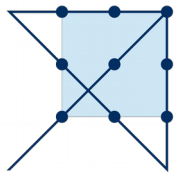


# Models and methods for disentangling the complexity of gene and species evolution

Benoit Morel

CMMS talks  
19.10.2023



**HITS**

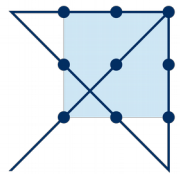
Heidelberg Institute for  
Theoretical Studies



Karlsruher Institut für Technologie

# About me

- Benoit Morel ([benoit.morel@h-its.org](mailto:benoit.morel@h-its.org))
- Studied computer science and mathematics
- Worked 4 years as a software developer
- PhD and postdoc:
  - Model and software development
  - In the field of phylogenetics



**HITS**

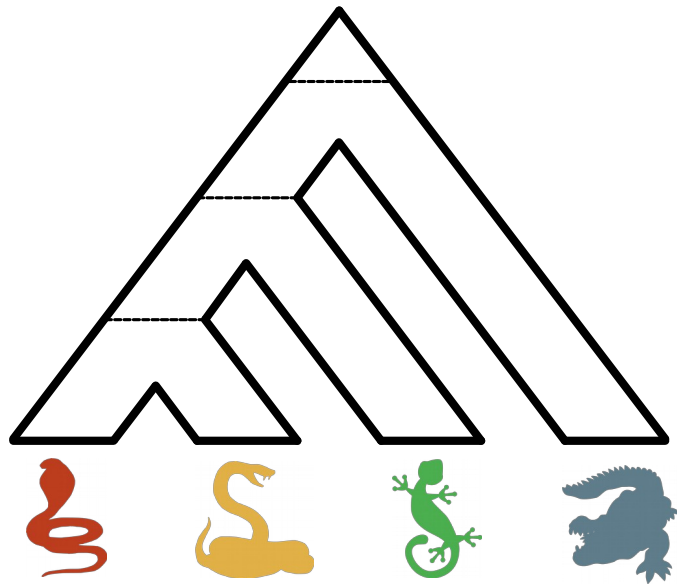
Heidelberg Institute for  
Theoretical Studies



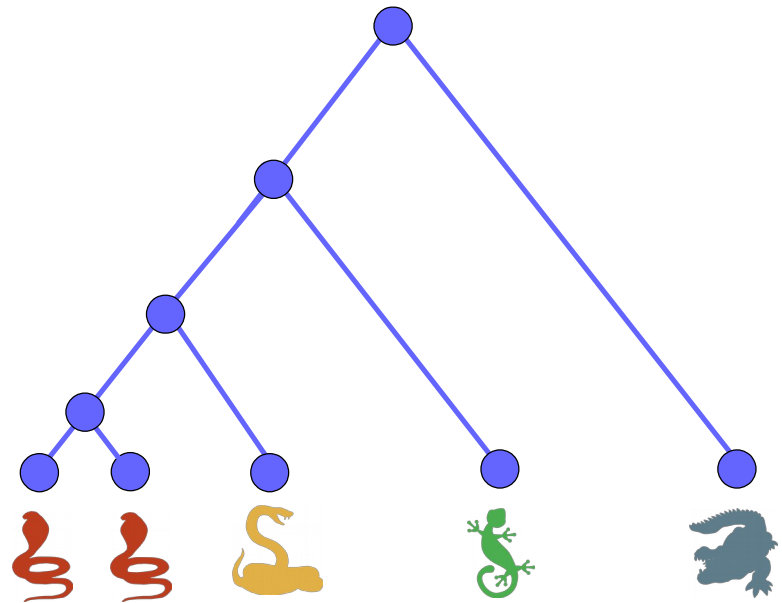
Karlsruher Institut für Technologie

# Phylogenetics

Study of evolutionary relationships among biological entities (genes or species)



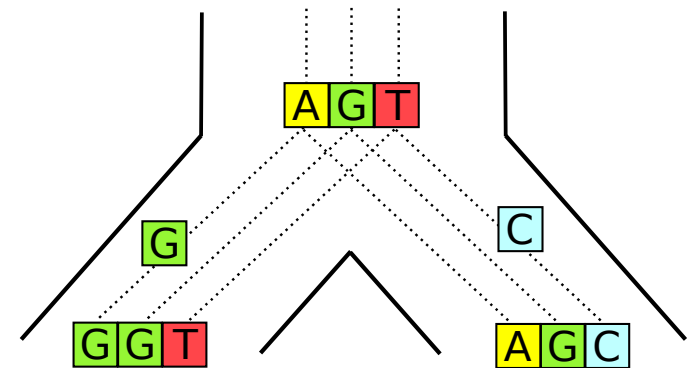
Species tree



Gene tree

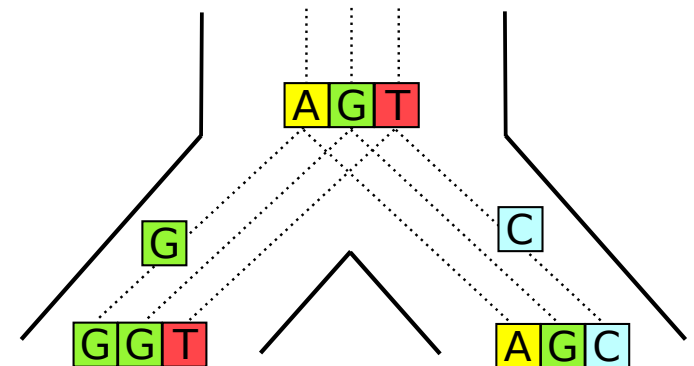
# To answer those questions...

- Describe evolution with a probabilistic model



# To answer those questions...

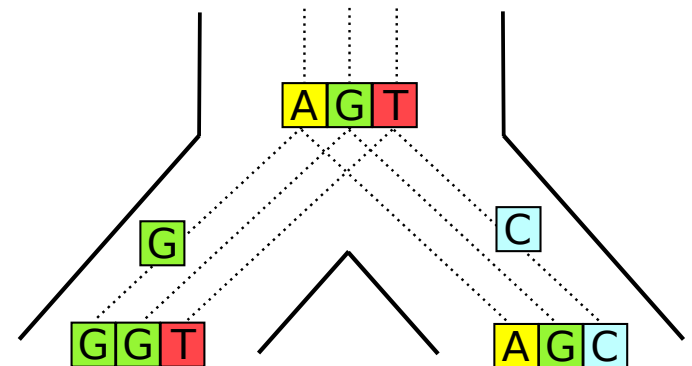
- Describe evolution with a probabilistic model
- Identify with parameters are known (data) and which ones have to be estimated



# To answer those questions...

- Describe evolution with a probabilistic model
- Identify with parameters are known (data) and which ones have to be estimated
- Use maximum likelihood to estimate the most likely unknown parameters by maximizing:

$P(\text{data} \mid \text{parameters})$



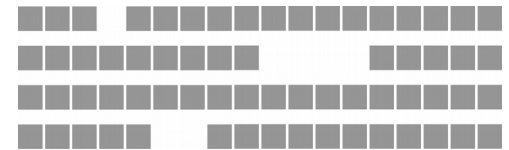
# Hierarchical model



Species tree

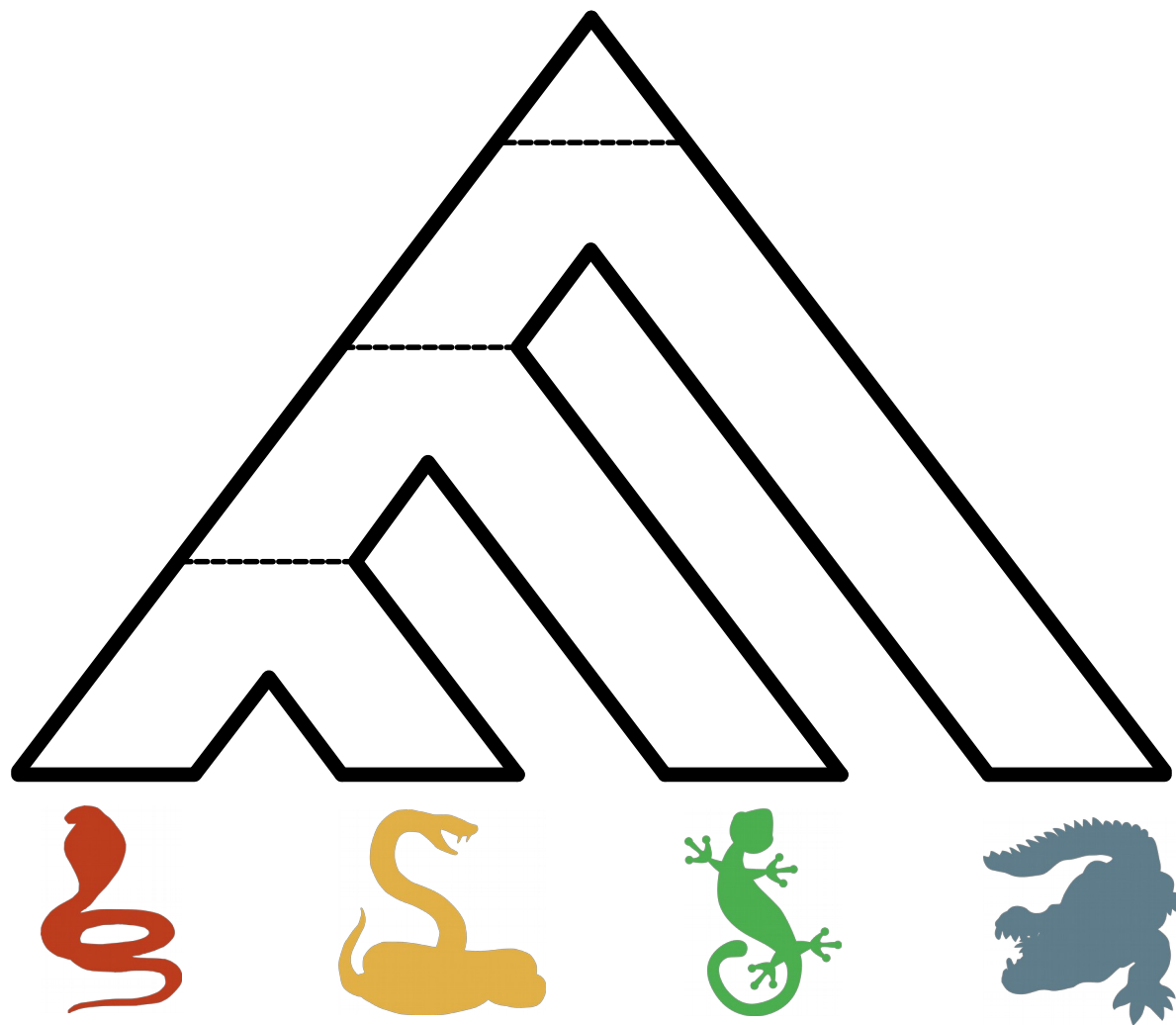


Gene tree



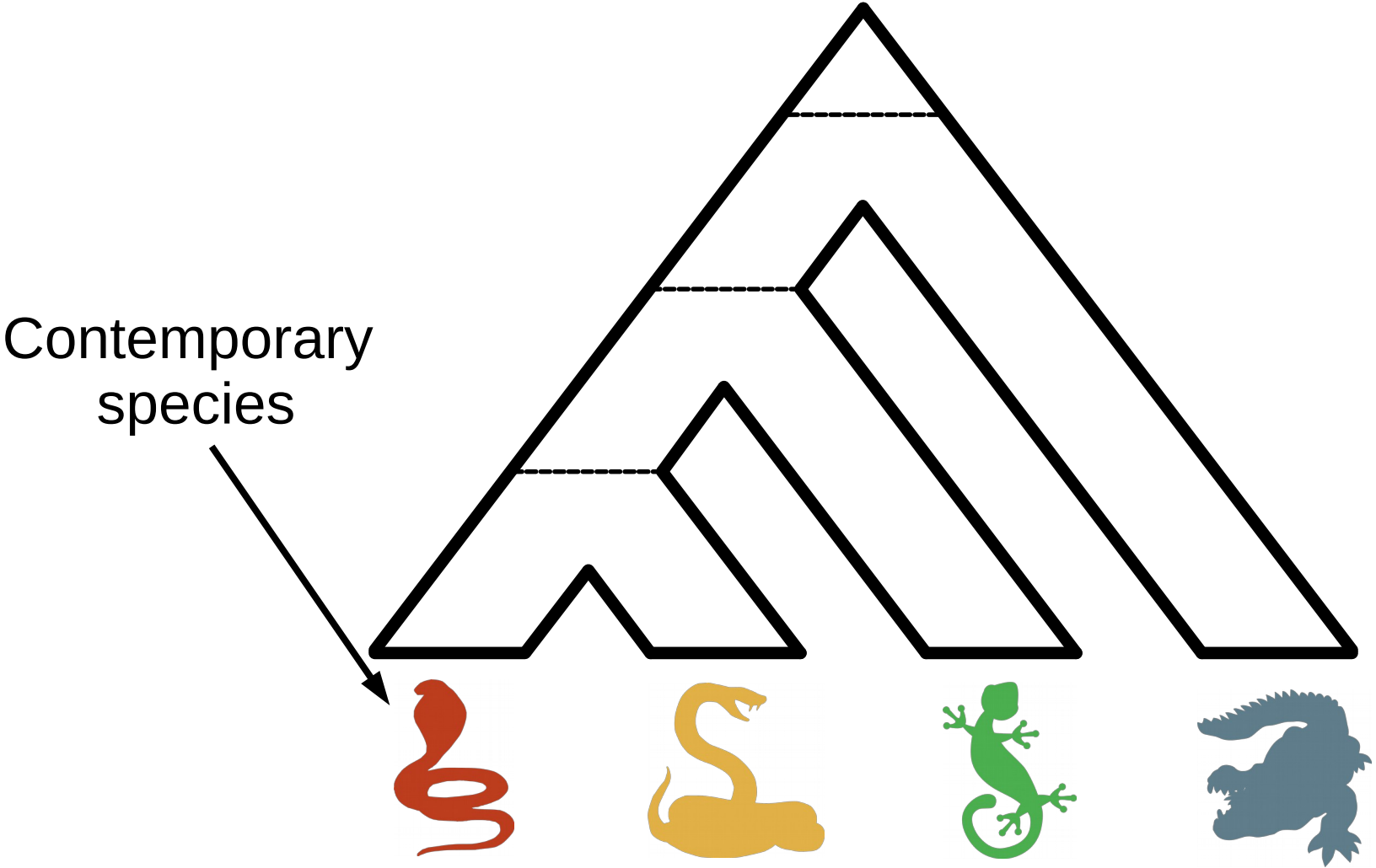
Gene sequences

# Species tree

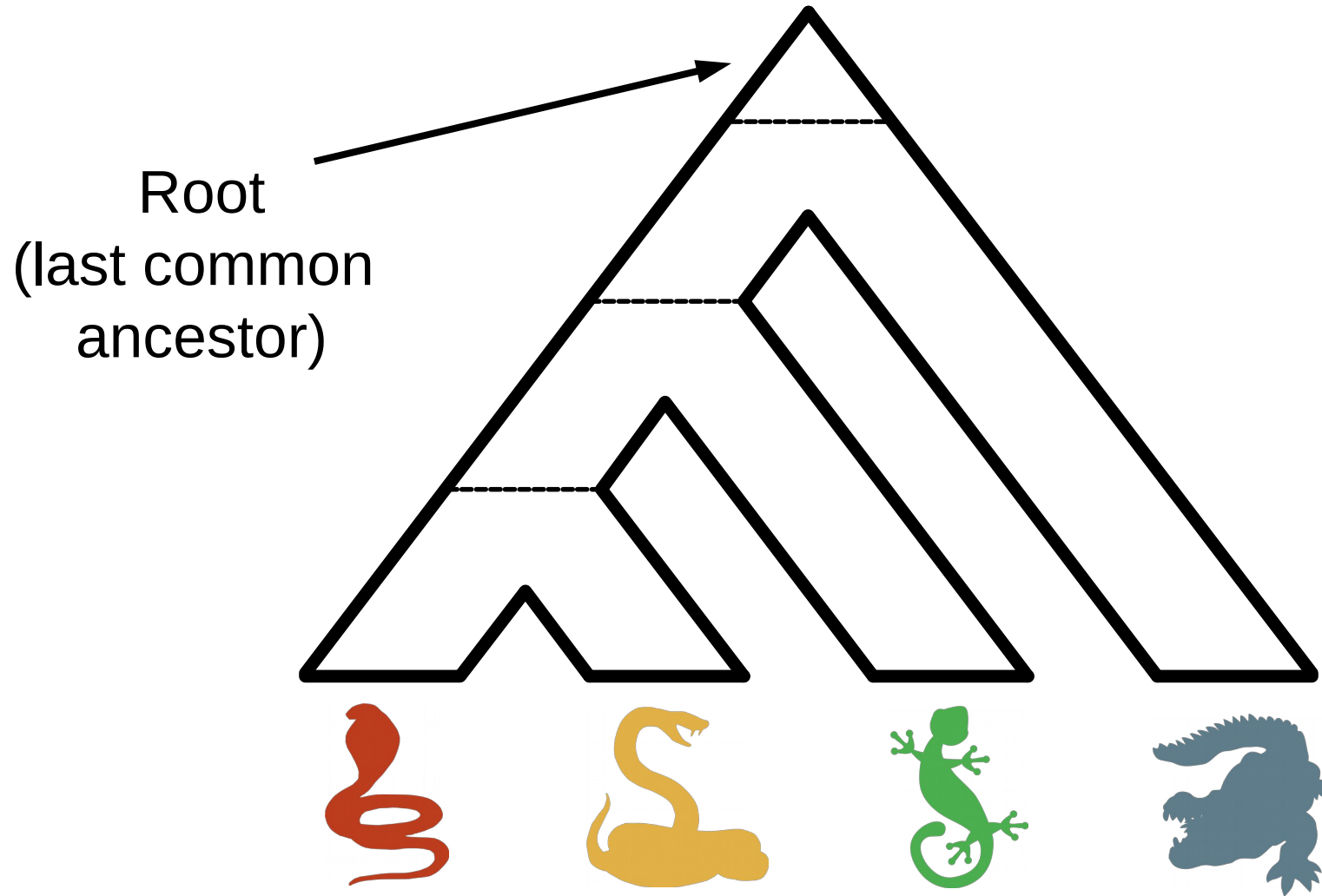




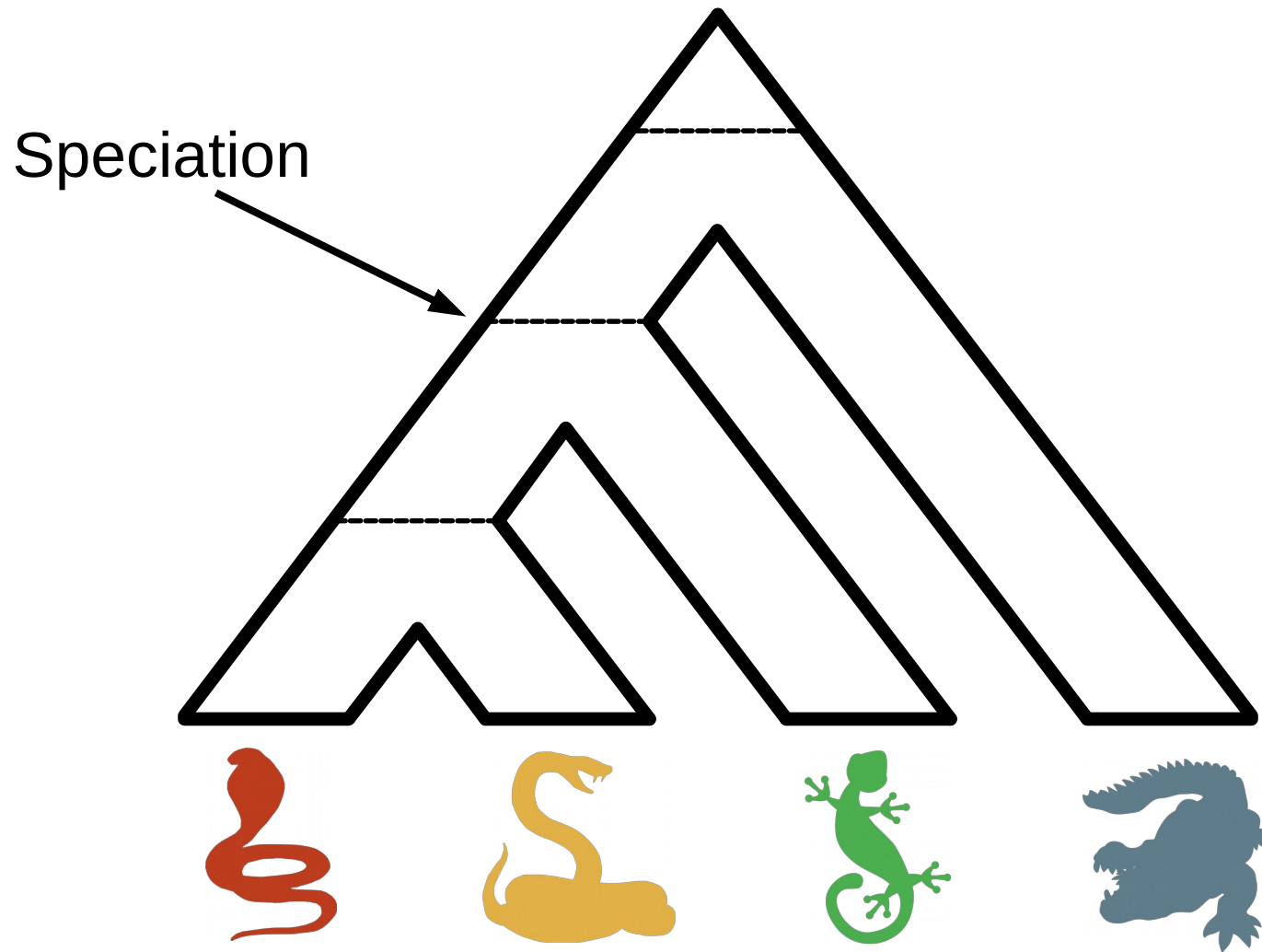
# Species tree



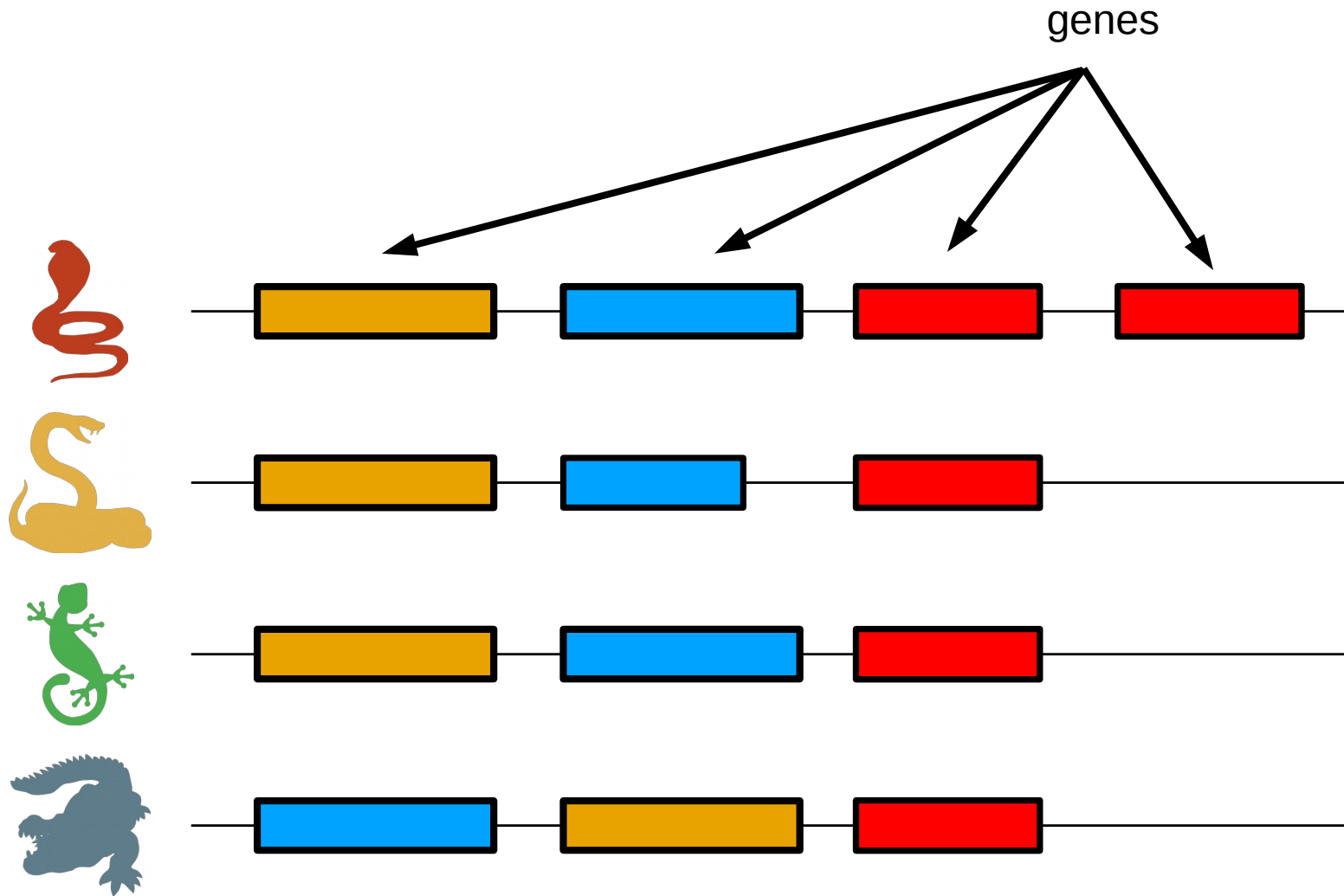
# Species tree



# Species tree

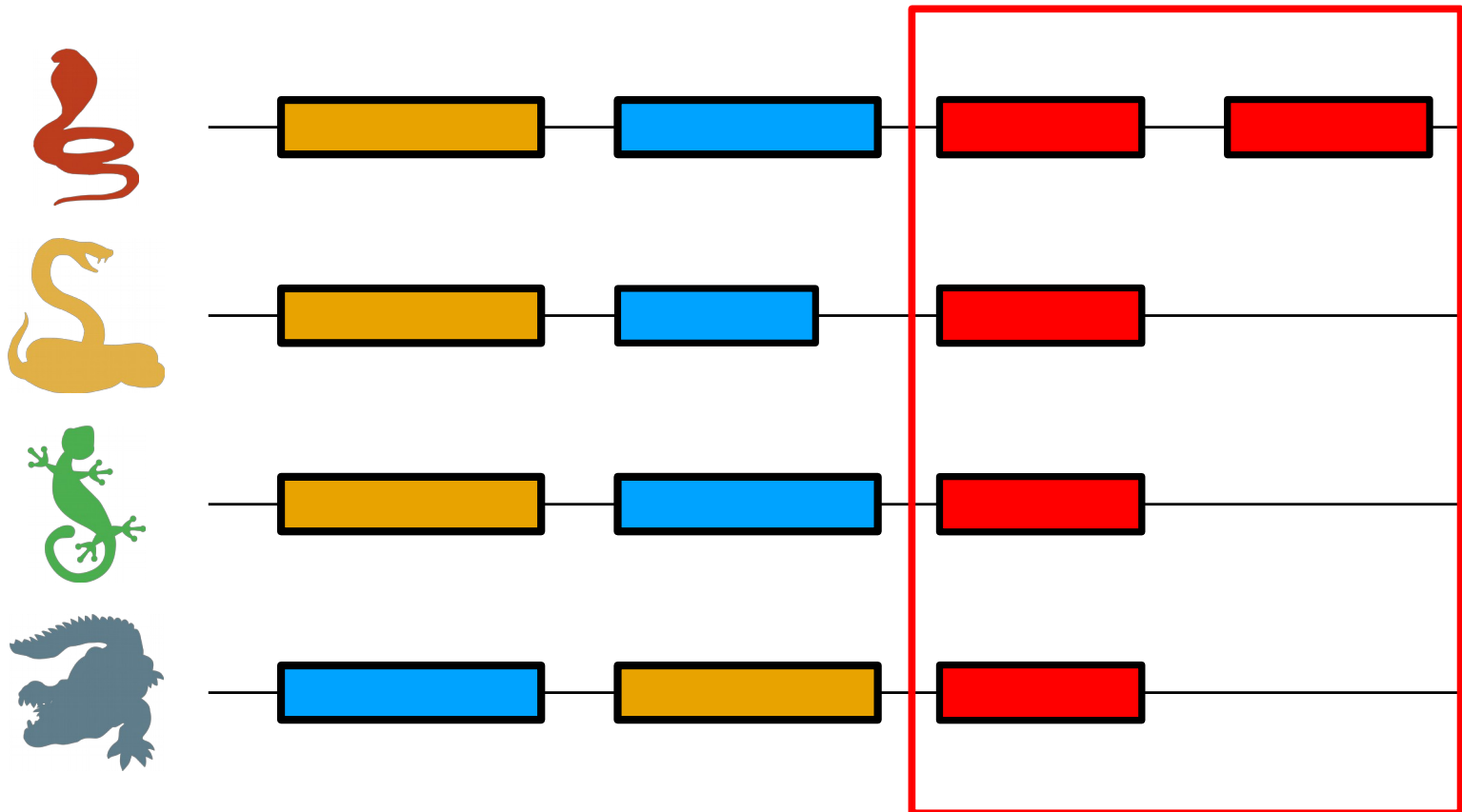


# Genomes

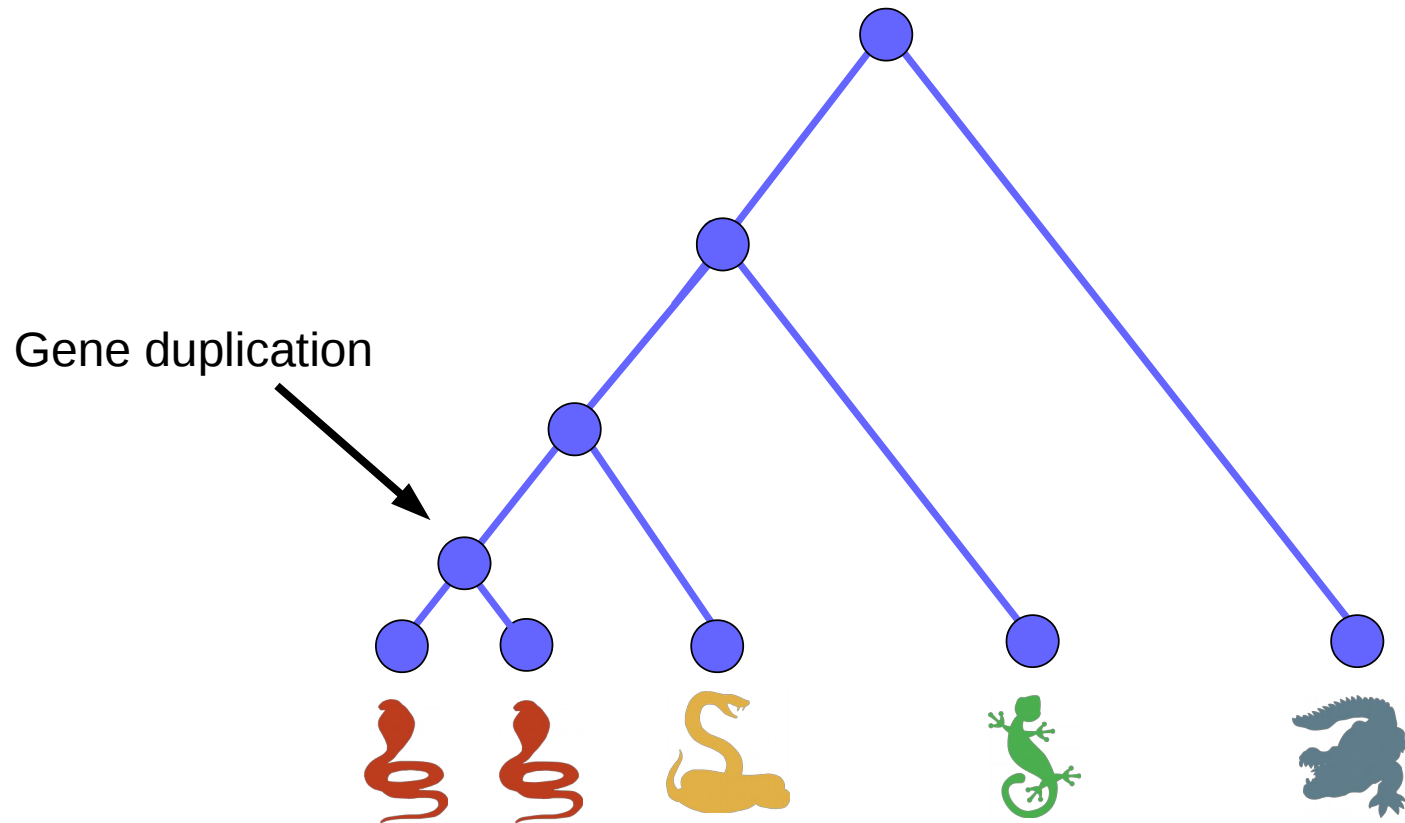


# Genomes

- **Gene family:** set of homologous genes (share a common ancestor)



# Gene (family) tree



# Gene sequences



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



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



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



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



A	A	-	-	C	T	T	C	A	T	T	-	A
---	---	---	---	---	---	---	---	---	---	---	---	---

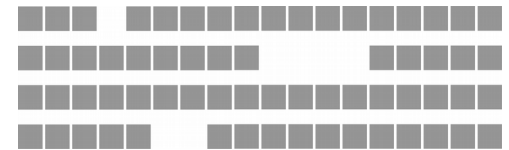
# Hierarchical model



Species tree



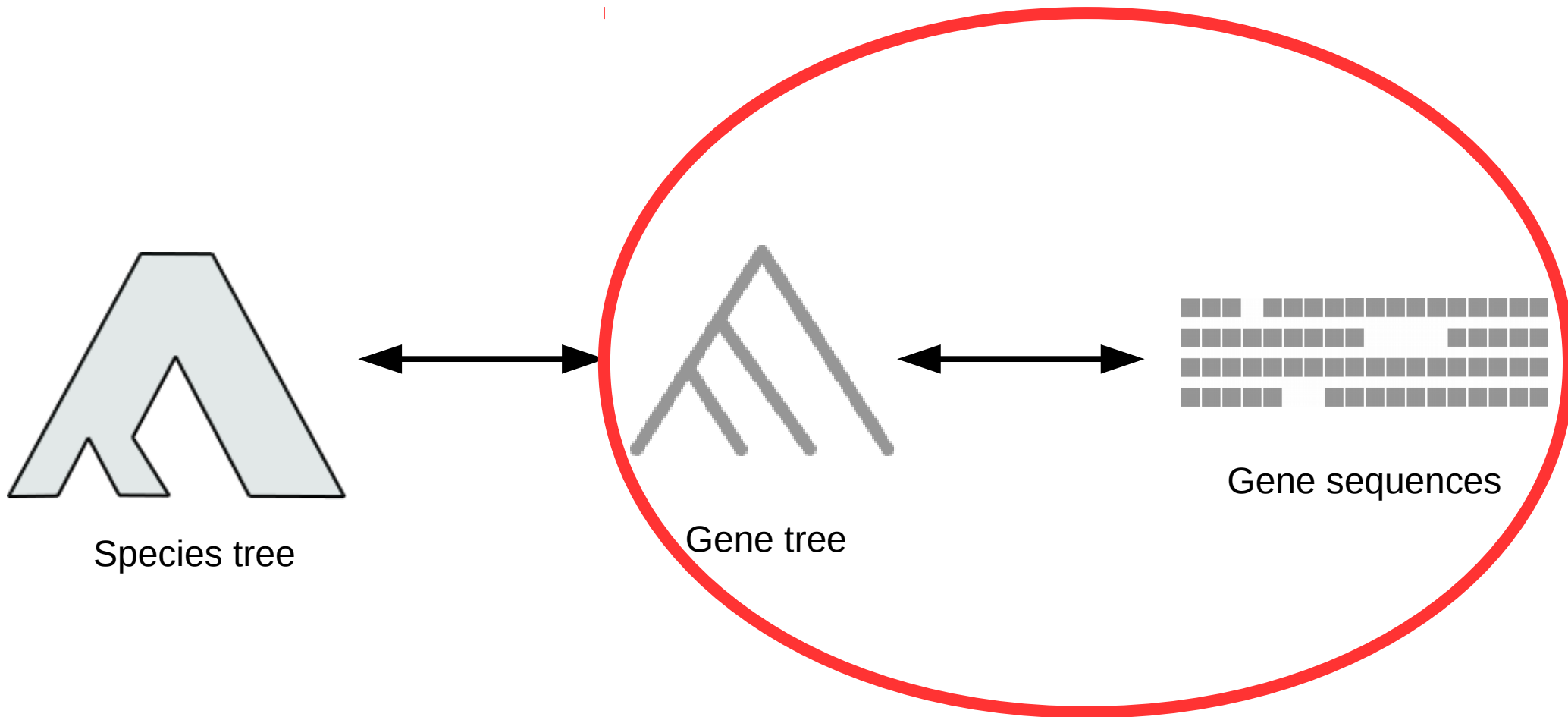
Gene tree



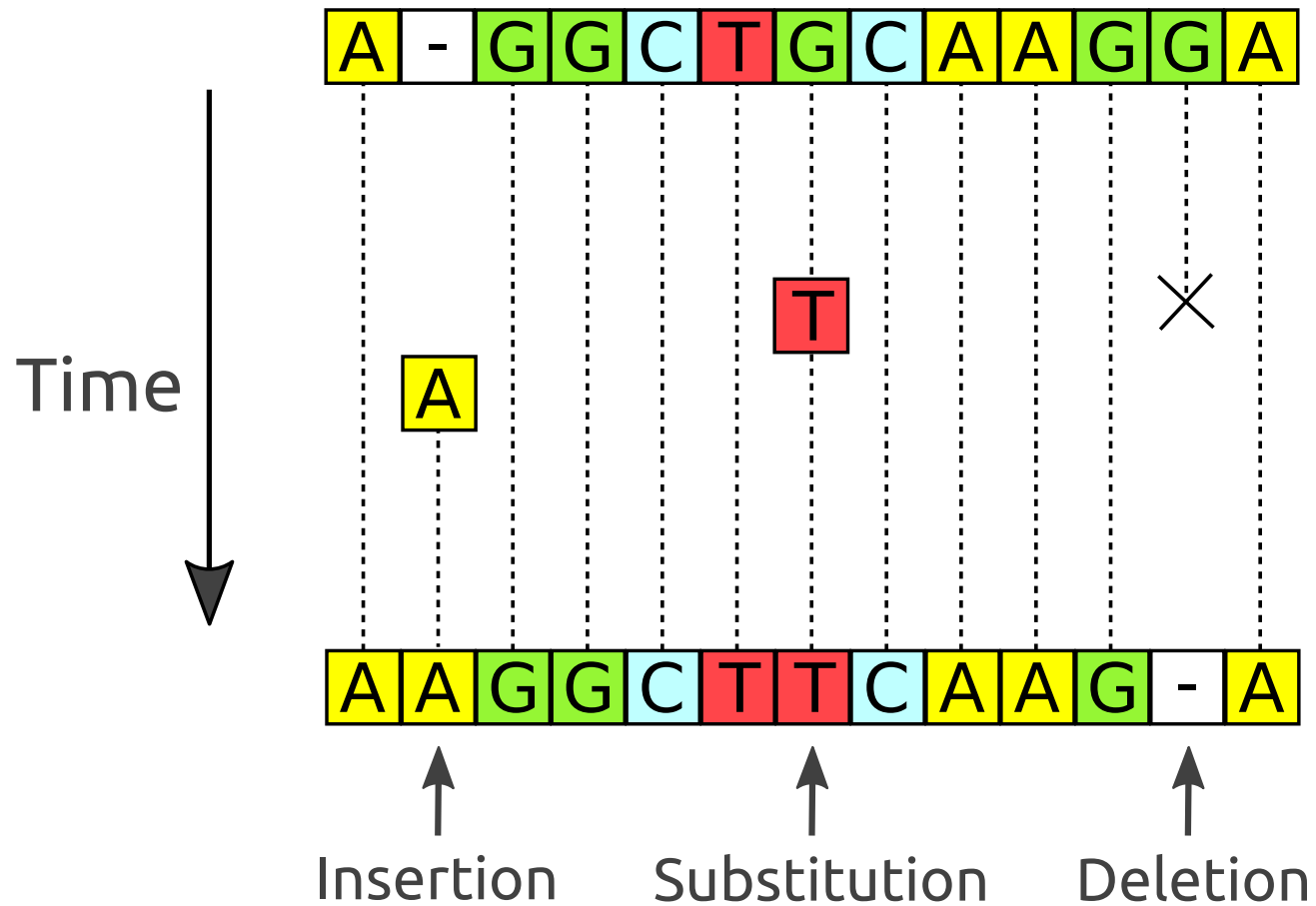
Gene sequences



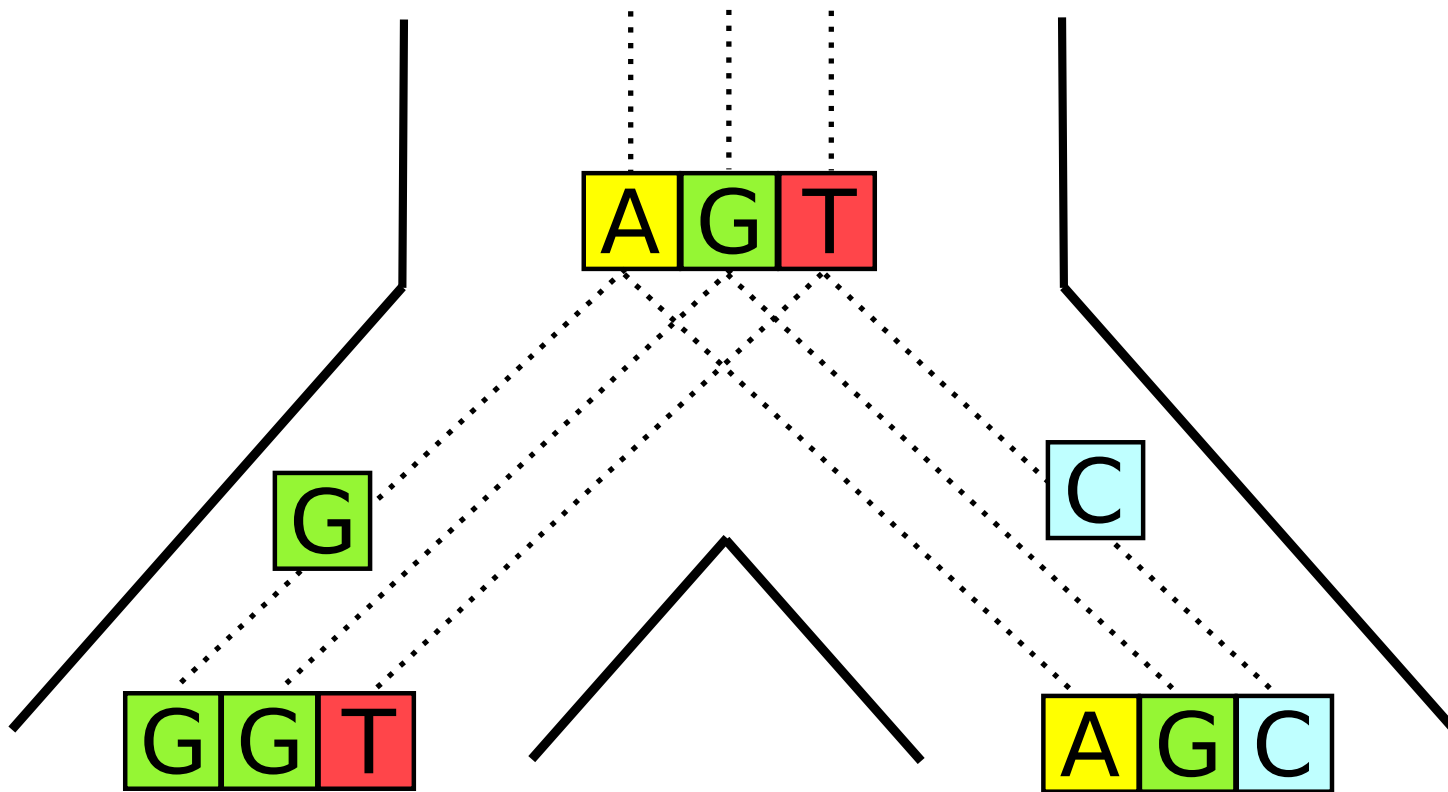
# Hierarchical model



# Sequence evolution



# Sequence evolution in a gene tree

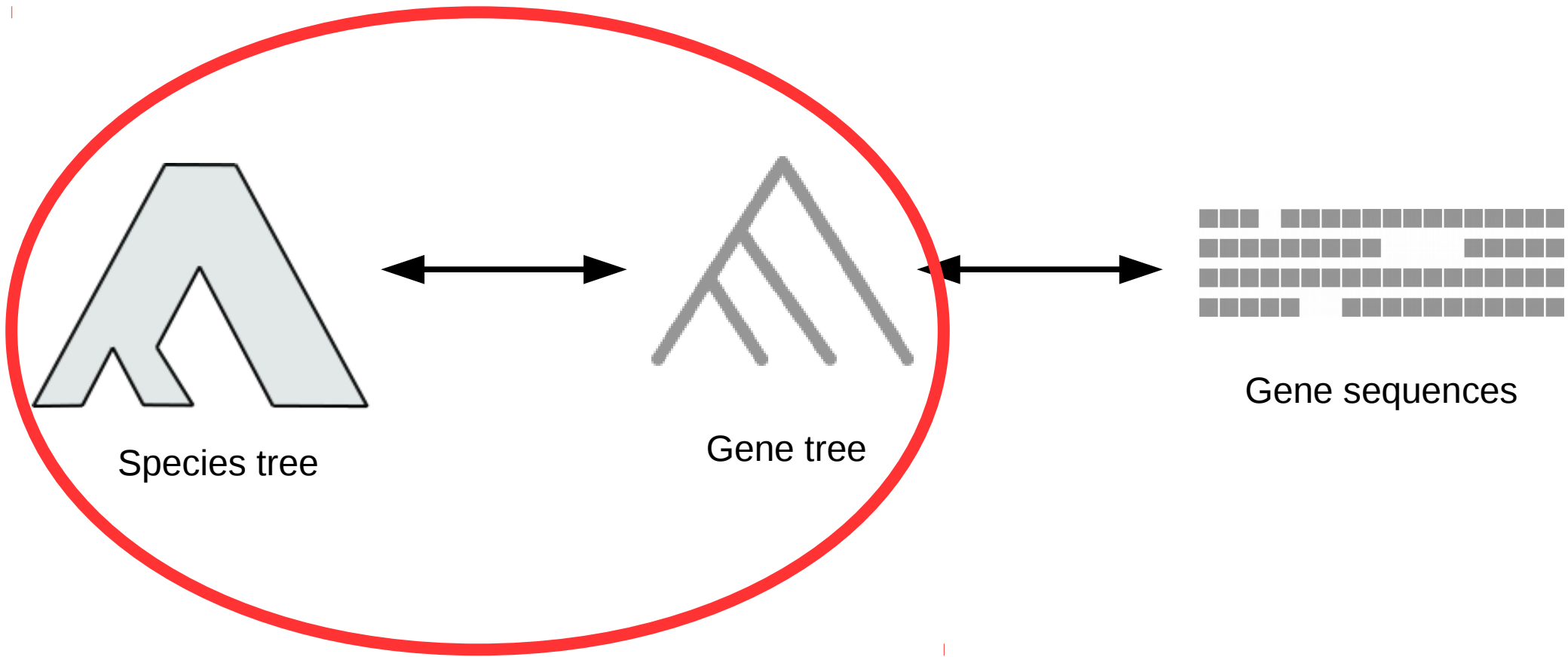


# Model of sequence evolution

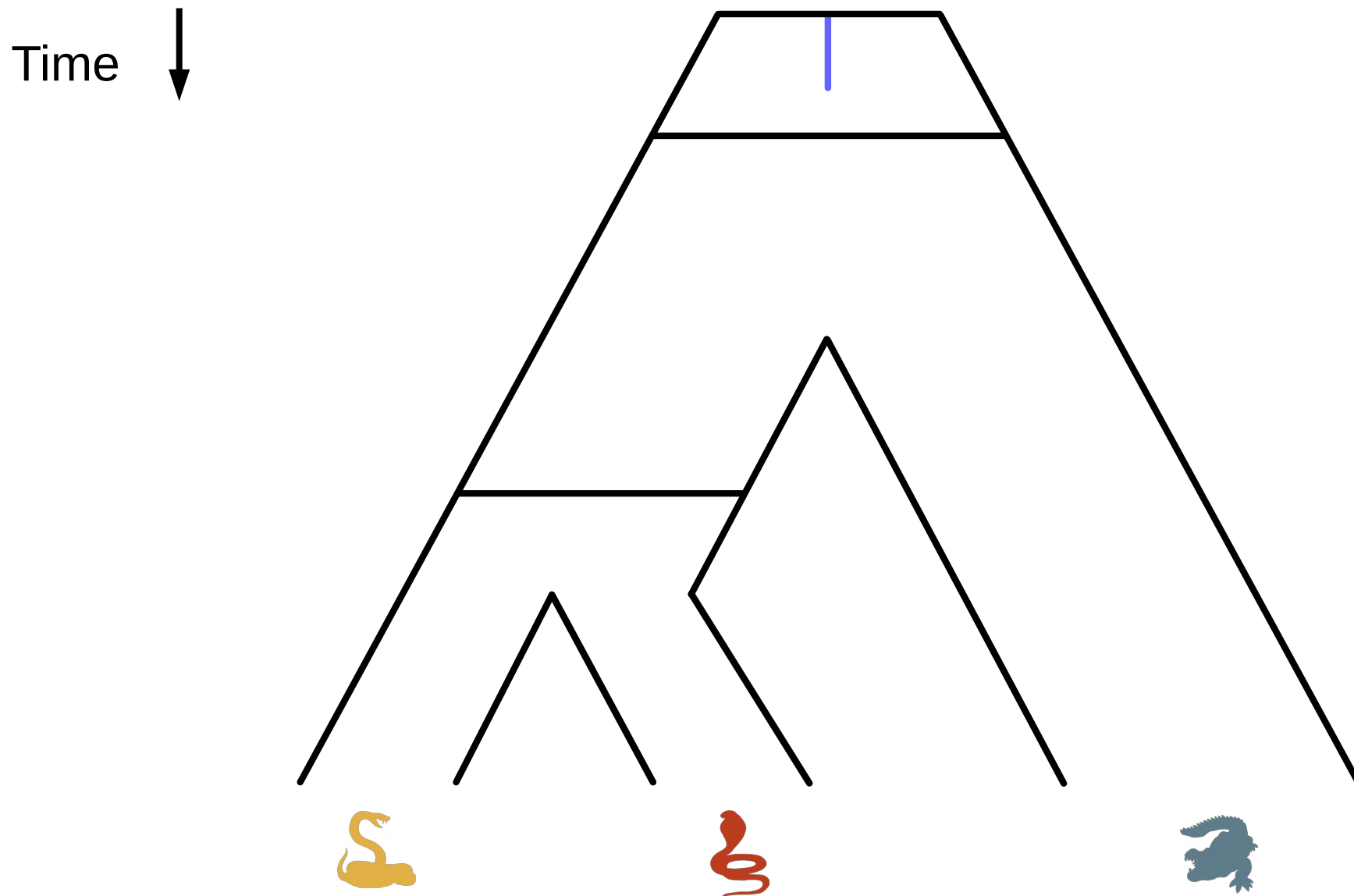
$$\text{Phylogenetic likelihood} = P(\text{sequences} \mid \text{gene tree})$$

The diagram illustrates the components of the phylogenetic likelihood equation. On the left, the text "Phylogenetic likelihood" is followed by an equals sign. To the right of the equals sign is a large letter "P" followed by a vertical bar. To the left of the vertical bar is a representation of "sequences" shown as four horizontal rows of gray bars of varying lengths, indicating gaps or missing data. To the right of the vertical bar is a representation of a "gene tree" shown as a simple branching structure with three tips. Below the sequences and gene tree icons are the labels "sequences" and "gene tree" respectively.

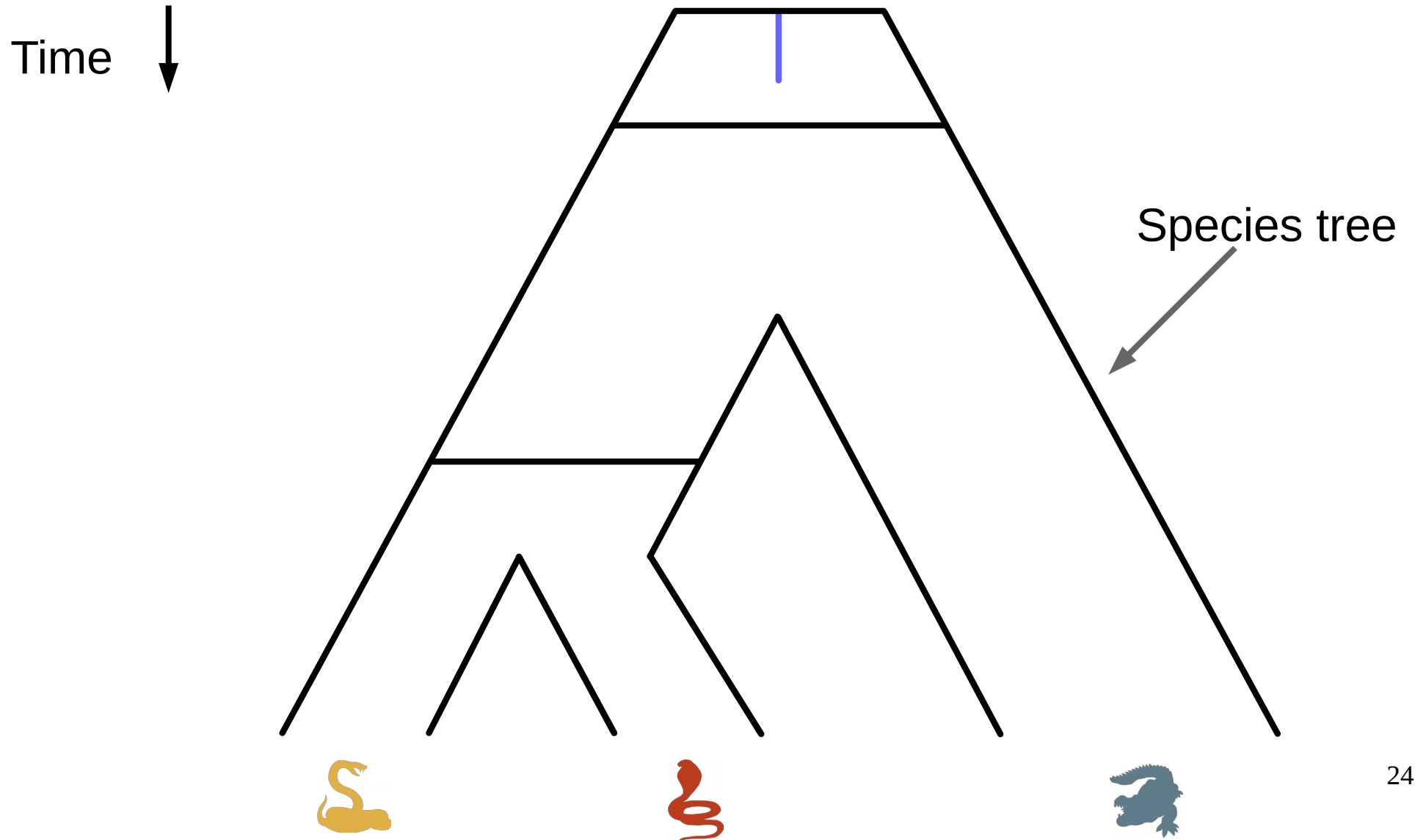
# Hierarchical model



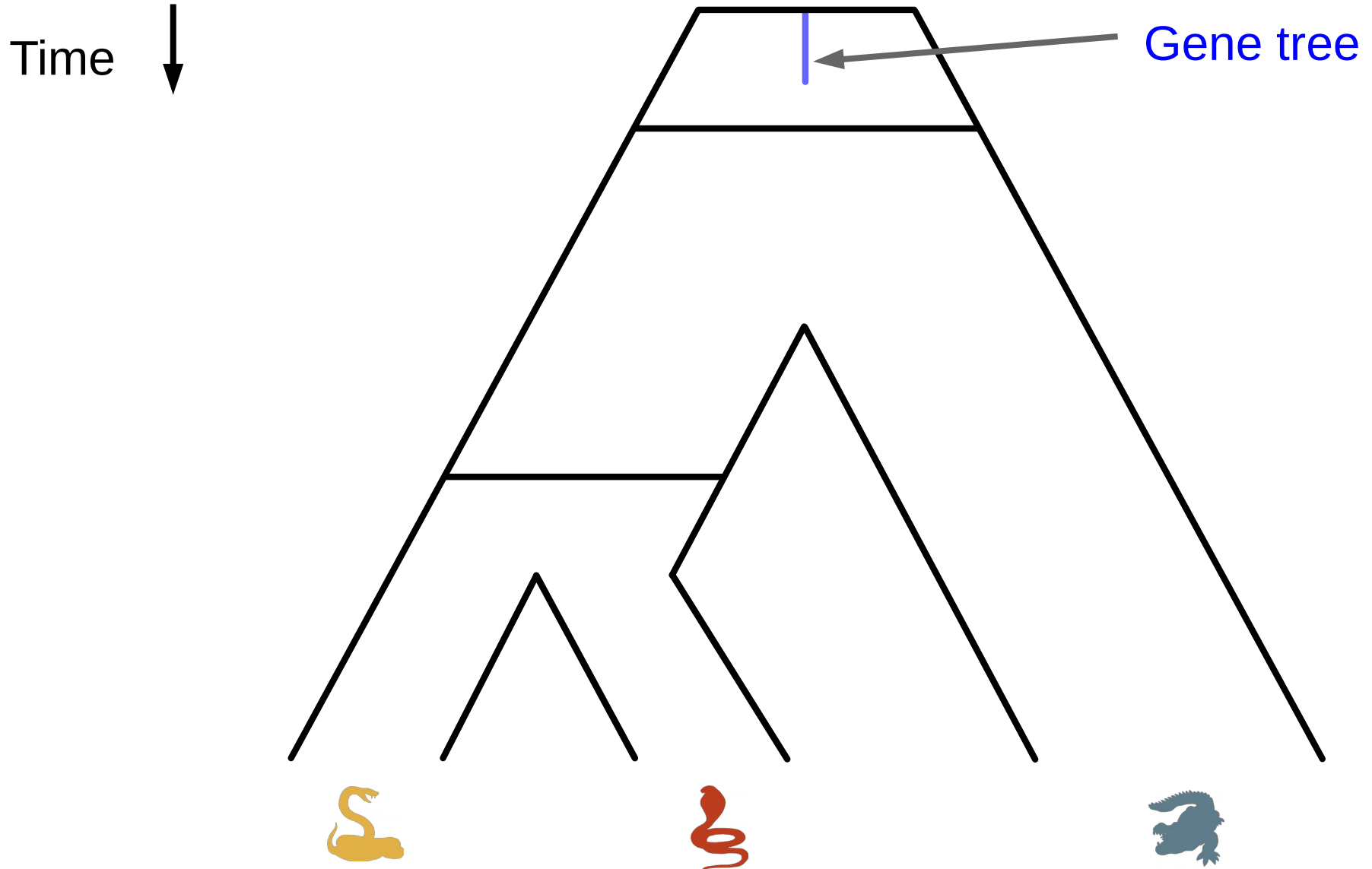
# Gene evolution



# Gene evolution

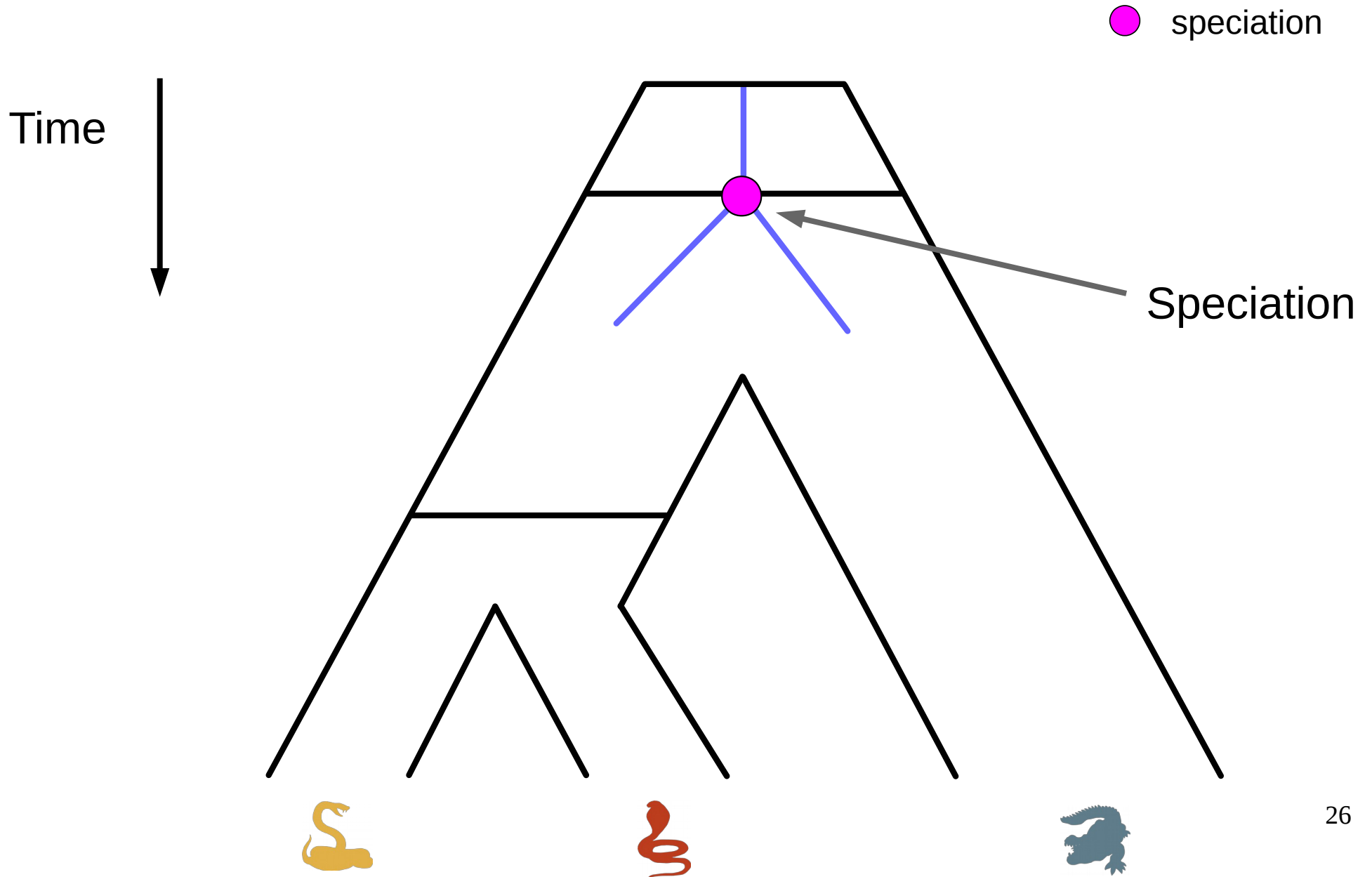


# Gene evolution

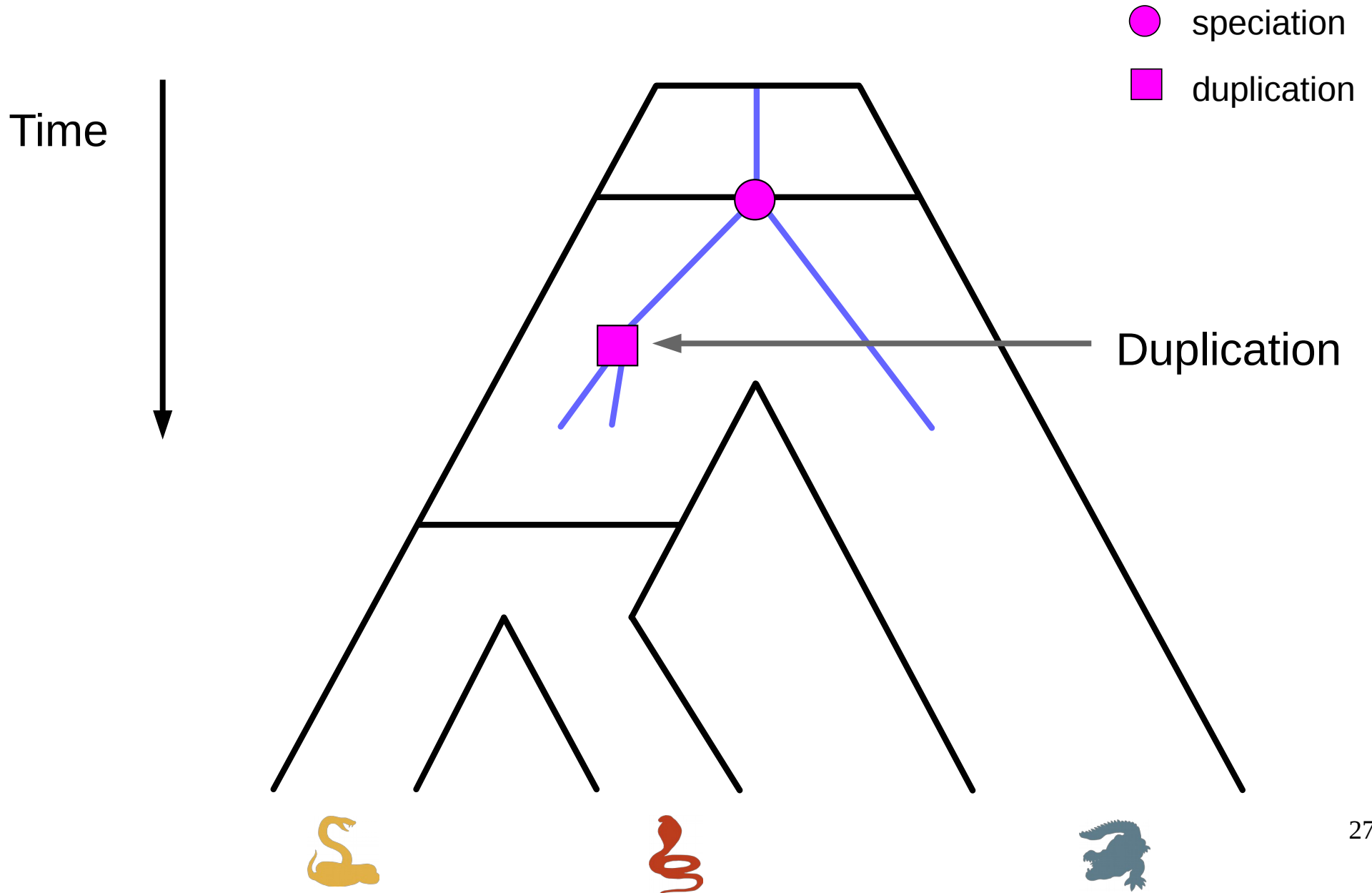




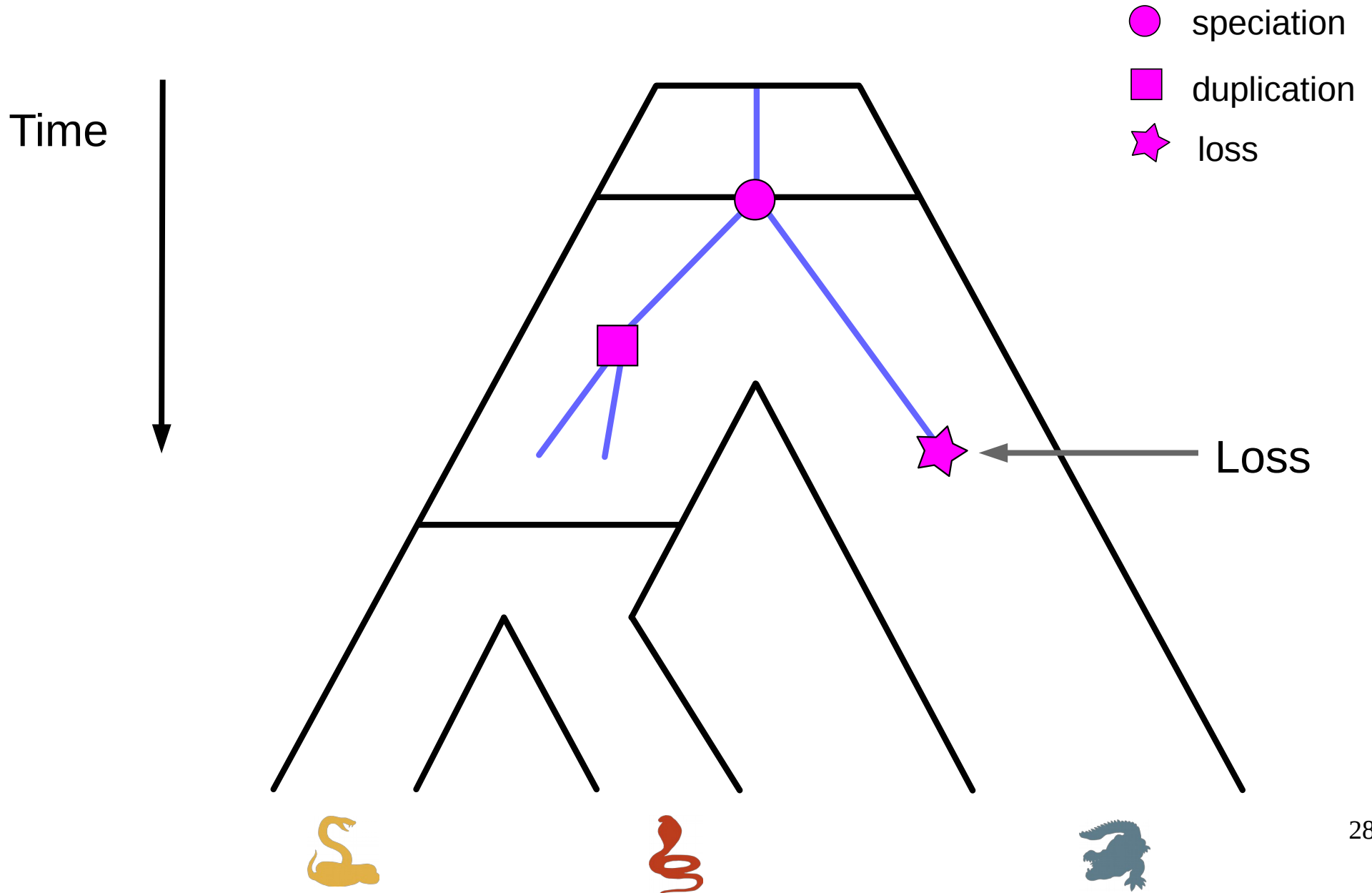
# Gene evolution



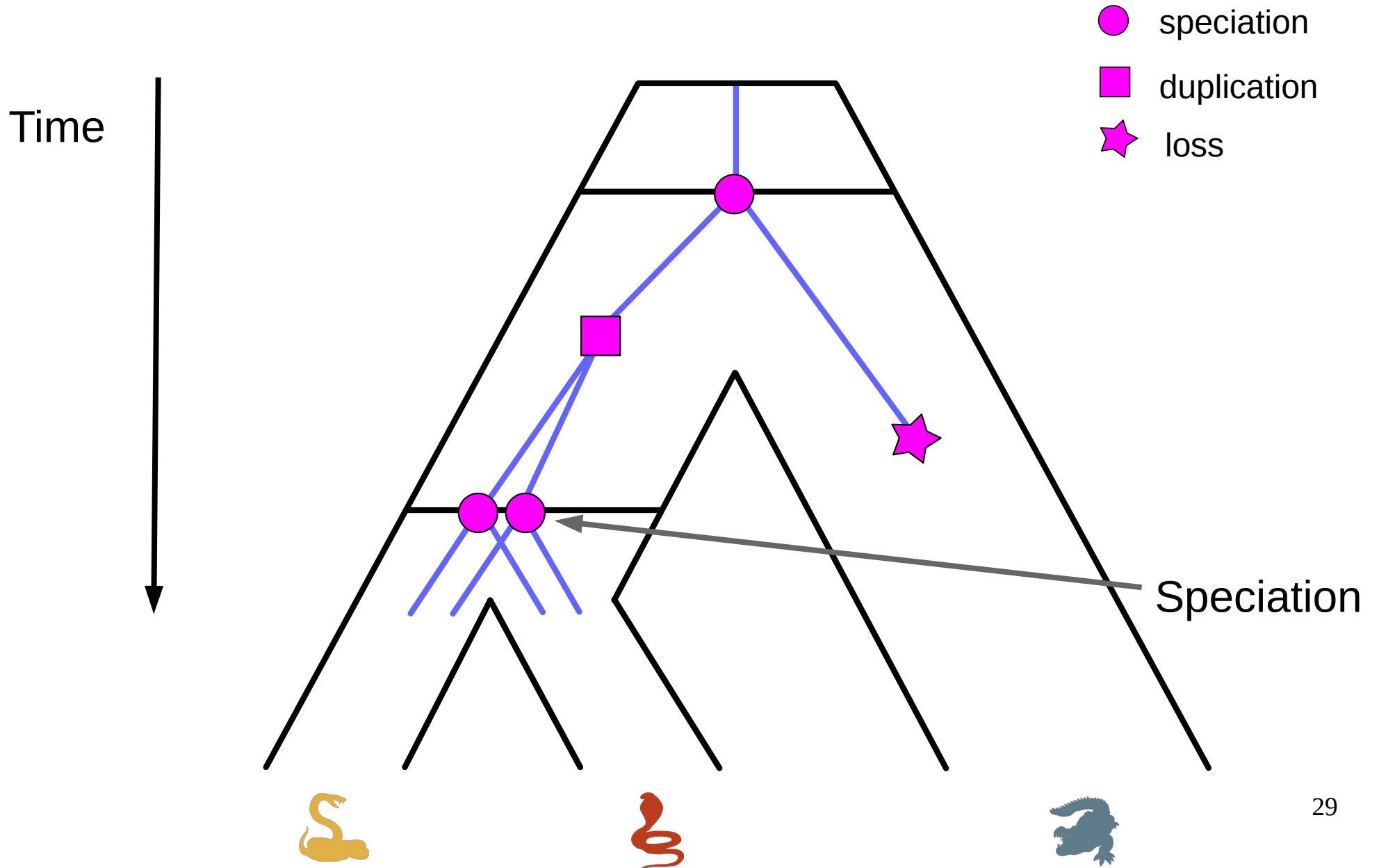
# Gene evolution



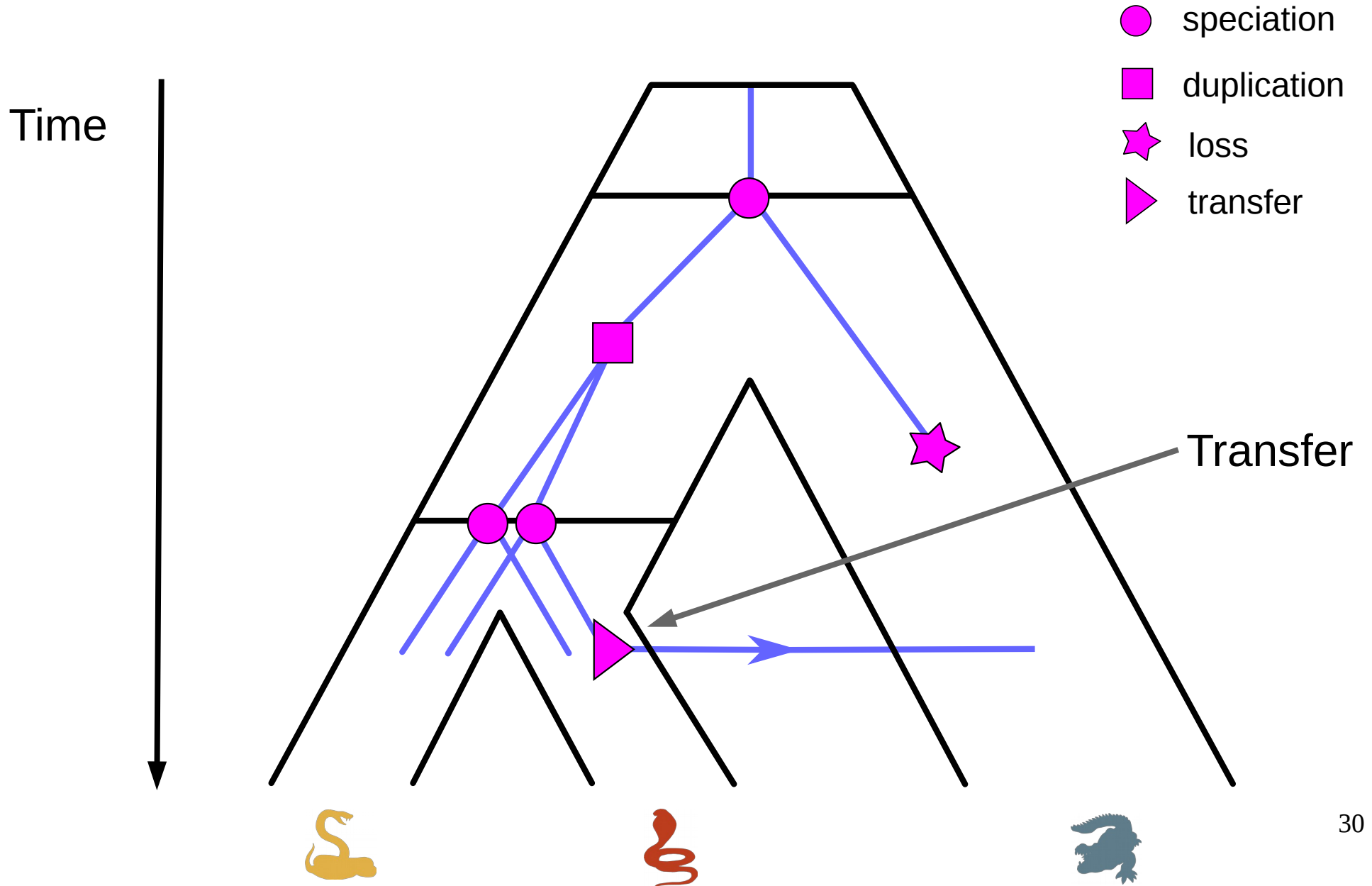
# Gene evolution



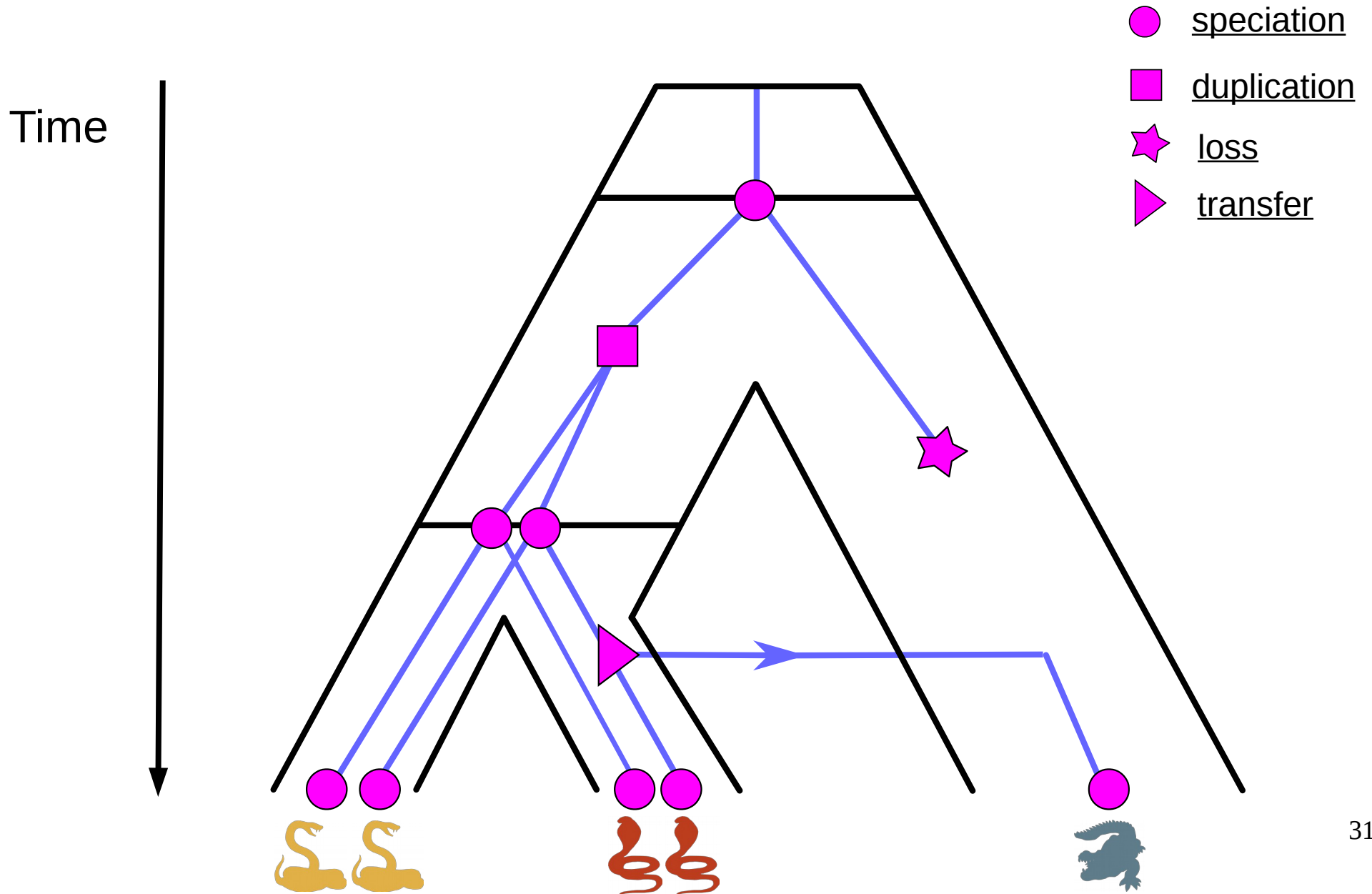
# Gene evolution



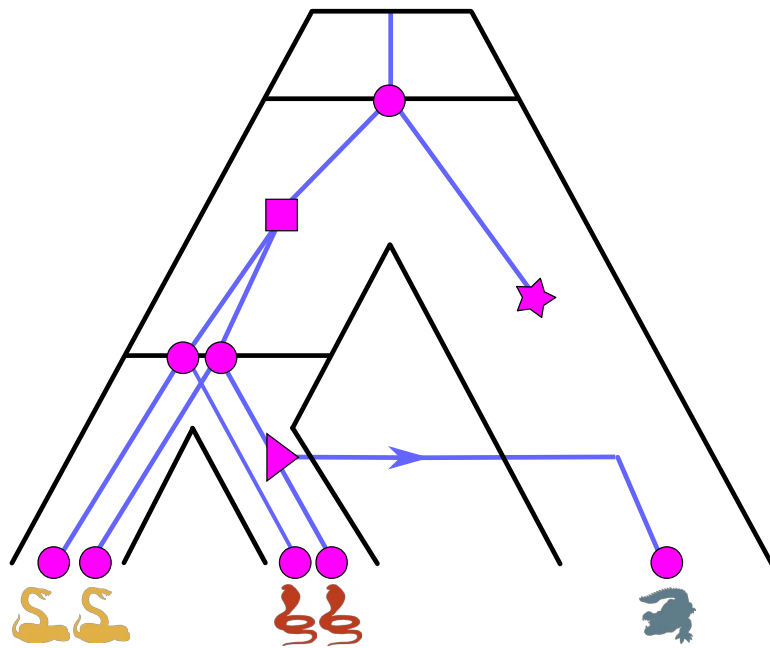
# Gene evolution



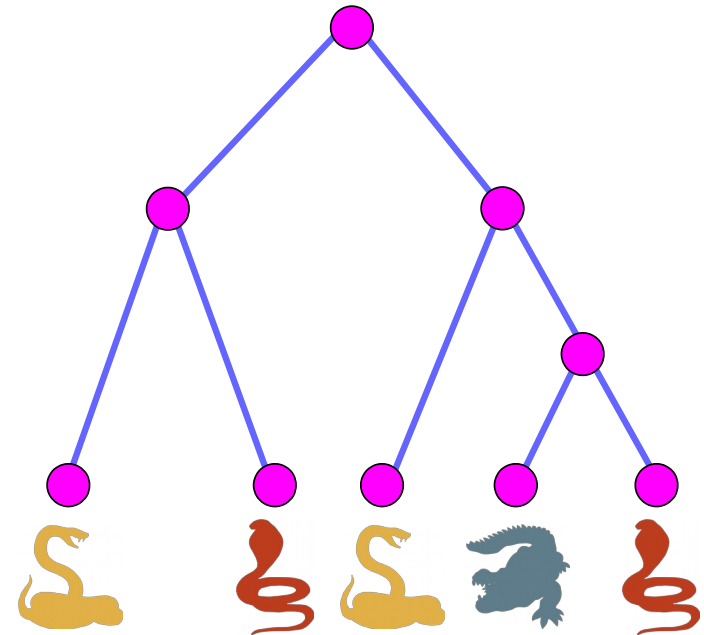
# Gene evolution



# Reconciliation scenario and gene tree

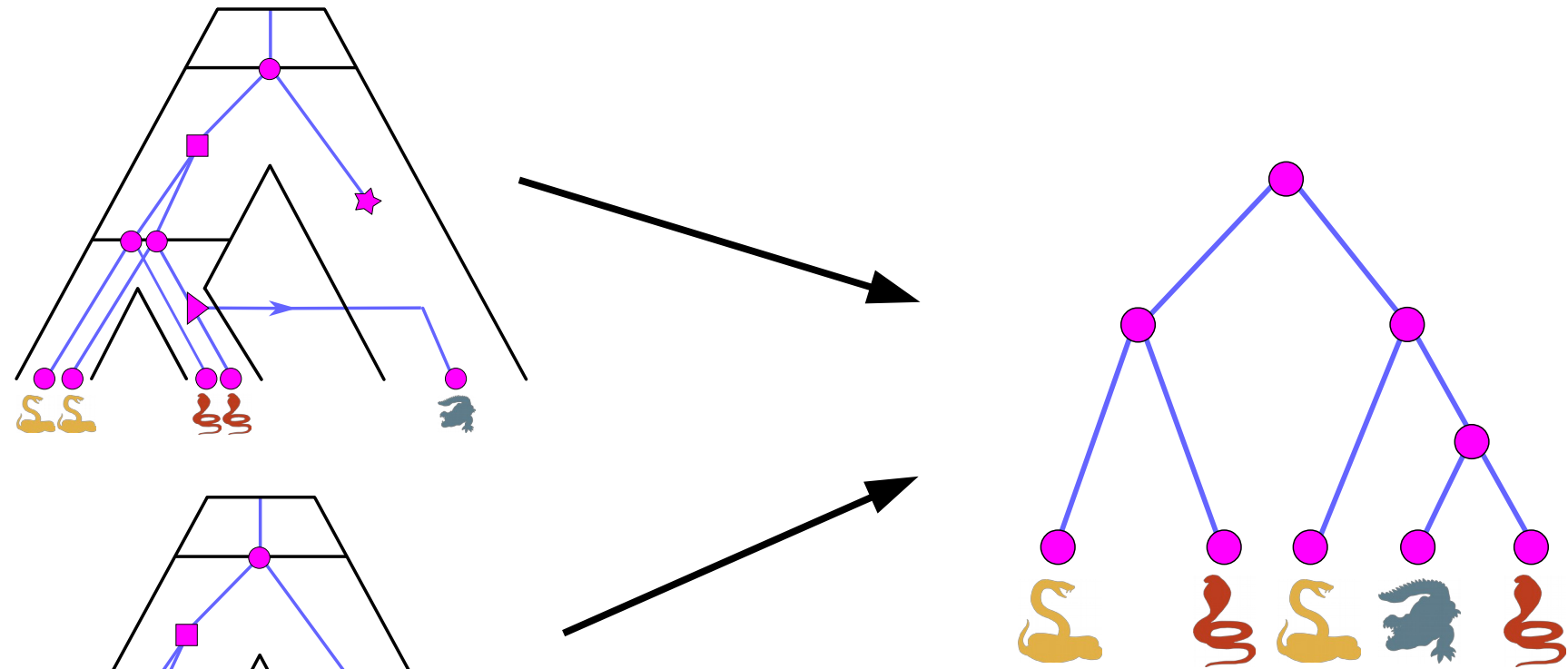


Reconciliation scenario



Gene tree

# Different scenarios can explain the same gene tree



Same gene tree




# The UndatedDTL model

- Describes gene tree evolution under a species tree
- Parametrized by the duplication, loss, and transfer rates
- Assumes that every species has the same chance of receiving a gene transfer

# Model of gene evolution

**Reconciliation likelihood =**  $P(\text{Gene tree} \mid \text{Species tree})$



The diagram illustrates the reconciliation likelihood formula. It shows a large 'P' followed by a large opening parenthesis. Inside the parenthesis, there is a simple branching tree (Gene tree) on the left, a vertical bar in the middle, and a more complex tree (Species tree) on the right. Below the Gene tree is the text 'Gene tree' and below the Species tree is the text 'Species tree'. The parenthesis closes with a large closing parenthesis.

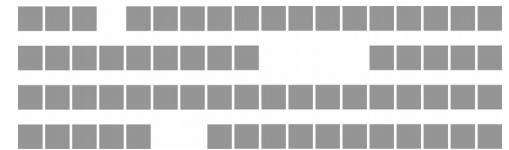
# Hierarchical model



Species tree



Gene tree



Gene sequences

# Hierarchical model

Reconciliation likelihood

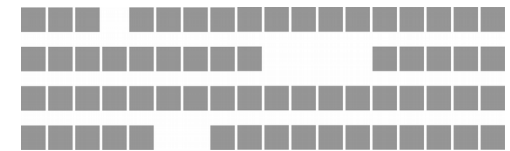
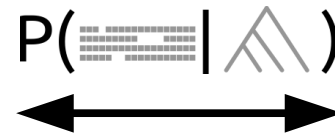
Phylogenetic likelihood



Species tree



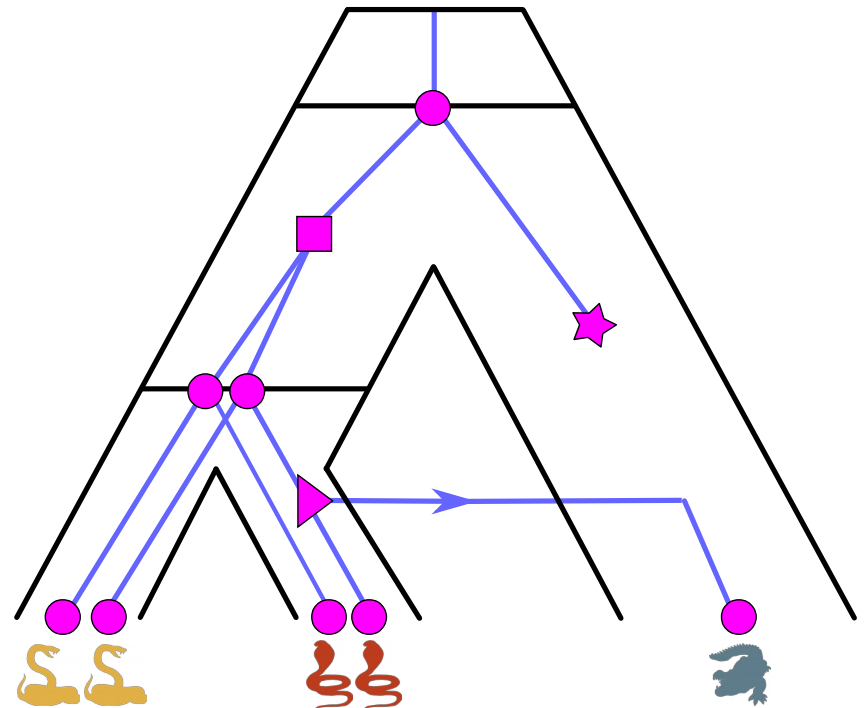
Gene tree



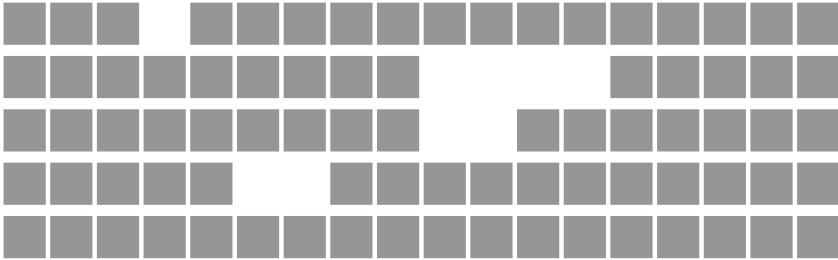
Gene sequences

# The goal

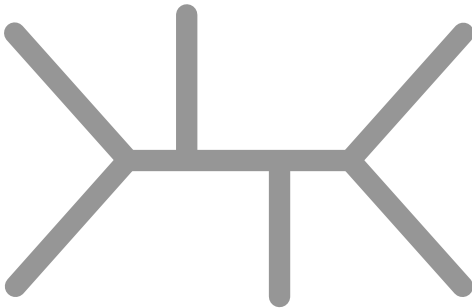
- We want to infer:
  - the gene tree
  - (the species tree)
  - their reconciliation
  - the DTL probabilities



# Gene tree inference



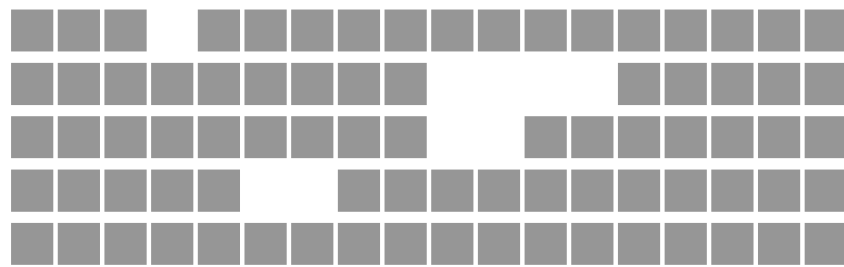
Gene sequence alignment



Gene tree

# Gene tree inference under maximum likelihood

- Search for the gene tree that maximizes the phylogenetic likelihood:

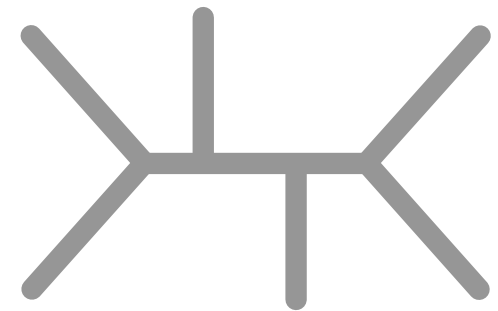


Gene sequence alignment

$$P(\text{alignment} \mid \text{tree})$$



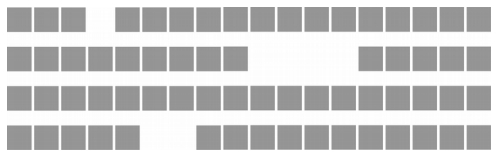
RAXML  
IQTREE



Gene tree

# Gene tree inference

- Gene sequences are short
- Not enough signal to resolve the gene tree



Gene sequence  
alignment

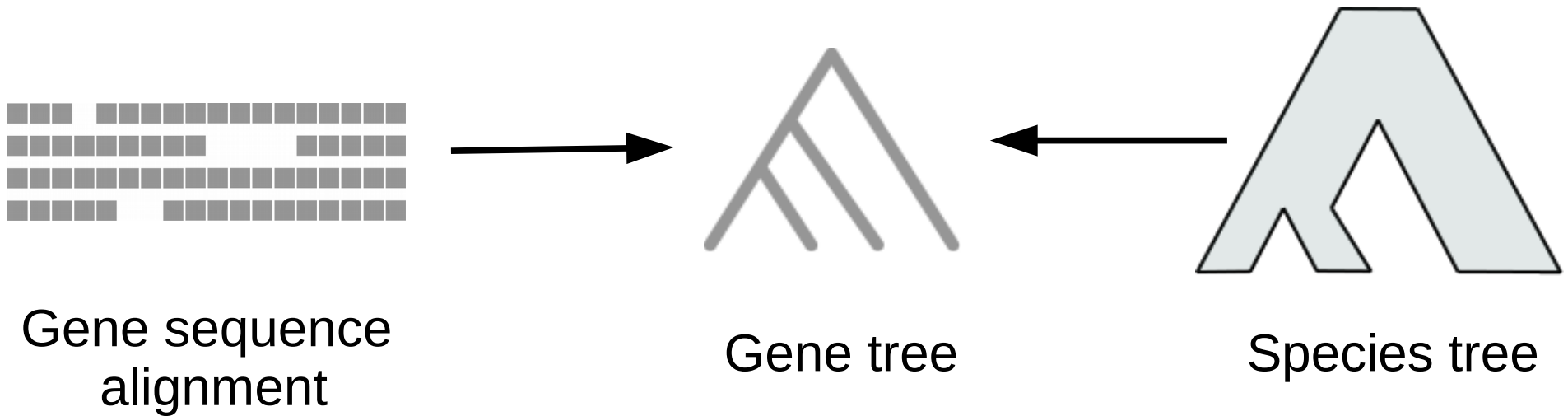


Inaccurate gene tree



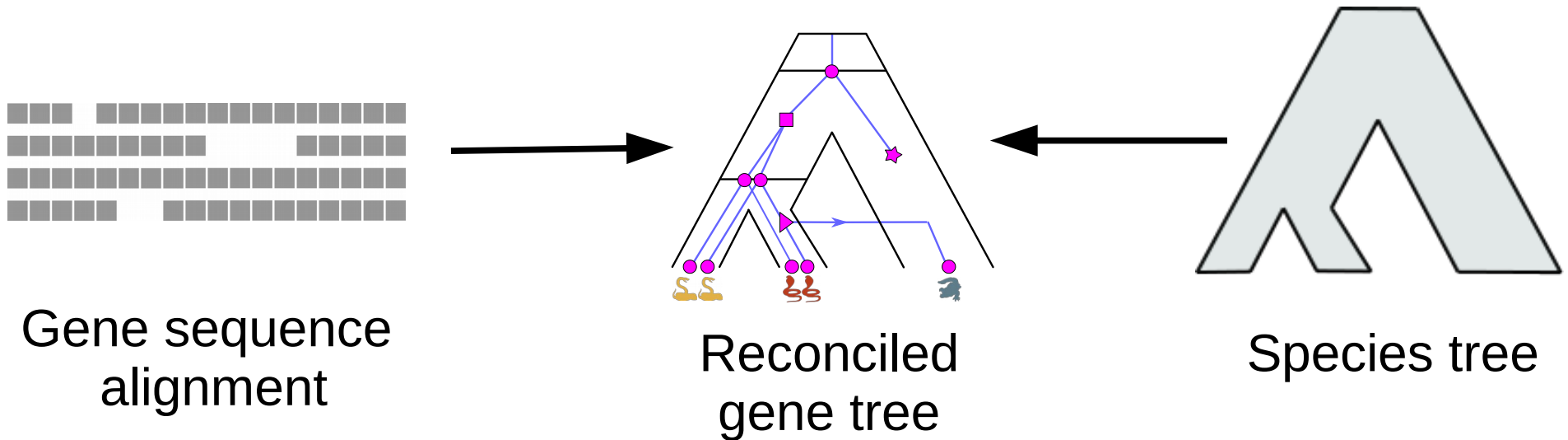
# Gene tree inference

- Solution: use the species tree



# Gene tree inference

- Solution: use the species tree



# Species tree aware gene tree inference

- GeneRax: optimize the gene tree
- AleRax: integrate over all gene trees

(there are many other interesting methods!)

# GeneRax

Find the gene tree that maximizes the joint likelihood:

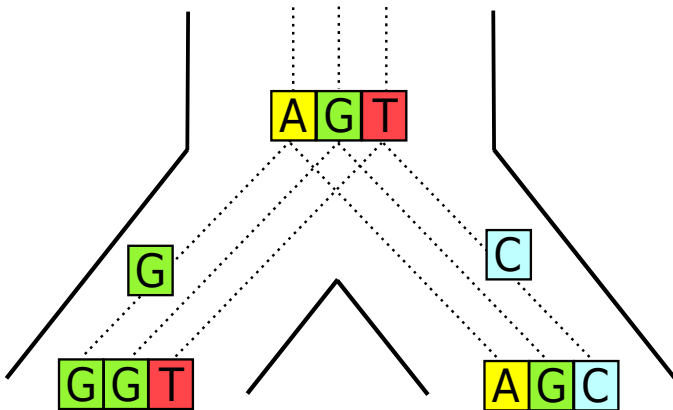
$$P(\text{alignments} \mid \text{tree}) \quad P(\text{tree} \mid \text{alignments})$$

# GeneRax

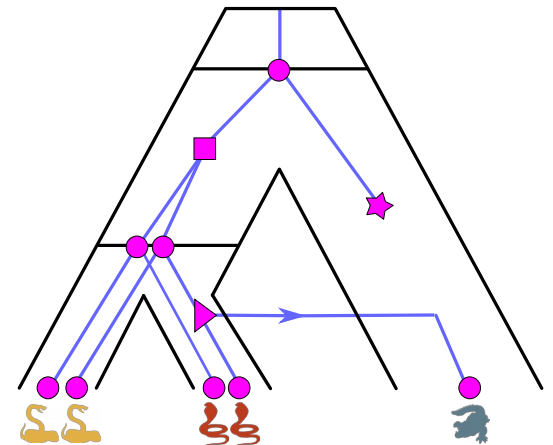
Find the gene tree that maximizes the joint likelihood:

$$P(\text{Sequences} \mid \text{Gene Tree}) \quad P(\text{Gene Tree} \mid \text{Species Tree})$$

Phylogenetic likelihood



Reconciliation likelihood



# How to find the maximum likelihood gene tree?

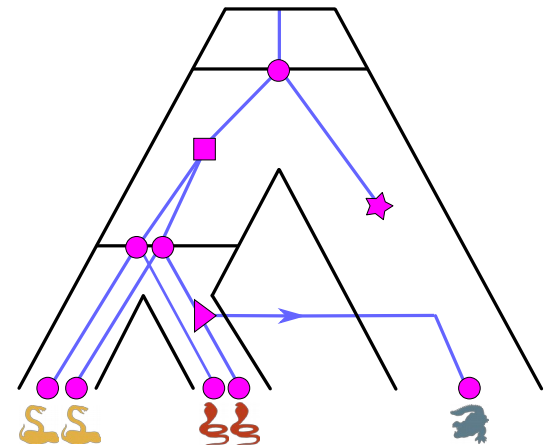
- Start from any gene tree  $G$
- For each “neighbor”  $G'$ :
  - estimate the likelihood of  $G'$
  - if  $G'$  has a higher likelihood than  $G$ , replace  $G$  with  $G'$
- Stop when no better gene tree can be found

# Model parameters

- Model parameters: D,T,L,S probabilities
- We optimize them to maximize the likelihood function after each round of tree search

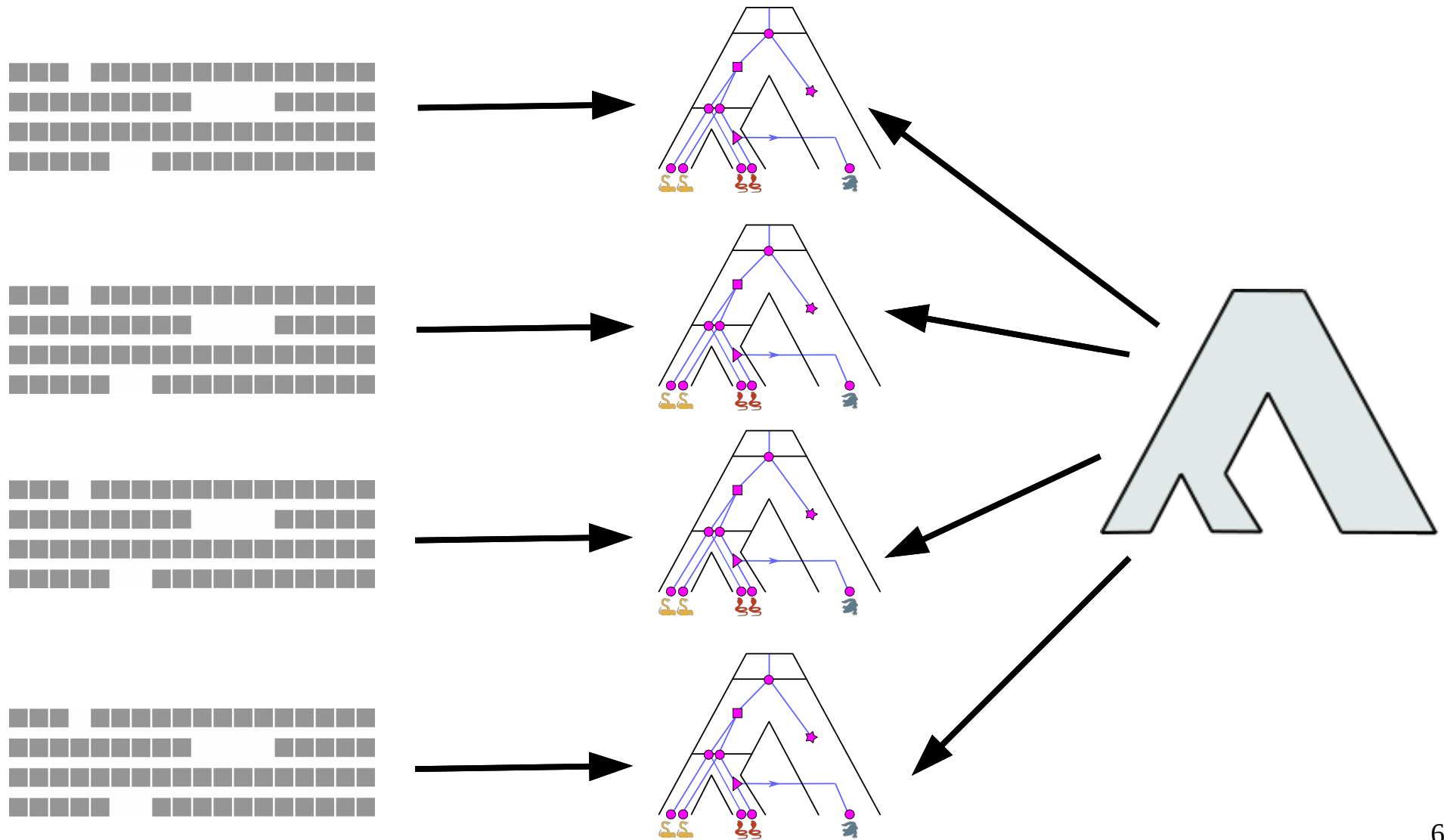
# Gene tree reconciliation

- Now we have the gene tree and the species tree
- We select the reconciliation scenario with the highest likelihood
- We use a recursive dynamic programming algorithm to iterate over all possible scenarios that are compatible with the gene tree





# GeneRax: thousands of families and hundreds of species



# Parallelization scheme

- Two-level parallelization:
  - We treat different gene families in parallel
  - We assign several cores to each individual gene families
- excellent parallel efficiency :-)

# GeneRax and AleRax

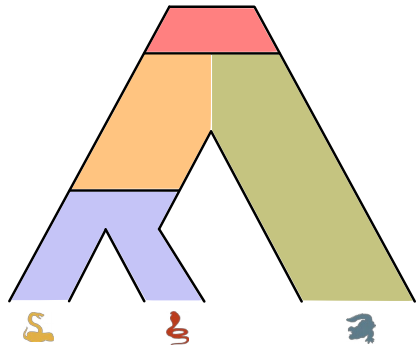
- GeneRax: co-estimate gene trees and model parameters
-

# GeneRax and AleRax

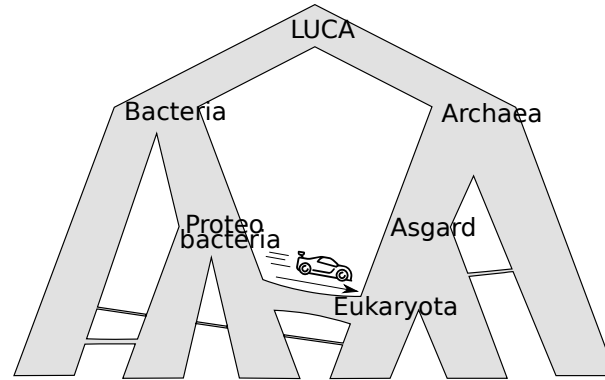
- GeneRax: co-estimate gene trees and model parameters
- AleRax: treat gene trees as latent variables, and integrate over them

$$\sum_{\mathcal{T}} P(\text{Sequences} | \mathcal{T}) P(\mathcal{T} | \mathcal{A})$$

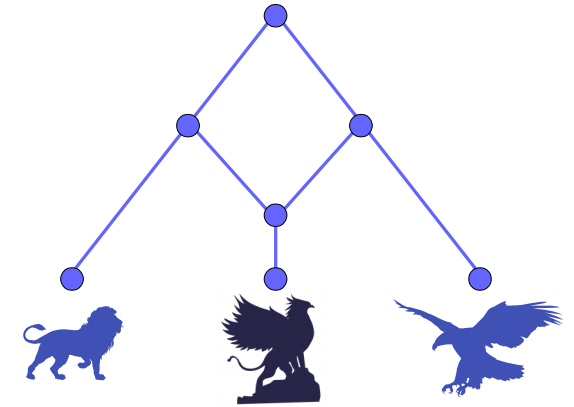
# Limitations and opportunities



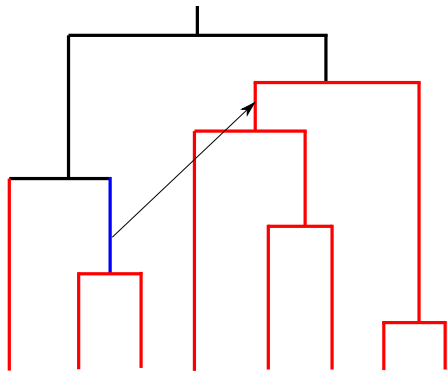
Heterogeneous DTL rates



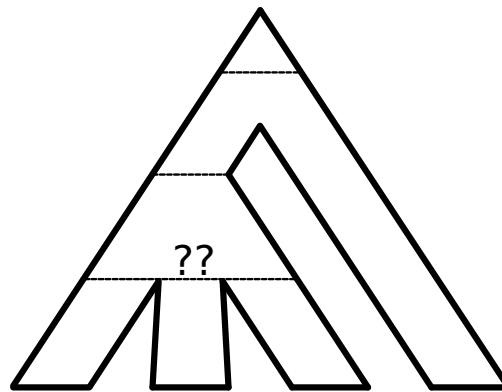
Transfer highways



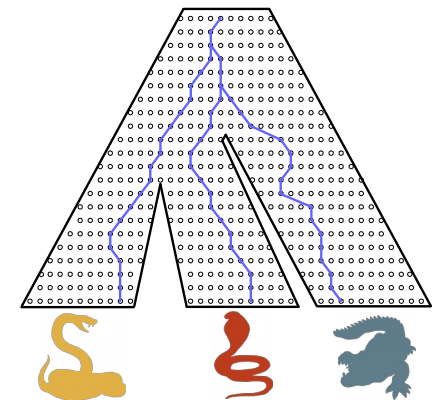
Model species networks



Time constraints



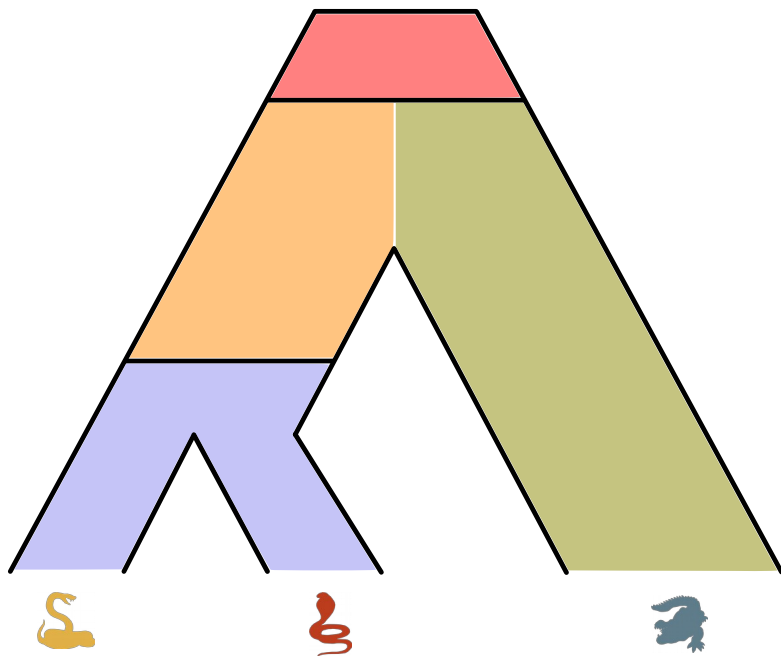
Species tree uncertainty



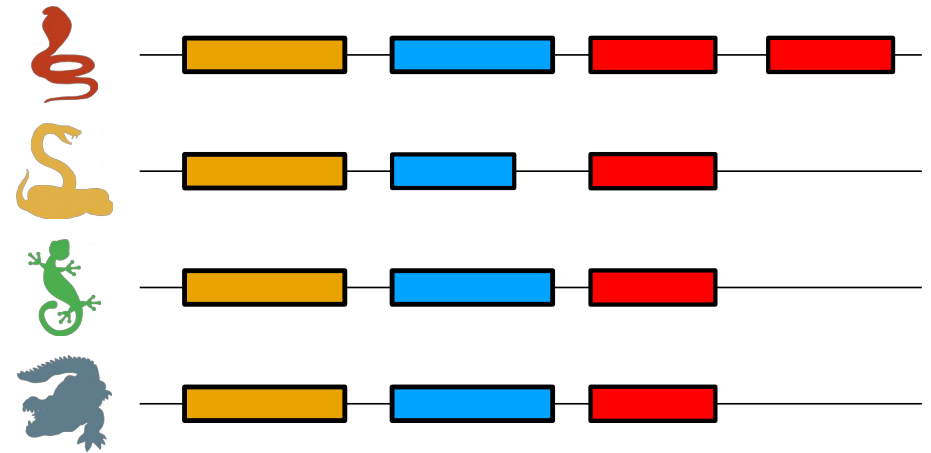
Incomplete lineage sorting

# Model parameter heterogeneity

- Different species and different families have different DTL probabilities



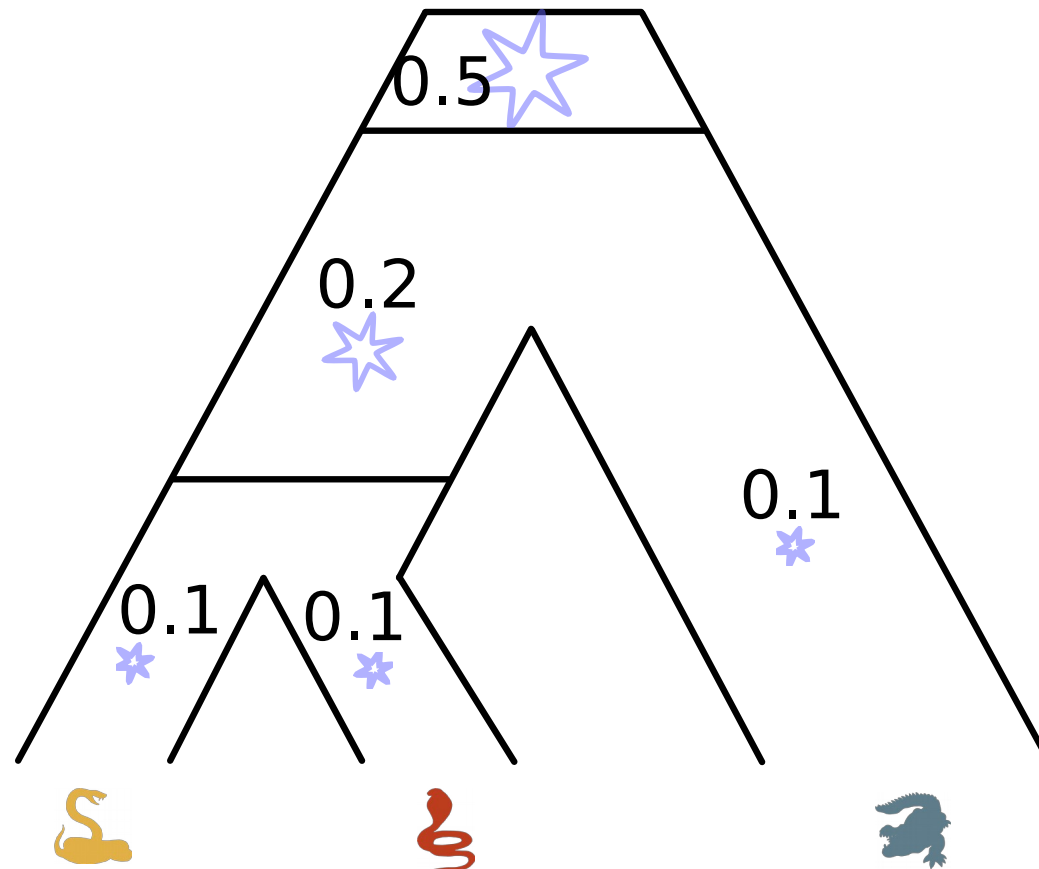
Per species heterogeneity



Per gene family heterogeneity

# Model parameter heterogeneity

- Origination (de novo gene) probabilities



# Transfer probabilities

- Transfer probabilities are not uniform

Cyanobacteria → another Cyanobacteria

Human → fig tree

Cyanobacteria → first plant



# Transfer probabilities

- Transfer probabilities are not uniform

Cyanobacteria → another Cyanobacteria

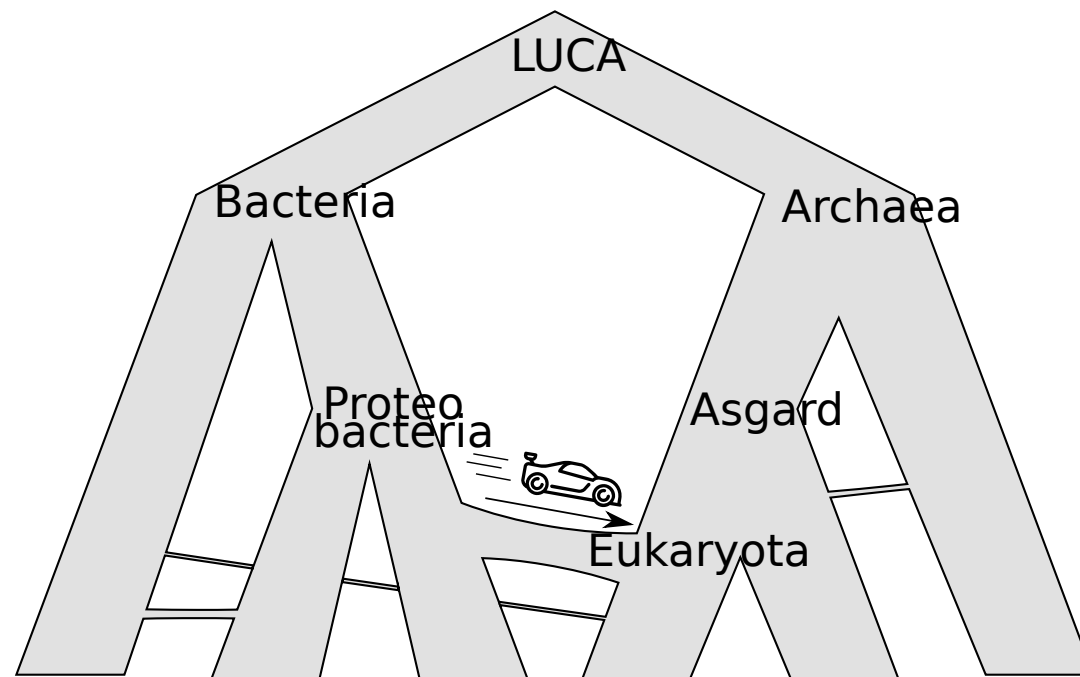
Human → fig tree

Cyanobacteria → first plant

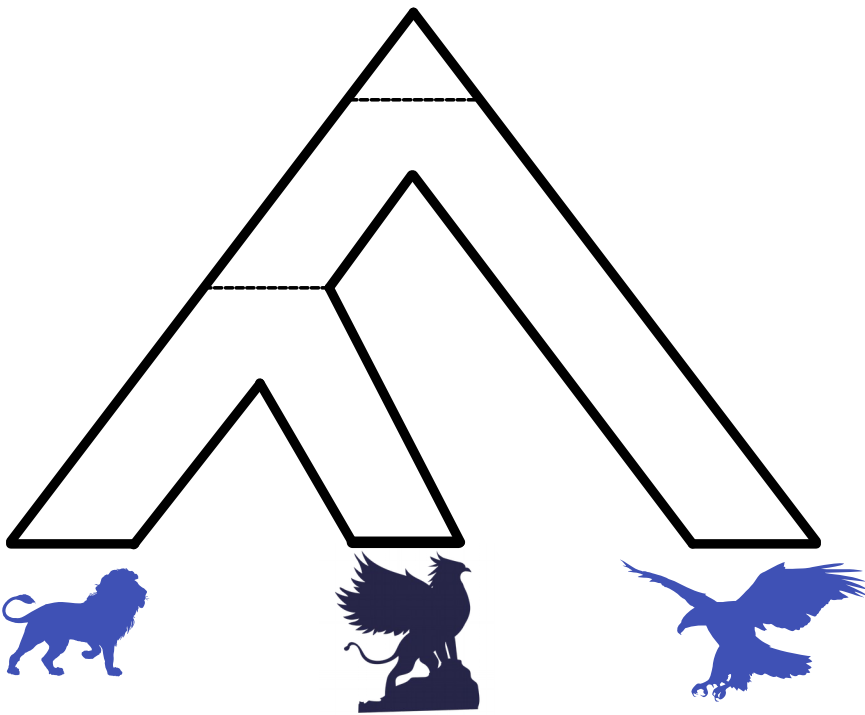
... But we can't estimate the  $N^2$  combinations

# Model horizontal gene transfer highways

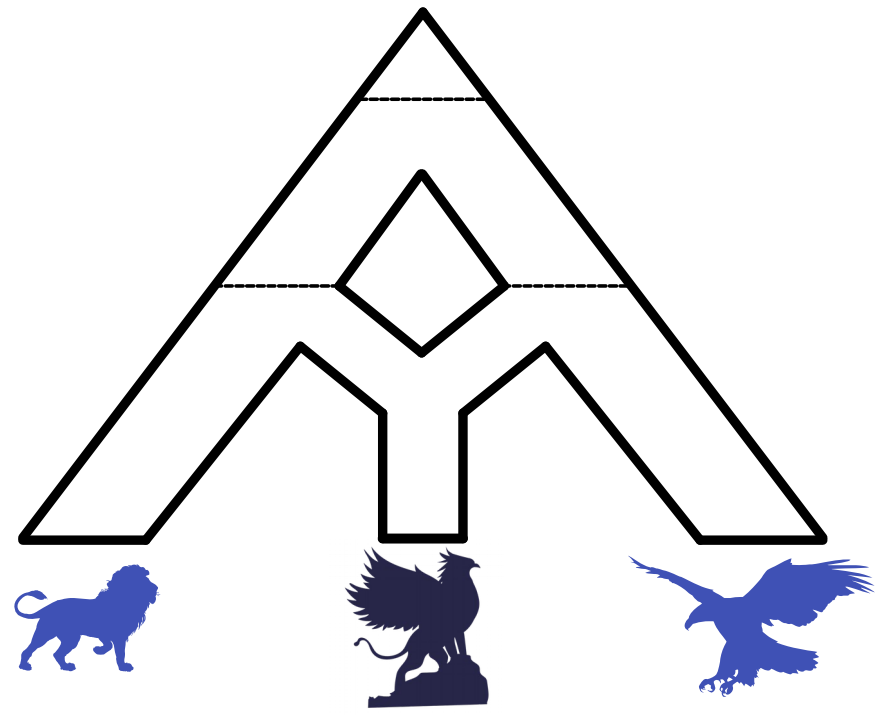
- Focus on the “interesting” pairs of species
- Transfer highway: pair of species that exchanged many genes



# Species networks

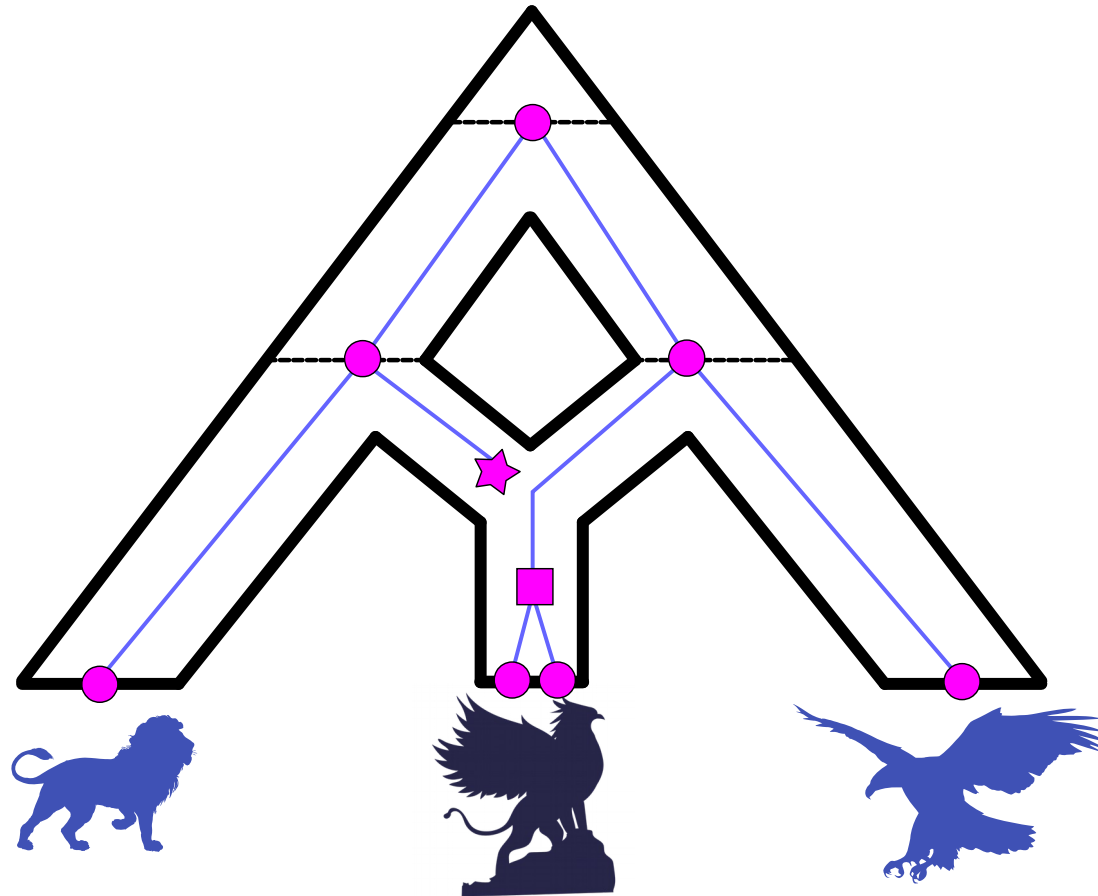


Species tree



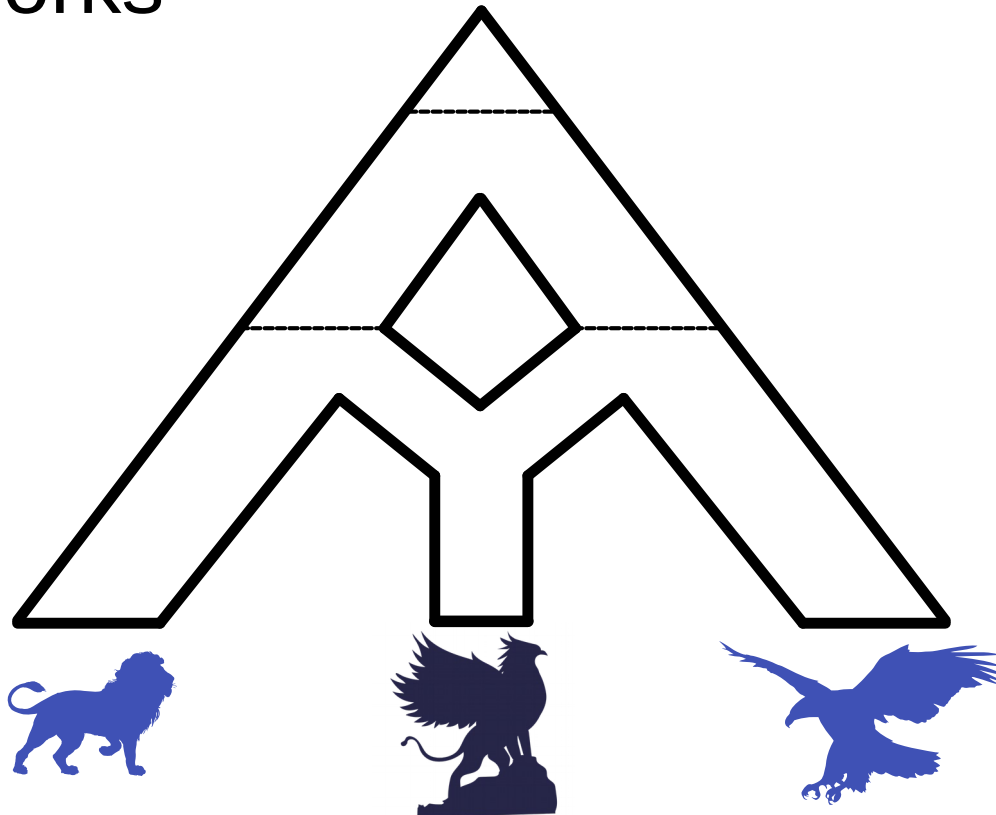
Species network

# Species networks

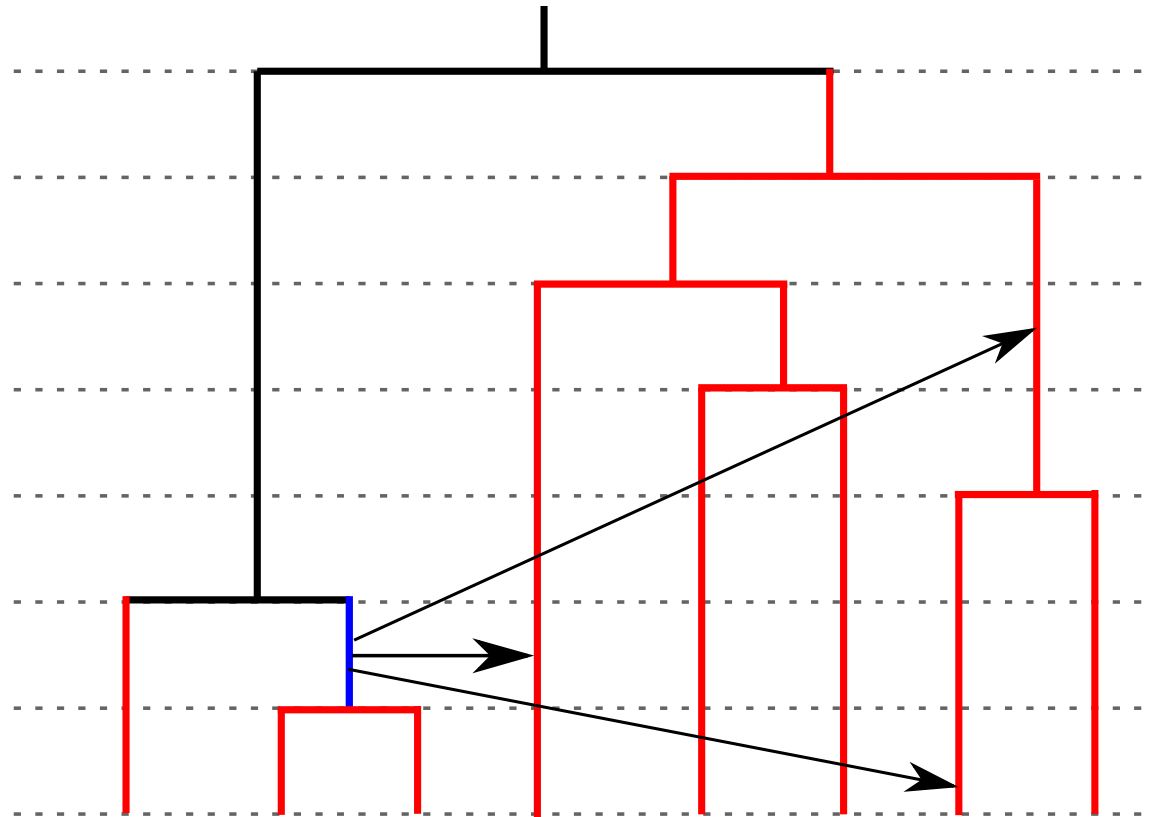


# Species networks

- Reconcile under networks
- Test network hypotheses
- Infer networks

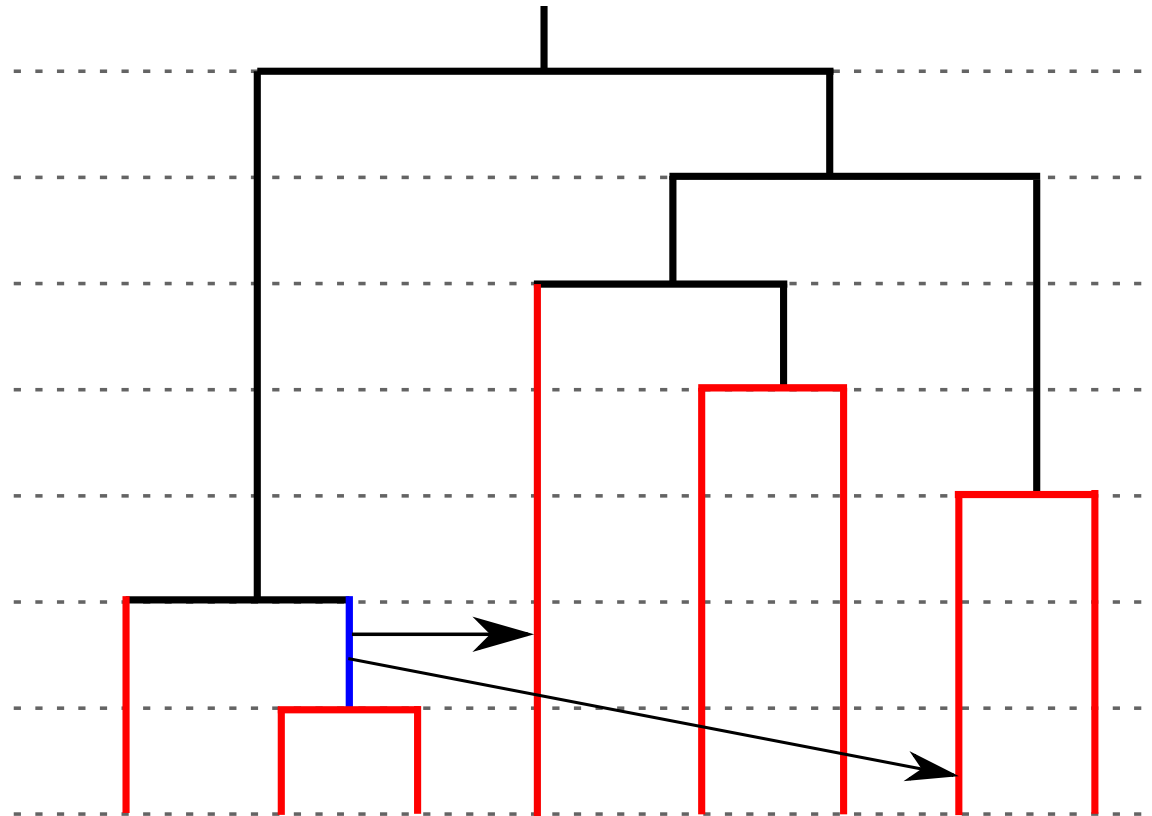


# Time constraints on the transfers



UndatedDTL model

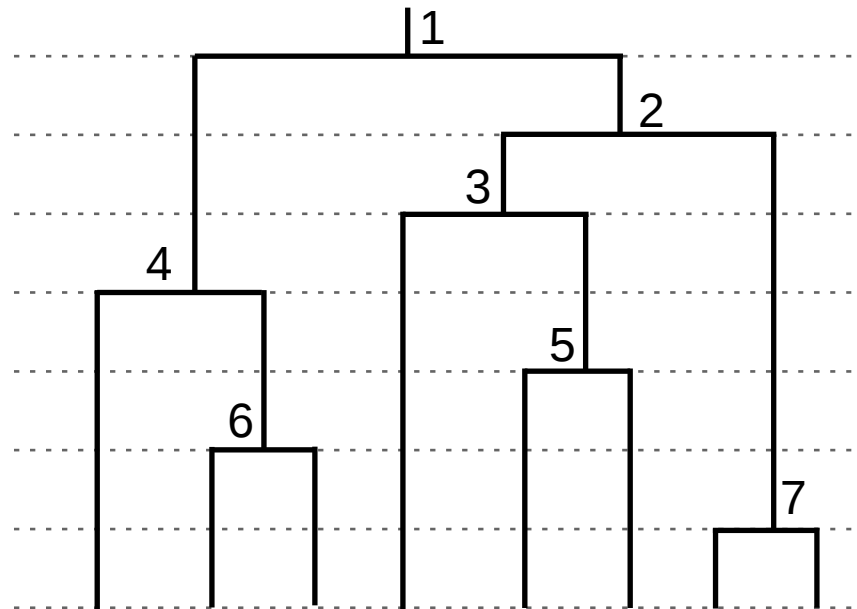
# Time constraints on the transfers



ReldatedDTL model

# Relative dating

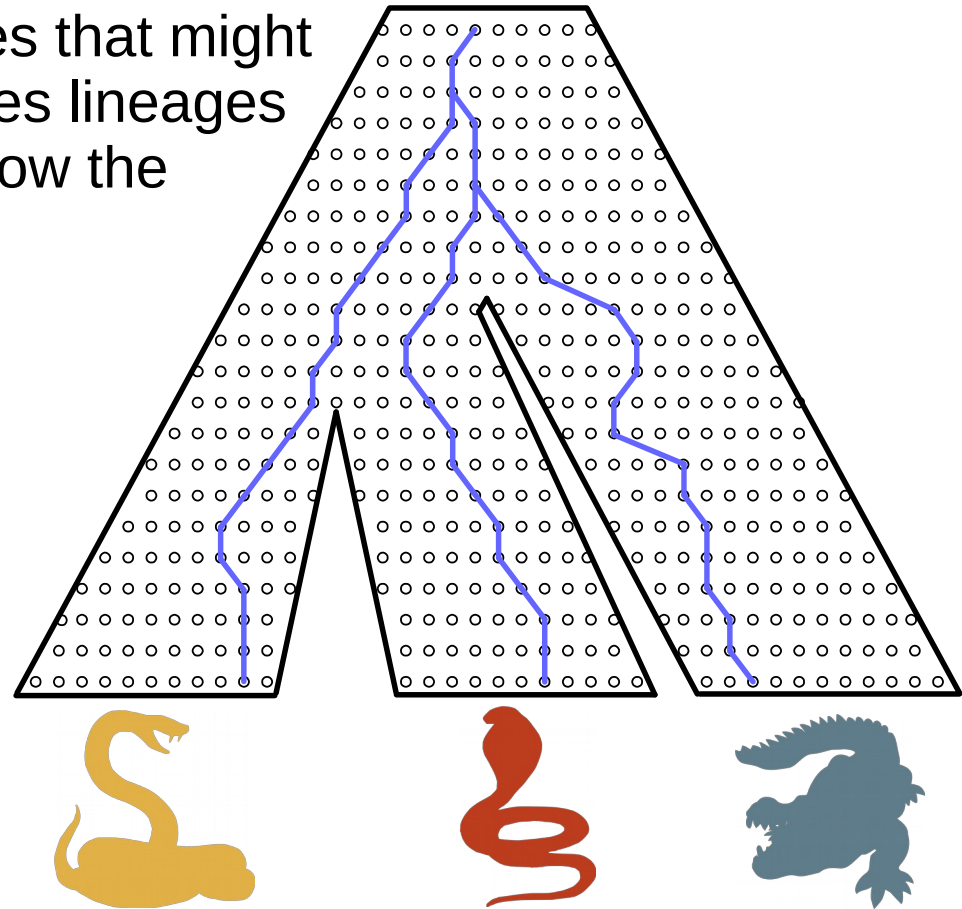
- Use the RelDated model to estimate the most likely order of speciation events



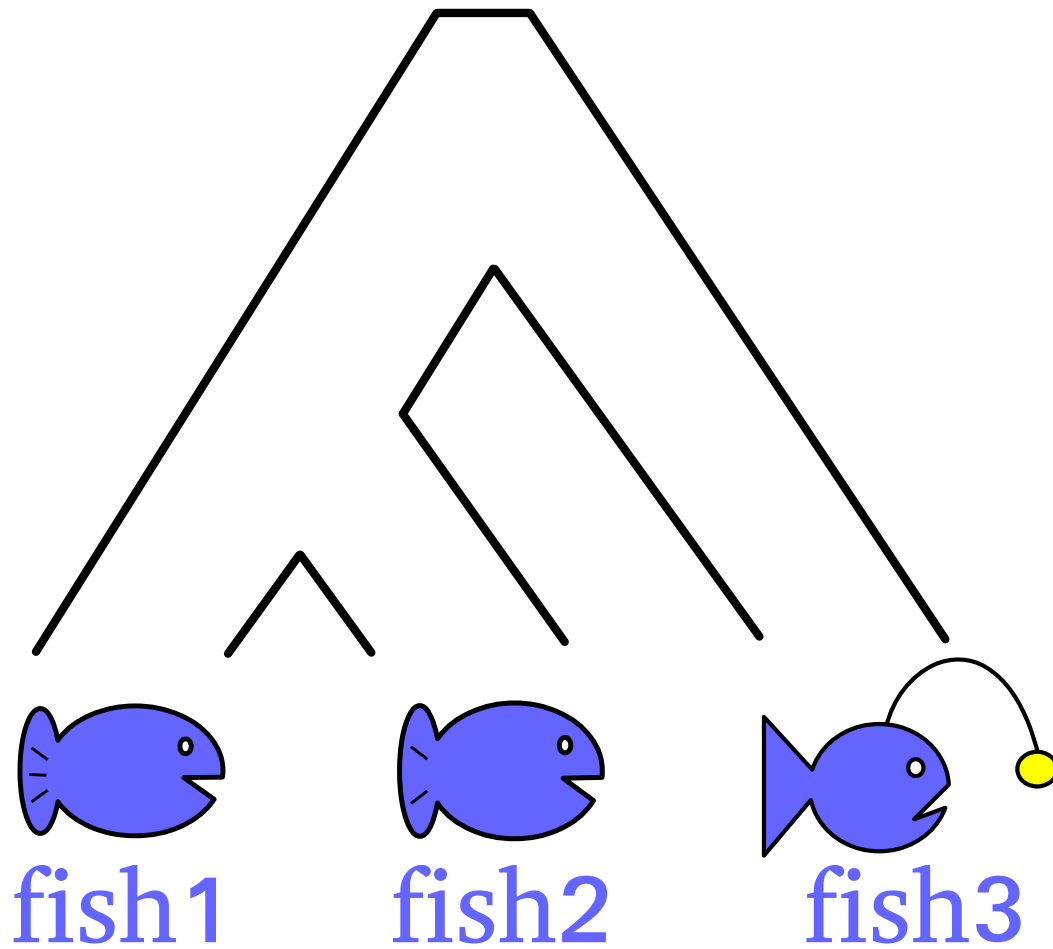


# Incomplete lineage sorting

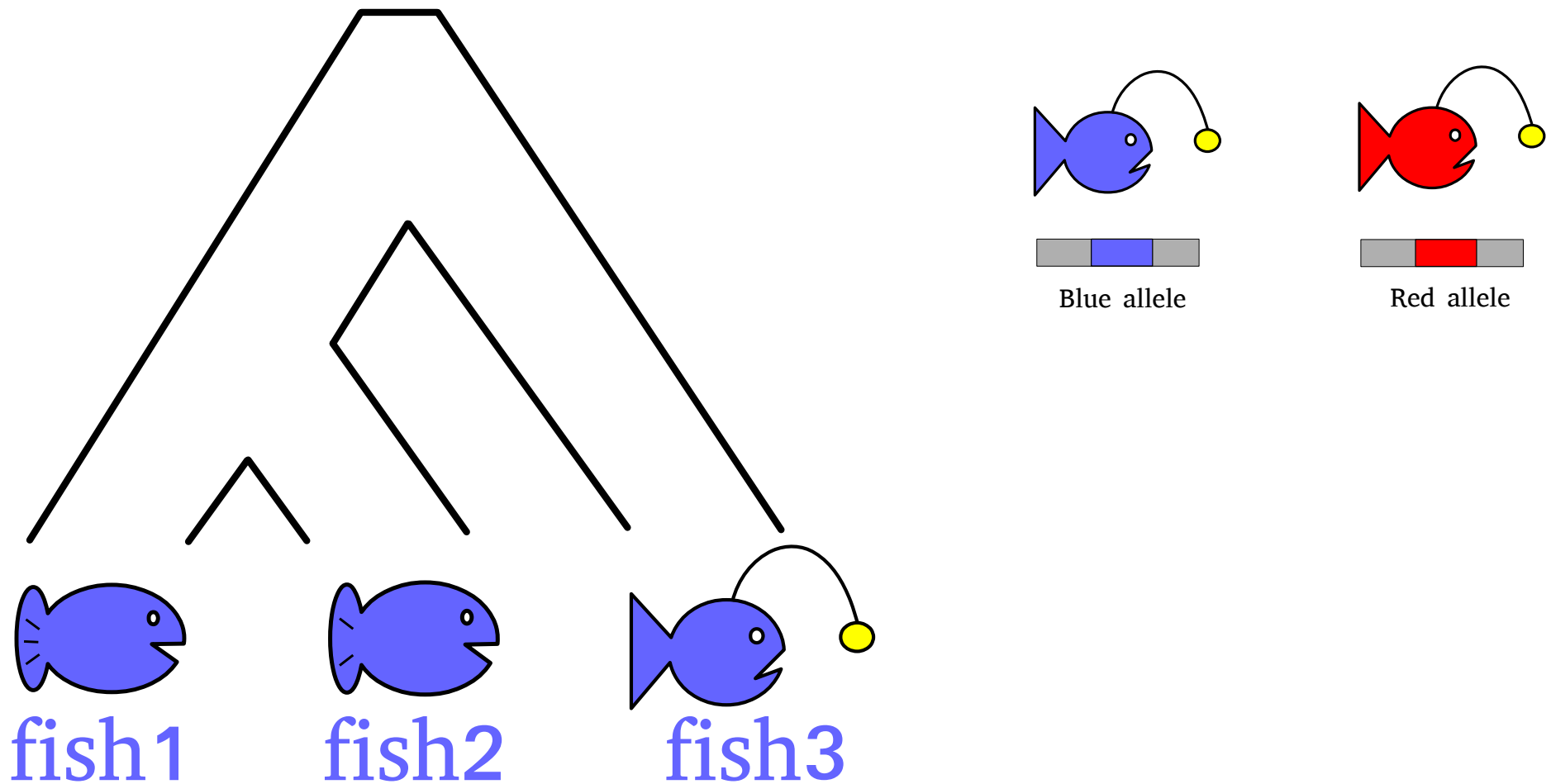
- Species are populations
- Gene can have multiple alleles that might co-exist across several species lineages
- Gene trees do not always follow the species tree structure



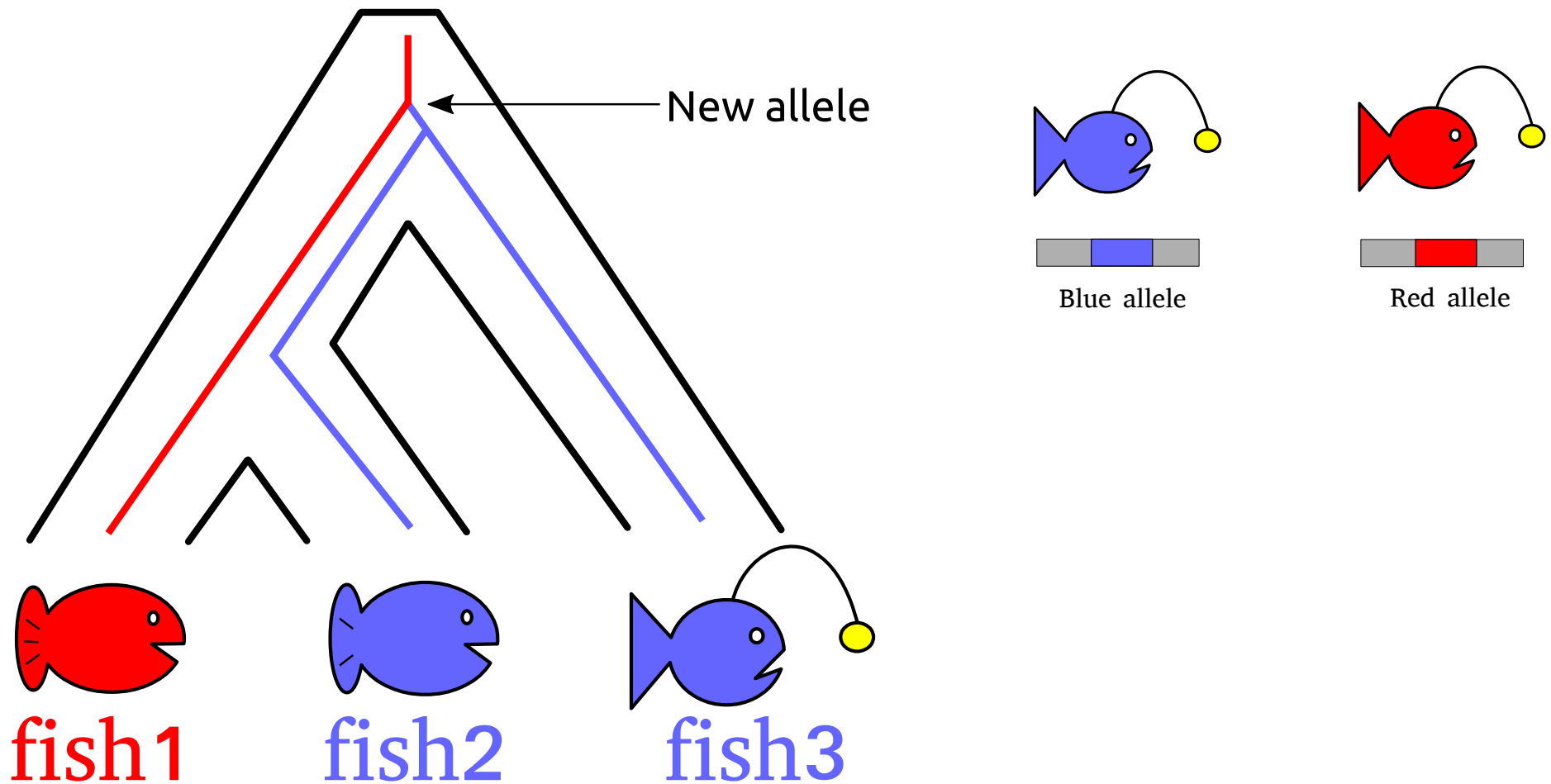
# Incomplete lineage fictional example



# Incomplete lineage fictional example

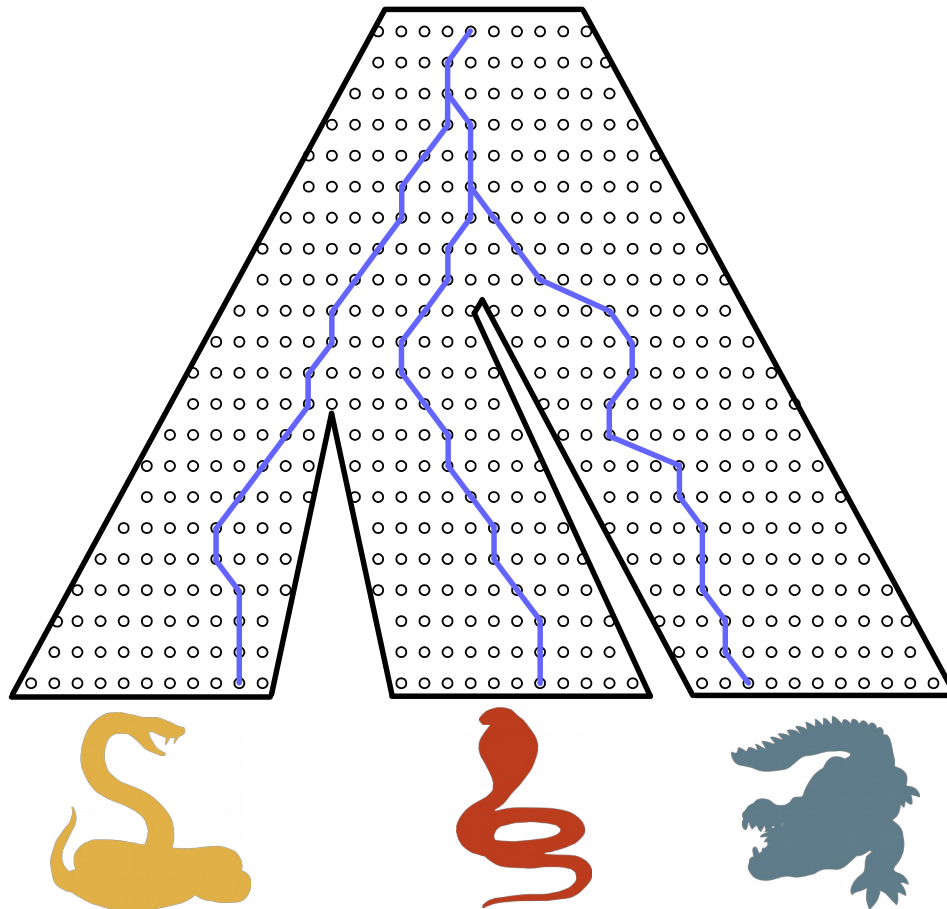


# Incomplete lineage fictional example



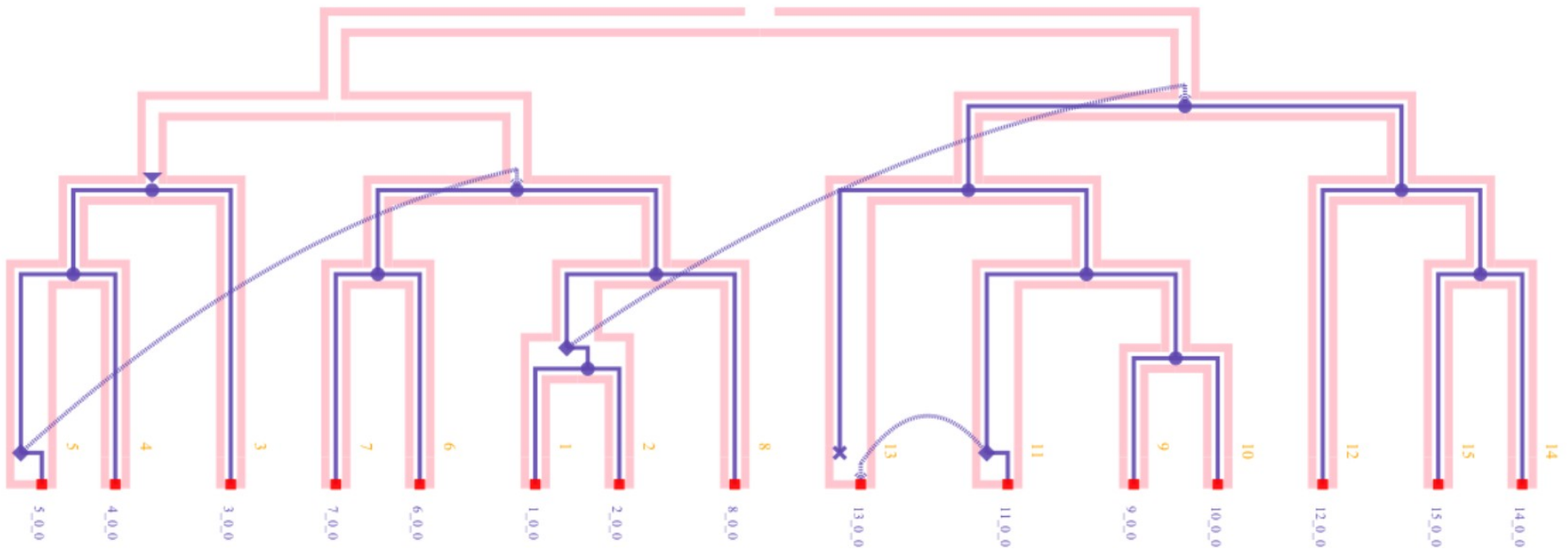
# ILS and reconciliation

- Reconciliation models assume no ILS
- What happens when ILS occurs nonetheless?



# ILS and reconciliation

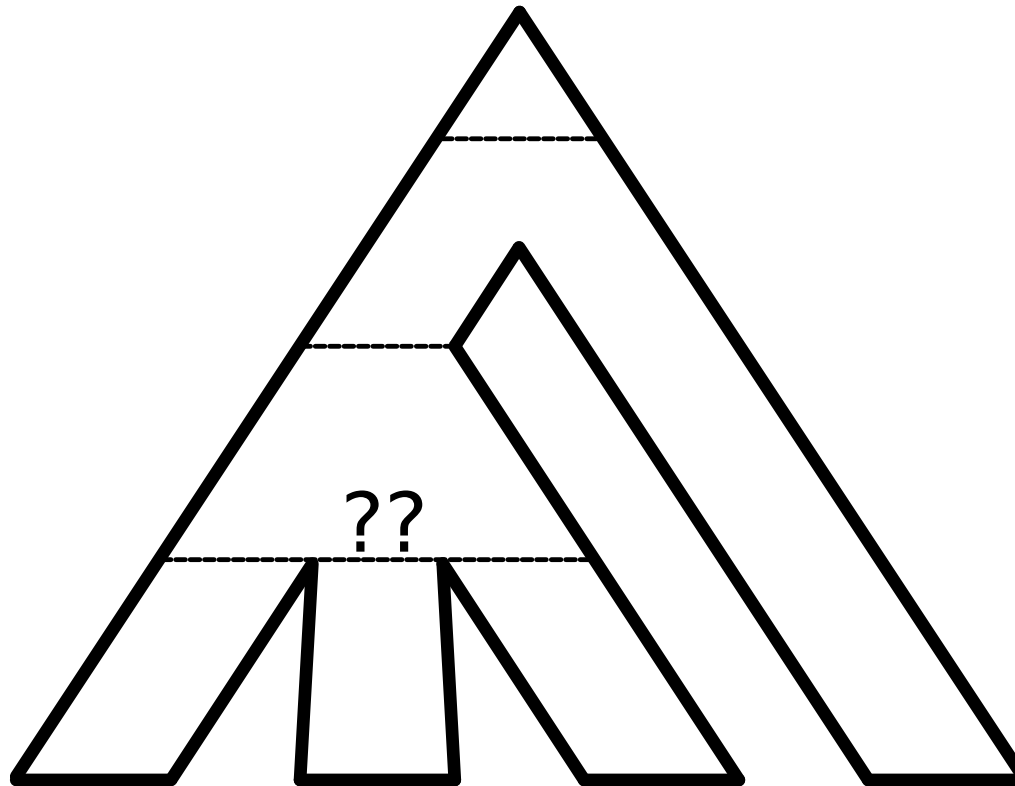
- Reconciliation models assume no ILS
- What happens when ILS occurs nonetheless?



Spurious horizontal gene transfers...

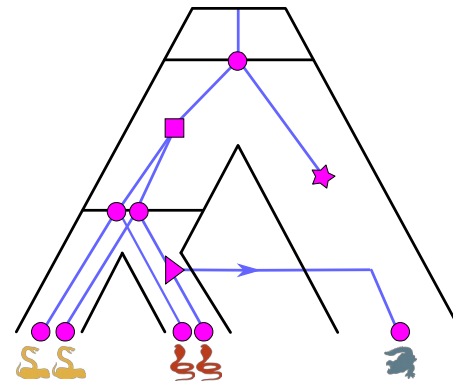
# Species tree uncertainty

- Instead of using a fixed unreliable tree
- Reconcile with a distribution of plausible species trees



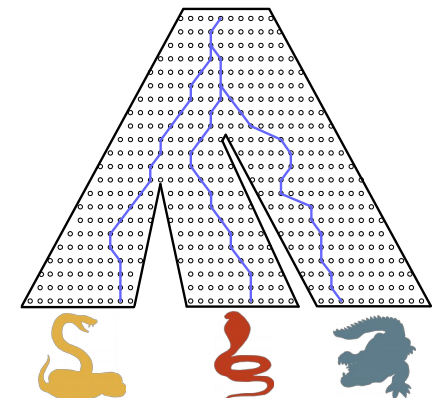
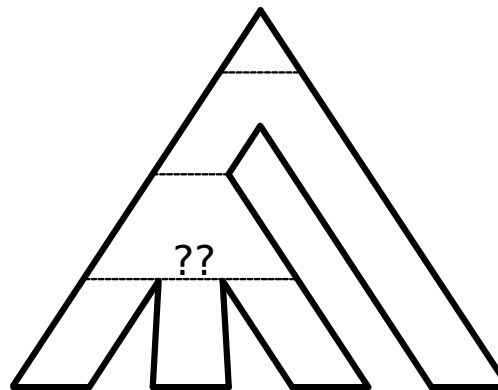
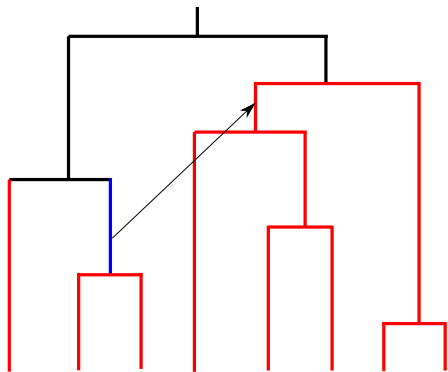
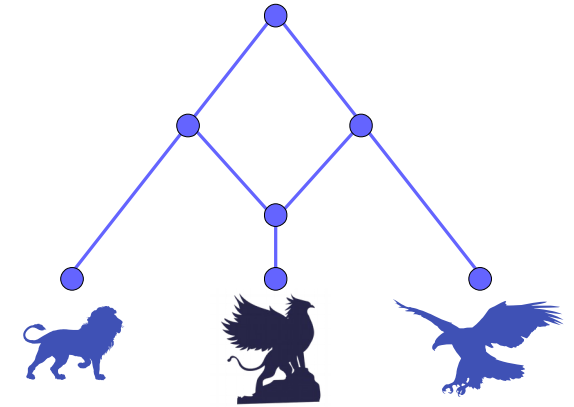
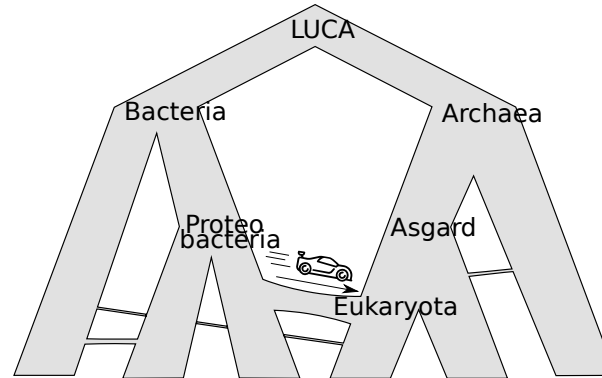
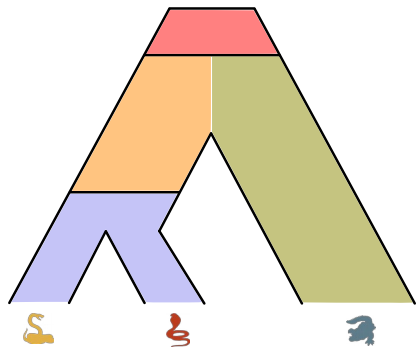
# Conclusion

- We have reconciliation methods that handle:
  - gene duplications, losses, and transfers
  - gene tree uncertainty
- ... but our models are too simple
- ... interesting computational challenges apply more complex models to large datasets



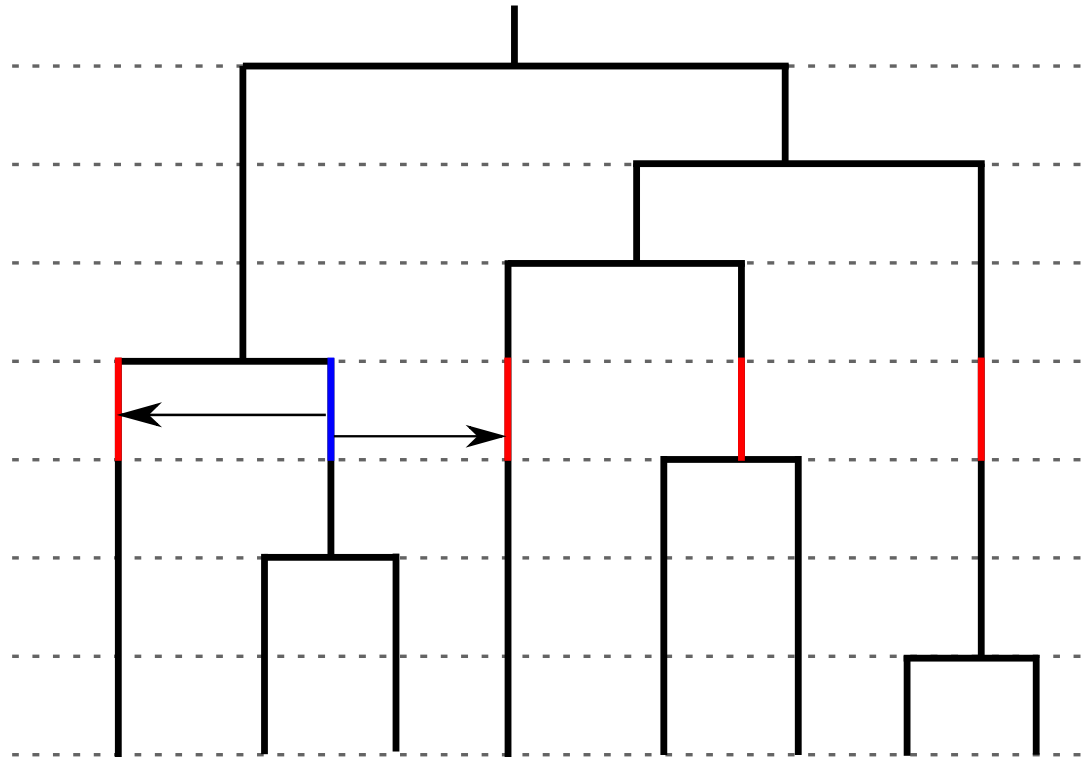


# Thank you!



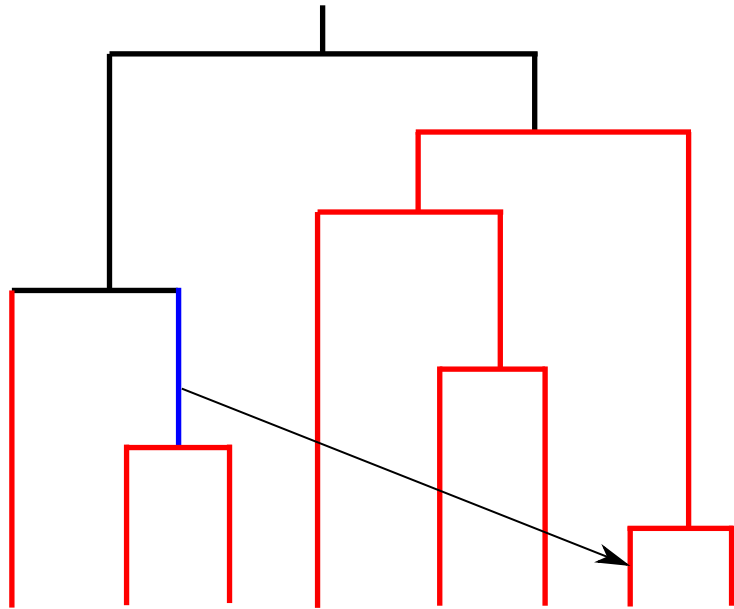
contact: [benoit.morel@h-its.org](mailto:benoit.morel@h-its.org)

# Time constraints on the transfers



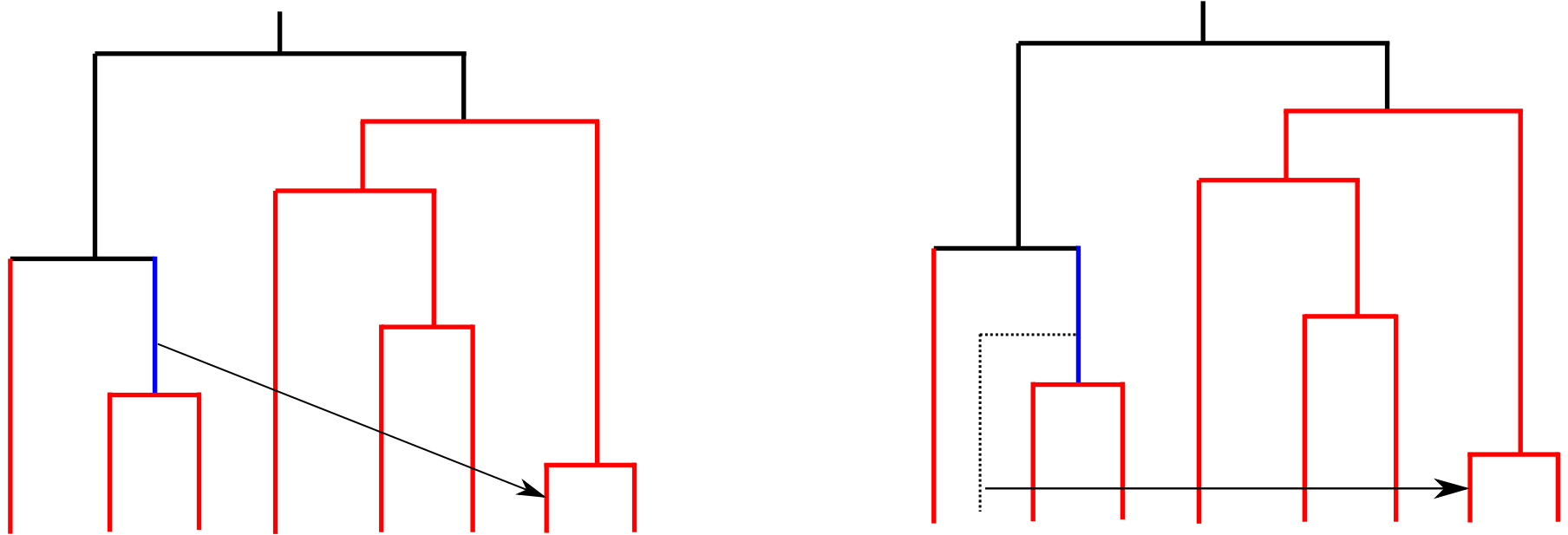
DatedDTL model  
(very slow likelihood computation)

# Time constraints on the transfers



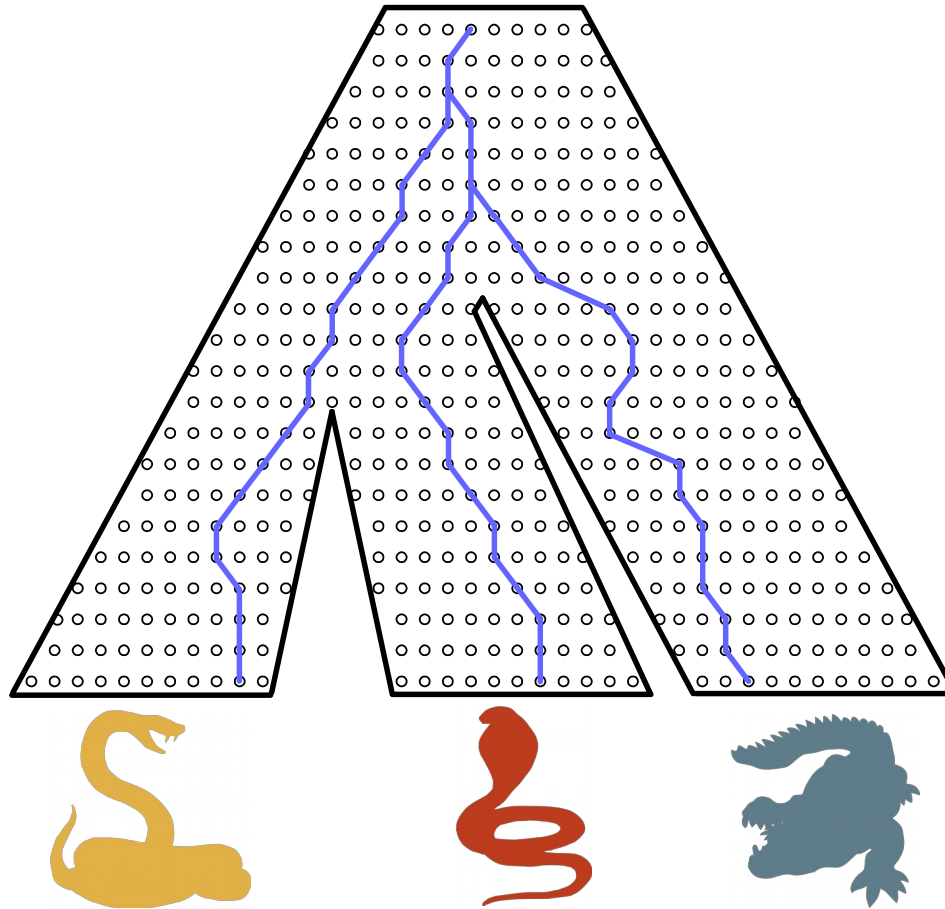
RelatedDTL model  
Allows transfers to the future

# Time constraints on the transfers



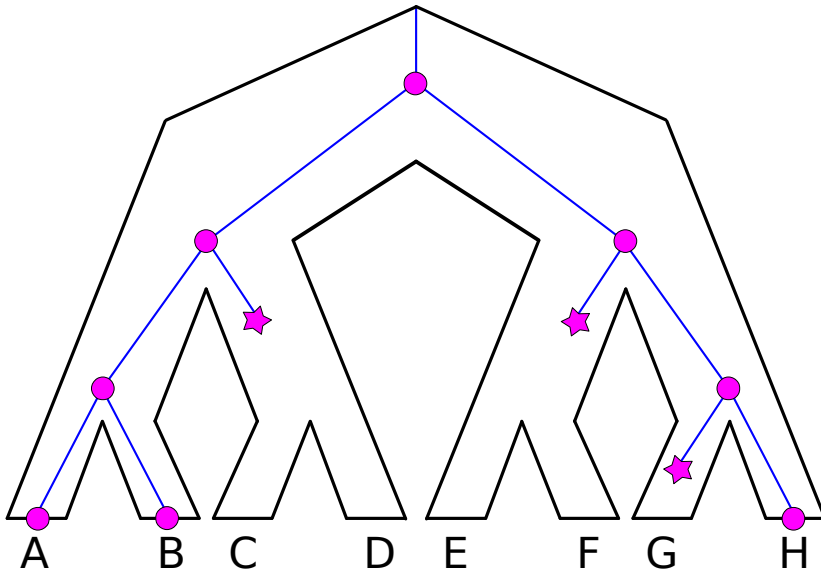
RelatedDTL model  
Allows transfers to the future

# Incomplete lineage sorting

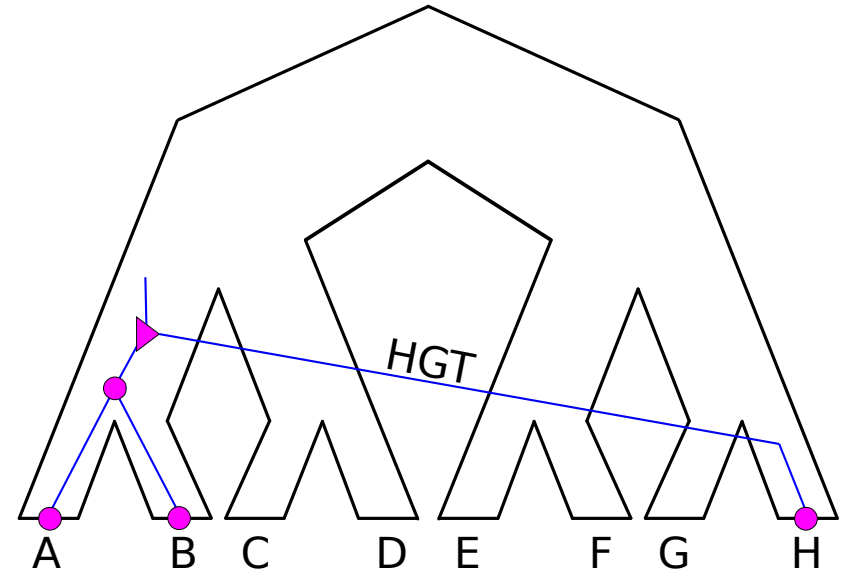


# Model parameter heterogeneity

- Crucial to assess competing scenarios



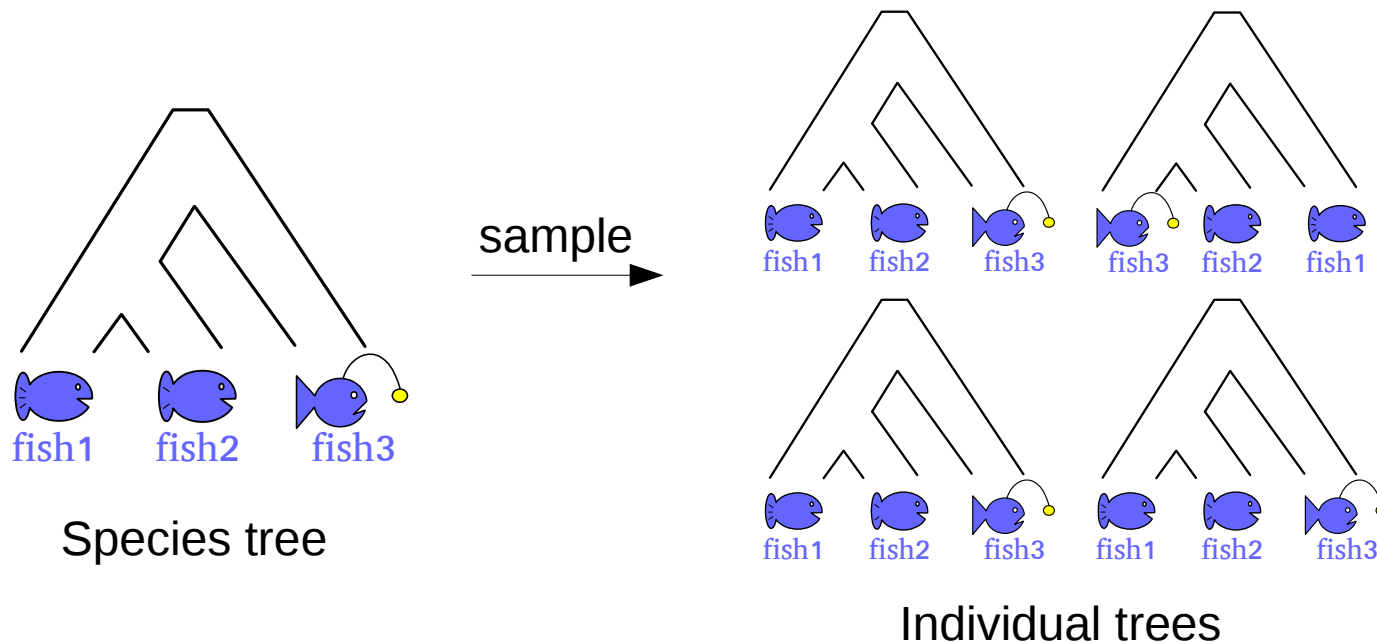
Scenario with 3 losses



Scenario with 1 HGT

# Solution

- Sample “individual trees” under the multi-species coalescent model
- Reconcile the distribution of gene trees with the distribution of species trees



# Species tree uncertainty

