# Algorithms for Computational Biology

**Zsuzsanna Lipták**

Masters in Molecular and Medical Biotechnology
a.a. 2015/16, fall term

Phylogenetic Trees II (Character Data)

# Character data

Now the input data consists of states of characters for the given objects, e.g.

- morphological data, e.g. number of toes, reproductive method, type of hip bone, . . . or
- molecular data, e.g. what is the nucletoide in a certain position.

# Character data

## Example

| | $C_1$ : # wheels | $C_2$ : existence of engine |
|---|---|---|
| bicycle | 2 | 0 |
| motorcycle | 2 | 1 |
| car | 4 | 1 |
| tricycle | 3 | 0 |

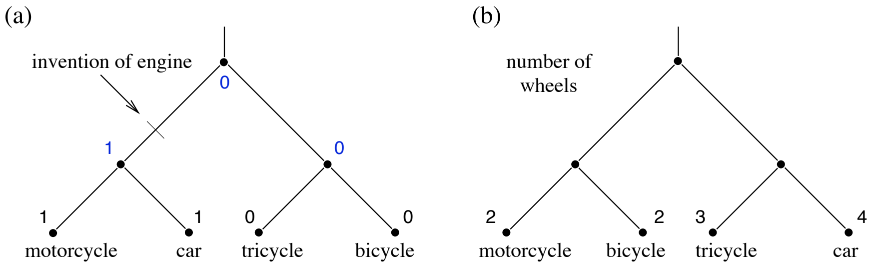
# Character data

## Example

| | $C_1$ : # wheels | $C_2$ : existence of engine |
|---|---|---|
| bicycle | 2 | 0 |
| motorcycle | 2 | 1 |
| car | 4 | 1 |
| tricycle | 3 | 0 |

- objects (species): Bicycle, motorcycle, tricycle, car
- characters: number of wheels; existence of an engine
- character states: $2, 3, 4$ for $C_1$;
  $0, 1$ for $C_2$ ($1 =$ YES, $0 =$ NO)
- This matrix $M$ is called a character-state-matrix, of dimension $(n \times m)$, where for $1 \leq i \leq n, 1 \leq j \leq m$: $M_{ij} =$ state of character $j$ for object $i$. (Here: $n = 4$, $m = 2$.)

# Character data




Two different phylogenetic trees for the same set of objects.

# Character data

We want to avoid

- parallel evolution (= convergence)
- reversals

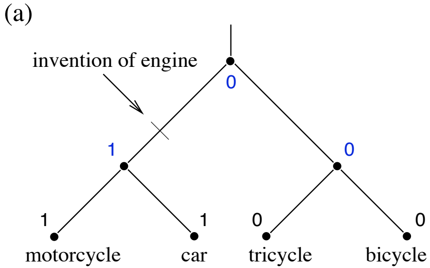
These two together are also called homoplasies.

Mathematical formulation: compatibility.

# Compatibility

## Definition

A character is compatible with a tree if all inner nodes of the tree can be labeled such that each character state induces one connected subtree.

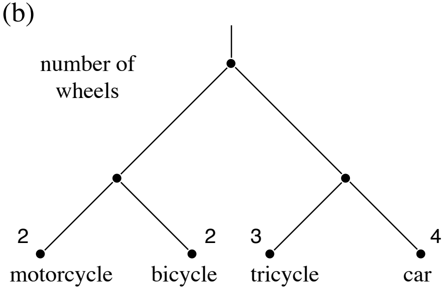



This tree is compatible with $C_2$, one possibility of labeling the inner nodes is shown.

# Compatibility

## Definition

A character is compatible with a tree if all inner nodes of the tree can be labeled such that each character state induces one connected subtree.

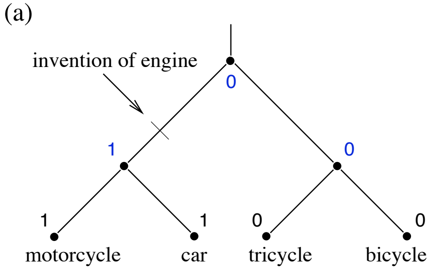



This tree is compatible with $C_1$. (We have to give a labeling of the inner nodes to prove this.) It is not compatible with $C_2$ (why?)

# Compatibility

## Definition

A character is compatible with a tree if all inner nodes of the tree can be labeled such that each character state induces one connected subtree.




This tree is also compatible with $C_1$: We have to give a labeling of the inner nodes (w.r.t. $C_1$) to prove this.

# Compatibility

Exercise:

The objects $\alpha, \beta, \gamma, \delta$ share three characters $C_1$, $C_2$, $C_3$. The following matrix holds their states:

| | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| $\alpha$ | $a$ | $c$ | $f$ |
| $\beta$ | $a$ | $d$ | $g$ |
| $\gamma$ | $b$ | $d$ | $h$ |
| $\delta$ | $b$ | $e$ | $f$ |

($C_1$ can have states $a, b$; $C_2$ states $c, d, e$; $C_3$ states $f, g, h$.)

Look at all possible tree topologies. Is there, among all these trees, a tree $T$ such that all characters are compatible with $T$? (Hint: It is enough to consider unrooted trees. Why?)

# Compatibility

Note that the question whether a character is compatible with a tree is independent of the other characters. Moreover, often all characters have the same states (typically $\{A, C, G, T\}$). Thus the previous problem is equivalent to this one:

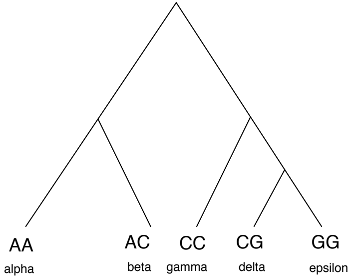
| | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| $\alpha$ | $A$ | $C$ | $A$ |
| $\beta$ | $A$ | $G$ | $C$ |
| $\gamma$ | $C$ | $G$ | $G$ |
| $\delta$ | $C$ | $T$ | $A$ |

# Perfect Phylogeny

## Definition

A tree $T$ is called a perfect phylogeny (PP) for $\mathcal{C}$ if all characters $C \in \mathcal{C}$ are compatible with $T$.

## Example



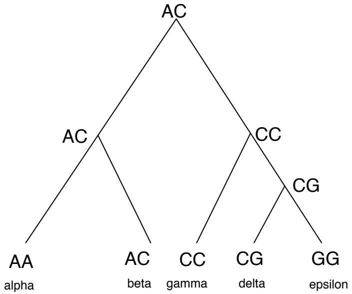

Why? We have to find a labeling of the inner nodes s.t. for both characters $C_1$ and $C_2$, each character induces a subtree.

# Perfect Phylogeny

## Definition

A tree $T$ is called a perfect phylogeny (PP) for $\mathcal{C}$ if all characters $C \in \mathcal{C}$ are compatible with $T$.

## Example

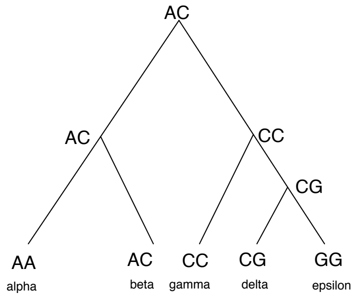


# Perfect Phylogeny

## Definition

A tree $T$ is called a perfect phylogeny (PP) for $\mathcal{C}$ if all characters $C \in \mathcal{C}$ are compatible with $T$.

## Example




Our first tree for the vehicles was also a PP, as well as the solution to the exercise.

# Perfect Phylogeny

- Ideally, we would like to find a PP for our input data.

# Perfect Phylogeny

- Ideally, we would like to find a PP for our input data.
- Deciding in general whether a PP exists is NP-hard.

# Perfect Phylogeny

- Ideally, we would like to find a PP for our input data.
- Deciding in general whether a PP exists is NP-hard.
- This is not really a problem, since most of the time, no PP exists anyway. Why: due to homoplasies; because our input data has errors; our evolutionary model probably has errors; and, and, and . . .

# Perfect Phylogeny

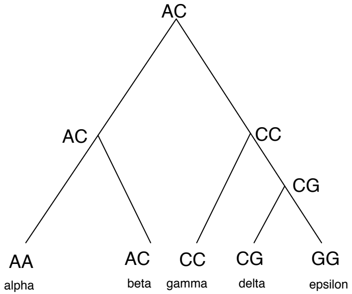
- Ideally, we would like to find a PP for our input data.
- Deciding in general whether a PP exists is NP-hard.
- This is not really a problem, since most of the time, no PP exists anyway. Why: due to homoplasies; because our input data has errors; our evolutionary model probably has errors; and, and, and ...
- Therefore we usually want to find a best possible tree.

# Parsimony

What is a best possible tree?

# Parsimony

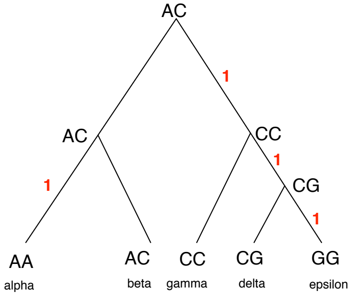
What is a best possible tree?




Why is this tree “perfect”?

# Parsimony

What is a best possible tree?




Why is this tree “perfect”?

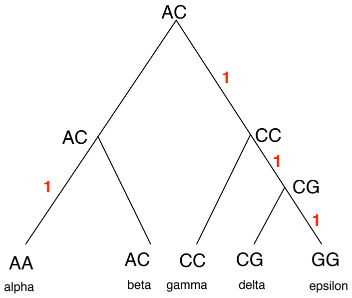
Because it has few state changes!

In red, we marked the edges where there are state changes (an evolutionary event happened), and how many (in this case, always 1).

# Parsimony

## Definition

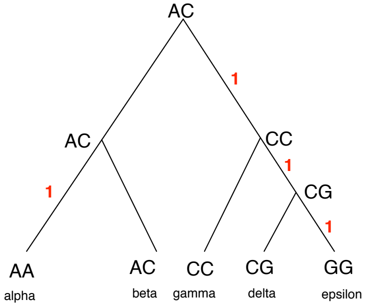
The parsimony cost of a phylogenetic tree with labeled inner nodes is the number of state changes along the edges (i.e. the sum of the edge costs, where the cost of an edge = number of characters whose state differs between child and parent).




The parsimony cost of this node-labeled tree is 4.

# Parsimony

## Same tree, different labelings

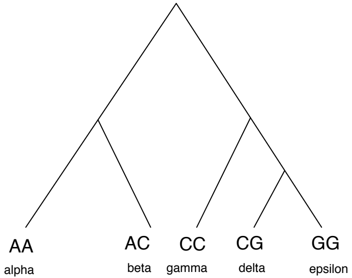



The parsimony cost of left node-labeled tree is 4, that of right node-labeled tree is 5.

# Parsimony

## Definition

The parsimony cost of a phylogenetic tree (without labels on the inner nodes) is the minimum of the parsimony cost over all possible labelings of the inner nodes.




The parsimony cost of this tree is 4, because the best labeling has cost 4.

# Small Parsimony

How can we find the best labeling of the inner nodes, given the tree? How can we find the parsimony cost of a given tree? This problem is called Small Parsimony, and it is polynomially solvable.

## Small Parsimony Problem

**Given:** a phylogenetic tree T with character-states at the nodes.
**Find:** a labeling of the inner nodes with states with minimum parsimony cost.

# Maximum Parsimony

**Definition**

The maximum parsimony problem is, given a character-state matrix, find a phylogenetic tree with lowest parsimony cost (= a "most parsimonious" tree).

# Maximum Parsimony

### Definition

The maximum parsimony problem is, given a character-state matrix, find a phylogenetic tree with lowest parsimony cost (= a "most parsimonious" tree).

- The underlying idea is (again) the Occam's razor principle: the simplest explanation is the best.

# Maximum Parsimony

### Definition

The maximum parsimony problem is, given a character-state matrix, find a phylogenetic tree with lowest parsimony cost (= a "most parsimonious" tree).

- The underlying idea is (again) the Occam's razor principle: the simplest explanation is the best.
- When a PP exists, then it is also the most parsimonious tree.

# Maximum Parsimony

### Definition

The maximum parsimony problem is, given a character-state matrix, find a phylogenetic tree with lowest parsimony cost (= a "most parsimonious" tree).

- The underlying idea is (again) the Occam's razor principle: the simplest explanation is the best.
- When a PP exists, then it is also the most parsimonious tree.
- In general, this problem is NP-hard.