtt{C}}(t_4)$$

root | left subtree | right subtree

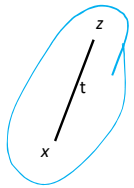
Extending the calculation obtain the probability of the column given the entire tree



The probability of this column given T_i is

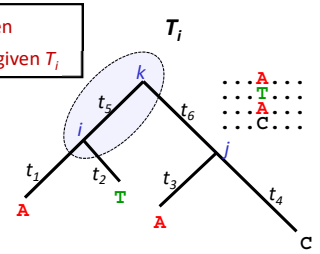
$$\sum_{i\in\{ACGT\}}\sum_{j\in\{ACGT\}}\sum_{z\in\{ACGT\}} p(z)p_{zx}(t_5)p_{zy}(t_6)\cdot p(x)p_{x\mathtt{A}}(t_1)p_{x\mathtt{T}}(t_2)\cdot p(y)p_{y\mathtt{A}}(t_3)p_{y\mathtt{C}}(t_4)$$

The entire calculation depends on terms like this where $p_{zx}(t)$ is the probability of starting with base k and observing base i after time t.



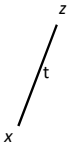
If we know how to calculate $p_{zx}(t)$, then we have everything needed to determine the probability of the MSA given T_i

If we know how to calculate $p_{kj}(t)$, then we can calculate the probability of the MSA given T_i



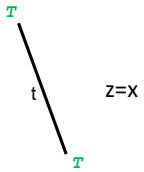
In order to calculate $p_{zx}(t)$ we need a model of substitutions during sequence evolution

$p_{ki}(t)$ = *the probability of starting with base z and observing base x after time t.*

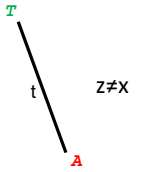


We discussed how to solve this problem with the Jukes Cantor model:

Given a site evolving according to Jukes Cantor with parameter *a*, what is the probability of observing *z* at time 0 and *x* at time *t*?



$$p_{xx}(\alpha, t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

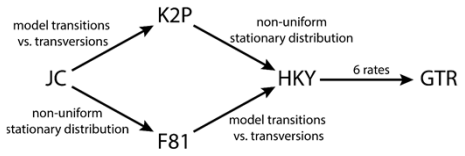


$$p_{xy}(\alpha, t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

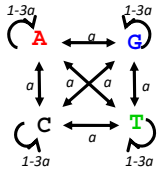
From last class

90

... but there are other models



Jukes-Cantor



p(A)=0.25
p(G)=0.25
p(C)=0.25
p(T)=0.25

To estimate the probability of an MSA given a tree requires a model of substitutions during sequence evolution

- Which evolutionary model?
  - Do all substitutions have the same rate?
- What rates?
- Do all sites have the same rate?
- Are there invariant sites that never change?

92

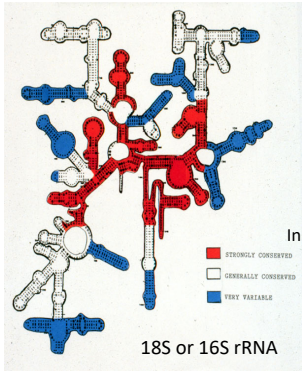
To estimate the probability of an MSA given a tree requires a model of substitutions during sequence evolution

- Which evolutionary model?
  - Do all substitutions have the same rate?
- What rates?
- ➢ Do all sites have the same rate?
- Are there invariant sites that never change?

93

Site heterogeneity:

Different sites may be changing at different rates



18S or 16S rRNA

Rates may be site specific, depending on functional and structural constraints (e.g., codon position, or alpha helix etc.)

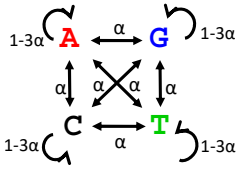
In this example, white sites are changing faster than red sites and more slowly than blue sites.

94

Modeling rate variation

It is not possible to assign a different rate to each position.
Instead, partition sites into a small number of rate categories, e.g., "slow, medium, fast, supersonic"

Use one model for the entire alignment, but different rates at different sites.

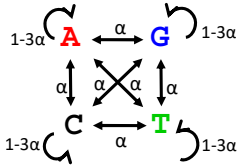


For example, all sites might be modeled with the Jukes Cantor model, but with different parameters: $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4$.

95

Modeling rate variation

In addition, some sites may not be changing at all.



Invariant sites: Sites that do not change…

- Examples: very recent divergence, purifying selection, parallel substitutions
- Incorporate invariable sites in model as an extra category. For every site, calculate the likelihood that it is an invariable site.

96

Discrete rates model

Given *k* rate classes,

- Determine the rate, r_k, for each class
- Determine the probability that site *i* is in class *k*.

Site class	*1*	*2*	*3*	…	*K*
Probability	p_1			…	p_K
Rate	r_1			…	r_k

	$r_?$	$r_?$		…		$r_?$		
. . .	A	G	G	G	A	T	T	C	A	. . .
. . .	A	C	G	G	T	T	T	-	A	. . .
. . .	A	C	G	G	A	T	T	-	A	. . .

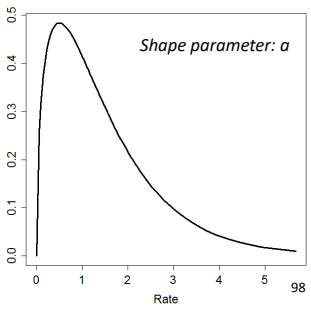
Problem: Too many parameters

97

Solution: Use a Gamma distribution

```
... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...
```

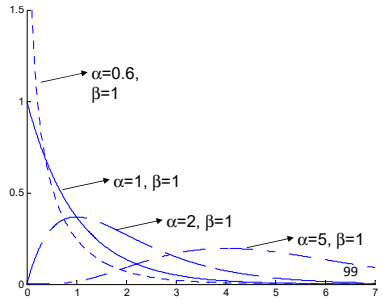
With a continuous Gamma distribution, each site has rate *r* with probability, *Gamma(r,a)*



Continuous Gamma distribution: A single parameter (α) captures a broad range of distributions with different shapes.

```
... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...
```

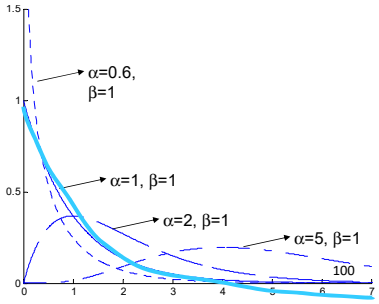
Each site has rate *r* with probability, *Gamma(r,a)*



Continuous Gamma distribution: A single parameter (α) captures a broad range of distributions with different shapes.

```
... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...
```

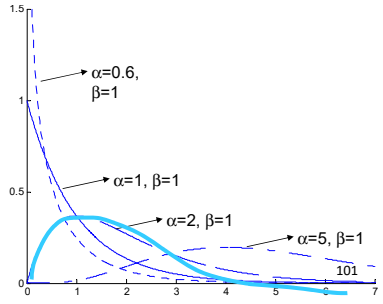
Each site has rate *r* with probability, *Gamma(r,a)*



Continuous Gamma distribution: A single parameter (α) captures a broad range of distributions with different shapes.

```
... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...
```

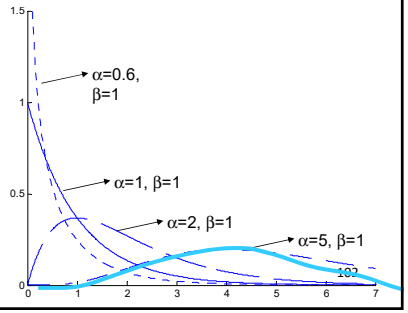
Each site has rate *r* with probability, *Gamma(r,a)*



Continuous Gamma distribution: A single parameter (α) captures a broad range of distributions with different shapes.

... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...

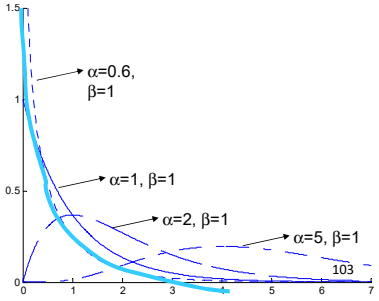
Each site has rate *r* with probability, *Gamma(r,a)*



Continuous Gamma distribution: A single parameter (α) captures a broad range of distributions with different shapes.

... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...

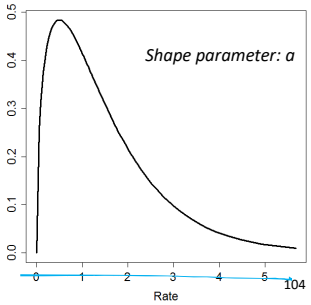
Each site has rate *r* with probability, *Gamma(r,a)*



Continuous Gamma distribution

... A G G G A T T C A ...
... A C G G T T T - A ...
... A C G G A T T - A ...

Each site has rate *r* with probability, *Gamma(r,a)*

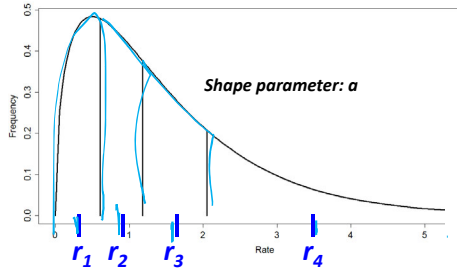


Problem:

Estimating the likelihood with a continuous function is too slow

Discrete Gamma distribution with *k* rate catories

- Divide area under the curve into *k* equal size regions
- Rates: midpoints of each region
- Only one parameter needed to describe all rate categories

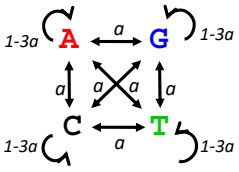


In most software, the default is *k = 4*

... A G G G A T T C
... A C G G T T T -
... A C G G A T T -
105

In summary, use models of sequence evolution to capture the following features:

Frequency of substitutions between bases/residues:



Base/residue frequencies:

p(A),p(C),p(G),p(T)

Invariant sites: sites that do not change

Multiple rate categories:

– Same model, but with different values of the parameter(s)

106

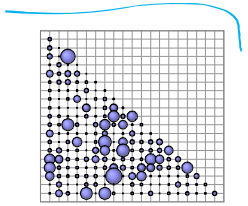
DNA substitution models

	C	G	T
A	●	●	●
C		●	●
G			●

DNA substitution model (e.g., JC, K2P, GTr)

- Four states (A, C, G, T)
- Model specifies the relative rates of substitution for all possible pairs of nucleotides

Amino acid substitution models

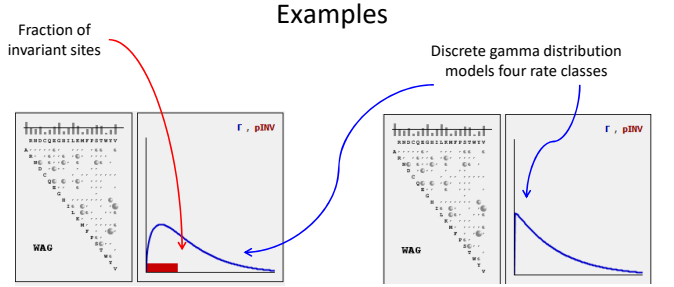


Amino acid substitution matrix (e.g., PAM, WAG, JTT, MtREV etc)

- Twenty states (A, C, … Y)
- Model specifies the relative rates of substitution for all possible pairs of amino acids

107

Examples



WAG amino acid substitution model with a gamma model of rate heterogeneity and invariant sites

WAG substitution model with a gamma model of rate heterogeneity only

Note that the shape parameter selected gives an increased peak at zero for the model without invariant sites (right). In this case, the lowest rate class is slower to compensate because the model does not account for invariant sites.

108

DISTANCE-BASED TREE RECONSTRUCTION

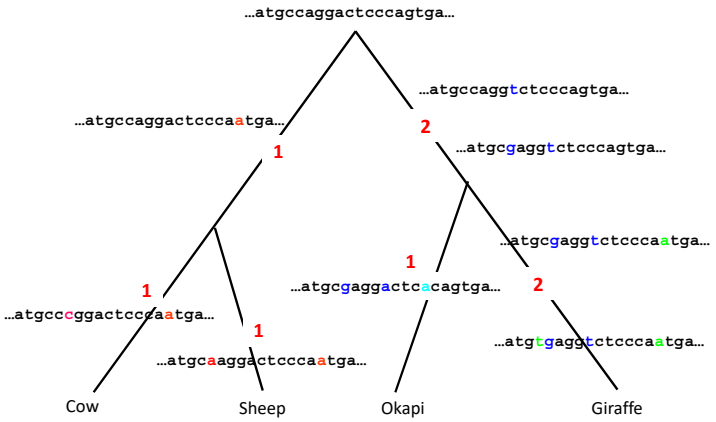
109

Phylogeny reconstruction outline

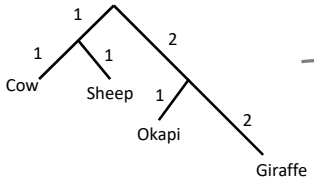
- Optimality criteria for scoring evolutionary trees
  - Maximum parsimony
  - Maximum likelihood estimation
  - **Distance-based evaluation**
- Find the *optimal* tree; the tree with the best score
  - How many hypotheses (trees with k leaves) are there?
  - Sampling hypotheses: how to search through the space of trees with k leaves?

110

If we could observe evolution in progress …we would know the number of changes on each branch…



.. and the distance between each pair of taxa

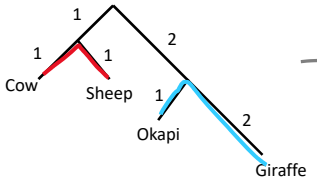


Change along tree branches can be expressed as pairwise distances.

	Sheep	Okapi	Giraffe
Cow	2	5	6
Sheep		5	6
Okapi			3

Tree distances

.. and the distance between each pair of taxa

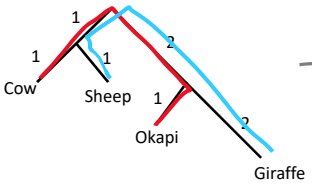


Change along tree branches can be expressed as pairwise distances.

	Sheep	Okapi	Giraffe
Cow	2	5	6
Sheep		5	6
Okapi			3

Tree distances

.. and the distance between each pair of taxa



Change along tree branches can be expressed as pairwise distances.

	Sheep	Okapi	Giraffe
Cow	2	5	6
Sheep		5	6
Okapi			3

Tree distances

Can we reverse this process?

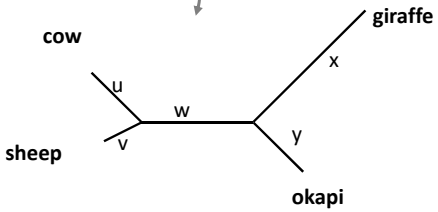
	Sheep	Okapi	Giraffe
Cow	2	5	6
Sheep		5	6
Okapi			3

Observed distances

Given observed distances between taxa…

…is there a tree with branch lengths that fit the observed distances?

Is it unique?



Outline

Distance-based phylogeny reconstruction

Convert multiple alignments into distances

Properties of pairwise distances between taxa

- Additive distances
- Ultrametric distances

Using pairwise distances to infer a tree: constructive methods

- UPGMA
- Neighbor Joining

If you've seen hierarchical clustering, this may seem similar

116

Calculating distances from MSAs

```
C  -  GCTTGTCCGTTACGAT
S  -  ACTTGACCGTTTCCTT
O  -  ACTTGTCCGAAACGAT
G  -  ACTTGTCTGTTACGAT
```

	Sheep	Okapi	Giraffe
Cow	2	5	6
Sheep		5	6
Okapi			3

For each pair of taxa

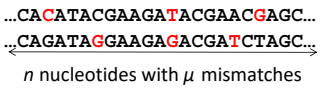
- Use the pairwise alignment *induced by the MSA.*
- Count substitutions to obtain pairwise distances
- Correct for multiple substitutions

Correcting for multiple substitutions with Jukes-Cantor

Given an alignment of *n* nucleotides that differs at *m* positions, the expected number of substitutions since the divergence of the two sequences is given by

$$D = \frac{-3}{4}\ln\left(1 - \frac{4}{3}\frac{\mu}{n}\right)n$$

From last lecture



For example, if we observe 200 mismatches in an alignment of 1000 nucleotides, then the number of actual substitutions is

$$\frac{-3}{4}\ln\left(1 - \frac{800}{3000}\frac{\mu}{n}\right)3000 = 233$$

Outline

Distance-based phylogeny reconstruction

Convert multiple alignments into distances

Properties of pairwise distances between taxa

- Additive distances
- Ultrametric distances

Using pairwise distances to infer a tree: constructive methods

- Neighbor Joining
- UPGMA

Distance Takeaways

- Several algorithms based on various assumptions about the data.
- If the data satisfy the assumptions, then distance-based algorithms will return the correct tree.
- If the data do not deviate too far from the assumptions, these methods give reasonable approximations
- They run in polynomial time.