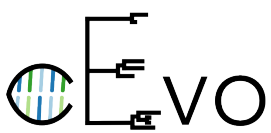
Computational Biology

Lecturers:
Tanja Stadler, Carsten Magnus & Tim Vaughan

Teaching Assistants:
Jūlija Pečerska, Jérémie Sciré,
Sarah Nadeau & Marc Manceau

Computational Evolution
Department of Biosystems Science and Engineering

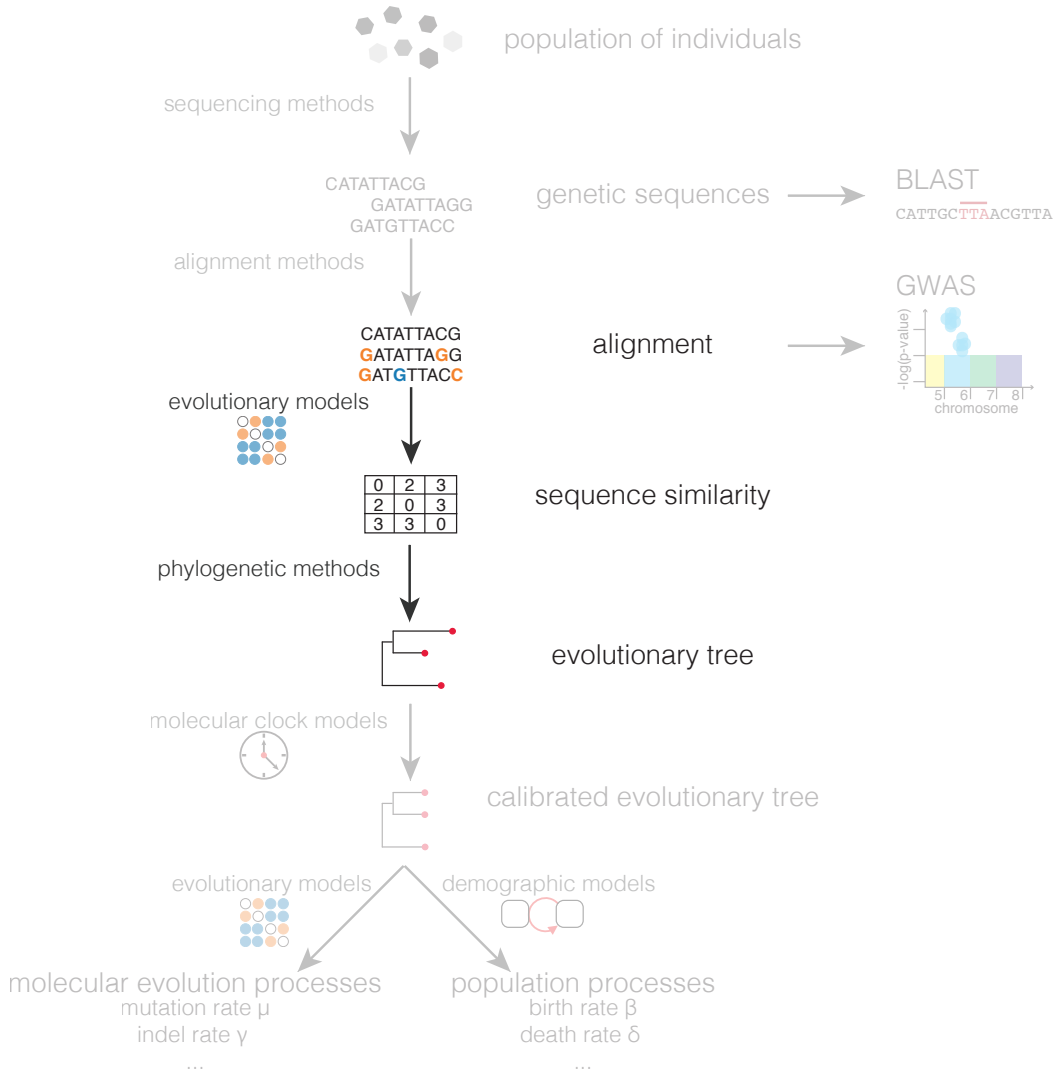
HS 2019



Definition of phylogenies & phenetic phylogenetic reconstruction: Questions

- What is the minimal number of cherries in a phylogenetic tree of 99 tips? What is the maximum number?
- In how many ways can you write the Newick string for a rooted tree with species A, B, C? In how many ways can you write the Newick string for a rooted tree with n species?
- Consider the least squares method. Why would we use weights $w_{i,j}$ which are not equal to 1?

Overview



Cladistic tree inference.

Phenetic vs. cladistic approach

- **phenetic approaches:**
  - similarity is measured by pairwise distances
  - overall similar sequences cluster
- **cladistic approaches:**
  - similarity is measured by shared characters in the alignment sequences (i.e. evolutionary process is accounted for implicitly)
  - parsimony method: find the tree that needs the minimal number of mutations to get the alignment

First cladistic tree

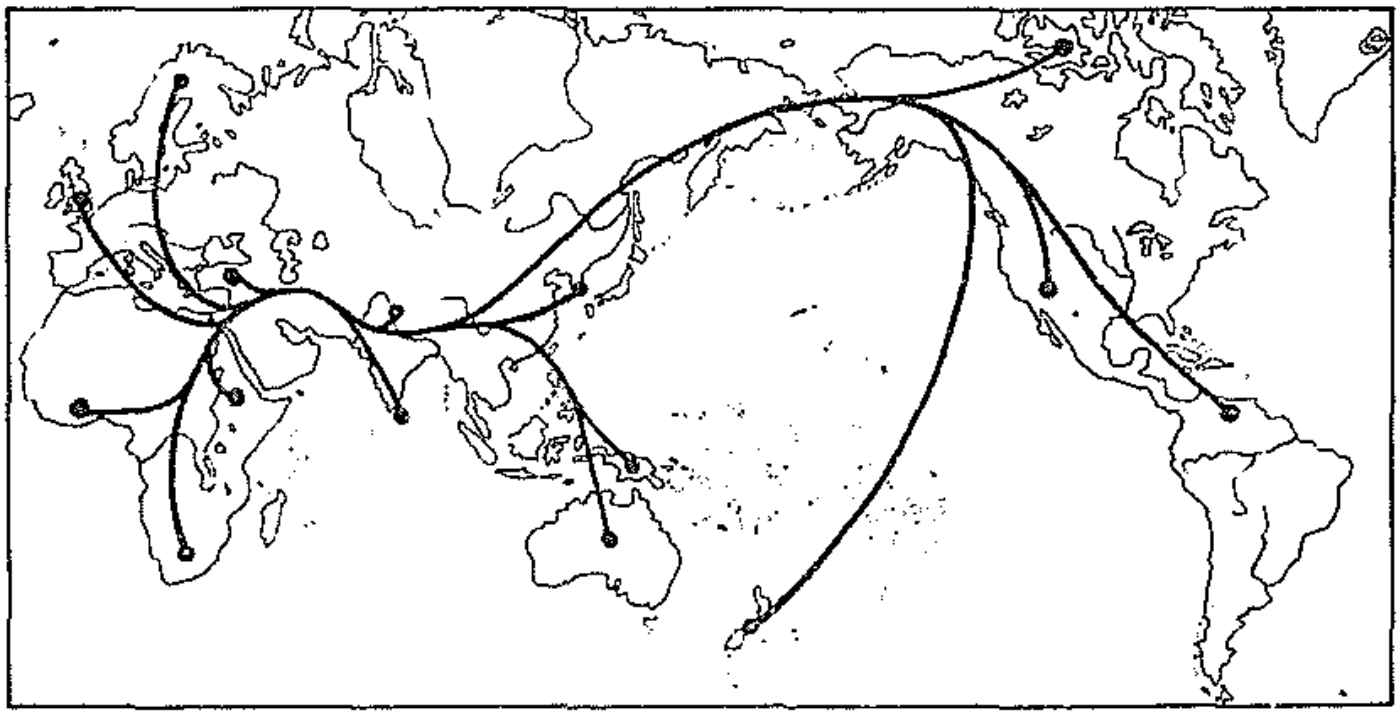


Figure adapted from [Sforza and Edwards, 1964]

First parsimony (minimum-evolution) tree: uniting human populations based on blood group gene frequencies.

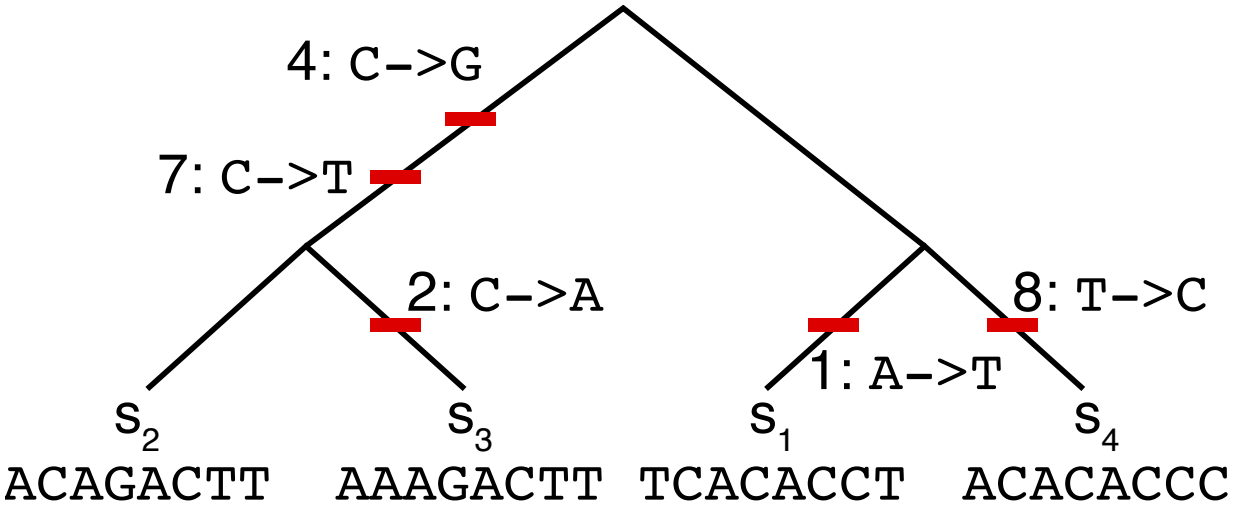
Parsimony method

- **parsimony score of a tree:** the minimal number of mutations required to explain the sequences at the tips of the tree
- **most parsimonious tree:** the tree with minimal parsimony score

Example: What is the minimal number of mutations needed to explain this alignment?

sequence 1: `T C A C A C C T`
sequence 2: `A C A G A C T T`
sequence 3: `A A A G A C T T`
sequence 4: `A C A C A C C C`

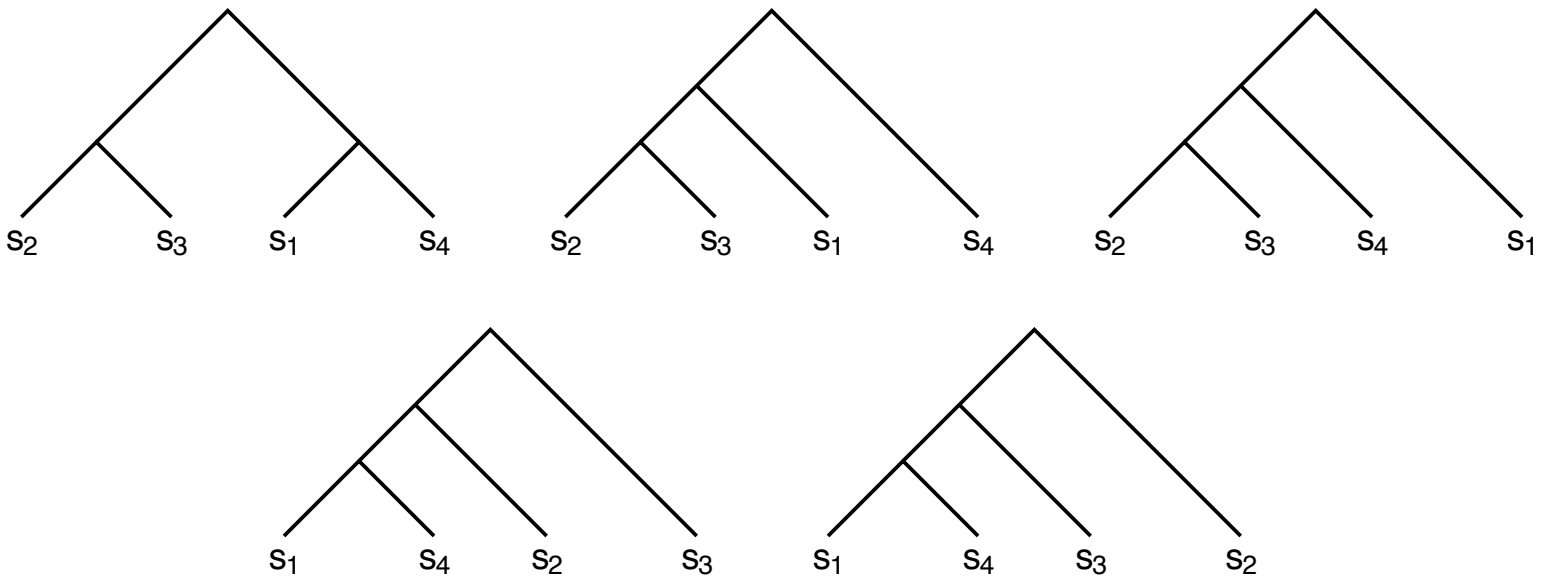
Parsimony score of our UPGMA tree



Rigorous: Label internal nodes with each possible ancestral sequence and then determine the number of mutations required for each assignment. The minimal number of mutations required is the parsimony score of the tree.

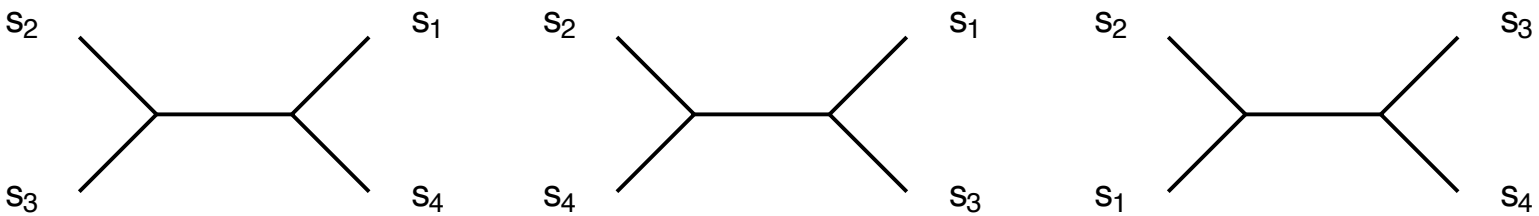
Question: How many possible internal sequence assignments exist?

Trees with same parsimony score



Rooted trees obtained from the same unrooted tree have the same parsimony score!

Trees with potentially different parsimony score



Parsimony tree reconstruction

- **Input:** Sequence alignment of $\mathfrak{n}$ sequences, with sequence length $\mathfrak{m}$
- **Iterate:**
  - Consider each unrooted tree (overall there are $(2\mathfrak{n}-5)!!$ unrooted trees). This step cannot be improved unless P=NP.
  - Calculate parsimony score for the unrooted tree (requires to consider $4^{\mathfrak{n}-1}\mathfrak{m}$ internal sequence assignments). This step can be improved considerably using the Fitch algorithm (next slide).
- **Output:** Unrooted tree with the lowest parsimony score

Fitch algorithm to quickly determine parsimony score

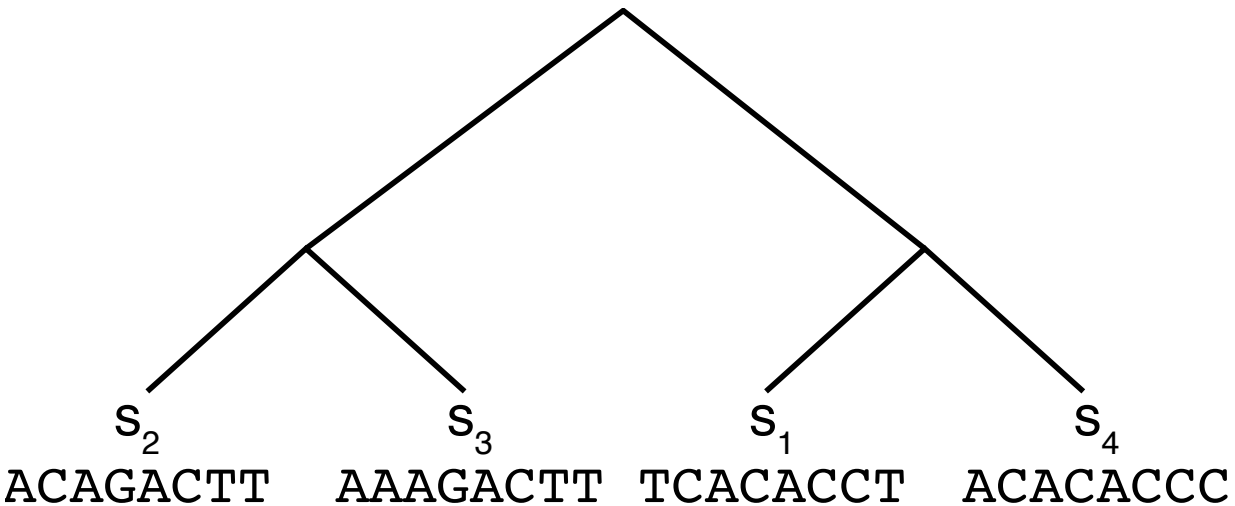
Input: Unrooted phylogenetic tree and an alignment of n sequences of length m, corresponding to the n tips of the tree.

Computational steps:

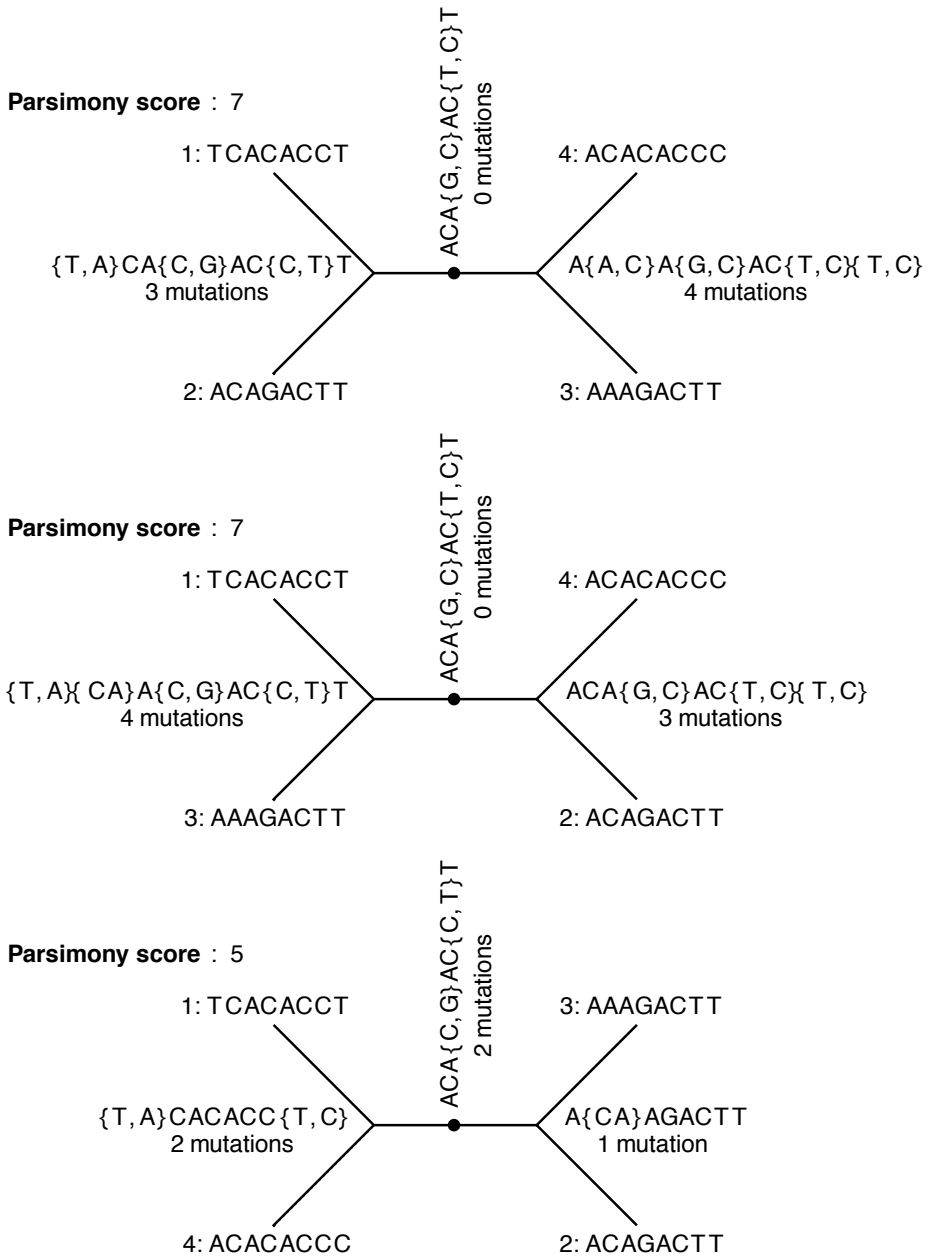
- root the tree at an arbitrary edge
- $k \leftarrow 0$
- while the root has no sequence assigned, iterate:
  1. choose a node in the tree where all descending nodes have sequences assigned
  2. assign a sequence to the chosen node:
     For $i = 1, \ldots, m$, do the following:
     Let C_l and C_r be the sets of nucleotides assigned to the two direct descendants of the chosen node for site i.
     If $C_l \cap C_r \neq \emptyset$, we assign $C_l \cap C_r$ to nucleotide i of the chosen node.
     If $C_l \cap C_r = \emptyset$, we assign $C_l \cup C_r$ to nucleotide i of the chosen node and set $k \leftarrow k + 1$.

Output: Parsimony score k of the tree, i.e. the minimal number of mutations required to explain the sequences at the tips.

Example - Parsimony score of our UPGMA tree



Parsimony score for the three unrooted trees

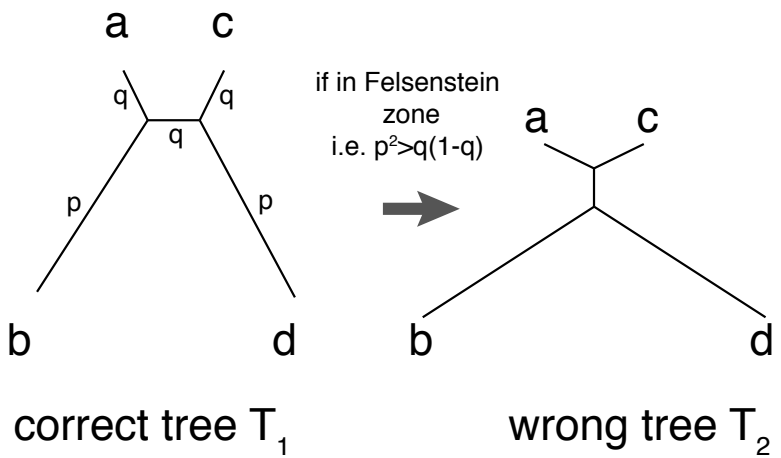


Computational effort

- We have to visit each internal node of the rooted tree.
- How many internal nodes does a rooted tree on n tips have? $n-1$
- Total number of steps: $(n-1)m$ with m sequence length. Thus, the Fitch algorithm improved the running time from $4^{n-2}m$ to $(n-1)m$ by using dynamic programming, i.e. solving the problem on subtrees (=solving subproblems) and then combining the solution for the bigger tree (=solving problem).
- Parsimony tree is found by calculating parsimony score for each unrooted tree
  ☞ The parsimony decision problem is NP-complete

Statistical inconsistency of parsimony: long branch attraction in the Felsenstein zone

Is parsimony statistically consistent?



Consider only two nucleotides (0 and 1) with same rate of transition for $0 \to 1$ and $1 \to 0$.
We denote the probability of change along the long branch with $\mathbf{p}$ and on the short branch with $\mathbf{q}$.
Parsimony infers the wrong tree if $\mathbf{p}, \mathbf{q}$ are in the Felsenstein zone, i.e. $\mathbf{p}^2 > \mathbf{q}(1-\mathbf{q})$.
E.g. $\mathbf{p} = 0.4, \mathbf{q} = 0.1$ is in the Felsenstein zone, as $0.16 > 0.09$ [Felsenstein, 1978].

UPGMA versus parsimony

Parsimony tree and UPGMA tree may be different, try yourself:

sequence 1: `AATAATT`
sequence 2: `ATATTAA`
sequence 3: `TAGGGAA`
sequence 4: `TGAAAGG`

Summary cladistic tree inference

- Uses full genetic sequences rather than the pairwise distances,

but:

- **statistically inconsistent:** no back substitutions or parallel substitutions are considered, which leads to long-branch attraction.

Nevertheless, can be used in certain situations for non-genetic character data.

Moreover, a relatively large community of loyal supporters continue to use cladistic methods for macroevolution.

Mechanistic tree inference: I Maximum likelihood method.

Example of maximum likelihood tree inference.

Use of ML methods: Origin of an infectious disease

The New York Times

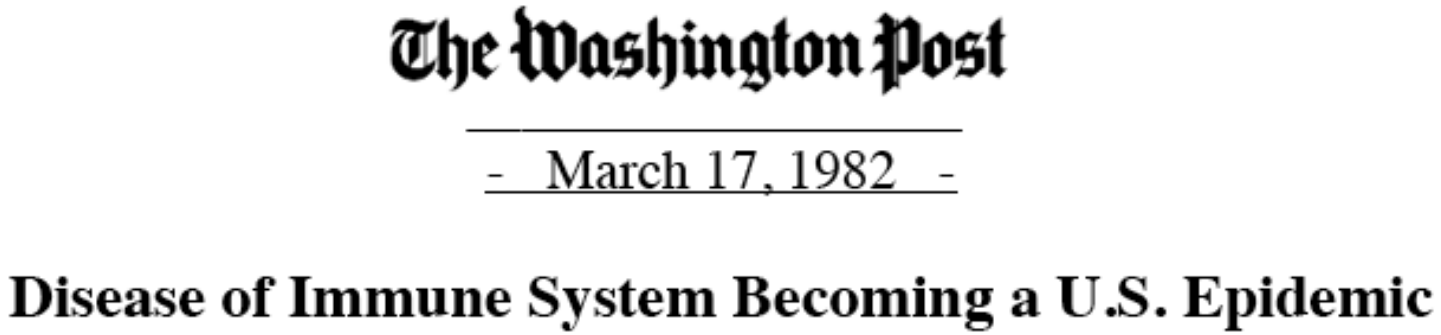
- July 3, 1981 -

RARE CANCER SEEN IN 41 HOMOSEXUALS

Outbreak Occurs Among Men in New York and California -- 8 Died Inside 2 Years

- "...have designated the immune disorder GRID, for gay-related immunodeficiency disease."

AIDS



The Washington Post

- March 17, 1982 -

Disease of Immune System Becoming a U.S. Epidemic

- "As of March 9, 1,145 Americans had contracted AIDS (acquired immunodeficiency syndrome); 428 of them are dead. Twenty percent of the cases appeared in the last two months. Half of the victims are under 35."

Cause of AIDS

April 22, 1984

FEDERAL OFFICIAL SAYS HE BELIEVES CAUSE OF AIDS HAS BEEN FOUND

By LAWRENCE K. ALTMAN , Special to the New York Times

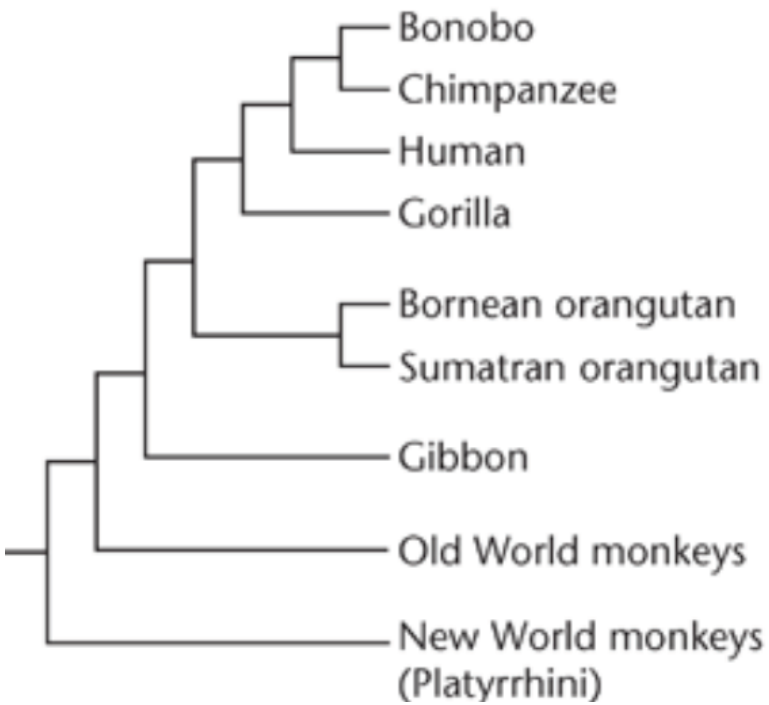
ATLANTA, April 21— Dr. James O. Mason, head of the Federal Centers for Disease Control, said today that he believed a virus discovered in France was the cause of acquired immune deficiency syndrome, or AIDS.

- "We now call this virus HIV (human immunodeficiency virus)."
- As of today, 0.5% of the human population is HIV positive.

Where did HIV come from?

- No similar pre-HIV virus known in human population;
- HIV must have evolved from some previously existing virus.

Simian immunodeficiency virus found in many simian species



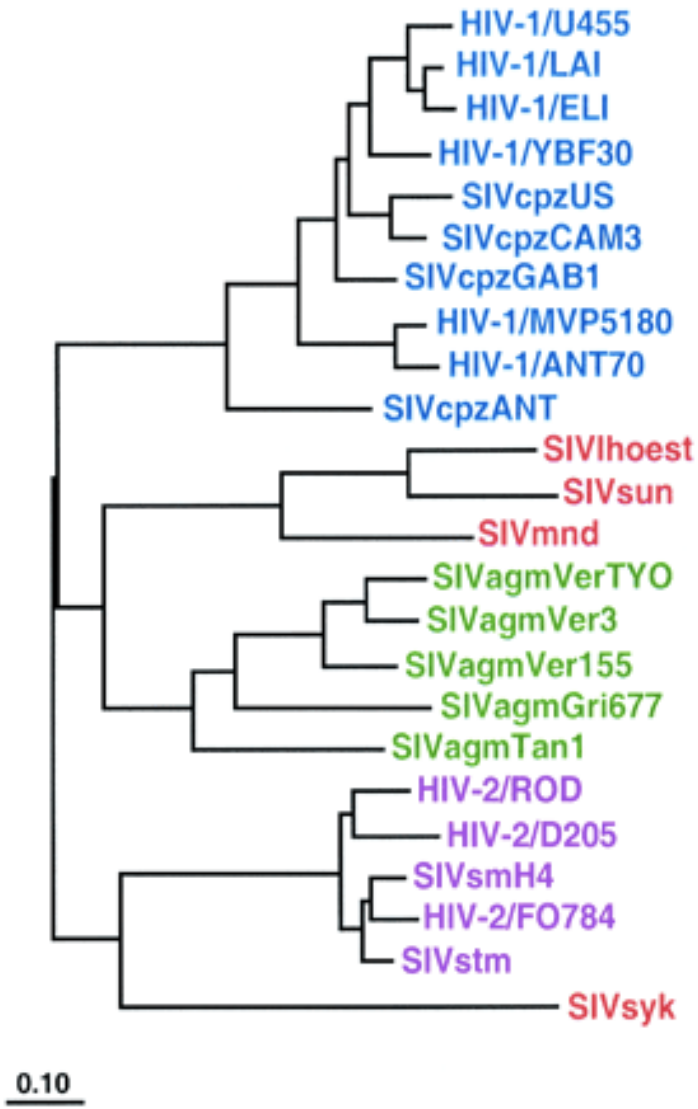
Species tree of simians based on simian species genetic data.

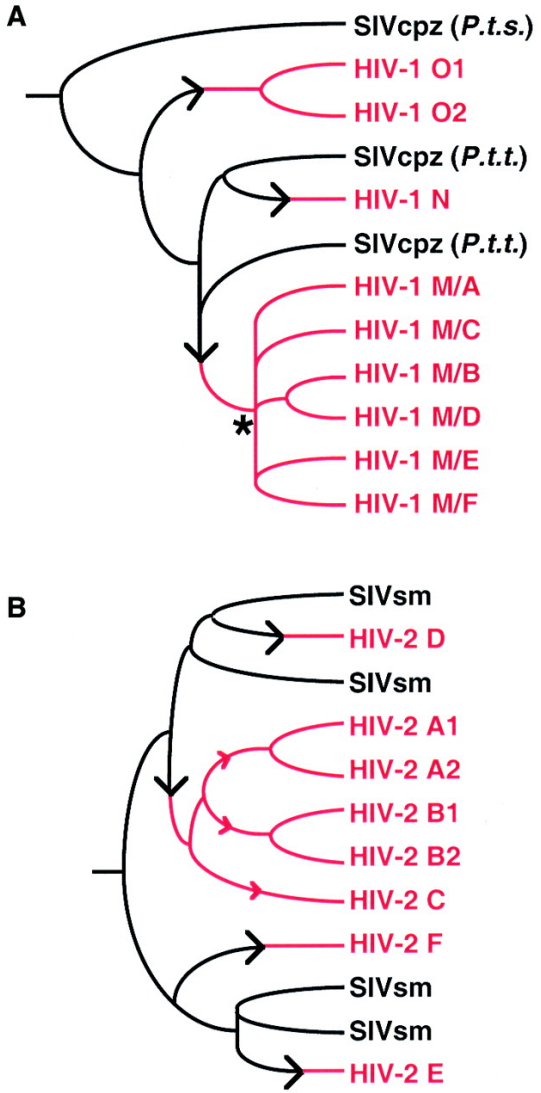
- Is HIV a zoonosis from simians to humans?

How do we determine the origin of HIV?

- Which data is required?
  - Incidence data tells us about the dynamics since this data was collected (i.e. post-1980)
  - Virus sequencing data from different host species collected post-1980 allows us to infer the phylogenetic tree informing us about pre-1980 (remember: branching events are transmission events, and these events may be way earlier than 1980!)
  - Huge efforts in 1990s to collect SIV sequencing data in Africa
- Maximum likelihood phylogenetic tree inference was used to investigate early HIV!

How do we determine the origin of HIV?





several host jumps have occurred

Maximum likelihood trees of full-length Pol sequences [Hahn et al., 2000]

Calculating the maximum likelihood tree.

Maximum likelihood tree inference

- **Input:** Sequence alignment
- **Output:** Tree which maximises the probability of the sequences given the tree & the sequence evolution parameters.
  - Requires an evolutionary model (JC69, HKY, GTR etc)
  - Parameters of the evolutionary model are co-estimated with the tree.

Maximum likelihood framework

- Lecture 4: Likelihood of a parameter is the probability of the observed data given the parameter.
- Maximum likelihood estimate is the parameter maximising the likelihood function given the data.
  - Assuming the number of 6 observed in repeated dice throws is binomially distributed, the fraction of dice throws where we observe a 6 is the maximum likelihood estimate for the probability p of observing a 6 (p is the parameter).
  - Assuming measurements are normally distributed, the average of the measurements is the maximum likelihood estimate for the mean m of the normal distribution (m is the parameter).

☞ What is the maximum likelihood phylogeny given the sequence data (the phylogeny is the parameter)?

Maximum likelihood in phylogenetics

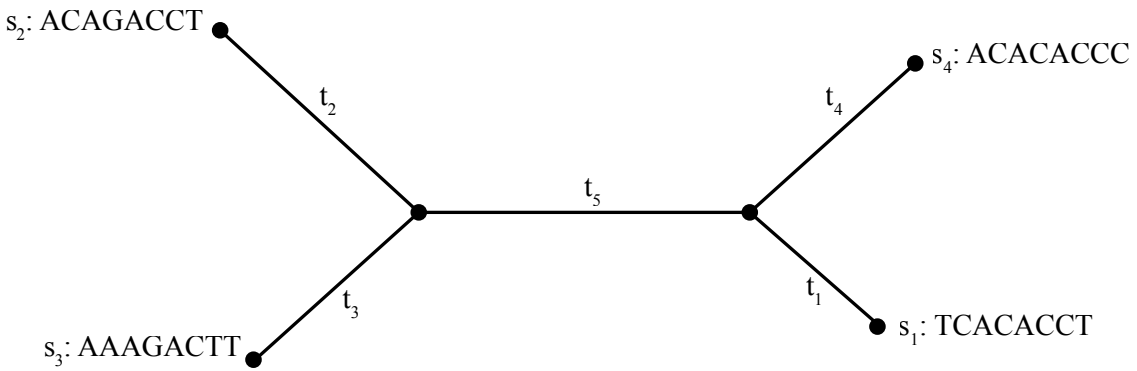
- Mechanistic model for evolution of the data (sequences):
  - Each unrooted tree $\mathcal{T}$ with branch lengths is a parameter
  - Sequences evolve on the tree according to parameters provided in the rate matrix Q

- $\mathrm{L}(\mathcal{T}, \mathrm{Q}; \mathrm{D}) := \mathrm{P}(\mathrm{D}|\mathcal{T}, \mathrm{Q})$ is called the *likelihood function* of the parameters $\mathcal{T}, \mathrm{Q}$ for the given sequence data
- **Inference:** Determine the $\mathcal{T}, \mathrm{Q}$ which best explain the alignment:

$$\max_{\mathcal{T}, \mathrm{Q}} \mathrm{L}(\mathcal{T}, \mathrm{Q}; \mathrm{D})$$

- We determine the best tree by evaluating the likelihood for "many" different proposed trees

Likelihood calculation for a given tree



- We assume that the sites in the alignment evolve independently from each other, i.e. we can consider each site separately.

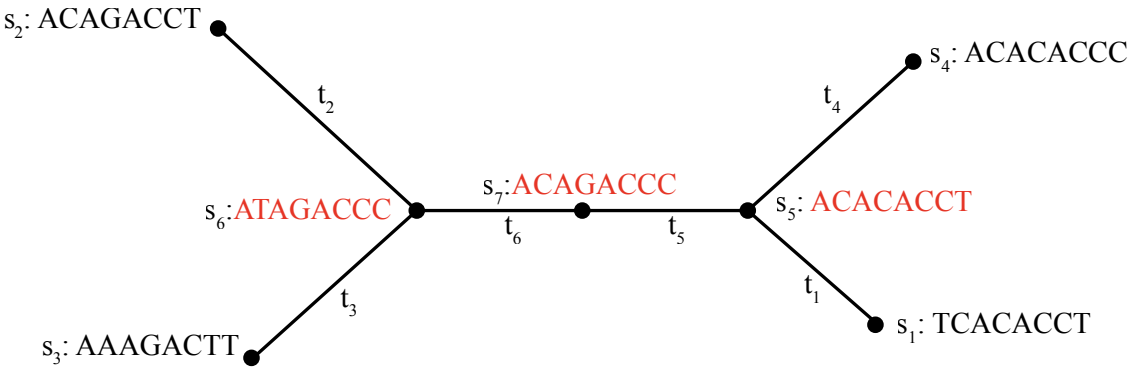
  Let the alignment consist of m sites, then:

$$P(s_1, \ldots, s_n | \mathcal{T}, Q) \qquad \leftarrow \text{this is the quantity we want!!}$$

$$= \prod_{j=1}^{m} P(s_{1,j}, \ldots, s_{n,j} | \mathcal{T}, Q)$$

  (where $s_{k,j}$ is the site j of sequence s_k).

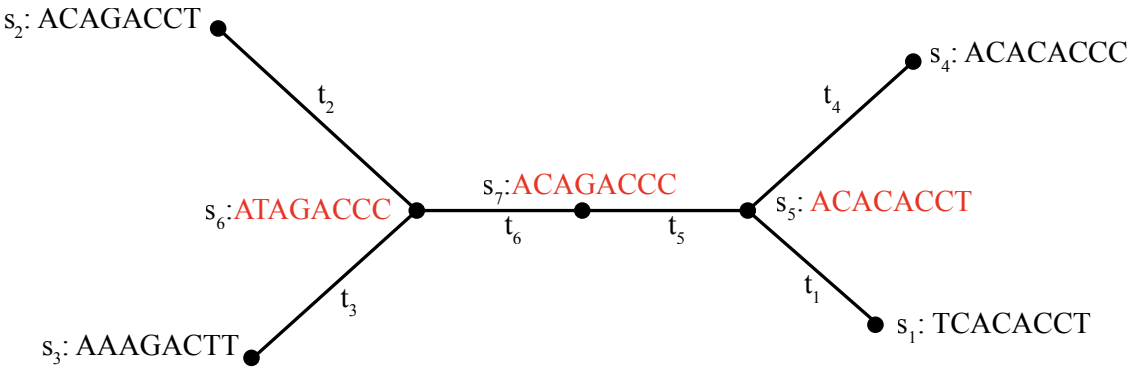
Likelihood calculation - sum over internal node sequences



- Typically, the substitution process is time-reversible.
- Thus for a proposed unrooted tree, subdivide an arbitrary edge to obtain a root, with (unknown) sequence s_{2n-1}.
- For n sequences, the rooted tree has $n-1$ internal nodes, with (unknown) sequences $s_{n+1}, \ldots, s_{2n-1}$.
- The probability of the nucleotides at the tips at site j is the sum over the probabilities of nucleotide states at the internal nodes and tips (where $s_{k,j}$ is the site j of sequence s_k),

$$P(s_{1,j}, \ldots, s_{n,j} | \mathcal{T}, Q) = \sum_{s_{n+1,j} \in \{A,C,G,T\}} \cdots \sum_{s_{2n-1,j} \in \{A,C,G,T\}} P(s_{1,j}, s_{2,j}, \ldots, s_{2n-1,j} | \mathcal{T}, Q)$$

Likelihood calculation - multiply all branches in tree

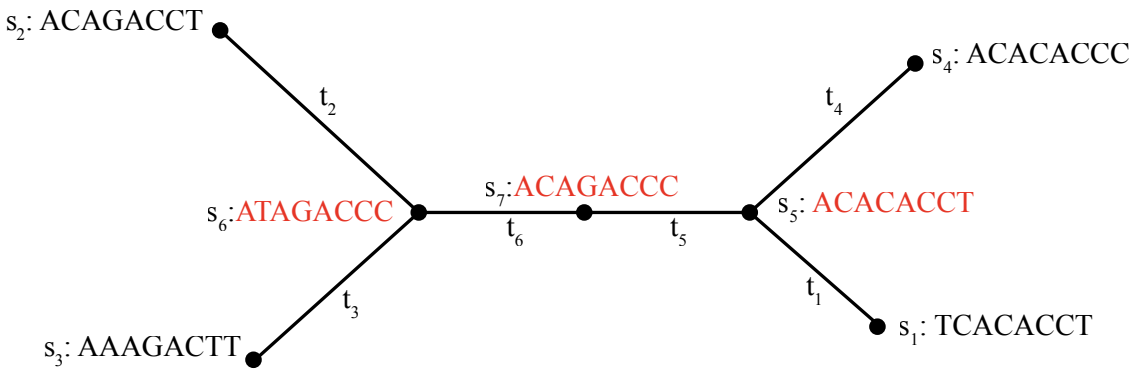


- Finally, $P(s_{1,j}, s_{2,j}, \ldots, s_{2n-1,j}|\mathcal{T}, Q)$ can be evaluated by calculating for each branch l (with starting sequence s_{l_1}, ending sequence s_{l_2}, and branch length t_l) the transition probability from the ancestral nucleotide to the descendant nucleotide.
- The root nucleotides (here s_7) are weighted by their equilibrium probabilities π, so overall,

$$P(s_{1,j}, s_{2,j}, \ldots, s_{2n-1,j}|\mathcal{T}, Q) =$$
$$\pi(s_{2n-1,j}) \prod_{l=1}^{2n-2} P_{s_{l_1,j}, s_{l_2,j}}(t_l)$$

(lecture 5: a rooted tree on n tips has $2n-2$ branches.)

Example



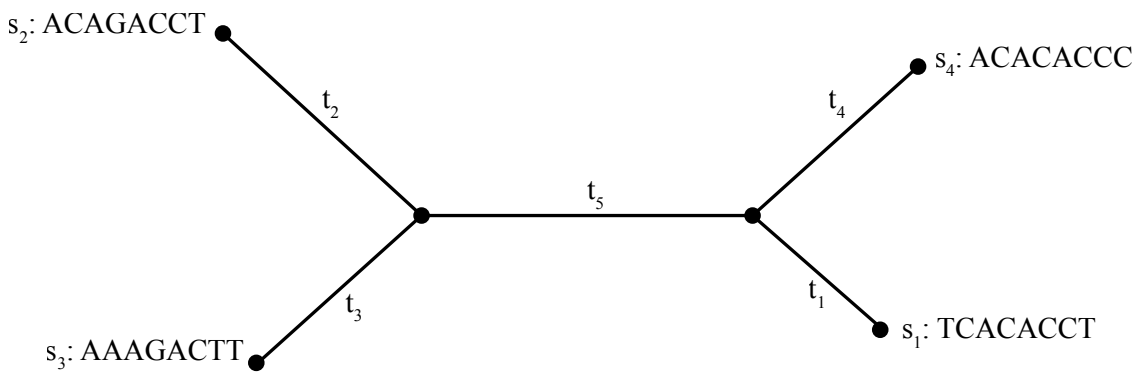
For site $j = 2$, we have,

$$P(s_{1,2}, s_{2,2}, \ldots, s_{2n-1,2} | \mathcal{T}, Q) = \pi(s_{2n-1,2}) \prod_{l=1}^{2n-2} P_{s_{l_1,2}, s_{l_2,2}}(t_l)$$

$$= \pi_C P_{C,C}(t_5) P_{C,C}(t_4) P_{C,C}(t_1) P_{C,T}(t_6) P_{T,C}(t_2) P_{T,A}(t_3)$$

We have $4 \times 4 \times 4 = 64$ possibilities for nucleotides at internal nodes of site j. Thus the sum in our tree with three internal nodes consists of 64 summands!

Taken together...



The probability to observe the data given a specific tree, $\mathcal{T}$ and a substitution rate matrix, Q, is

$$P(s_1, \ldots, s_n | \mathcal{T}, Q) =$$

$$\prod_{j=1}^{m} \left[\sum_{s_{n+1,j} \in \{T,C,A,G\}} \cdots \sum_{s_{2n-1,j} \in \{T,C,A,G\}} \left(\pi(s_{2n-1,j}) \prod_{l=1}^{2n-2} p_{s_{l_1,j}, s_{l_2,j}}(t_l) \right) \right]$$

where $p_{s_{l_1,j}, s_{l_2,j}}(t_l)$ is the transition probability from the nucleotide at the start of branch l,
$s_{l_1,j}$, to the nucleotide at the end of branch l,
$s_{l_2,j}$, at site j, with branch length t_l.
As derived in lecture 3, the substitution rate matrix Q defines the transition probabilities.

Running time

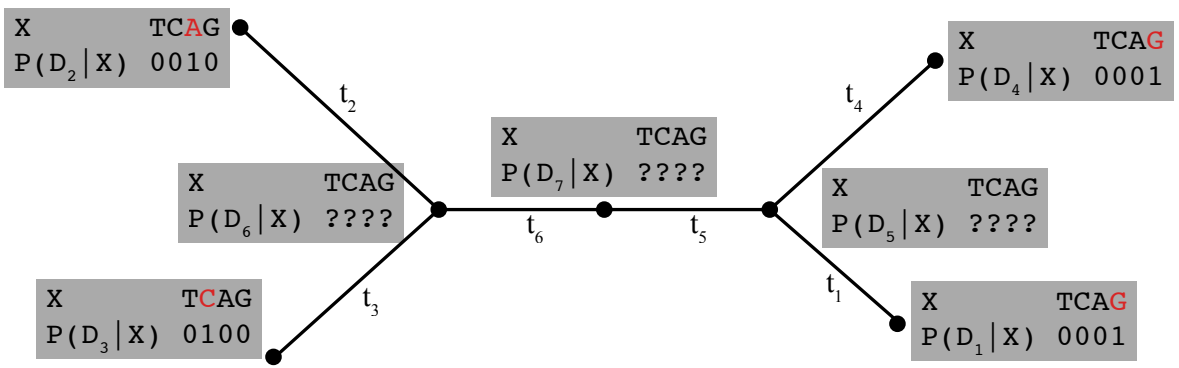
1. we need to visit each single tree in the tree space
2. for each tree we need to calculate the likelihood:
   - multiply over all sites ($\mathbf{O(m)}$)
   - sum over internal nucleotides at $\mathbf{n-1}$ internal nodes ($\mathbf{O(4^{n-1})}$)
   - multiply over $\mathbf{2n-2}$ branches ($\mathbf{O(2n-2)}$)

☞ running time of likelihood calculation is $\mathbf{O(m4^n n)}$ – very slow

☞ Felsenstein's pruning algorithm speeds up this calculation using a clever way to store intermediate steps that are reused in the summation [Felsenstein, 1981]

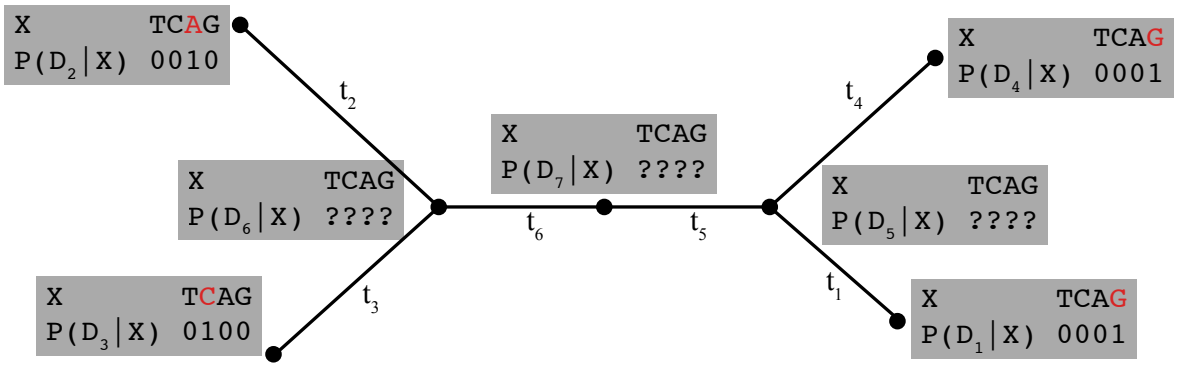
Improving the running time of ML tree inference: Felsenstein's pruning algorithm.

Felsenstein's pruning algorithm



- Make summation efficient (polynomial in n) through dynamic programming [Felsenstein, 1981]!
- We consider site j (and later multiply all sites as before).
- **Let the probability of the tip nucleotides descending from the node k be $P(D_k|X)$, given the nucleotide at node k is X.**
- The tree is traversed as in the Fitch algorithm to determine $P(D_k|X)$ for all nodes.

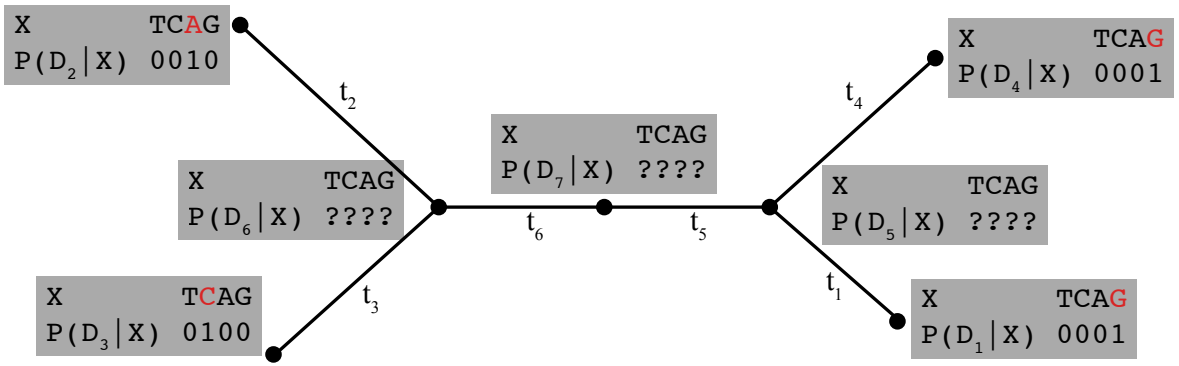
Felsenstein's pruning algorithm



- At a tip k, $P(D_k|X) = 1$ iff X is the observed nucleotide; $P(D_k|X) = 0$ otherwise ($X \in \{A, C, G, T\}$).
- "Cherries" are pruned recursively towards the root, let k be a node with the descendants l, m:

$$P(D_k|X) = \left(\sum_{Y \in \{A,C,G,T\}} P_{X,Y}(t_l) P(D_l|Y) \right) \times \left(\sum_{Z \in \{A,C,G,T\}} P_{X,Z}(t_m) P(D_m|Z) \right).$$

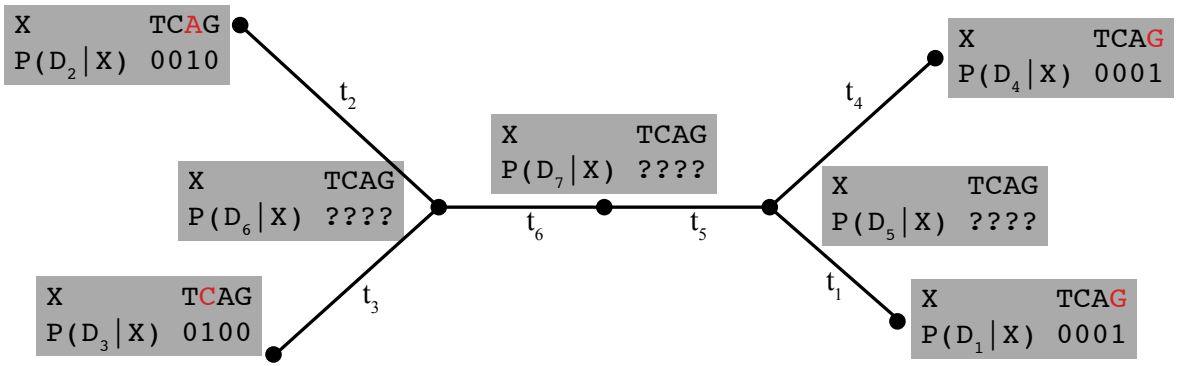
Felsenstein's pruning algorithm



- Thus for the root r, we calculated $P(D_r|X)$ where $X \in \{A, C, G, T\}$.
- Finally, the probability of the sequences at site j is

$$P(s_{1,j}, \ldots, s_{n,j}|\mathcal{T}, Q) = \sum_{X \in \{A,C,G,T\}} P(D_r|X)\pi_X.$$

Felsenstein's pruning algorithm - an example



- Example: evaluate the "?" representing $P(D_6|T)$:

$$\begin{aligned} P(D_6|T) &= \left(\sum_{Y\in\{A,C,G,T\}} P_{T,Y}(t_2)P(D_2|Y)\right) \\ &\times \left(\sum_{Z\in\{A,C,G,T\}} P_{T,Z}(t_3)P(D_3|Z)\right) \\ &= P_{T,A}(t_2) \times P_{T,C}(t_3) \end{aligned}$$

Felsenstein's pruning algorithm

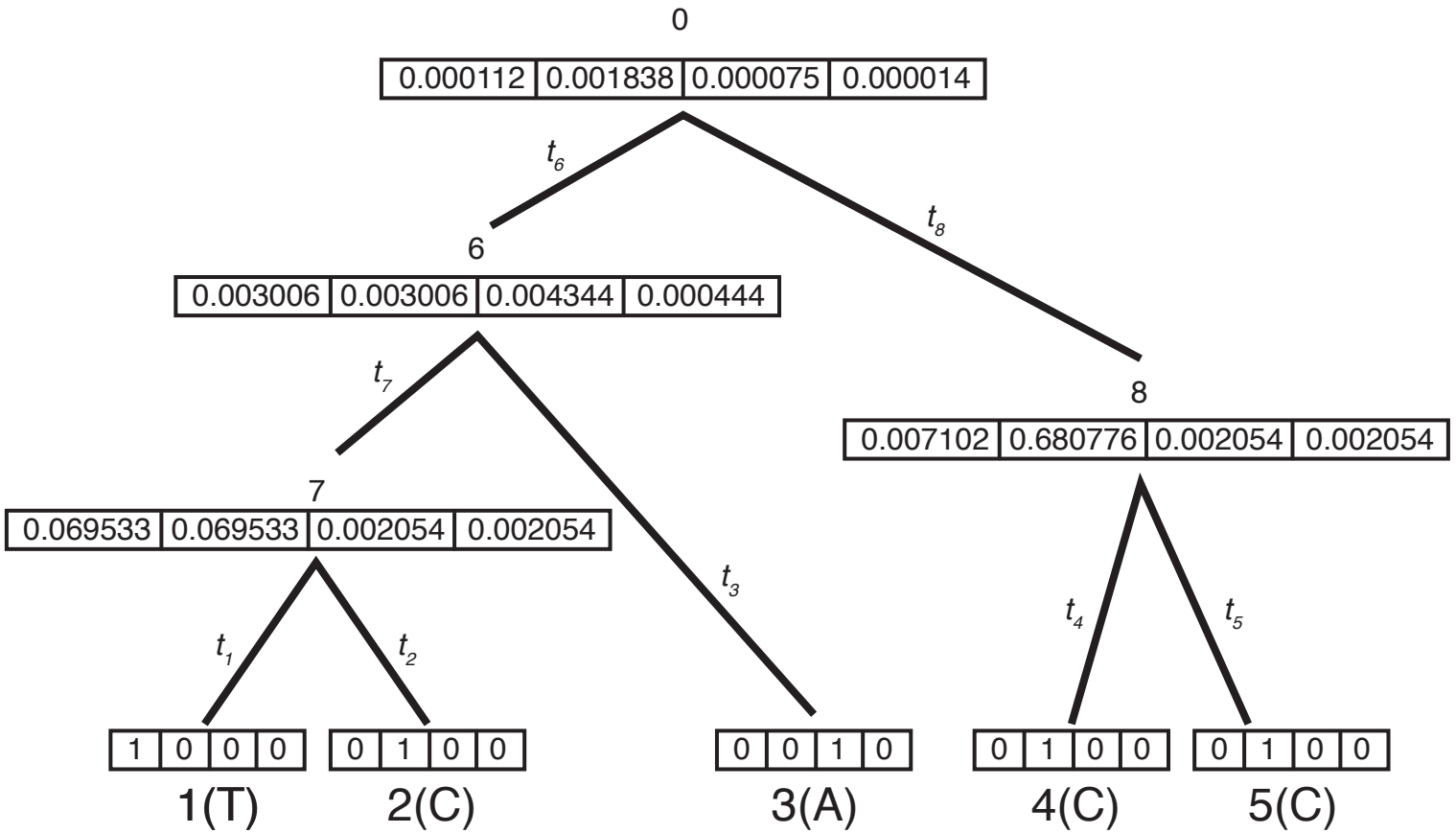


Figure adapted from [Yang, 2014]

The "?" evaluated for an example in Figure 4.2 of [Yang, 2014] (page 105), using the K80 model, pendant branch lengths are $t_1, t_2, t_3, t_4, t_5 = 0.2$ and internal branch lengths are $t_6, t_7, t_8 = 0.1$.

Running time of Felsenstein's pruning algorithm

The running time of Felsenstein's likelihood calculation can be determined as follows:

- each recursion step is a summation over two times four states, i.e. constant; we have $\mathrm{O}(\mathrm{n})$ nodes and thus the recursion has running time $\mathrm{O}(\mathrm{n})$
- the recursion has to be performed for each of the m sites, $\mathrm{O}(\mathrm{m})$
- thus, in total the running time is $\mathrm{O}(\mathrm{nm})$

☞ Felsenstein's pruning algorithm speeds up the likelihood step from $\mathrm{O}(\mathrm{m}4^{\mathrm{n}}2\mathrm{n})$ (brute force) to linear, $\mathrm{O}(\mathrm{nm})$

☞ The problem of finding a tree and branch lengths with likelihood value $\leqslant \mathrm{L}$ is NP-complete

Maximum likelihood inference in a nutshell

- Goal is to maximize the likelihood formula:

$$\max_{\mathcal{T},\mathrm{Q}} \mathrm{L}(\mathcal{T}, \mathrm{Q}; \mathrm{D})$$

  to obtain the maximum likelihood parameter estimate (i.e. the ML tree and the ML rates).
- We have to consider all unrooted trees (very slow)
- Additionally we have to try out "all" realistic branch lengths (slow)
- We have to integrate (sum) over all internal node sequences (fast with Felsenstein's pruning algorithm)

Summary.

Summary inference methods

phenetic approaches:

- algorithmic methods:
  - UPGMA: assumes strict molecular clock, i.e. branch lengths correspond to calendar time
    we obtain a rooted, ultrametric tree
  - neighbour-joining algorithm [Saitou and Nei, 1987] (not explained here): branch lengths correspond to number of mutations
    output tree will be unrooted
  - polynomial running time, statistically consistent
- optimality methods:
  - least squares methods
  - NP-complete, statistically consistent
- phenetic approaches disregard information beyond pairwise distances

Summary inference methods

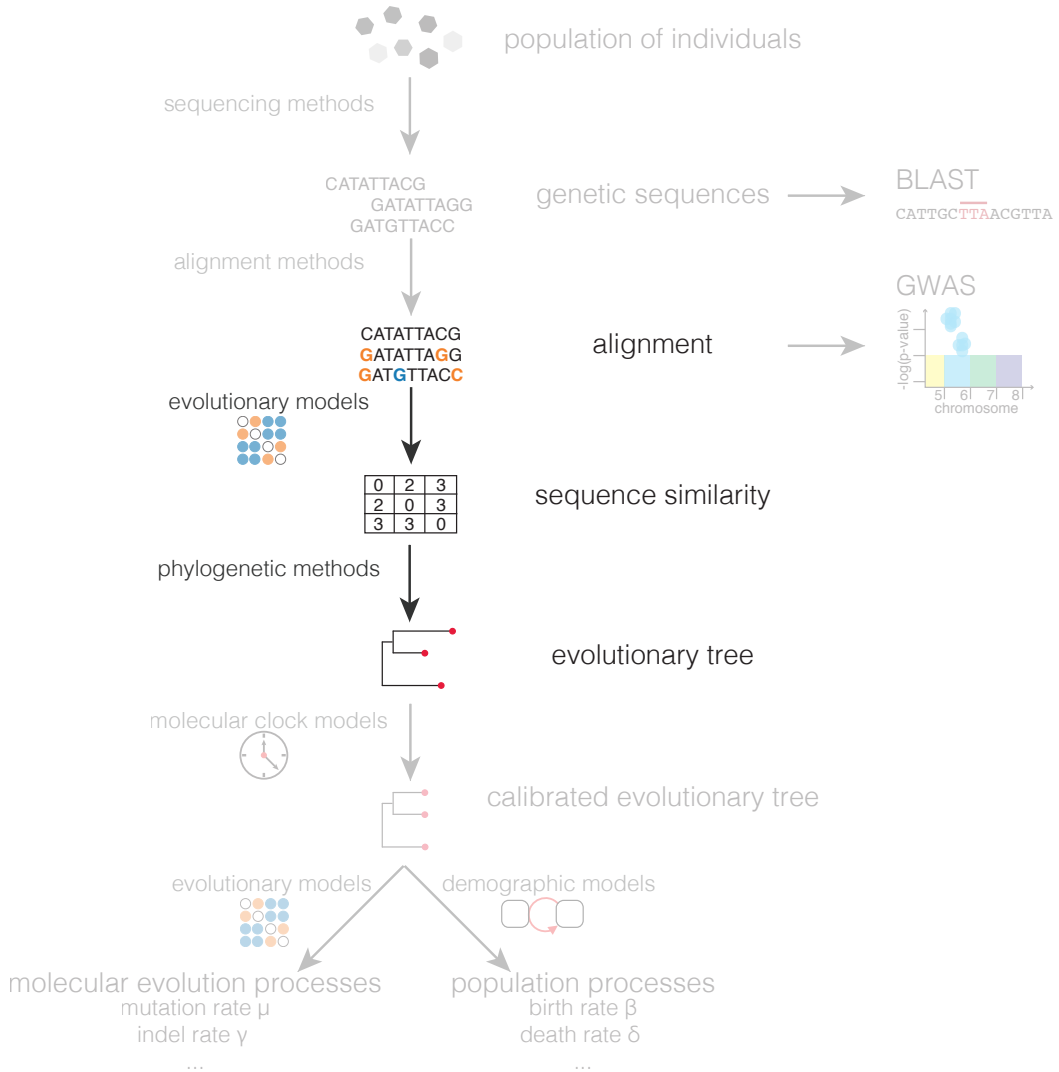
cladistic approaches:

- parsimony:
  - returns the tree that requires the least amount of mutations
  - statistically inconsistent
  - NP-complete

mechanistic approaches:

- maximum likelihood method:
  - returns the maximum likelihood tree with branch lengths corresponding to number of mutations alongside evolutionary model parameters
  - explicitly accounts for evolution
  - statistically consistent
  - NP-complete

Overview



Parsimony & Maximum Likelihood: Questions

- Consider the Fitch algorithm. Do you obtain all most parsimonious ancestral sequences when choosing the different nucleotides in the curly brackets?
- Does the maximum likelihood tree reconstruction method return estimates for the internal sequences? Give a reason for your answer.
- Does the Fitch algorithm return the parsimony score for any phylogenetic tree and any sequence alignment? Or are there situations when the Fitch algorithm does not return the smallest number of mutations required?

References I

- Felsenstein, J. (1978). The number of evolutionary trees. *Systematic Biology*, 27(1):27–33.
- Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.
- Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic zoology*, pages 406–416.
- Hahn, B. H., Shaw, G. M., De, K. M., Sharp, P. M., et al. (2000). Aids as a zoonosis: scientific and public health implications. *Science*, 287(5453):607–614.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.
- Sforza, C. L. and Edwards, A. W. F. (1964). Reconstruction of evolutionary trees. *in: Phenetic and Phylogenetic Classification*, pages 67–76.
- Yang, Z. (2014). *Molecular Evolution – A Statistical Approach*. Oxford University Press.