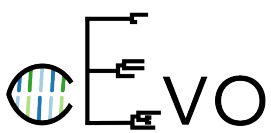
Computational Biology

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



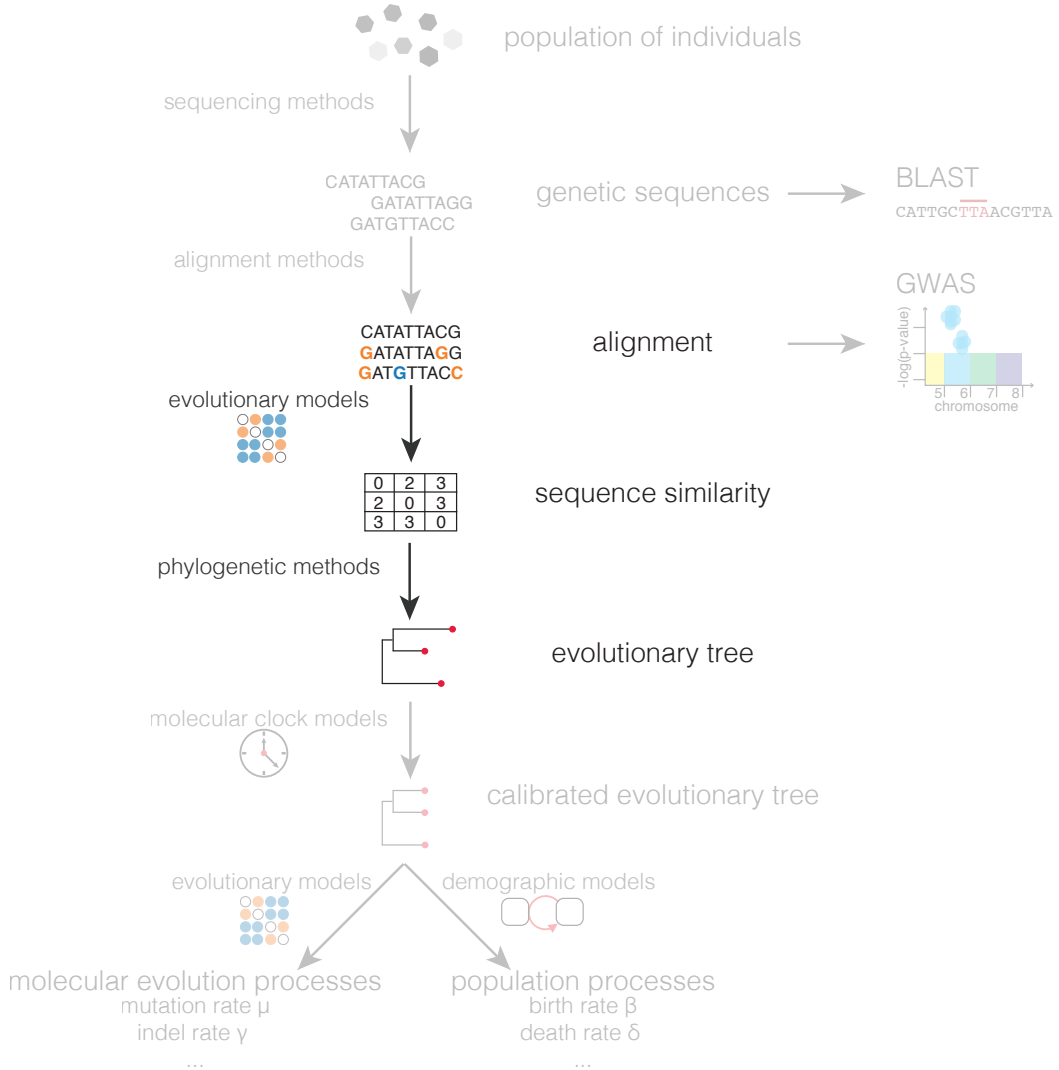
05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Molecular evolution, AA and codon substitution models: Questions

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

- What are the differences between the JC69 and the TN93 models?
- Which of the two models would you chose if you were to perform a phylogenetic analysis based on sequence distances?
- In lecture 3 we tried to naively reconstruct a phylogeny based on three sequences (the same tree and sequences appear on slide 33 of this lecture). The pairwise Hamming distances were all the same. Would you expect that any of the presented nucleotide sequence models would result in different trees?

Overview



05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Overview

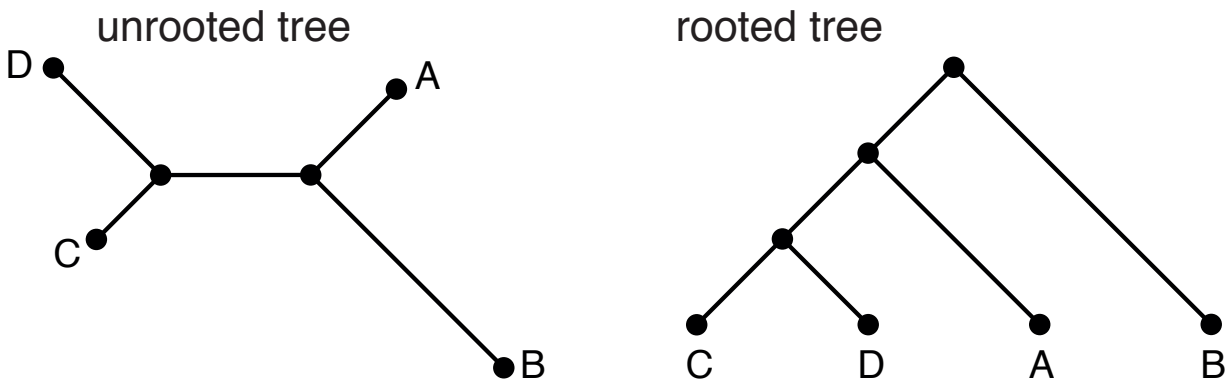
- phylogenies:
  - definition
  - vocabulary
  - visualisation
  - examples
- how to infer phylogenies:
  - phenetic approaches: UPGMA, least squares
  - cladistic approaches in lecture 6
  - mechanistic approaches in lecture 6 & 7 (maximum likelihood) and lecture 11 (Bayesian)

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Introduction to phylogenetic trees.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

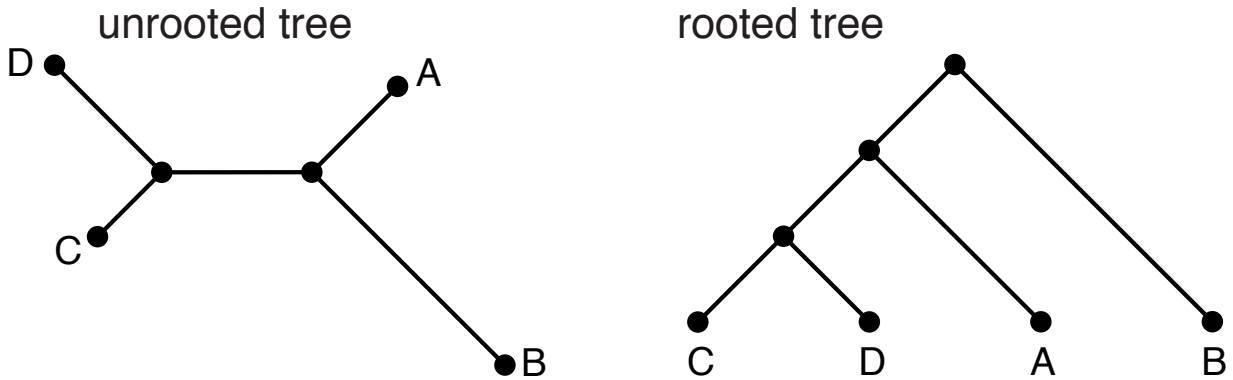
Definition of a phylogenetic tree



- a **tree** is a graph of nodes and branches without loops
- an **unrooted phylogenetic tree** is a tree with two types of nodes:
  - **tip/leaf:** node with 1 branch attached
  - **internal node:** node with 3 branches attached
- a **rooted phylogenetic tree** is an unrooted tree in which one branch is divided by a new node (root)
- unrooted trees can be **rooted** with an **outgroup** (here B), which means that the branch ending in B is subdivided by the root node. B is chosen by the user as a very distantly related organism to the remaining organisms in the tree.
- each branch may have a length $\geqslant 0$ assigned

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

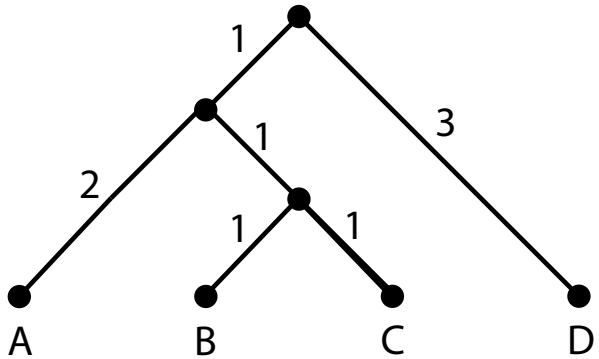
Phylogenetic terms



- **pendant branch:** a branch attached to a tip
- **cherry:** a pair of tips which are only separated by one internal node
- **caterpillar tree:** a rooted tree with only one cherry
- **monophyletic group/clade:** all descendants of an internal node in a rooted tree
- **ultrametric tree:** rooted tree where the sum of branch lengths from any tip to the root is the same
- **polytomy:** the definition of a phylogenetic tree is extended such that internal nodes may have more than 3 branches attached. Such a node is a polytomy. It can be represented as a classic phylogenetic tree with branch lengths of 0.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

String representation (Newick) of rooted trees



Recursively:

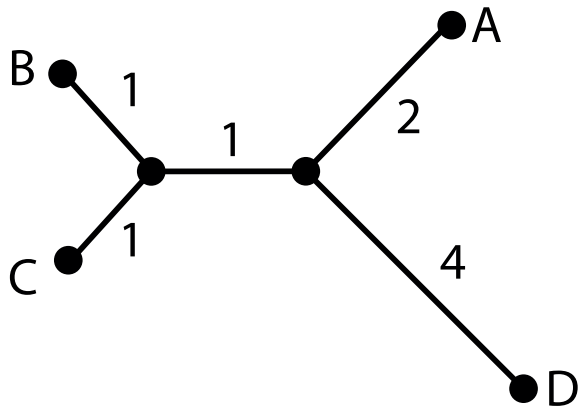
- choose two tips (e.g. C and D) that form a cherry;
- replace the two tips by the new tip $(\mathrm{C} : \mathrm{t_C}, \mathrm{D} : \mathrm{t_D})$, where $\mathrm{t_X}$ is the length of the branch ancestral to node X;
- the length of the branch ancestral to the new tip is the branch length ancestral to the cherry.

What is a Newick format for the rooted tree above?
$(((\mathrm{B} : 1, \mathrm{C} : 1) : 1, \mathrm{A} : 2) : 1, \mathrm{D} : 3)$

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

String representation (Newick) of unrooted trees

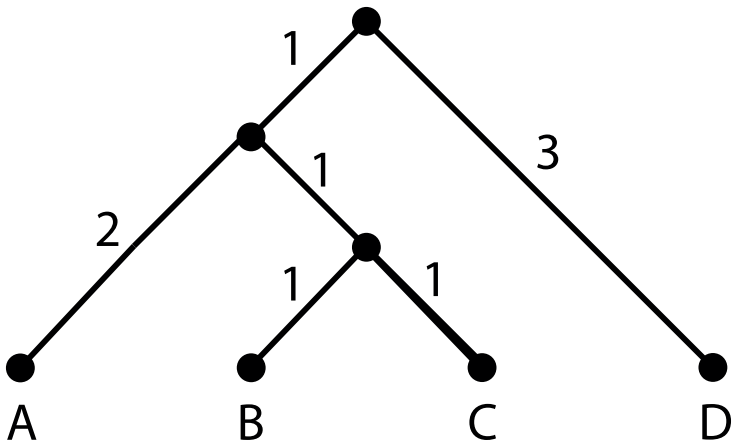
05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References



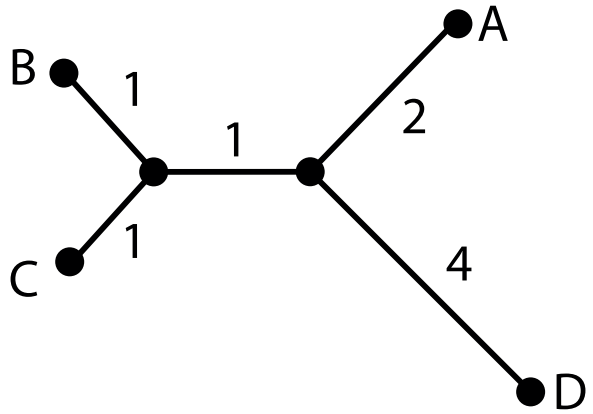
- choose an internal node arbitrarily
- proceed as in the rooted tree towards the chosen internal node.
- connect the three last tips X, Y, Z to $(X : t_X, Y : t_Y, Z : t_Z)$

What is a Newick format (depends on the chosen internal node) for the unrooted tree above? $((A : 2, D : 4) : 1, B : 1, C : 1)$

Equivalent Newick representations



$((((B:1,C:1):1,A:2):1,D:3)$
$((A:2,(C:1,B:1):1):1,D:3)$



$((A:2,D:4):1,B:1,C:1)$
$((B:1,C:1):1,A:2,D:4)$

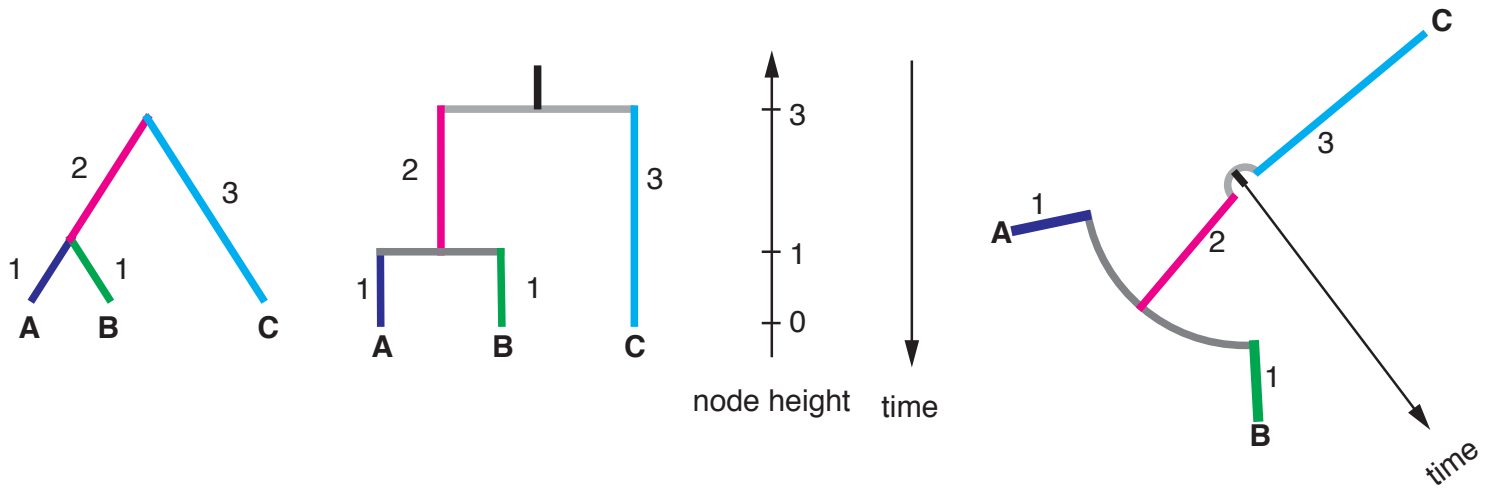
A tree can have different but equivalent Newick representations.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Graphical visualisation of rooted trees

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Phylogenetic trees can be graphically represented in different ways:

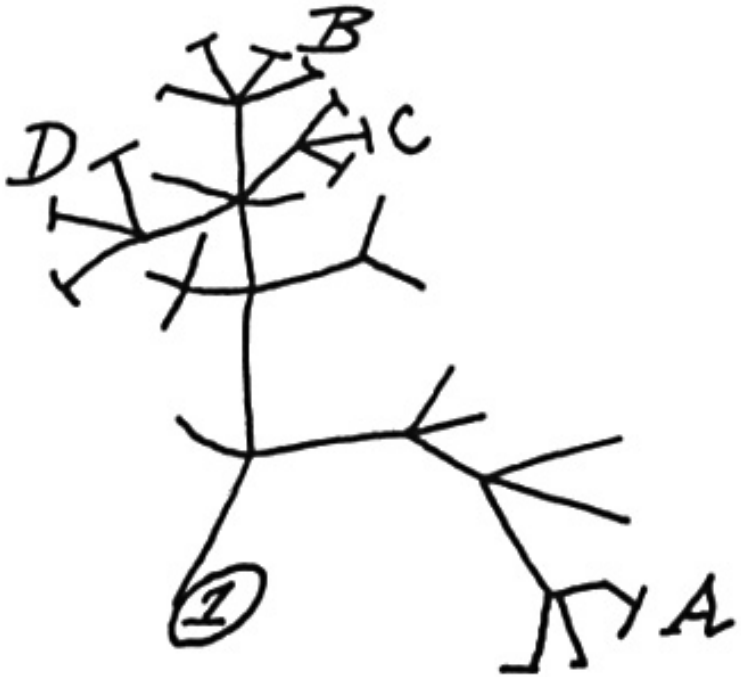


Only the distance along the evolutionary time axes contributes to branch lengths!

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Examples of phylogenies -- and what we can learn from them.

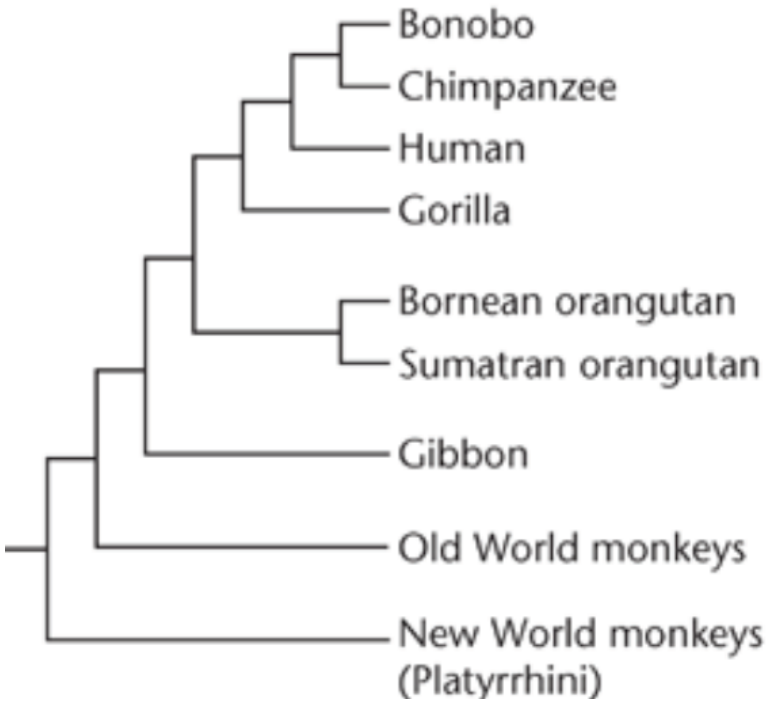
First phylogenetic tree



- phylogenetic tree drawn by Charles Darwin in his 1837 notebook
- graphic representation of how species evolved
  - **tips:** species
  - **branches:** ancestry

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Phylogeny of species



- species phylogeny of simians
  - **tips:** simian species consisting of apes, old world monkeys (78) and new world monkeys (53)
  - **branching events:** speciation events
  - **branch lengths:** time between speciation events
- phylogenies contain information about species relationships

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Tree of Life

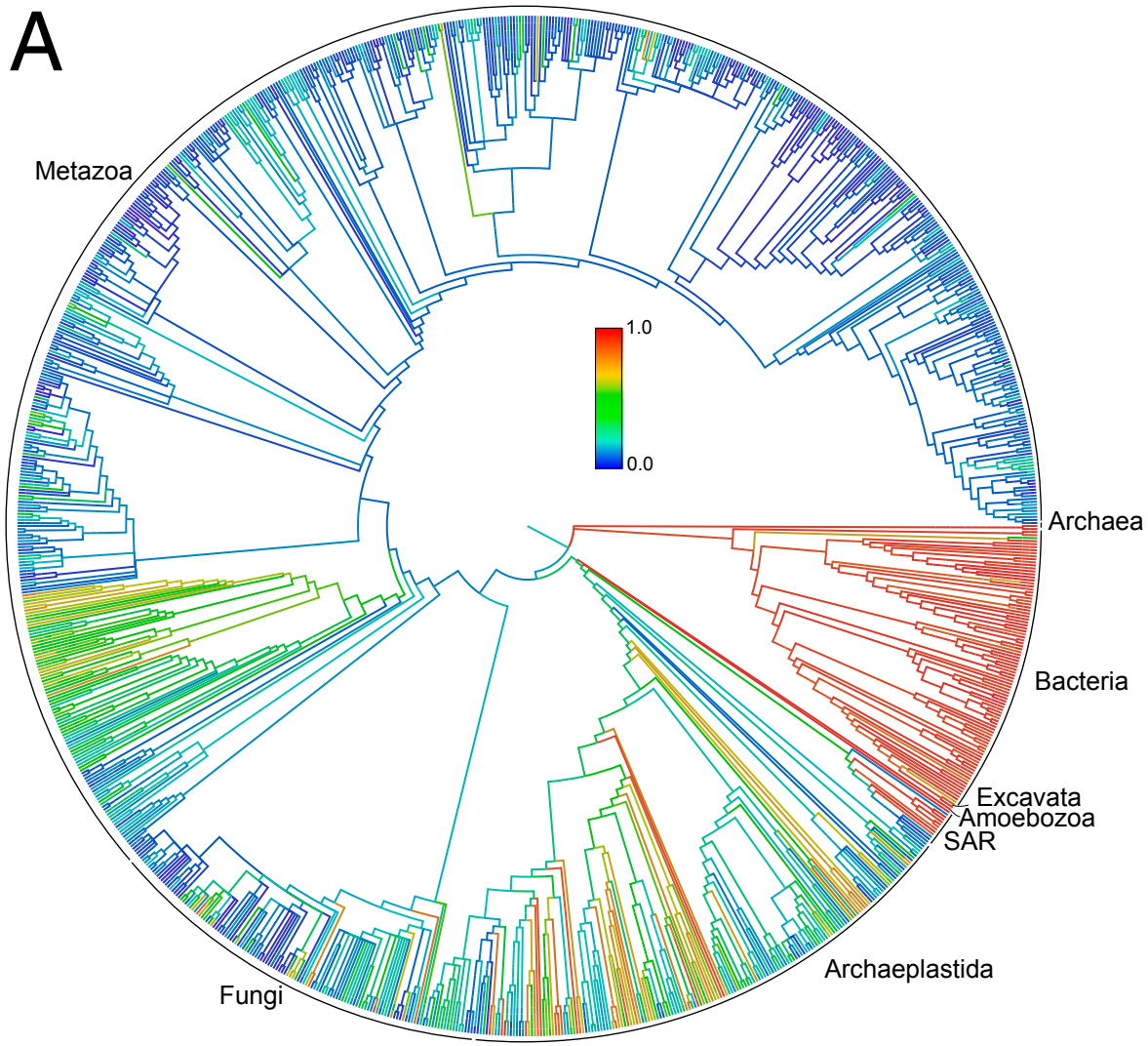
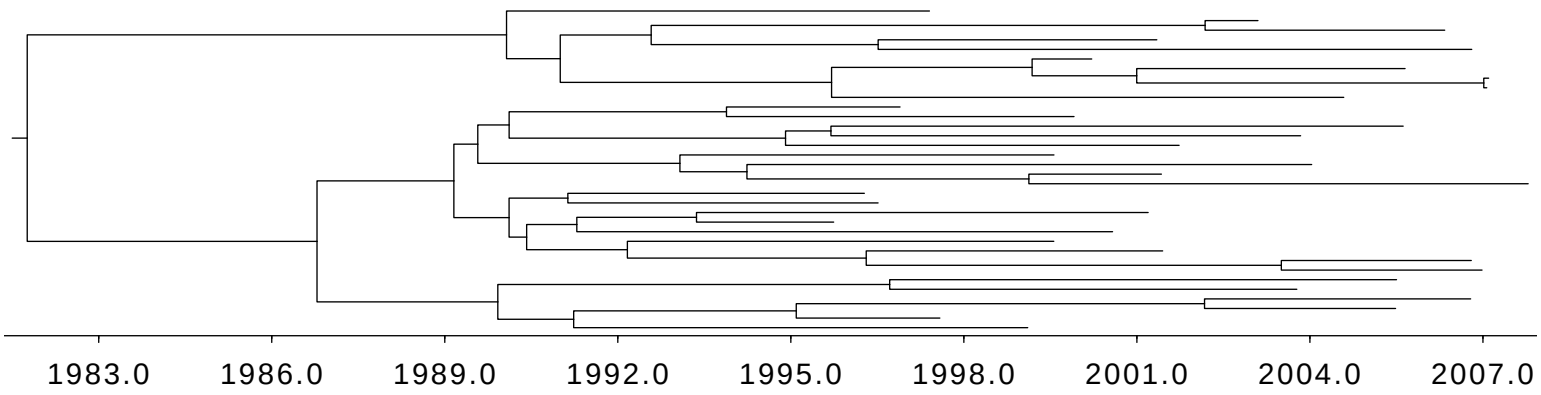


Figure adapted from [Hinchliff et al., 2015]

Tree of life on 2.3 million tips.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Phylogeny of pathogens



- pathogen phylogeny of HIV epidemic
  - **tips:** different infected hosts
  - **branching events:** transmission events
  - **branch lengths:** time between transmission events
- pathogen phylogeny is an approximation of part of the transmission chain, and thus contains information about transmission dynamics (e.g. origin of epidemics, transmission group interaction, criminal cases)

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Phylogenetic inference -- how to reconstruct phylogenetic trees.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

How would you build a phylogeny?

finch phylogeny inferred in [Sato et al., 1999]

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

How are the phylogenies reconstructed?

Similar individuals are clustered together. Data for measuring similarity used to be morphology for species. Now it is typically sequencing data for species or pathogens or B-cells etc.
Similarity may be defined in different ways:

- **phenetic:**
  - based on overall similarity
  - pairwise distance-based
  - methods: UPGMA, least squares algorithm
- **cladistic:**
  - based on shared characteristics
  - character-based
  - method: parsimony
- **mechanistic:**
  - based on evolutionary model
  - character-based
  - methods: maximum-likelihood, Bayesian inference

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Basis of inference tools: sequence alignment

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

sequence 1: `TCACACCT`
sequence 2: `ACAGACTT`
sequence 3: `AAAGACTT`
sequence 4: `ACACACCC`

- reminder (lecture 2):
  - alignment: each site is a *homolog*
  - alignment is obtained from raw sequencing reads by putting reads together such that number of mutations, insertions and deletions is minimized
- phylogenetic inference methods infer the phylogenetic tree based on such an alignment

At this point we assume the alignment is given.

Phenetic approaches: Distance-based methods

- **Basic idea:**
  1. define how to measure distance between sequences (e.g. Hamming, JC69, HKY distance from lecture 3)
  2. calculate distances between all pairs of sequences
  3. find a tree where the distances, i.e. the branch lengths, between the pairs of tips "most closely" follow the sequence distance matrix

- **Two strategies:**
  1. **algorithmic approach:** sequences separated by the smallest distance are clustered iteratively in a tree (e.g. via the UPGMA algorithm, Neighbor joining algorithm)
  2. **optimality approach:** minimise the difference of the pairwise sequence distances to the pairwise distances between the tips in the tree (calculated by summing the length of branches between two tips)

☞ Only distances between pairs of sequences are used but not any higher order correlations between sequences.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

The sequence distance matrix

sequence 1 (s_1): TCACACCT
sequence 2 (s_2): ACAGACTT
sequence 3 (s_3): AAAGACTT
sequence 4 (s_4): ACACACCC

Hamming distance

H	s_1	s_2	s_3	s_4
s_1	-	3	4	2
s_2		-	1	3
s_3			-	4
s_4				-

JC69 distance

$\hat{d} = -\frac{3}{4}\log(1 - \frac{4}{3}\hat{p})$

JC	s_1	s_2	s_3	s_4
s_1	-	0.52	0.82	0.30
s_2		-	0.14	0.52
s_3			-	0.82
s_4				-

R code:

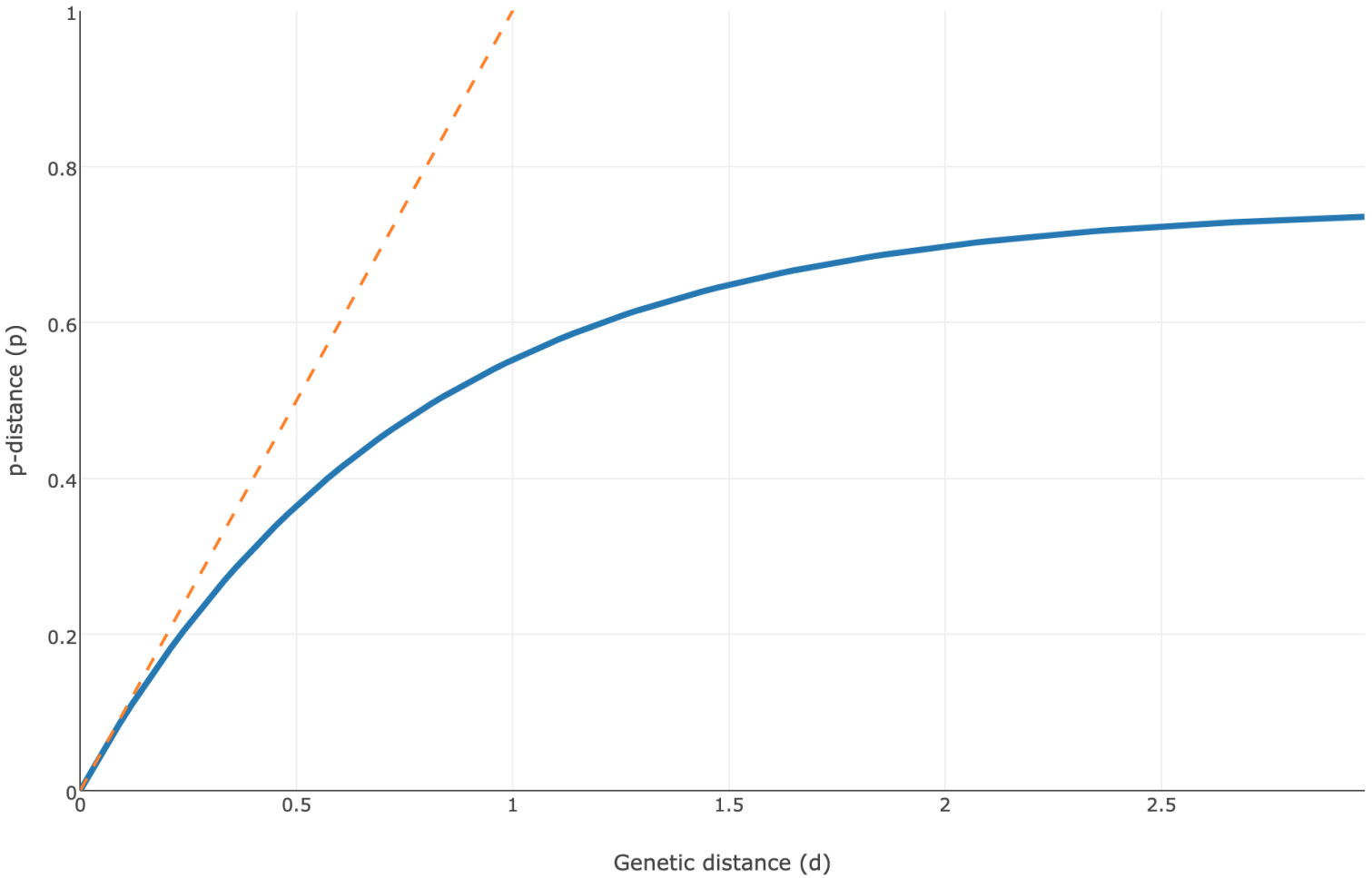
```
 p<-c(3,4,2,1,3,4)/8
djukescantor <- function(p){-3/4*log(1-4/3*p)}
djukescantor(p)
```

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Relationship between p-distance and genetic distance

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Using genetic distance for tree building is preferred!



Phenetic inference I: UPGMA, algorithmic

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

UPGMA algorithm

Unweighted Pair-Group Method using Arithmetic Averages (UPGMA) algorithm [Sokal and Michener, 1958]

Input: distance matrix

Output: ultrametric phylogenetic tree

- all sequences must come from the same time point
- the algorithm assumes evolution according to a **strict molecular clock:**

  the rate of DNA/RNA/protein sequence evolution is constant over time [Zuckerkandl and Pauling, 1965]
- the output is an **ultrametric tree:**

  rooted tree, the total branch lengths from all tips to the root are equal

Neighbour-joining algorithm [Saitou and Nei, 1987] relaxes these assumptions:
Branch lengths correspond to number of mutations; output tree will be unrooted.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

UPGMA algorithm

Input: Distance matrix.

Computational steps:

Initialize the size of each node s_i as $n_i = 1$.

While the distance matrix is not empty, iterate:

1. Choose nodes s_i and s_j such that $d(s_i, s_j)$ is the smallest entry in the distance matrix (in case of several mimima choose one uniformly at random).
2. Coalesce s_i and s_j to node $s_{i,j}$, with size $n_{i,j} = n_i + n_j$. The branch length between $s_i, s_{i,j}$ and between $s_j, s_{i,j}$ is chosen such that all tips descending from $s_{i,j}$ have the same distance $d(s_i, s_j)/2$ to $s_{i,j}$.
3. If the distance matrix includes more than 2 nodes, include $s_{i,j}$ into the distance matrix, with $d(s_m, s_{i,j}) = \frac{n_i\, d(s_i,s_m)+n_j\, d(s_j,s_m)}{n_i+n_j}$ where s_m is a node in the distance matrix.
4. Delete nodes s_i and s_j from the distance matrix.

Output: Ultrametric phylogenetic tree (obtained from the last performance of step 2.).

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

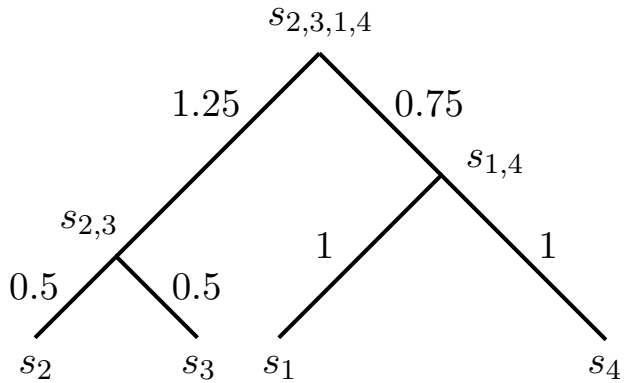
UPGMA tree from Hamming distance matrix

	s_1	s_2	s_3	s_4
s_1	-	3	4	2
s_2		-	1	3
s_3			-	4
s_4				-

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

UPGMA tree from Hamming distance matrix

we obtain the tree:



original distance matrix:

	s_1	s_2	s_3	s_4
s_1	-	3	4	2
s_2		-	1	3
s_3			-	4
s_4				-

tree distance matrix:

	s_1	s_2	s_3	s_4
s_1	-	3.5	3.5	2
s_2		-	1	3.5
s_3			-	3.5
s_4				-

- if there is a tree with its tree distance matrix being equal to the sequence distance matrix, this tree will be the UPGMA tree of the sequence distance matrix (proof by induction, see script)

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Phenetic inference II: Least squares, optimality

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Least squares methods

Least squares methods use an optimality criterion instead of an algorithmic approach:

Squared difference between the sequence distance matrix and the tree distance matrix is

$$S := \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}(D_{i,j} - d_{i,j})^2$$

where D is the between sequence distance matrix, d the tree distance matrix for the proposed tree, and w weights ($w_{i,j}$ may be 1 or e.g. inverse proportional to $D_{i,j}$) [Felsenstein, 2004].

Algorithm

Input: distance matrix

Repeat until all tree topologies were proposed:

1. propose an unrooted tree topology (without branch lengths)
2. minimise S by optimising the branch lengths

Output: tree with the smallest S

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Quality assessment of phylogeny reconstruction methods.

I Runtime.

Runtime of UPGMA

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

To get an estimate of the number of computation steps in UPGMA on $\mathbf{n}$ tips, we need to know:

- How many times do we prune nodes?
  $\mathbf{O(n)}$
- How many calculations do we perform per pruning?
  $\mathbf{O(n^2)}$

- UPGMA has the runtime $\mathbf{O(n^3)}$ for $\mathbf{n}$ sequences

Runtime of least squares

For each possible tree we need to optimise
$S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}(D_{i,j} - d_{i,j})^2$.
Thus, we need to visit each tree in the space of trees.

☞ How many trees on n tips exist?

- number of unrooted trees on n tips
- number of rooted trees on n tips

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Counting trees on $n = 2, n = 3, n = 4$

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

We need a strategy to derive the number of trees:

1. count number of branches in an unrooted tree with n tips, b_n
2. use b_n to calculate the number of unrooted trees with n tips, τ_n
3. use τ_n to calculate the number of rooted trees with n tips, τ_n^r

Step 1: Counting branches

- How many branches b_n does an unrooted tree on n tips have?
  - Consider a tree on $n = 2$ tips:
    $b_2 = 1$
  - Consider an unrooted tree on n tips. How many branches are added when adding an additional tip?
    2 branches, thus:
    $b_{n+1} = b_n + 2$

In general: $b_n = ?$
$b_n = b_2 + 2 \times (n - 2) = 2n - 3.$

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Toolbox: Proof by induction

Theorem: An unrooted tree with n tips has $b_n = 2n - 3$ branches.

Proof: Induction over n.

Hypothesis to prove: $b_n = 2n - 3$.
Base step: Check that the hypothesis holds for $n = 2$.
Yes: $b_2 = 1 = 2 \times 2 - 3$.
Induction hypothesis: We suppose the formula holds for all $k < n$.
In particular: $b_{n-1} = 2(n-1) - 3$.
Inductive step: Given the induction hypothesis, show that the formula holds for n.
This is the case because:
$b_n = b_{n-1} + 2 = 2(n-1) - 3 + 2 = 2n - 3$, where the second equality follows from the induction hypothesis.

q.e.d. (quod erat demonstrandum)

□

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Step 2: Counting unrooted trees

We now know that a tree with n tips has $b_n = 2n - 3$ **branches**. With this we can derive how many unrooted **trees** with n tips exist:

- How many unrooted trees on 2 tips exist?
  $\tau_2 = 1$
- Given a tree on n tips, in how many ways can we add the $(n+1)$ tip?
  We can add the tip to any branch in the n-tip tree, i.e.
  $\tau_{n+1} = \tau_n \times b_n$.
- For $n \geqslant 3$ holds:

$$\begin{aligned} \tau_n &= \tau_2 \times b_2 \times b_3 \ldots \times b_{n-1} \\ &= 1 \times 1 \times 3 \times 5 \times 7 \times \ldots (2n-5) \\ &= (2n-5)!! \end{aligned}$$

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Step 3: Counting rooted trees

How many **rooted** trees on n tips exist?

- We obtain a rooted tree on n tips by choosing an unrooted tree on n tips and picking one branch which is divided by a root node.
- $\tau_n^r = \tau_n \times b_n = (2n-5)!!(2n-3) = (2n-3)!!$

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Runtime: Least squares method vs. UPGMA

Compare the number of operations (only the order, not exact) that these algorithms need to perform:

# of tips n	least squares # of unrooted trees τ_n	UPGMA n^3
4	3	64
5	15	125
6	105	216
7	945	343
8	10395	512
9	135135	729
10	2027025	1000
⋮	⋮	
20	221 643 095 476 699 771 875	8000
⋮	⋮	
50	10^{74}	125 000

The least square decision problem is an NP-complete problem [Day, 1987].

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

NP completeness of a problem

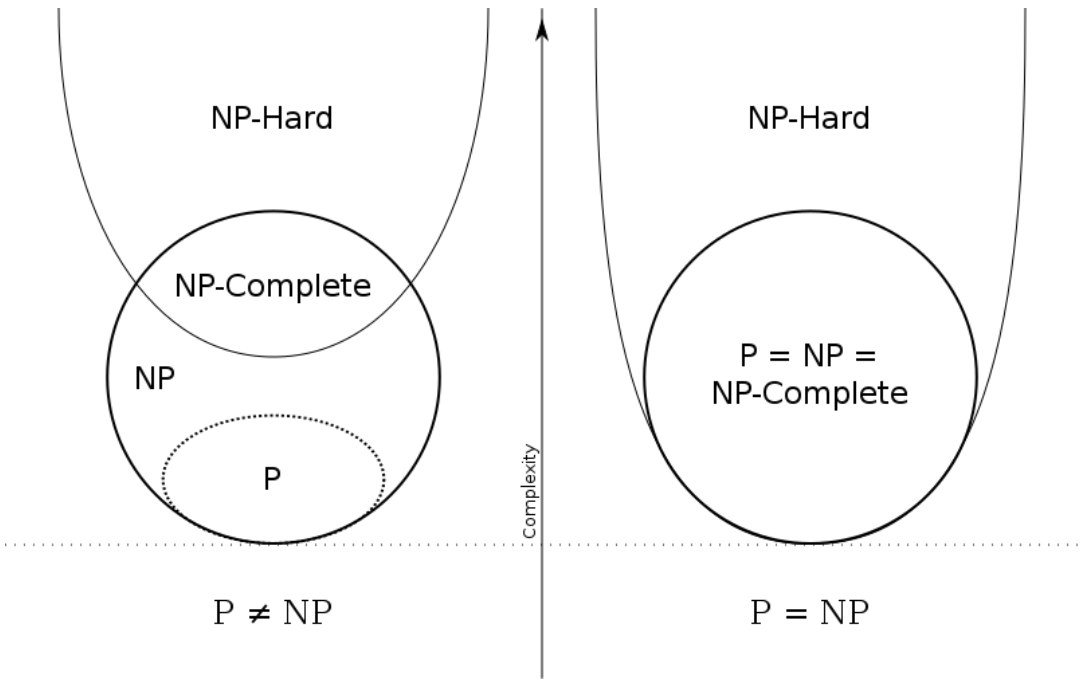
- P = polynomial time.
  A problem is in P if the runtime until a solution with input size n is found is n^k where k is some input-independent number (e.g. n^3 for UPGMA)

- NP = non-deterministic polynomial time
  - Consider a decision problem X (e.g. Is there a tree with a least squares difference of less than x?)
  - A decision problem is **in NP** if it can be verified in polynomial time (e.g. it is easy to determine the least squares difference for a given tree).
  - A decision problem in NP is **NP complete** if the *travelling salesman problem* can be solved using an algorithm to solve the decision problem X together with potentially a polynomial time transformation algorithm.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Your chance to win 1'000'000 US Dollars

- No polynomial time algorithm is known for NP-complete problems. If one is found for any NP-complete problem, we can solve all NP complete problems in polynomial time!
- Prove or disprove $\mathbf{P} = \mathbf{NP}$ is one of 7 Millennium Prize Problems - you can win 1'000'000 US Dollars!



05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Quality assessment of phylogeny reconstruction methods.
II Statistical consistency.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Statistical consistency

A phylogenetic reconstruction method is **statistically consistent** if the true tree is returned for an infinite amount of data (i.e. infinitely long sequences).

Formally:

A phylogenetic reconstruction method is **statistically consistent** if for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} P(||\hat{T} - T|| < \epsilon) = 1$$

where n is the sequence length, T the true tree, and $\hat{T}$ the inferred tree.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Statistical consistency

- Let sequences evolve on a fixed tree with some model $\mathbf{M}$, and distances in the tree in units of substitution (i.e. the expected number of substitutions is 1 within distance 1). Then the maximum likelihood distance matrix based on the simulated sequences approaches the distances in the tree with increasing sequence length, since maximum likelihood estimators are statistically consistent.
- **UPGMA is consistent**: The distance matrix tends towards the tree distances. If the sequence distance matrix equals the tree distance matrix, then the true tree is reconstructed. A rigorous proof of consistency is more involved, but you can numerically check consistency with the developed algorithms in HW2 and HW3.
- **Least squares method is consistent**: Squared difference between the calculated distance matrix and the tree distance tend towards 0, thus the true tree is *a* least squares tree! Further each tree has a unique distance matrix, thus the true tree is *the* least squares tree.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Summary.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Summary phenetic inference

Quality assessment:

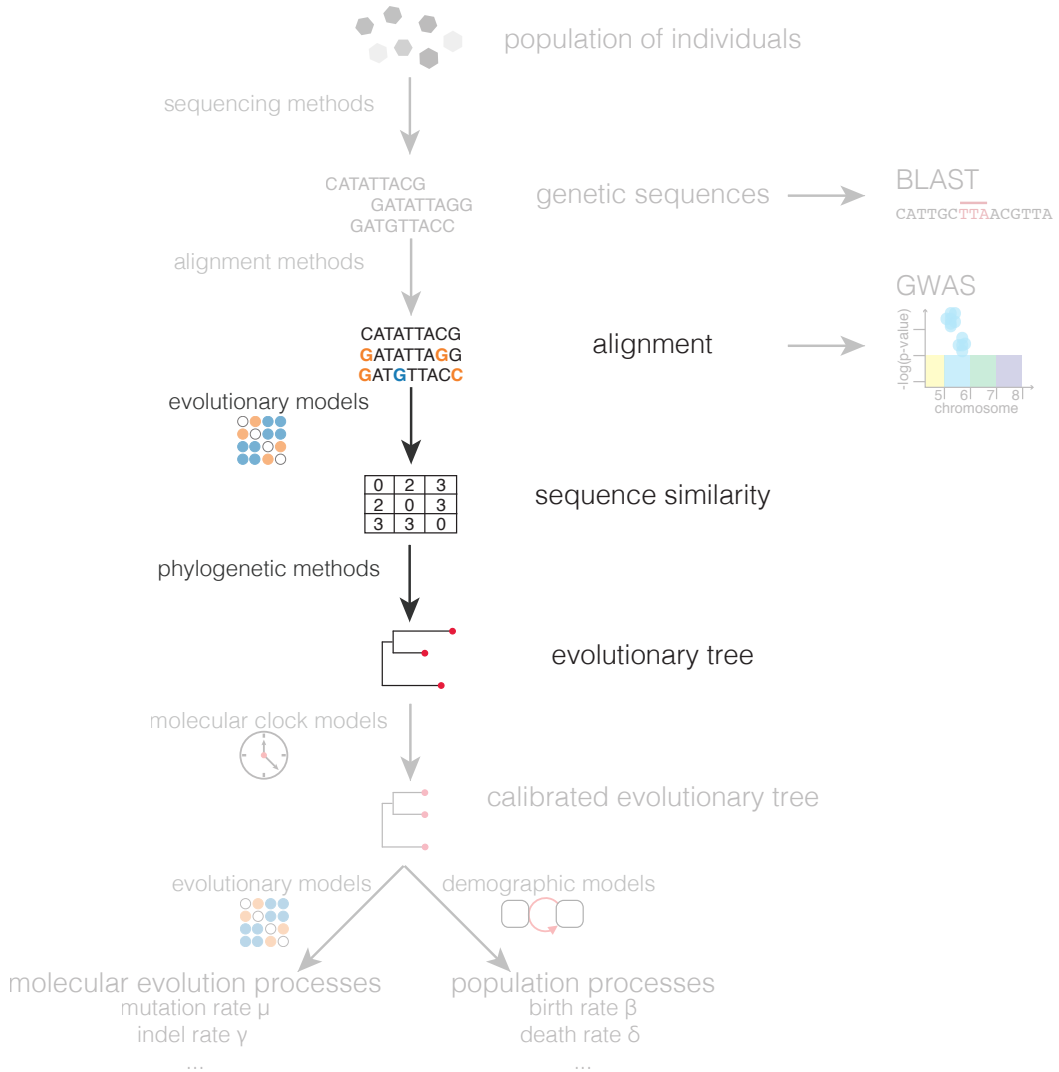
- Runtime
  - UPGMA & neighbour joining are algorithmic "clustering methods": polynomial runtime
  - Least squares methods are optimality-based methods: NP complete
- Consistency
  - UPGMA, neighbour joining and least squares methods are statistically consistent.

Problem of phenetic approaches:

- disregard information beyond pairwise distances;
- large distances come with large variances, which are typically ignored.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Overview



05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

Definition of phylogenies & phenetic phylogenetic reconstruction: Questions

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References

- 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 $w_{i,j} \neq 1$?

References I

- Day, W. H. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4):461–467.
- Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer associates Sunderland.
- Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D., McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T., and Cranston, K. A. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*.
- 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.
- Sato, A., O'hUigin, C., Figueroa, F., Grant, P. R., Grant, B. R., Tichy, H., and Klein, J. (1999). Phylogeny of Darwin's finches as revealed by mtDNA sequences. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 96(9):5101–5106.
- Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.
- Zuckerkandl, E. and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. In *Evolving Genes and Proteins*, pages 97–166. Elsevier.

05: Phylogenetics
Introduction to phylogenetic trees
Formal definitions
Tree visualisation
Examples of phylogenies
Phylogenetic inference
Phenetic approach
UPGMA
Least squares methods
Quality of reconstruction methods
Runtime
Statistical consistency
Summary
References