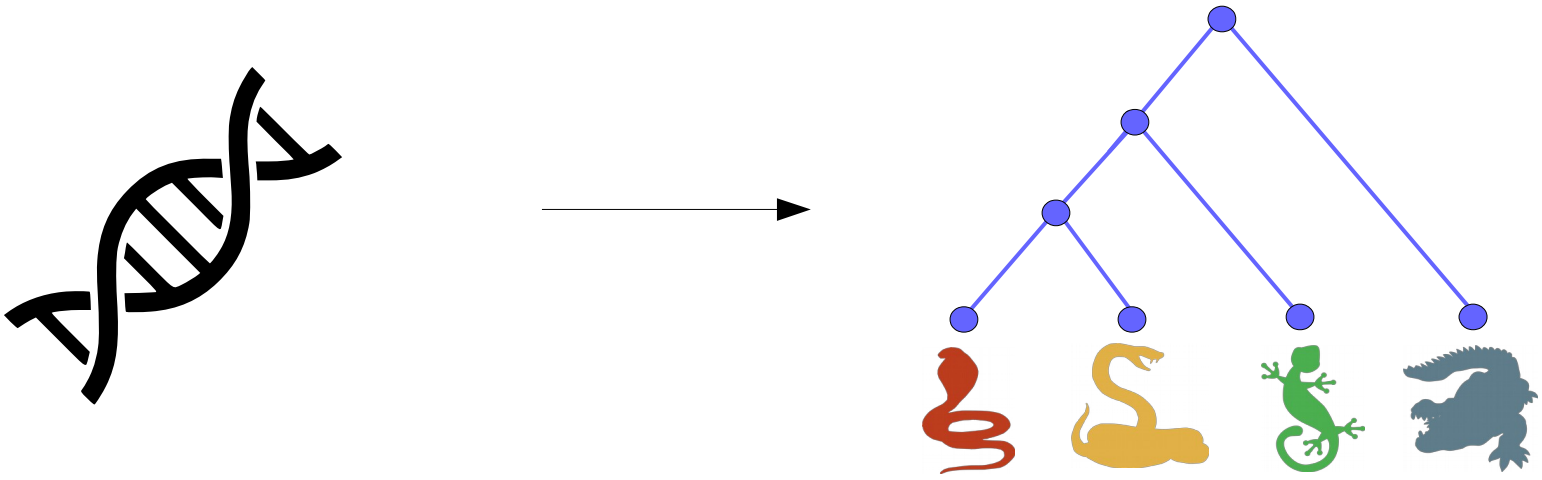
Species tree inference from genome-scale datasets

*ITA*PHY WORKSHOP, June 8 2023*

Benoit Morel (benoit.morel@h-its.org)



About me

- Postdoc at the Heidelberg Institute for Theoretical Studies (Germany)

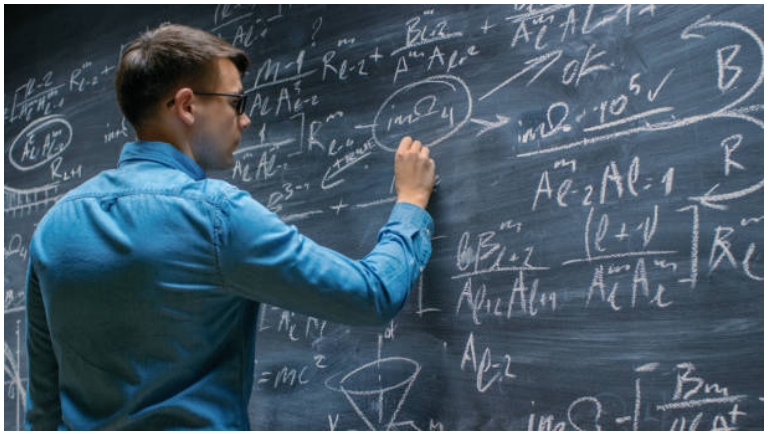
About me

- Postdoc at the Heidelberg Institute for Theoretical Studies (Germany)
- I am a computer scientist

About me

- Postdoc at the Heidelberg Institute for Theoretical Studies (Germany)
- I am a computer scientist
- But don’t be afraid

How do we work?



Mathematician

Software developer

Biologist

You said

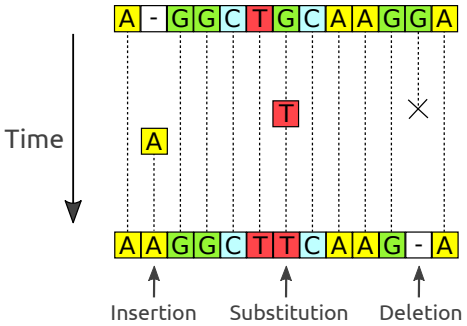
- “I am not confident with phylogenetics”
- “I don’t know which tool to use”
- “Evolution is a mess”

Today

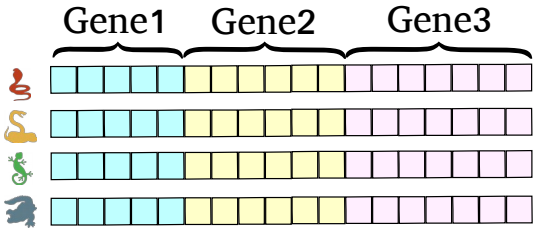
- We will describe different tree inference methods and their limitations

- Please ask (stupid) questions!

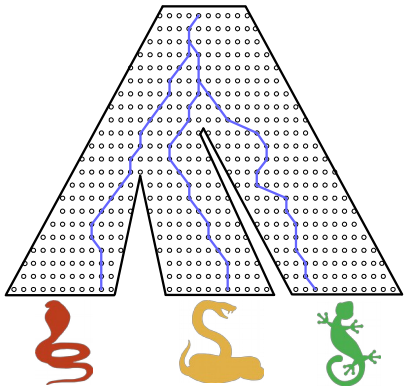
Structure of the lecture



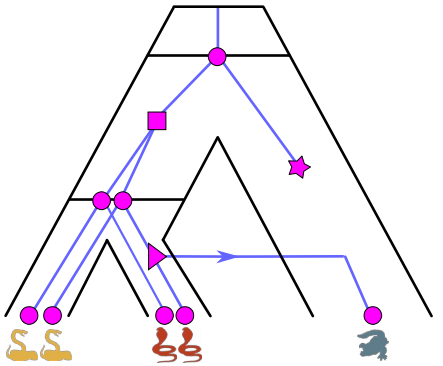
Single-gene



Multiple genes

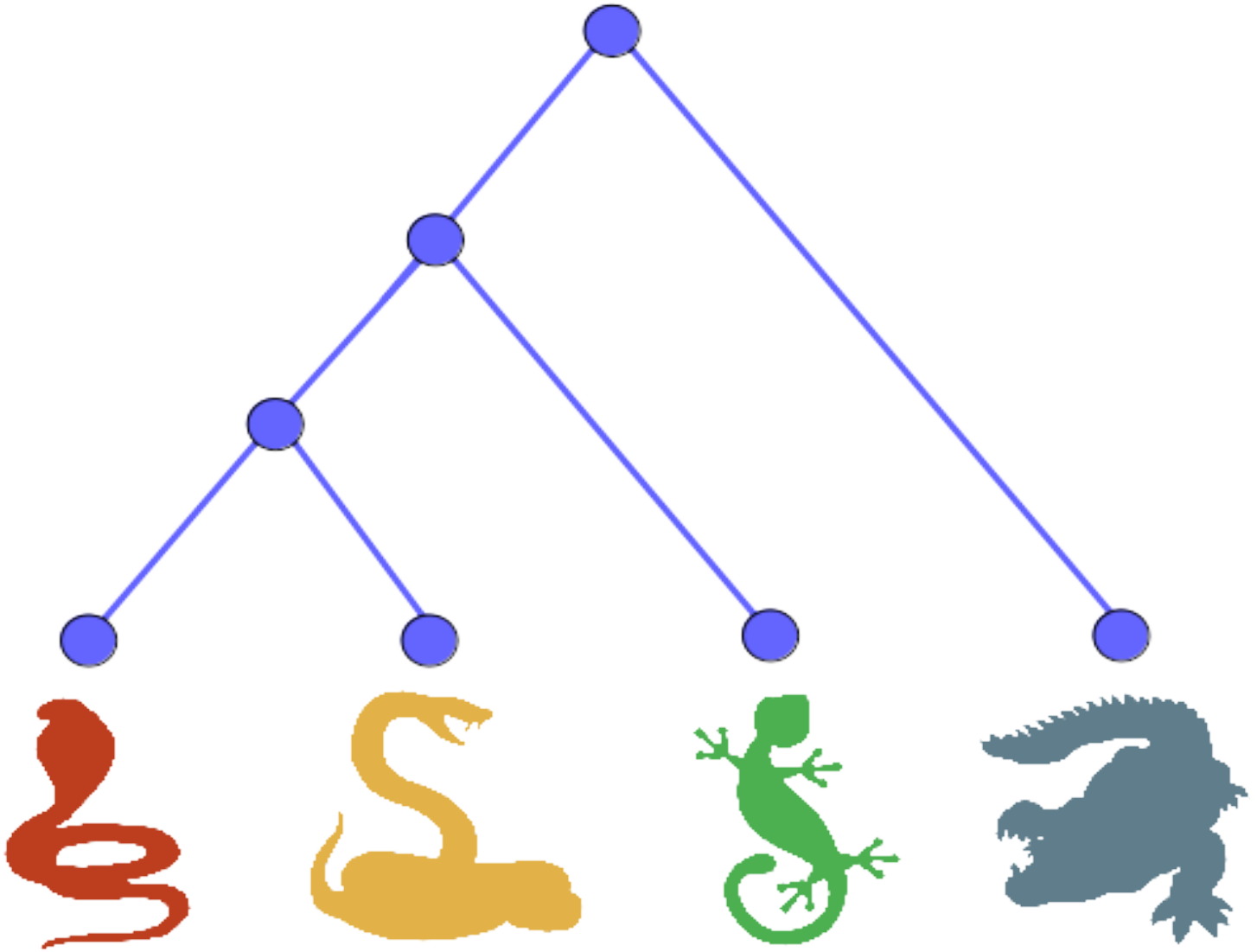


Gene conflicts

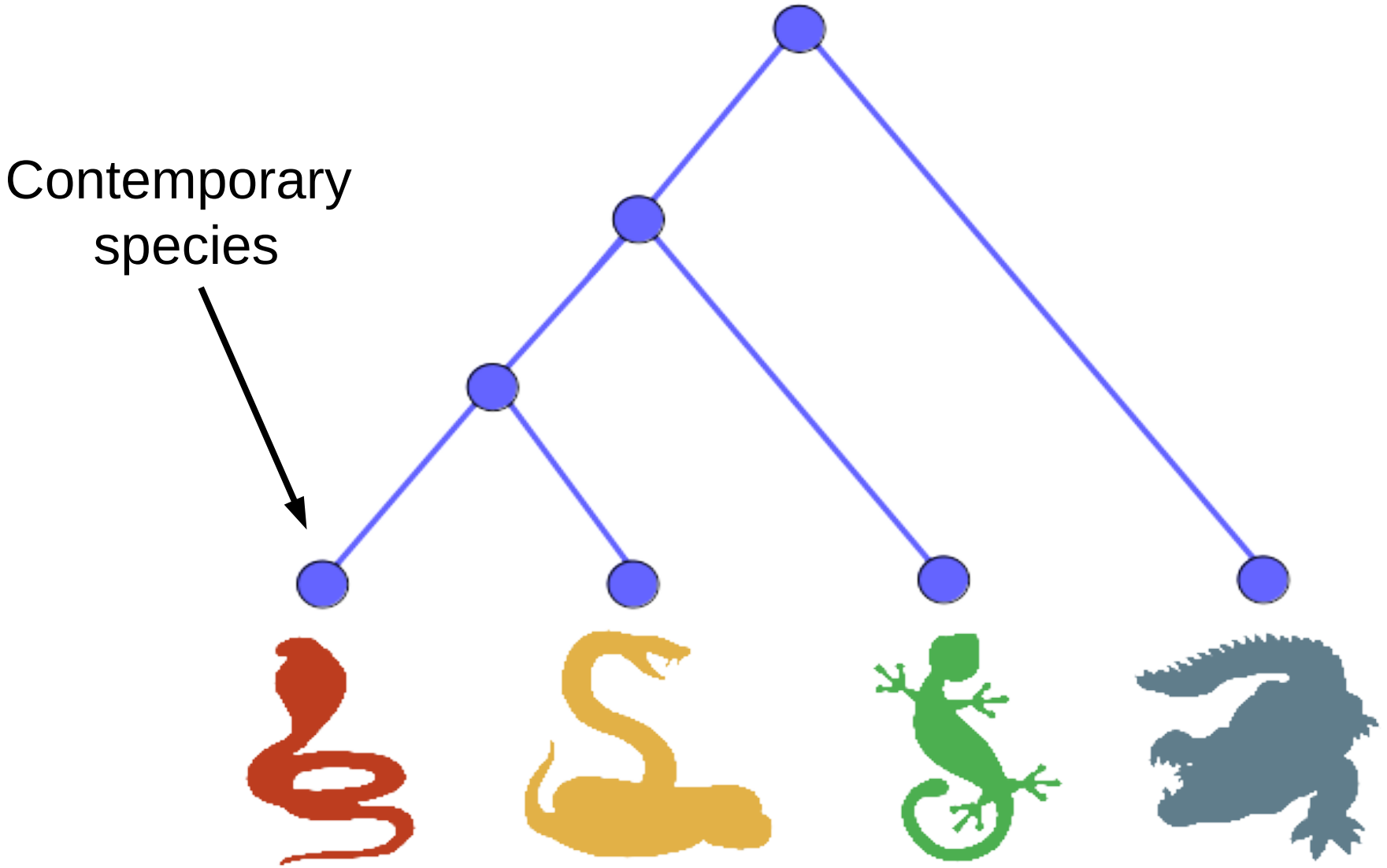


Multi-copy families

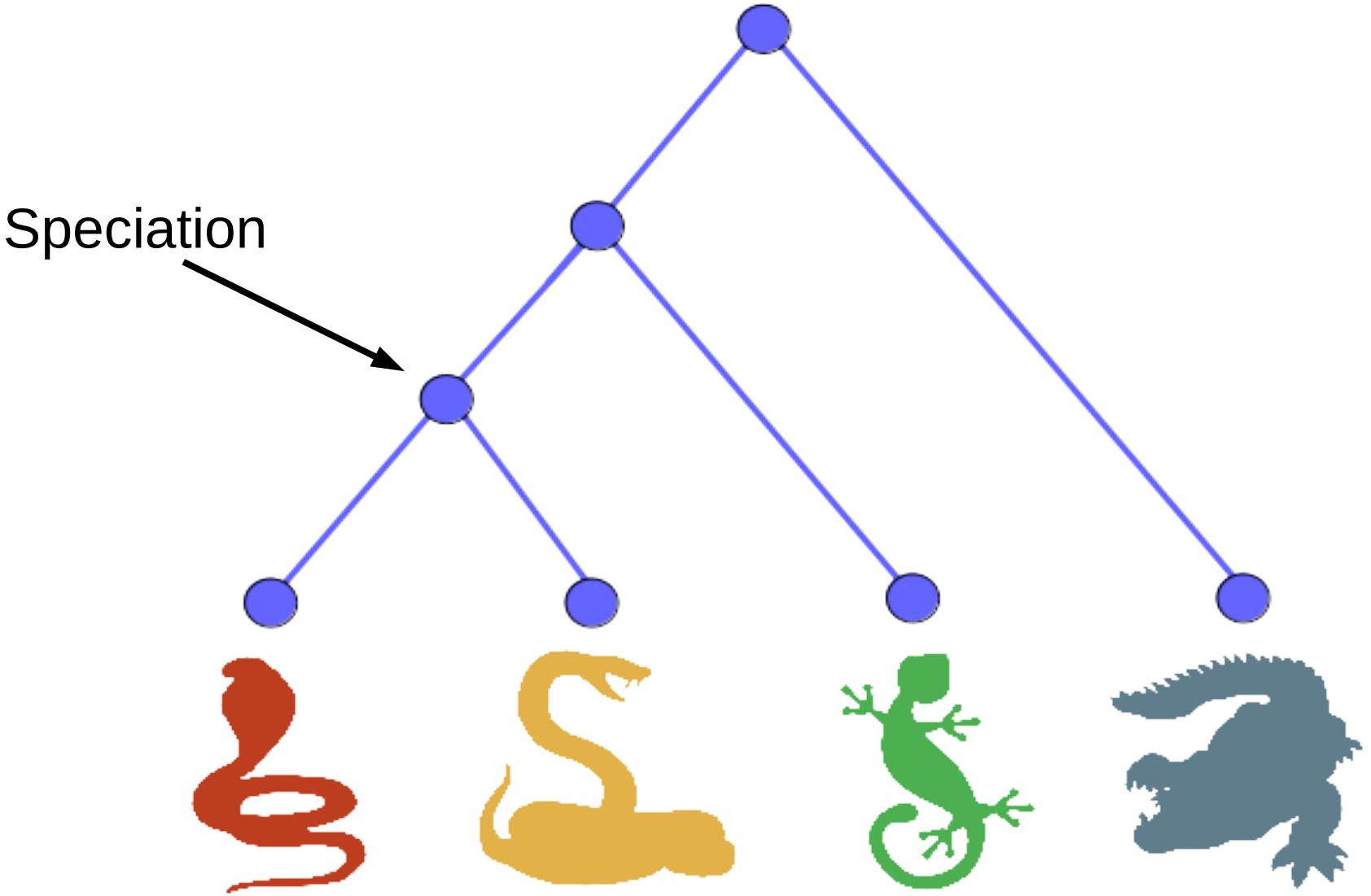
Species tree



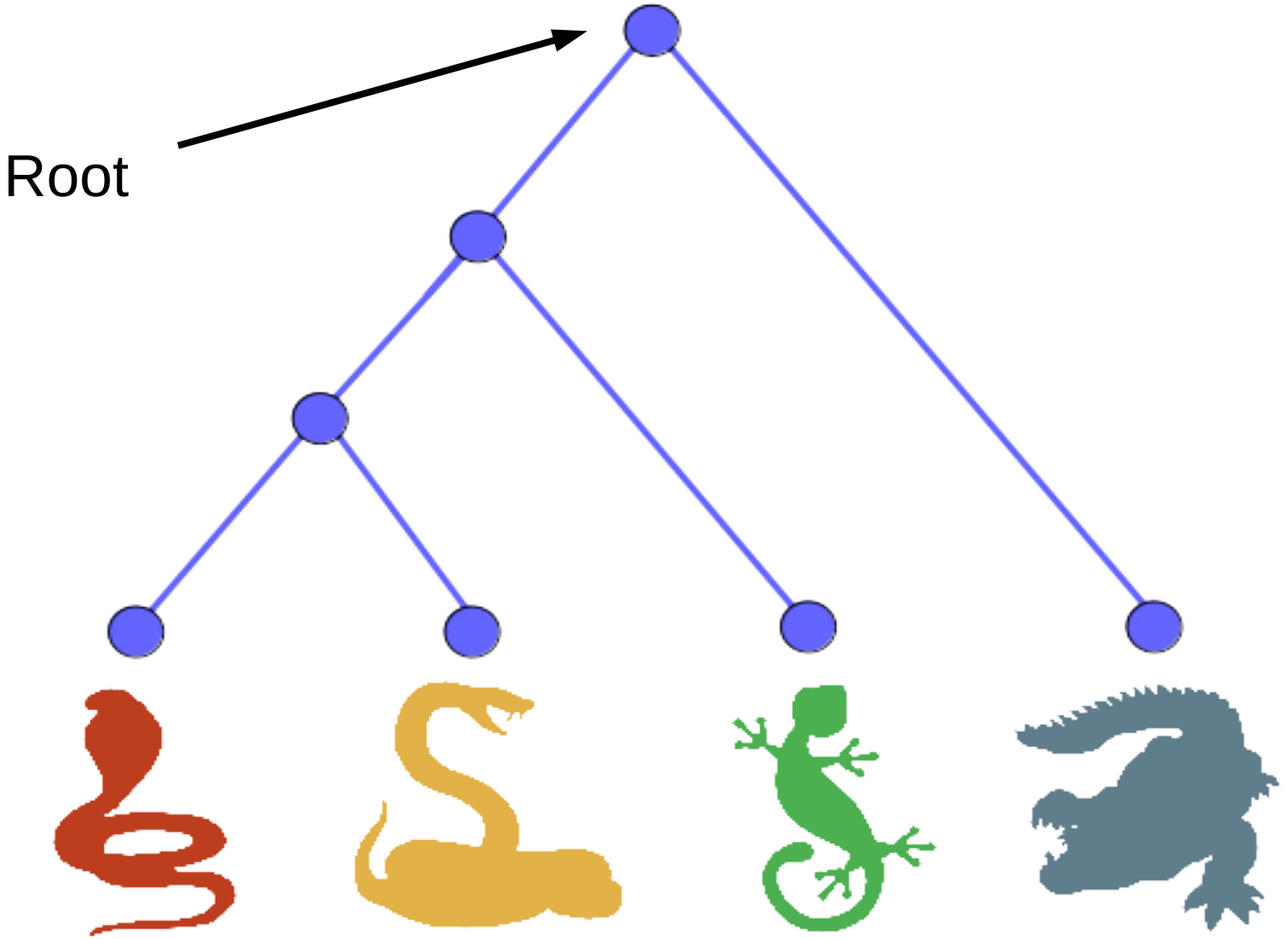
Species tree



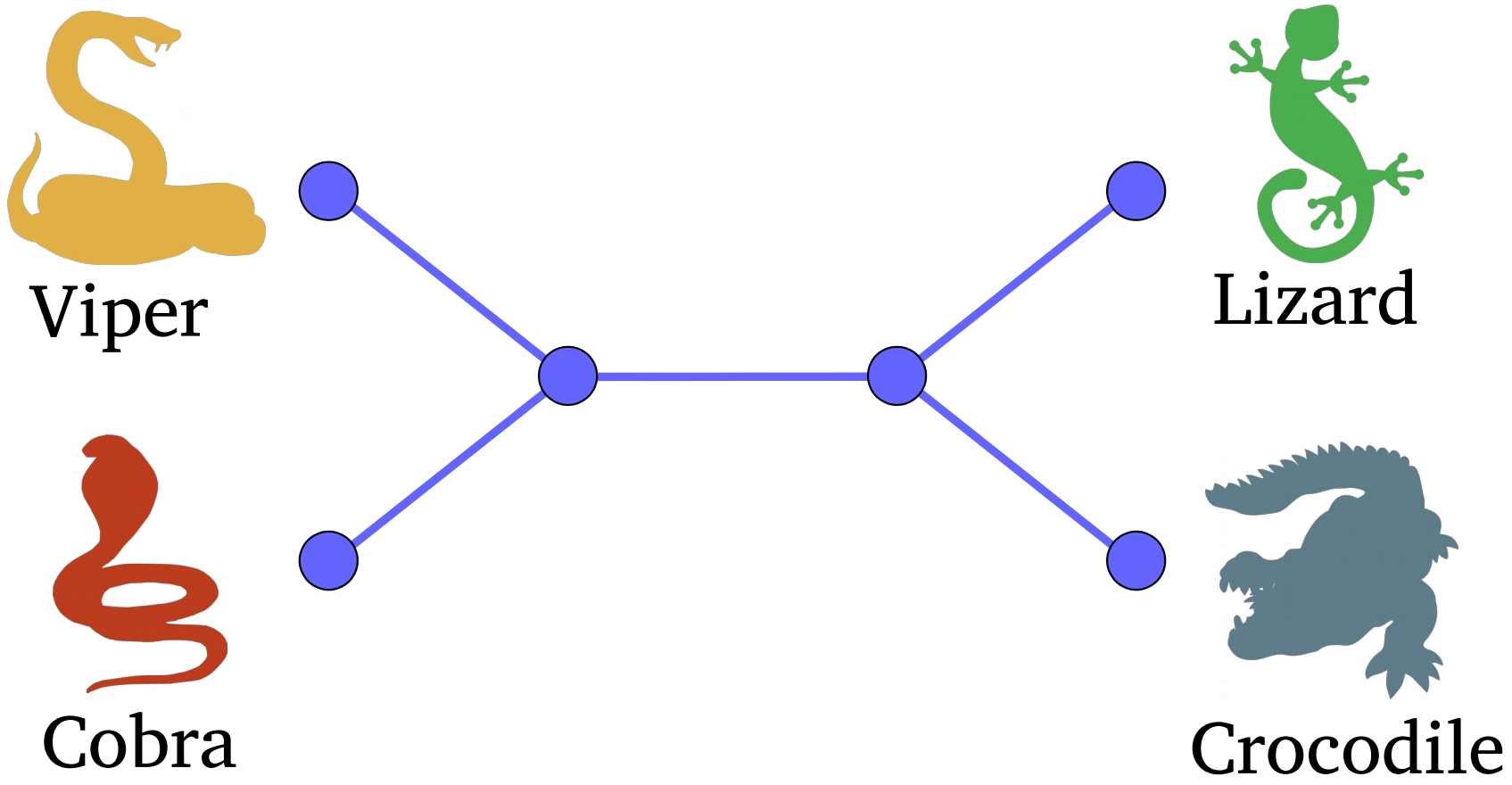
Species tree



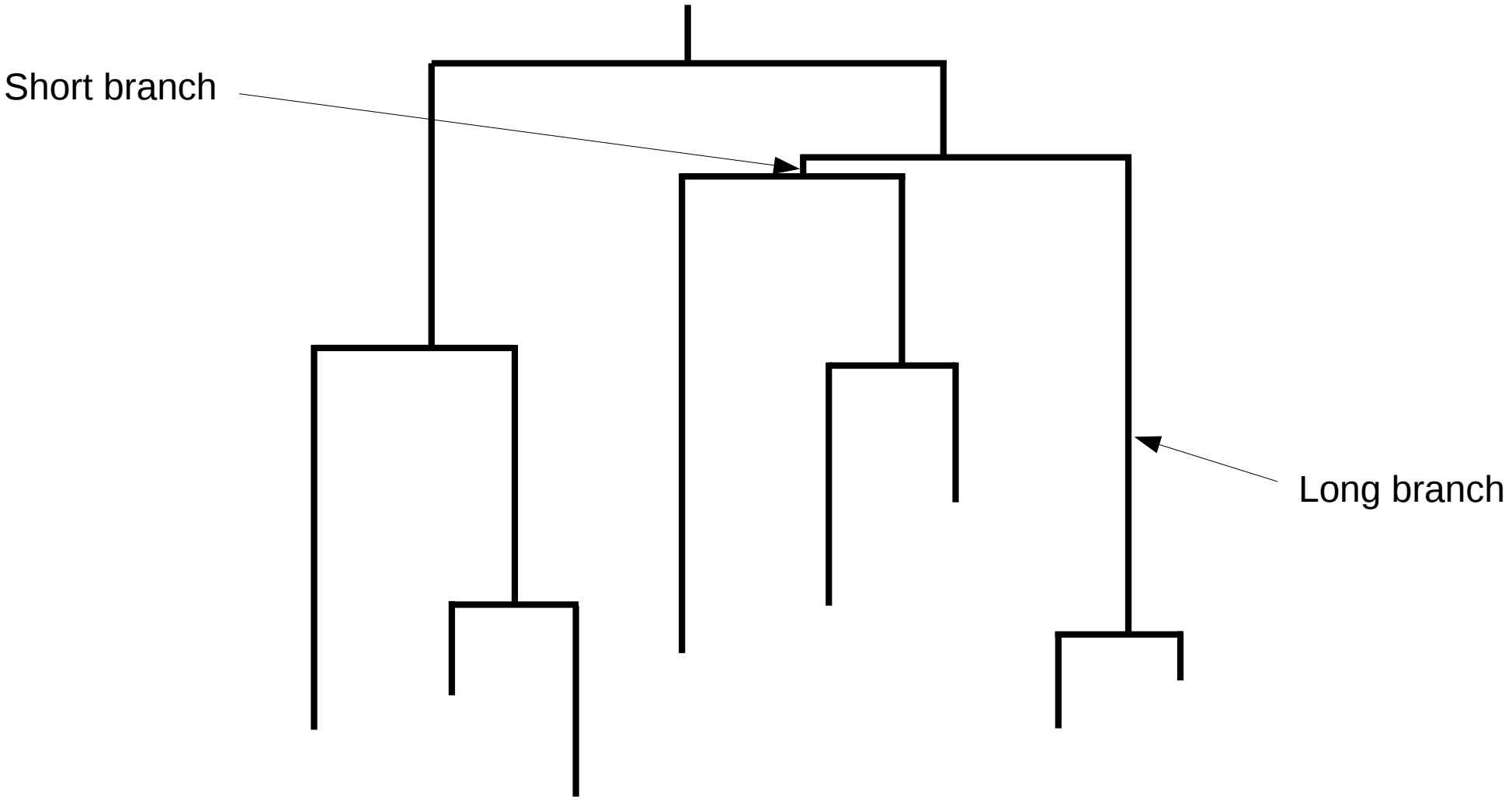
Species tree



Unrooted species tree



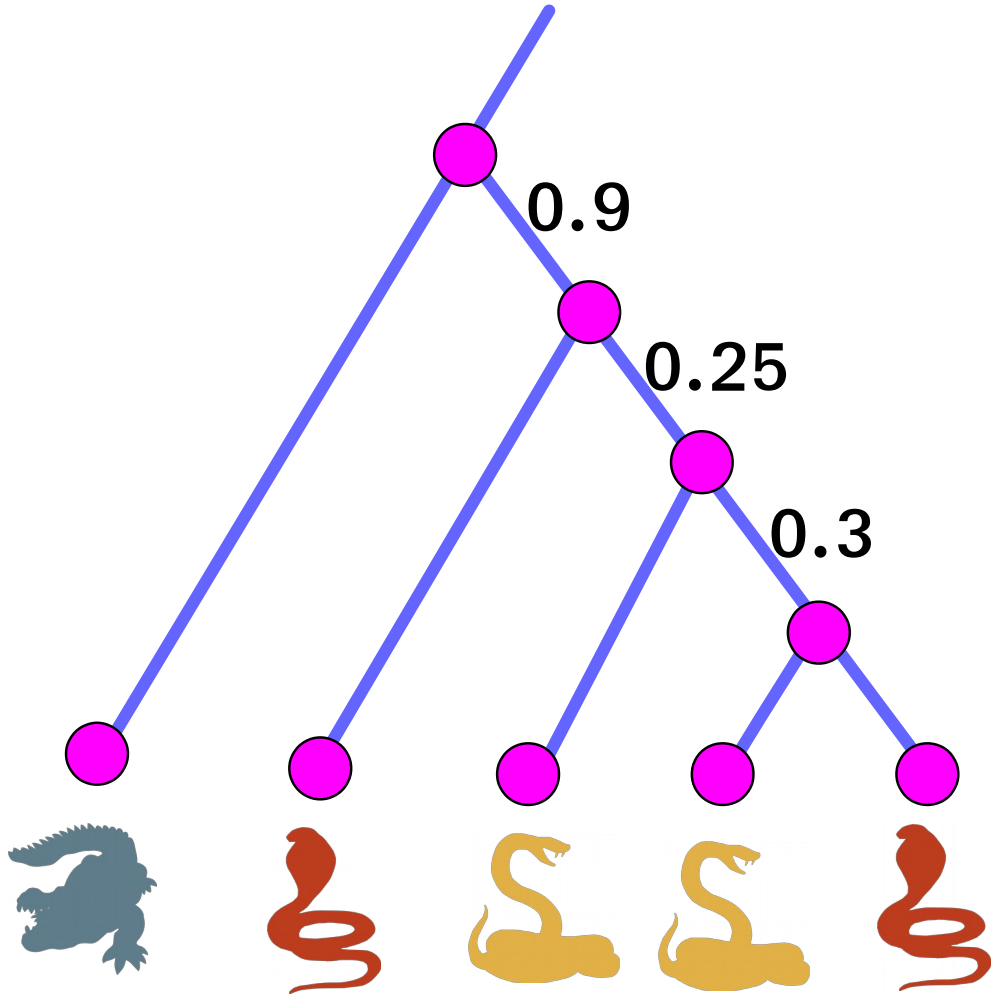
Branch lengths



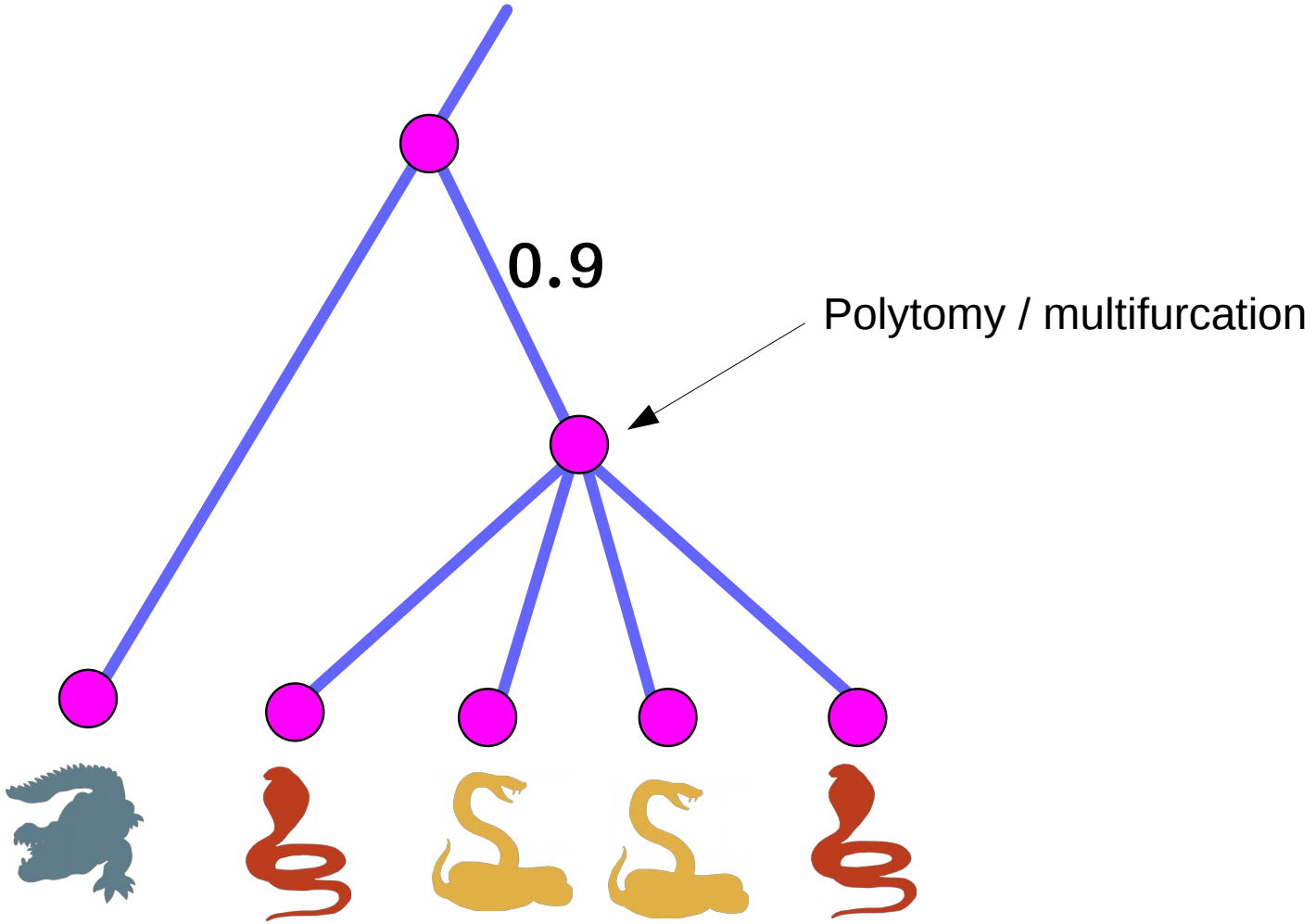
Branch lengths

- Different units:
  - Average number of substitutions per site
  - Coalescent units (number of generations / effective population size)
  - Time
  - …

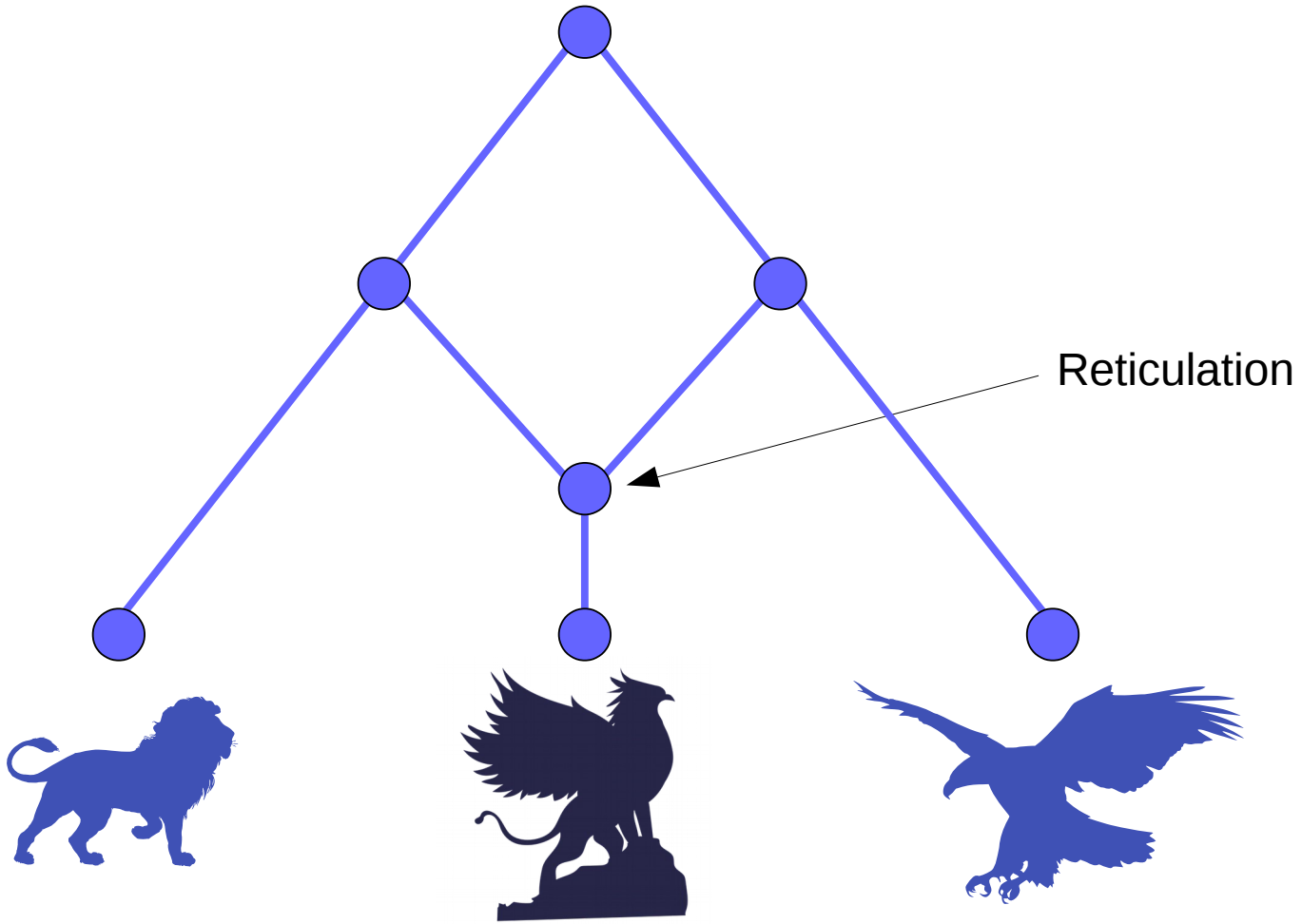
Support values: measure of confidence



Polytomies: represent uncertainty

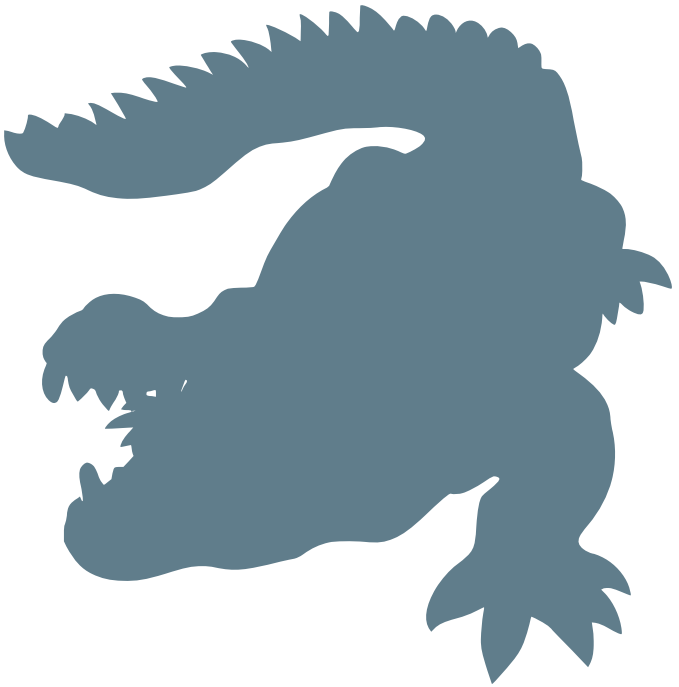


Species networks: nodes can have two parents



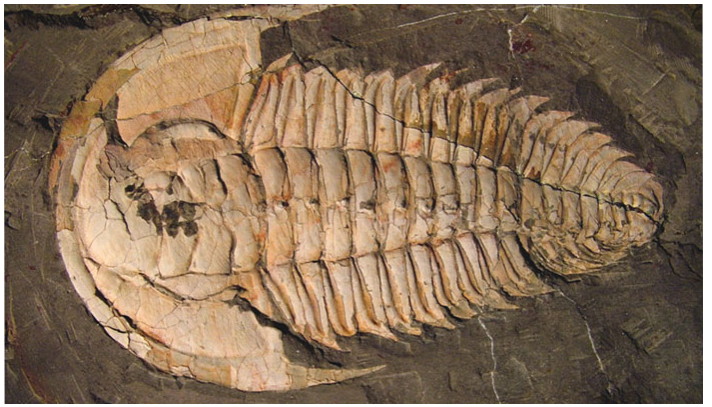
What do we have?

- Contemporary species that we can observe and sequence



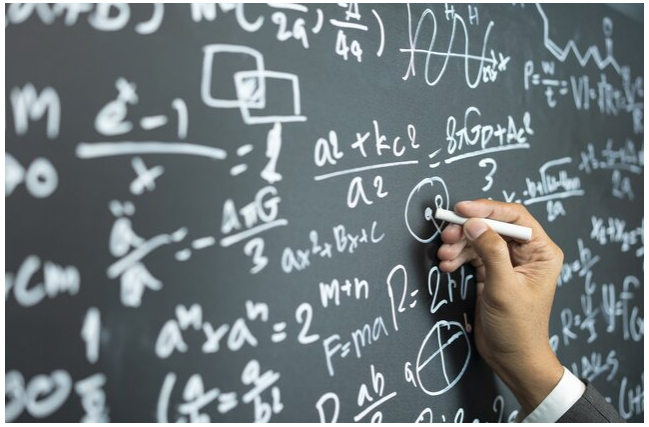
What do we have?

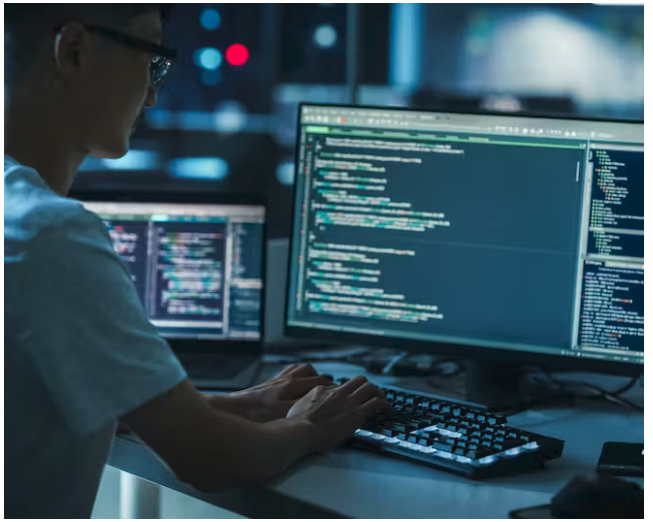
- Contemporary species that we can observe and sequence
- Fossils



What do we have?

- Contemporary species that we can observe and sequence
- Fossils
- Great mathematical models and software





What do we have?

- Contemporary species that we can observe and sequence
- Fossils
- Great mathematical models and software
- Prior knowledge about the problem

What do we have?

- Contemporary species that we can observe and sequence
- Fossils
- Great mathematical models and software
- Prior ~~knowledge~~ **beliefs** about the problem

What do we have?

- Contemporary species that we can observe and sequence
- Fossils
- Great mathematical models and software
- Prior ~~knowledge~~ **beliefs** about the problem
- Personal expectations

In this lecture

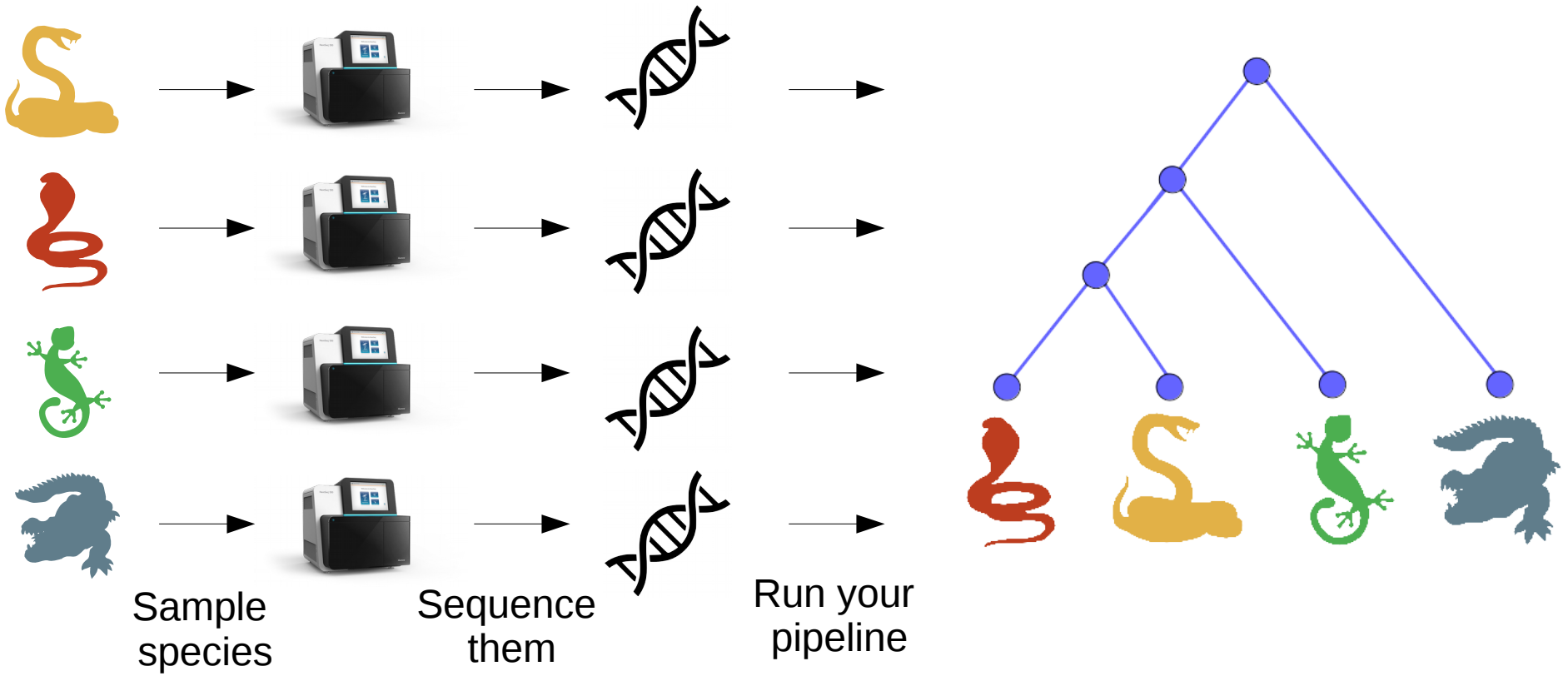
- Species tree inference methods
  - Parsimony, distance, max. likelihood, Bayesian
  - Concatenation method
  - Single-copy gene tree methods
  - Multi-copy gene tree methods

- Models of sequence and gene evolution

Disclaimer

- There is no “best inference method”
- It depends on the data
- Different experts have different opinions on the matter
- We won’t cover all the excellent methods

Tree inference from molecular data



From a single gene

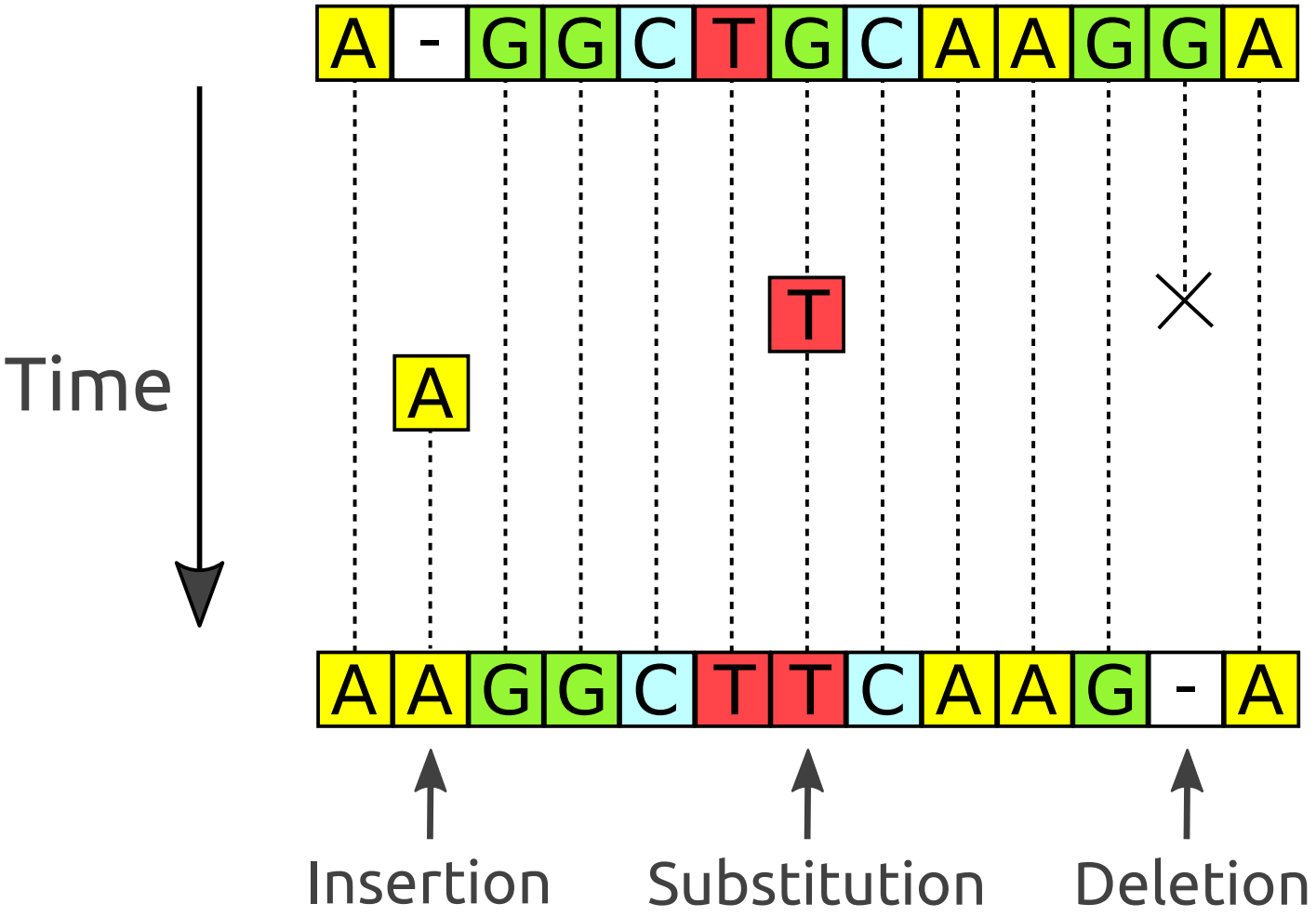
- Select a specific gene (e.g. 16s)
- Sequence it from each species
- Align the sequences
- Infer a tree

Homologous genes

- Genes that evolved from the same ancestral sequence

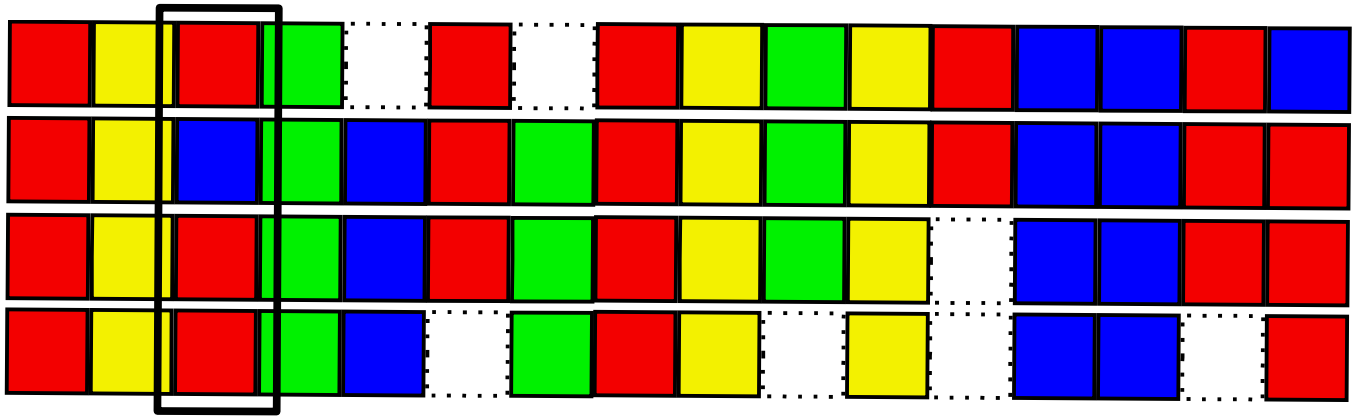
- Their evolution can be described with a tree

Sequence evolution



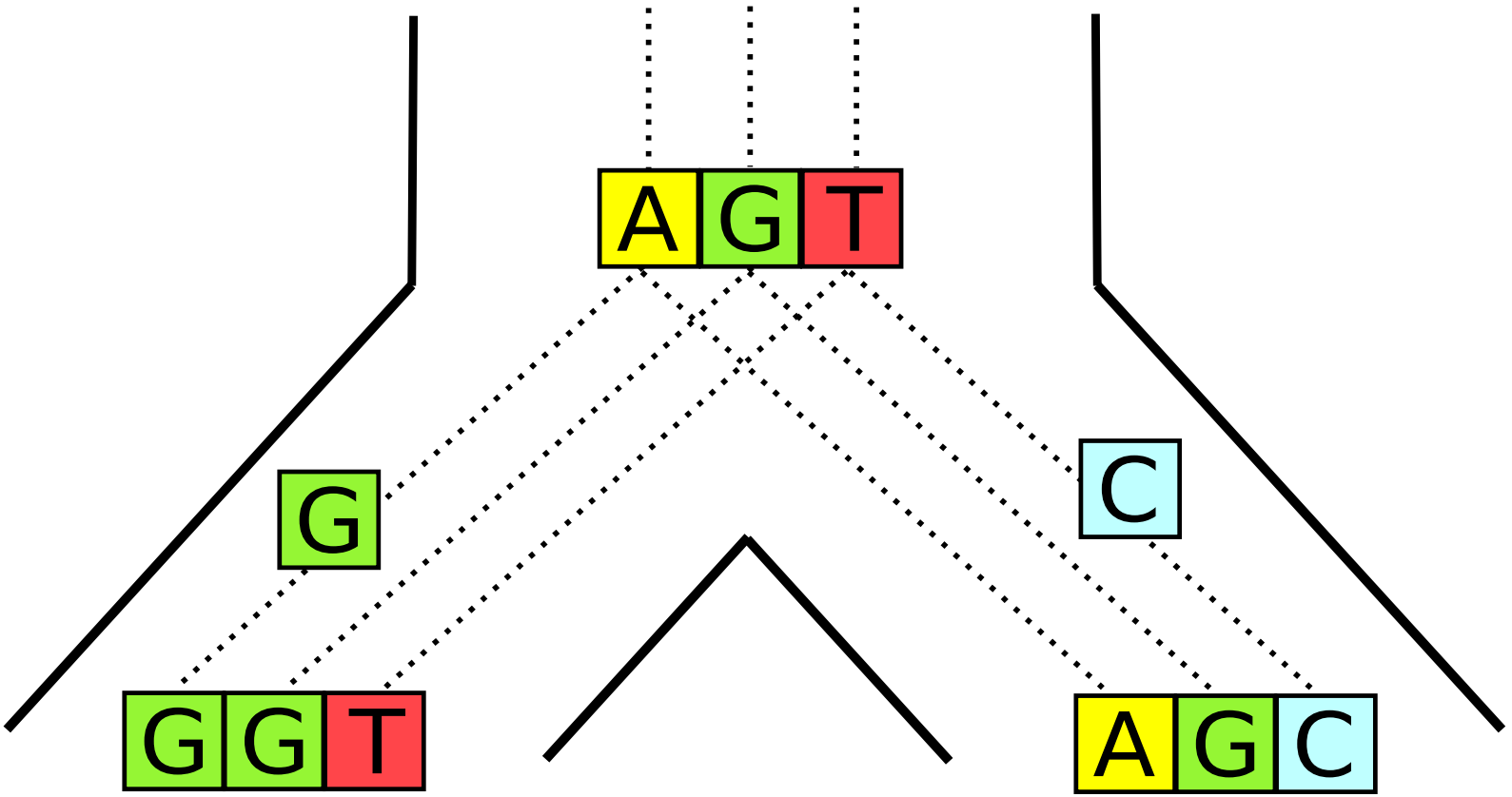
Sequence alignment

- Insertions and deletions are difficult to model
- Instead we align the sequences
- We assume that each column is homologous (evolved from a common ancestral state)



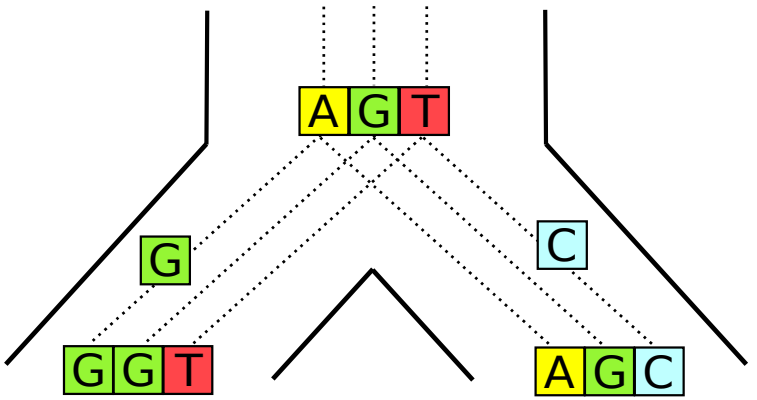
Homologous site

Sequence evolution and trees



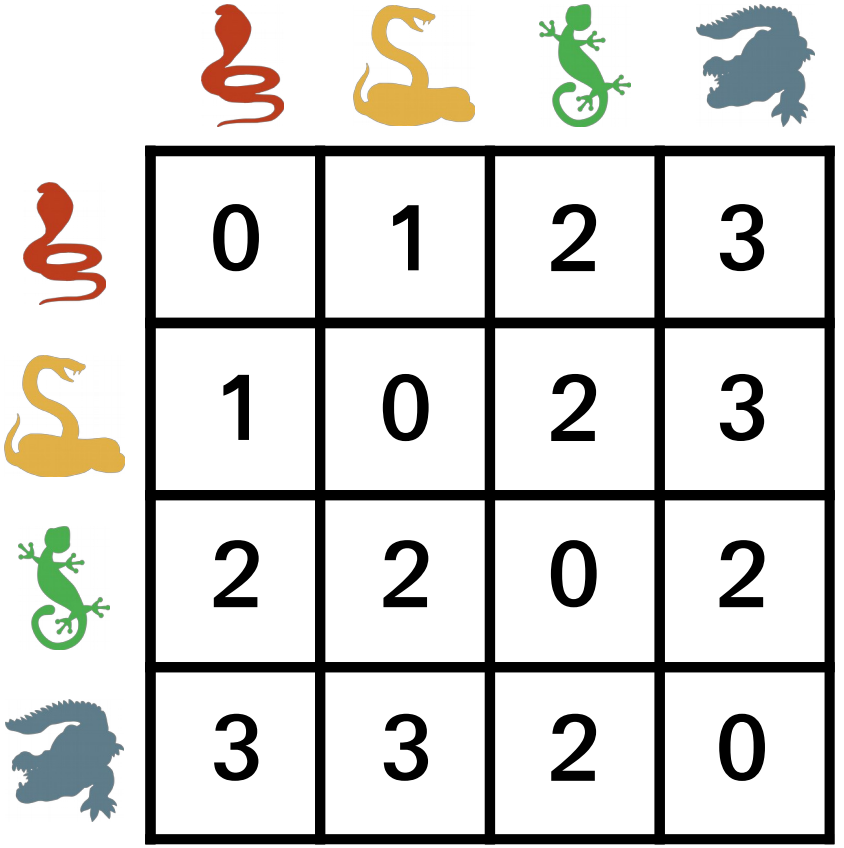
Distance methods

- After a split in the tree, sequences diverge
- More recent split → more similar sequences
- Similar sequences should be grouped together in the tree



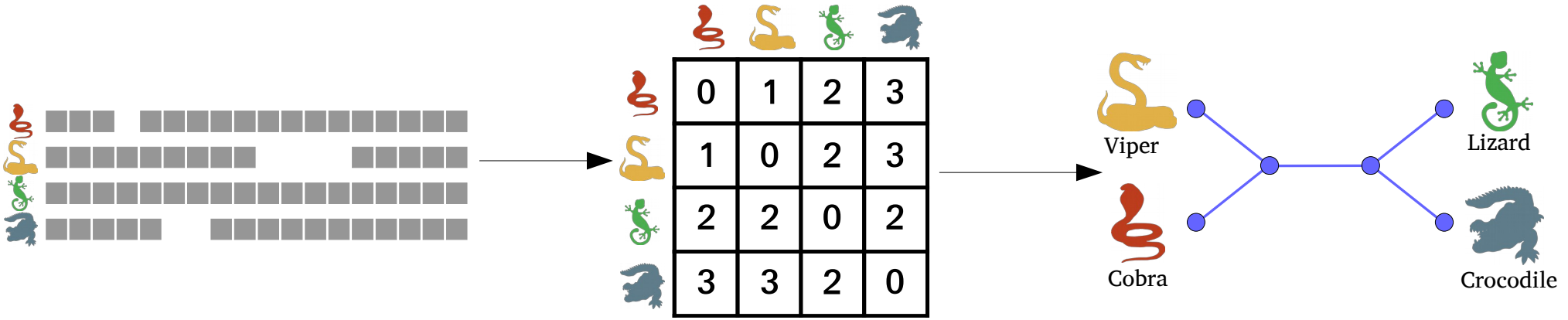
Distance matrix

Pairwise distances



	0	1	2	3
	1	0	2	3
	2	2	0	2
	3	3	2	0

Distance methods

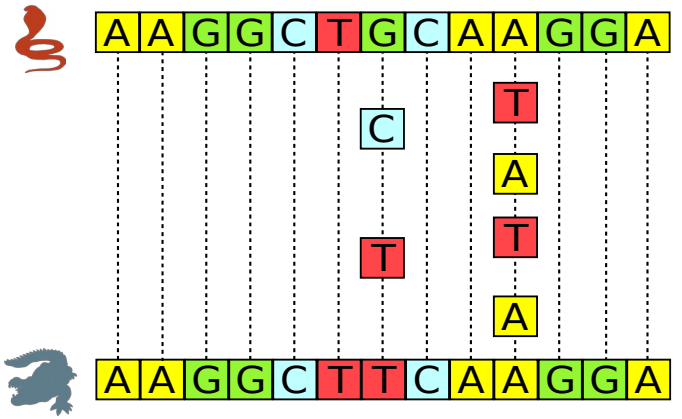


Neighbor joining
FastMe
UPGMA

Distance methods

- Advantages:
  - Fast!

- Weaknesses:
  - Pairwise distances can saturate for distant species
  - Does not account well for multiple hits



Occam’s razor / parsimony principle

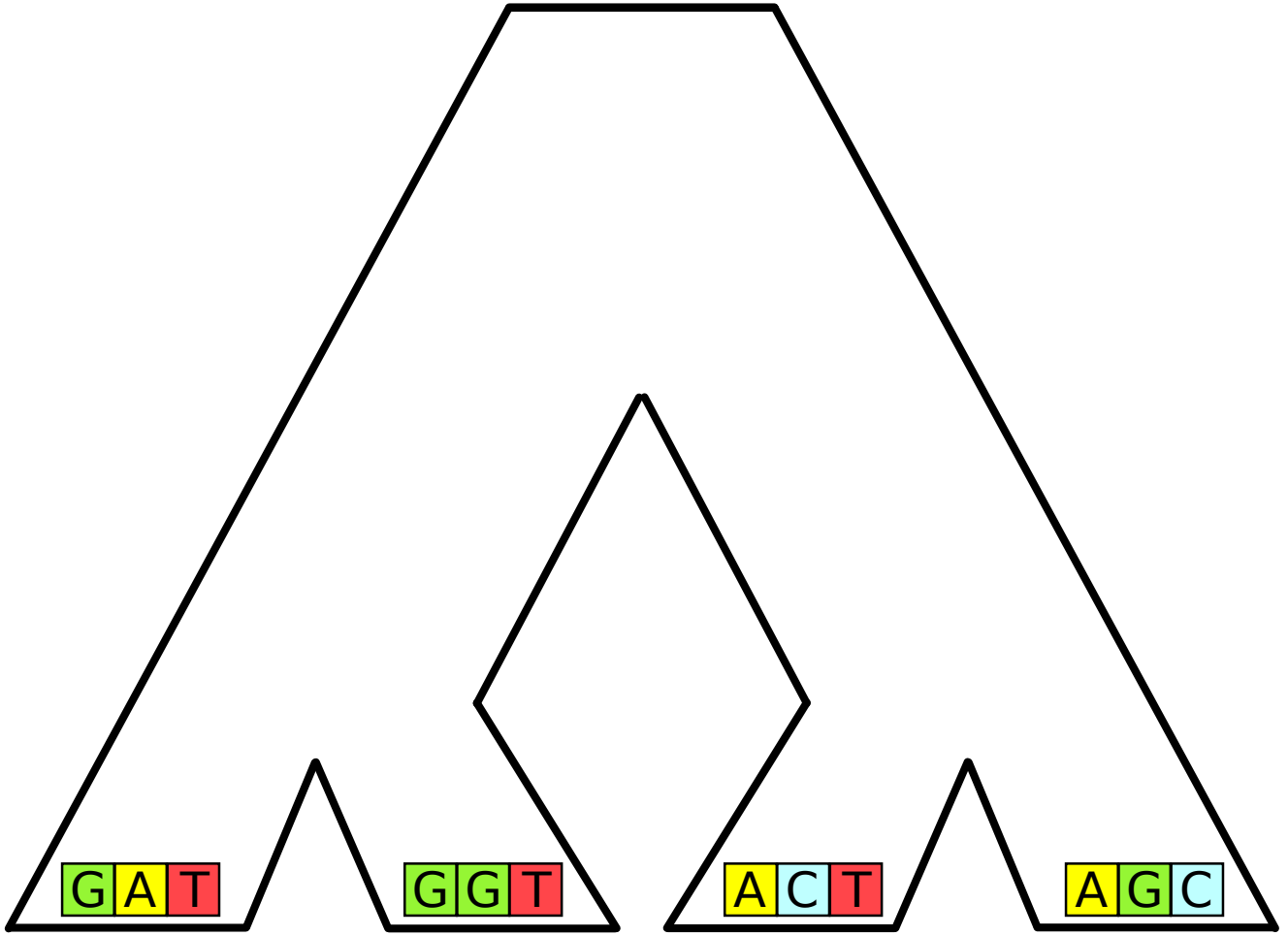
“*The simplest explanation is usually the best one*”

Parsimony methods

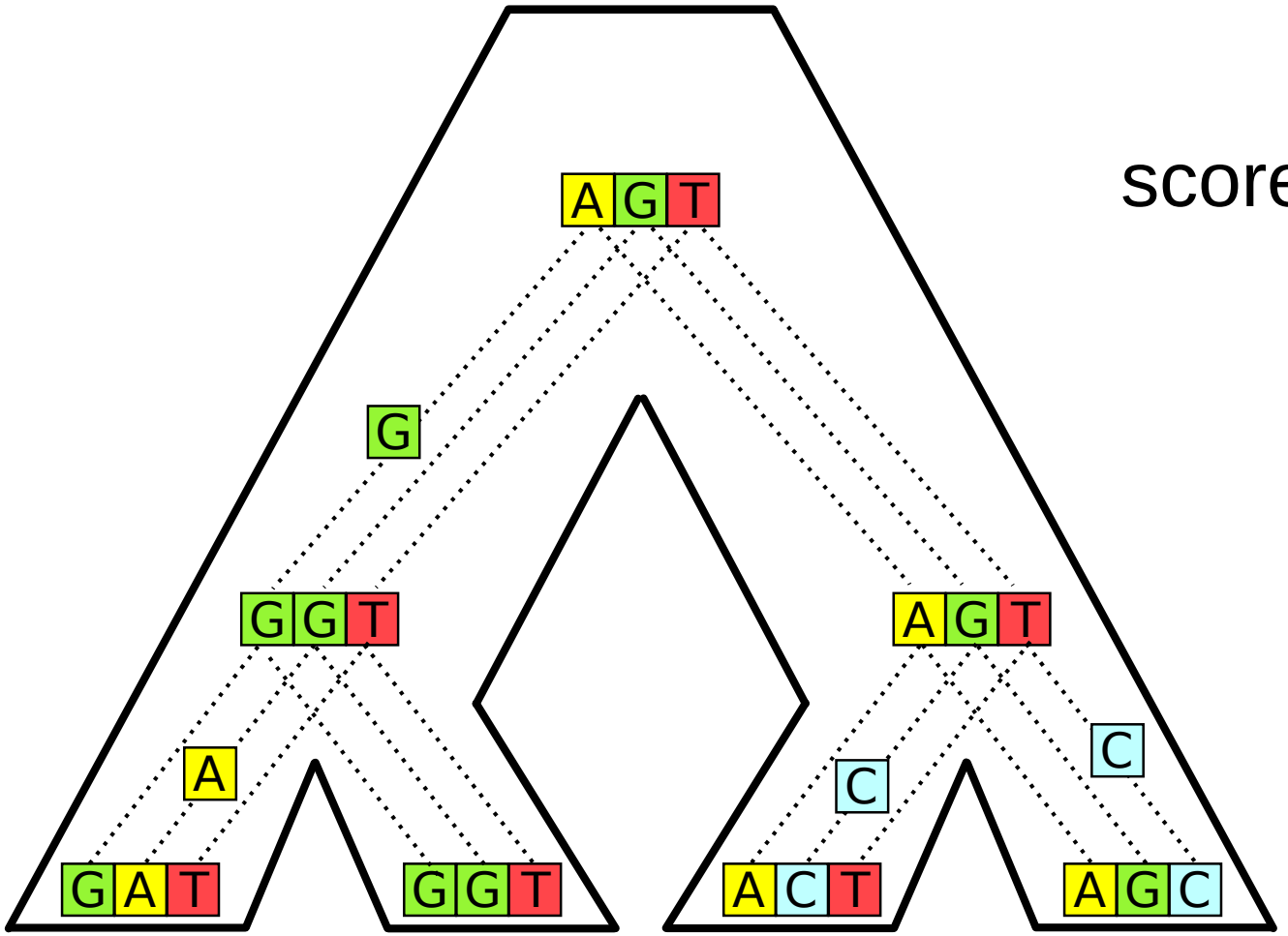
- The tree that requires the least number of mutations is the most plausible

- Parsimony score = number of mutations

Parsimony score



Parsimony score



score = 4

Tree search algorithms

- Tree search: tries to find the tree that maximizes a score
- Often: no guarantee to converge to the best solution
- Tree search is expensive (needs to evaluate the score of many candidate trees)

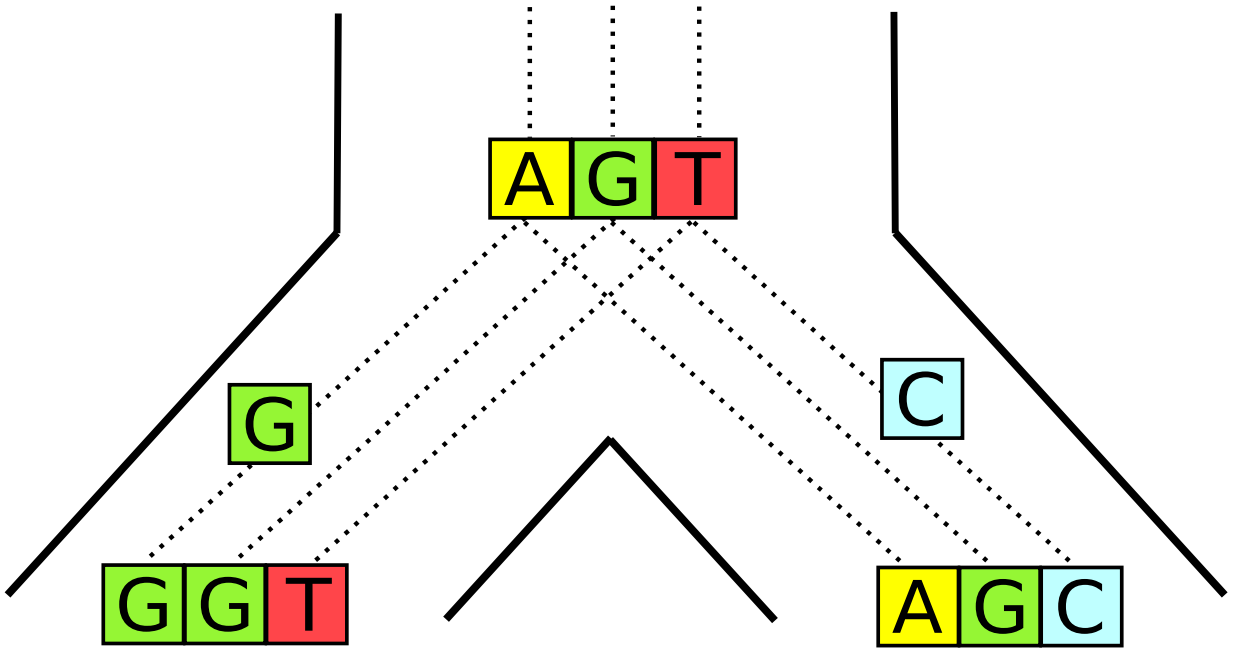
Parsimony methods

- Advantages:
  - Faster than probabilistic methods

- Weaknesses:
  - Slower than distance methods
  - Does not count multiple hits

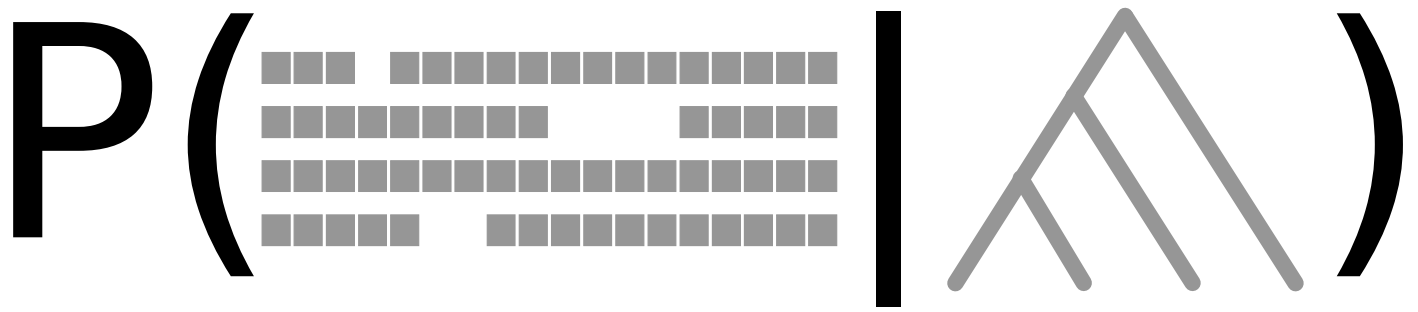
Tree likelihood

Sequence evolution can be described with a probabilistic model (e.g. GTR, LG, etc.)



Tree likelihood

Phylogenetic likelihood: probability of observing the alignment given the tree



Tree likelihood

Phylogenetic likelihood: probability of observing the alignment given the tree

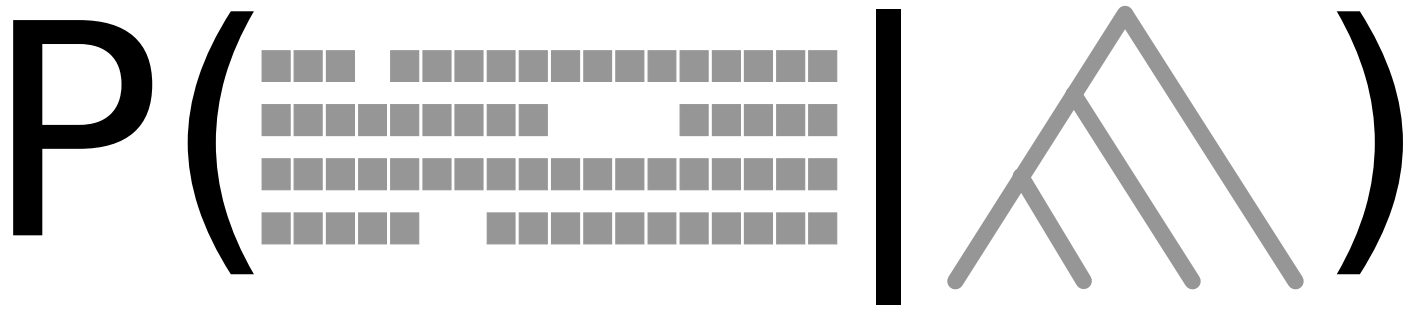
In other words: given a tree, its branch lengths, and the subsitution rates. If I start from a random ancestral sequence, what is the probability to exactly obtain the observed sequences by chance?

Tree likelihood

- In phylogenetics, likelihoods can be insanely small numbers
- We usually use log likelihoods:
  - L(tree) = 0.000000000000169
  - LL(tree) = -29.40

Tree likelihood

- It is not the probability of the tree
- Integrates over all possible scenarios of mutations that would generate the alignment
- Depends on the branch lengths and model parameters



Maximum likelihood methods

- Given an alignment, search for:
  - the tree
  - the branch lenghts
  - the model parameters (e.g. substitution rates)

  that maximize the likelihood

- Tree search algorithm

Maximum likelihood tree inference

Standard tools:

- RAxML-NG *[Kozlov et al. 2019]*
- IQTREE-2 *[Minh et al. 2020]*
- (FastTree-2) *[Price et al. 2010]*

Maximum likelihood tree inference

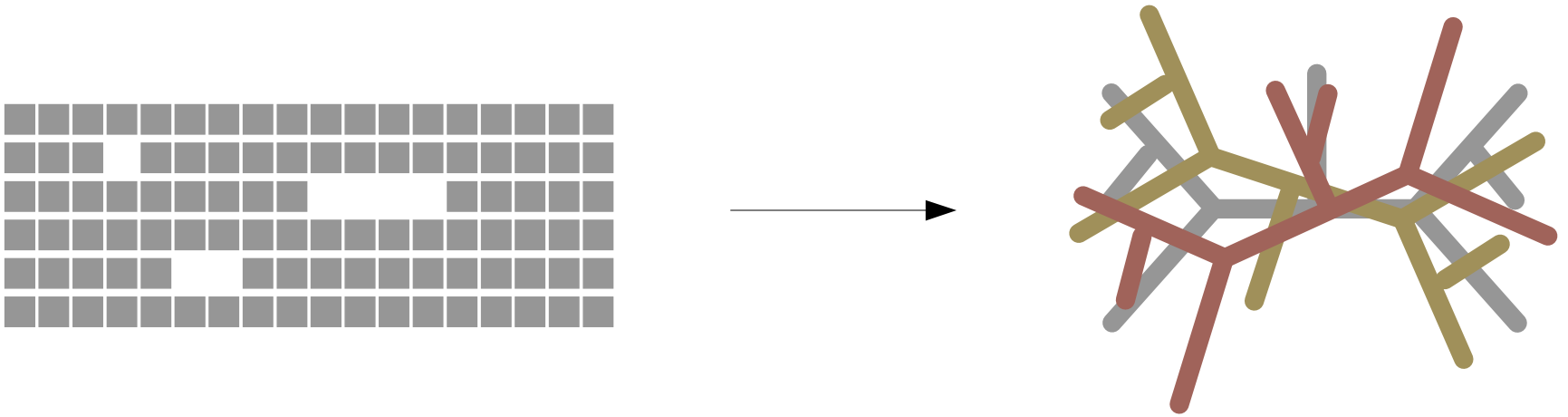
- Advantages:
  - More accurate
  - Allows to estimate model parameters from the data
  - Allows to use complex models of evolution

- Weaknesses:
  - Slower than the other methods

Bayesian tree inference

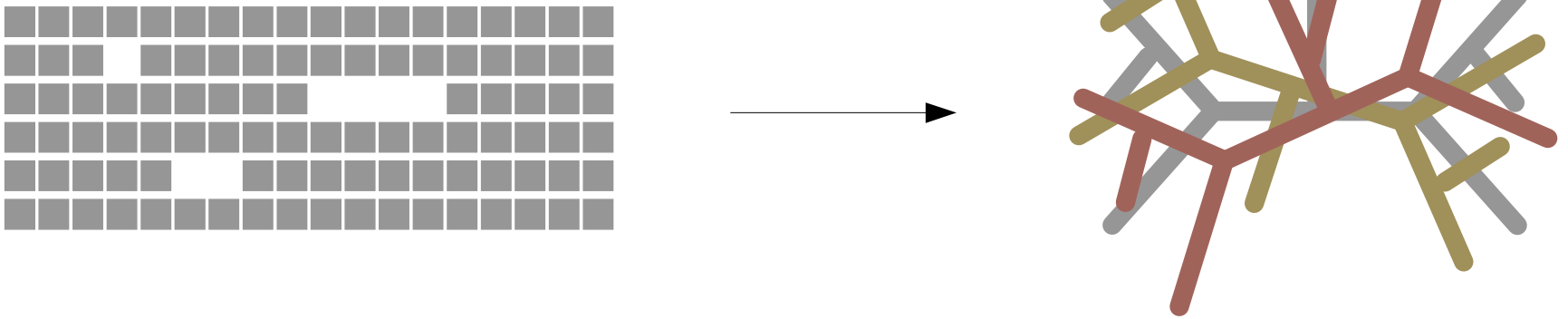
- Infers a distribution of trees

- In practice: a set of trees. Each tree appears with a frequency proportional to its probability



Bayesian tree inference

- Advantages:
  - Probabilistic method
  - Great to deal with uncertainty
- Weaknesses:
  - Expensive
  - Assessing convergence is difficult
  - Requires arbitrary priors



Bayesian tree inference

- Standard tools:
  - MrBayes 3.2 *[Ronquist et al. 2012]*
  - PhyloBayes 3 *[Lartillot et al. 2009]*
  - RevBayes *[Höhna et al. 2016]*
  - ExaBayes *[Aberer et al. 2014]*

Which method is the best?

- **There is no best method**
- Maximum likelihood and Bayesian inference are usually considered to be more accurate
- Bayesian inference does more than inferring just one tree
- Distance and parsimony methods can deal with larger datasets

Tree inference from a single gene

- Very distant or very close speciation events are difficult to resolve (LBA, multiple hits, saturation etc.)
- We need as much data as possible to resolve the difficult nodes
- The length of a single gene sequence is limited...

Phylogenomics

Draw information from whole genomes.

How can we combine phylogenetic signals from different genes families?

Phylogenomics

Obstacles:

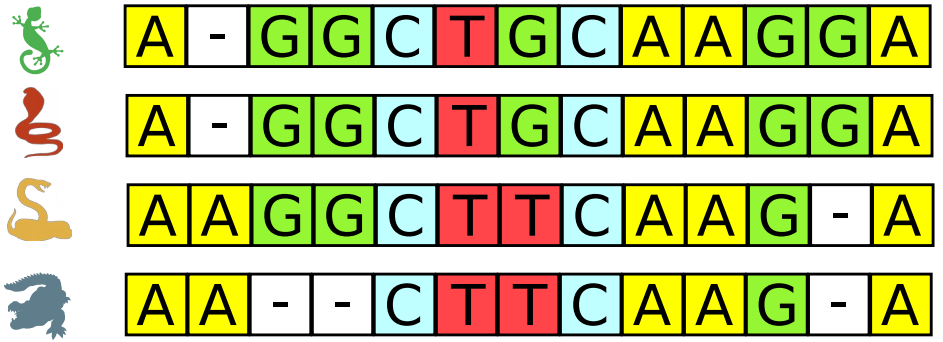
- Genome translocation
- Incomplete lineage sorting
- Gene duplication, loss, transfer
- … and more :-(

Phylogenomics

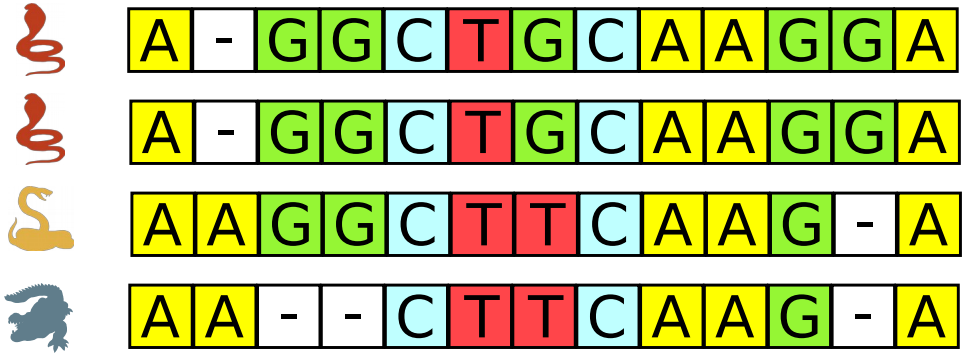
- Identify homologous genes
- Group them into gene families
- Align each gene family separately

Single-copy VS multi-copy gene families

- Single-copy: at most one gene per species



- Multi-copy: potentially more than one gene per species



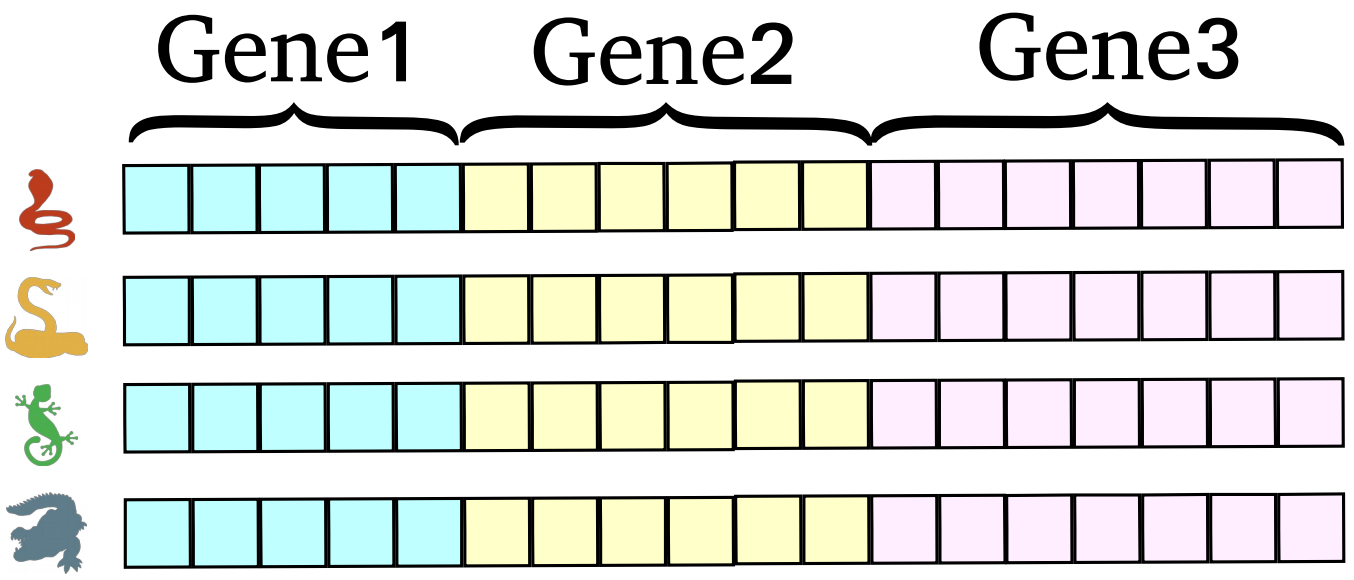
Phylogenomics

We will first assume single-copy gene families.

(no duplication, no horizontal gene transfer)

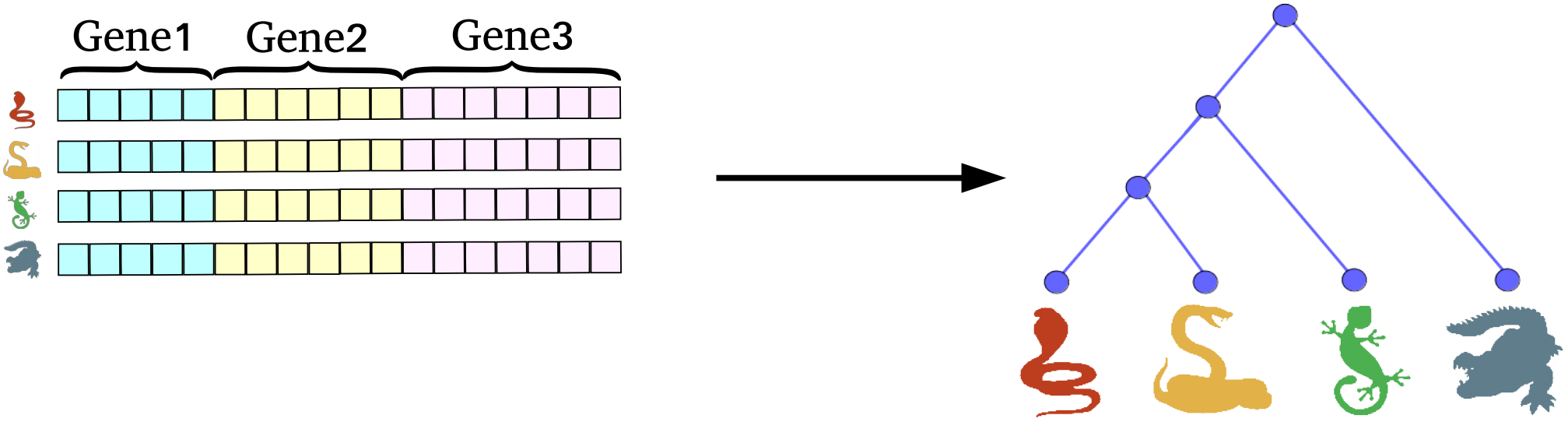
Concatenation methods

Concatenate gene alignments into a supermatrix:



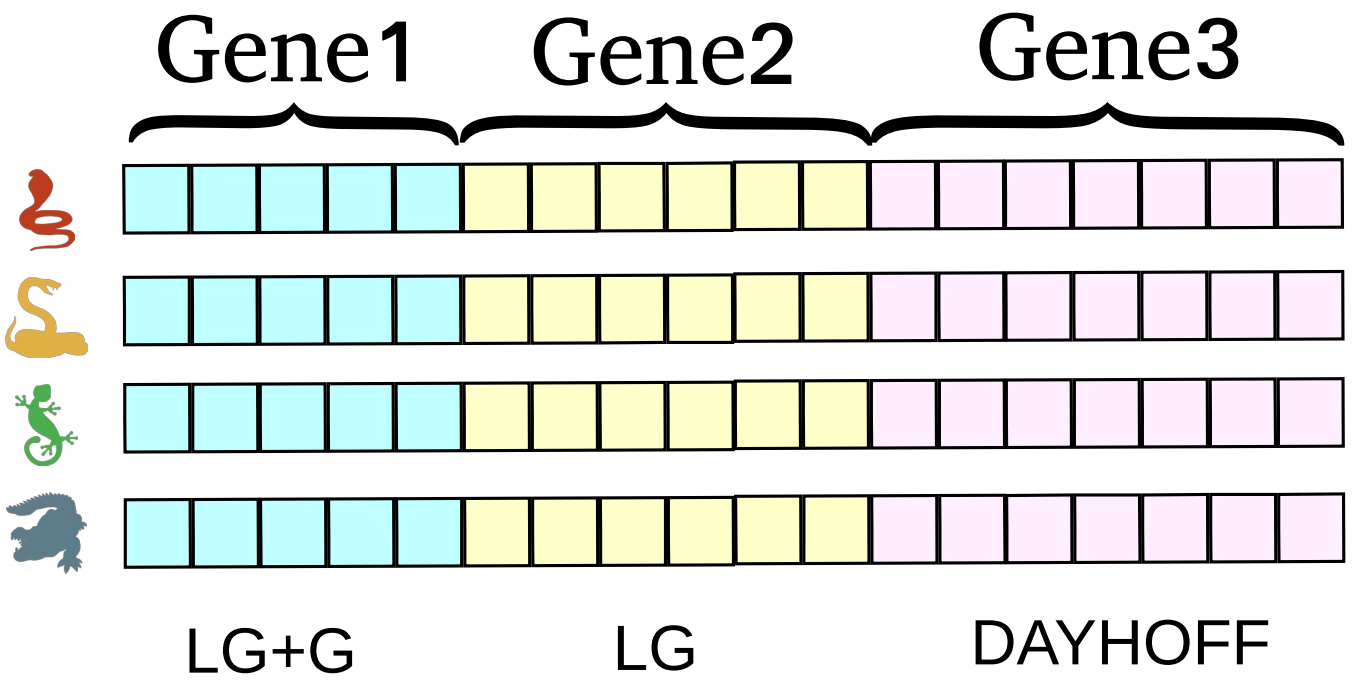
Concatenation methods

Infer the tree from the supermatrix



Concatenation methods

Each gene can have different mutation rates and different models

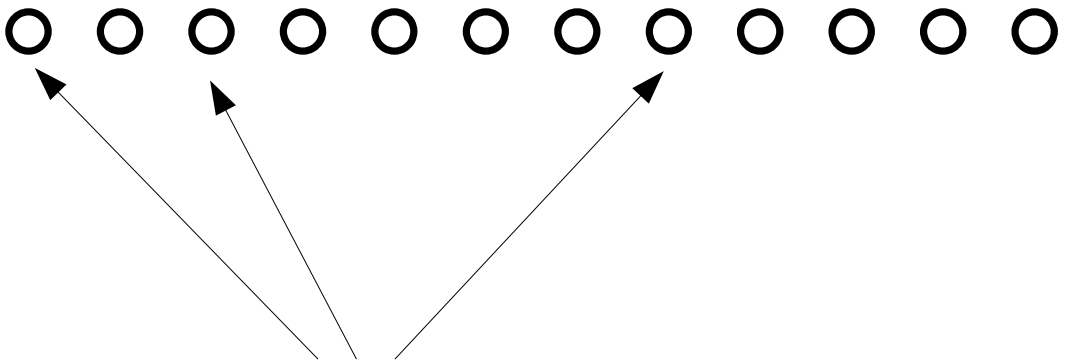


Gene trees

- A gene tree describes the evolutionary history of a gene family
- Different gene families can have different evolutionary histories
- Gene trees often conflict with the species tree

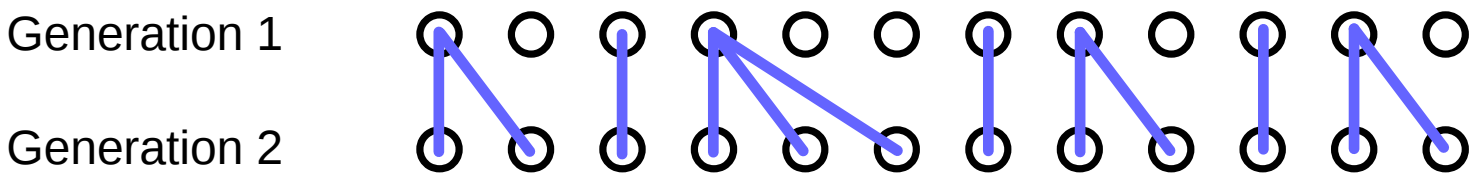
Populations

Generation 1



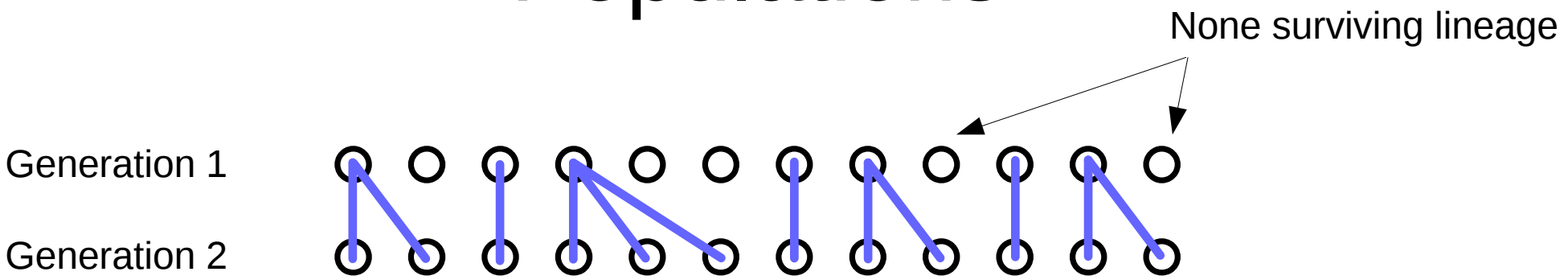
Different gene alleles in a population

Populations



Each gene in Generation 2 has one parent in Generation 1

Populations



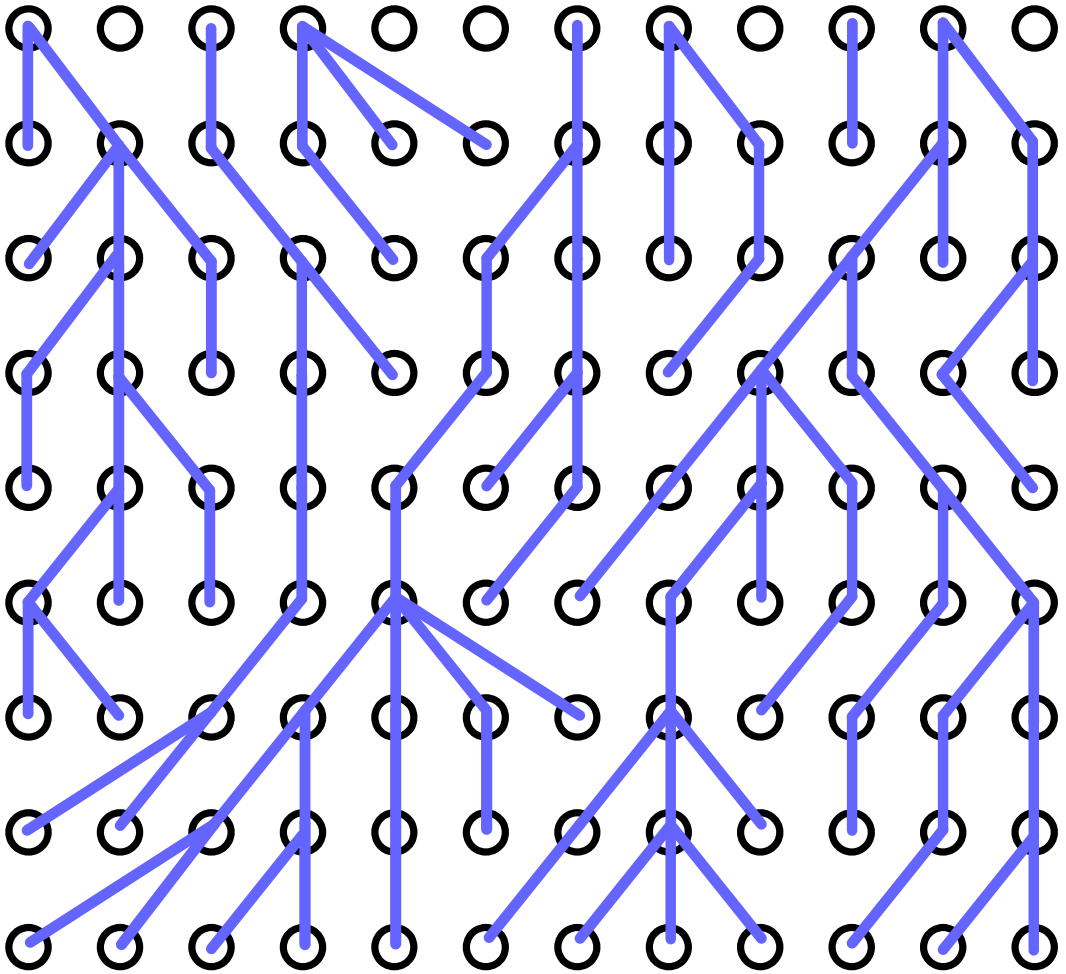
Each gene in Generation 2 has one parent in Generation 1

Populations

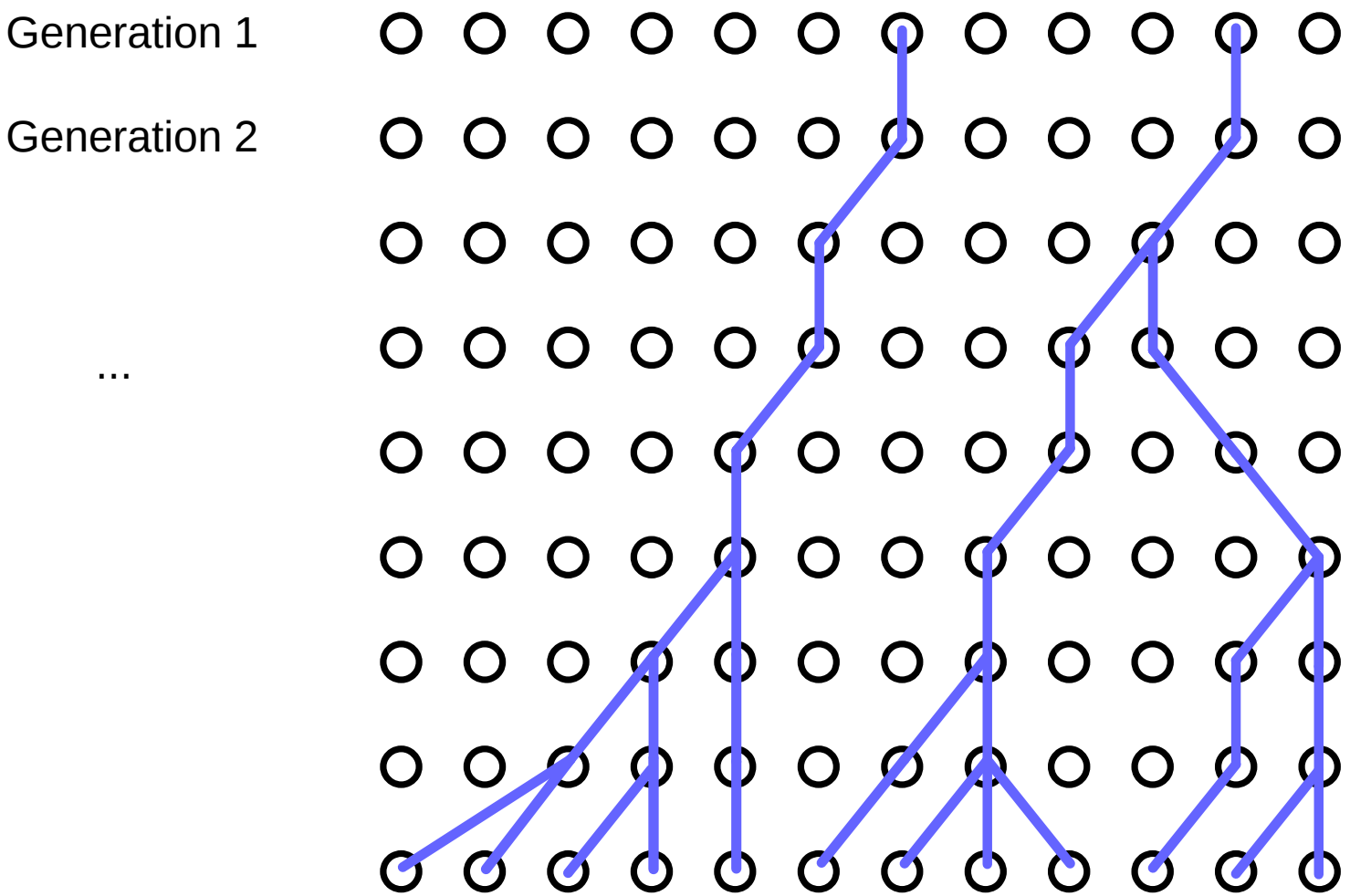
Generation 1

Generation 2

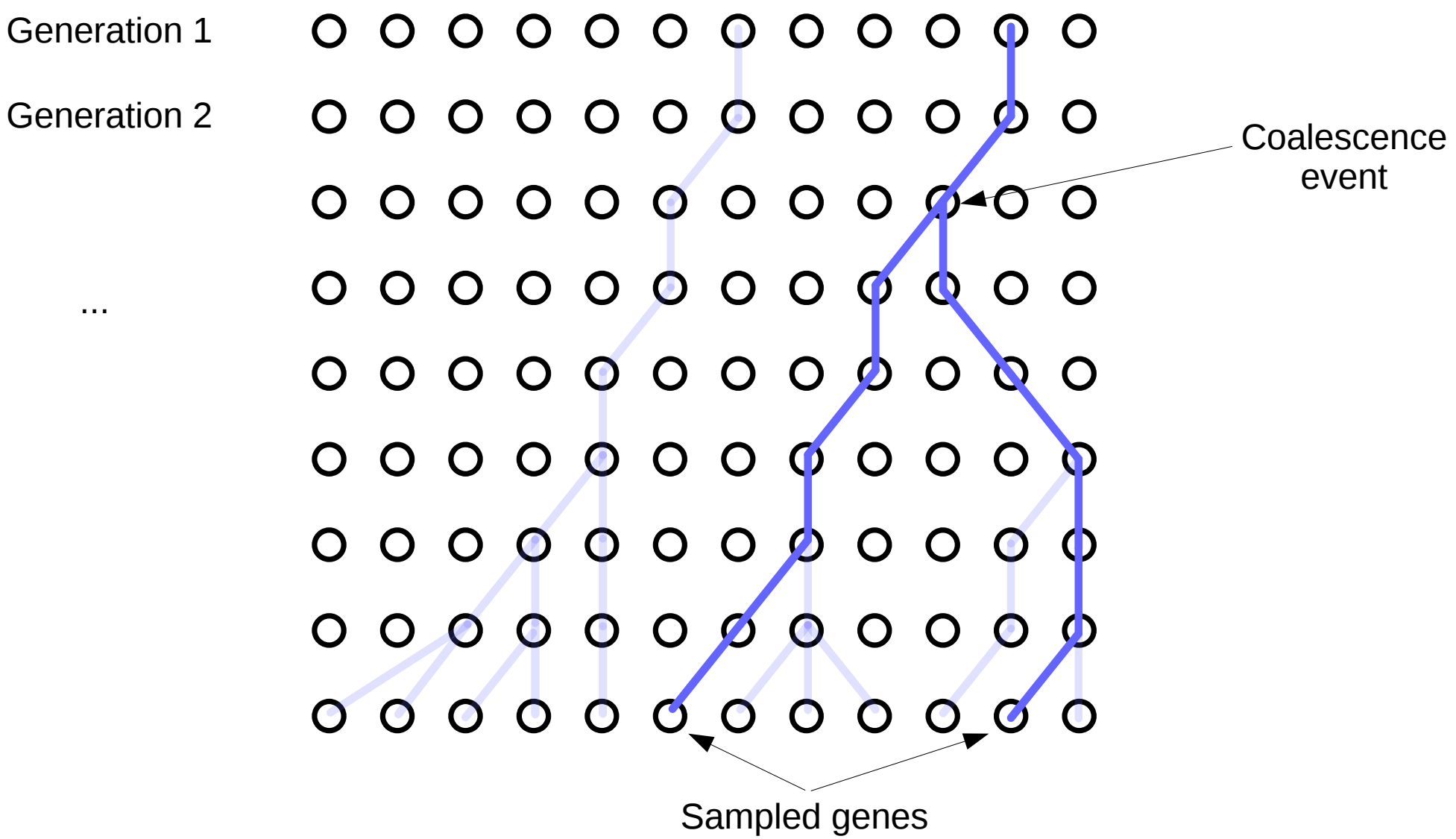
...



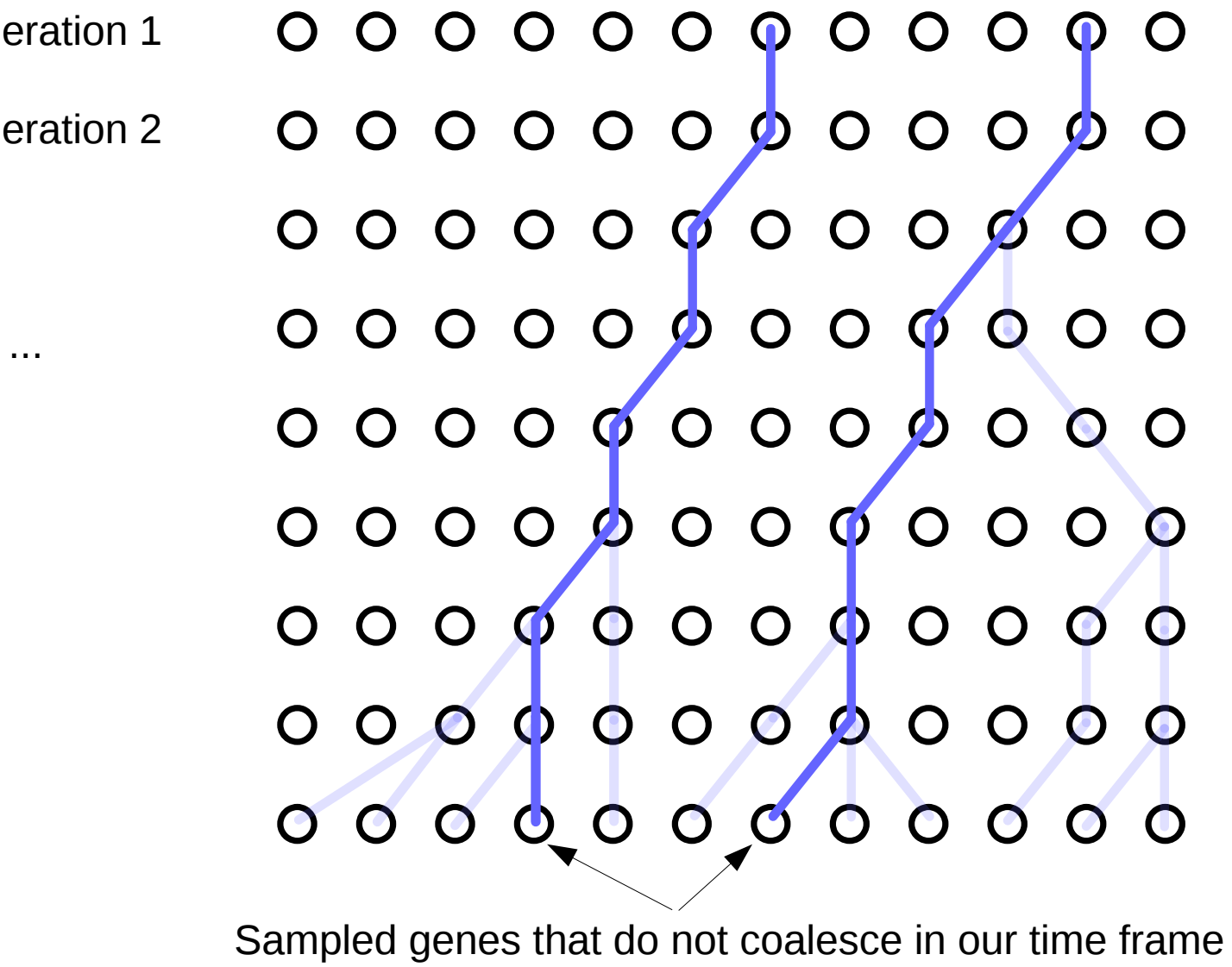
Surviving lineages



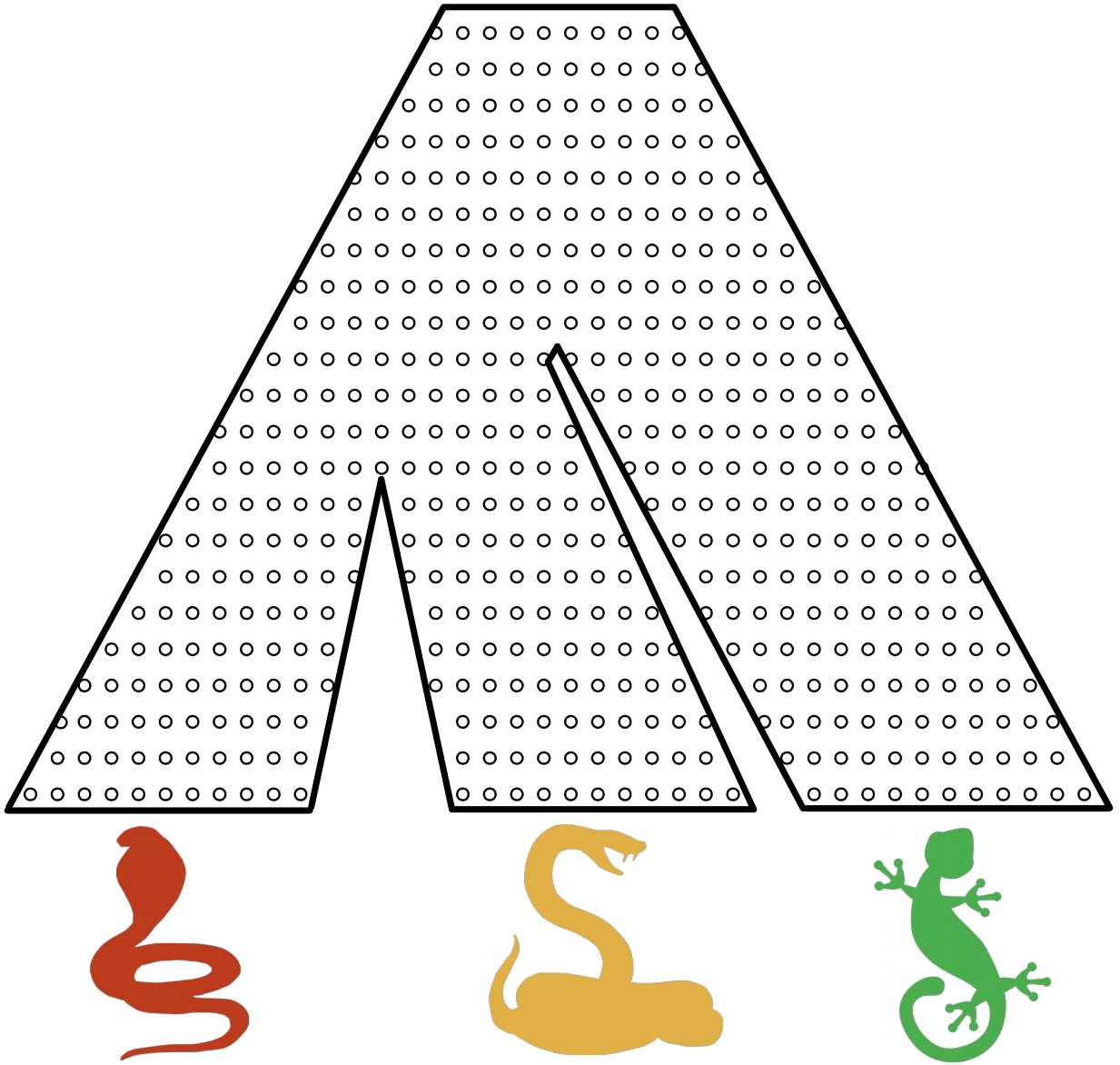
Coalescence



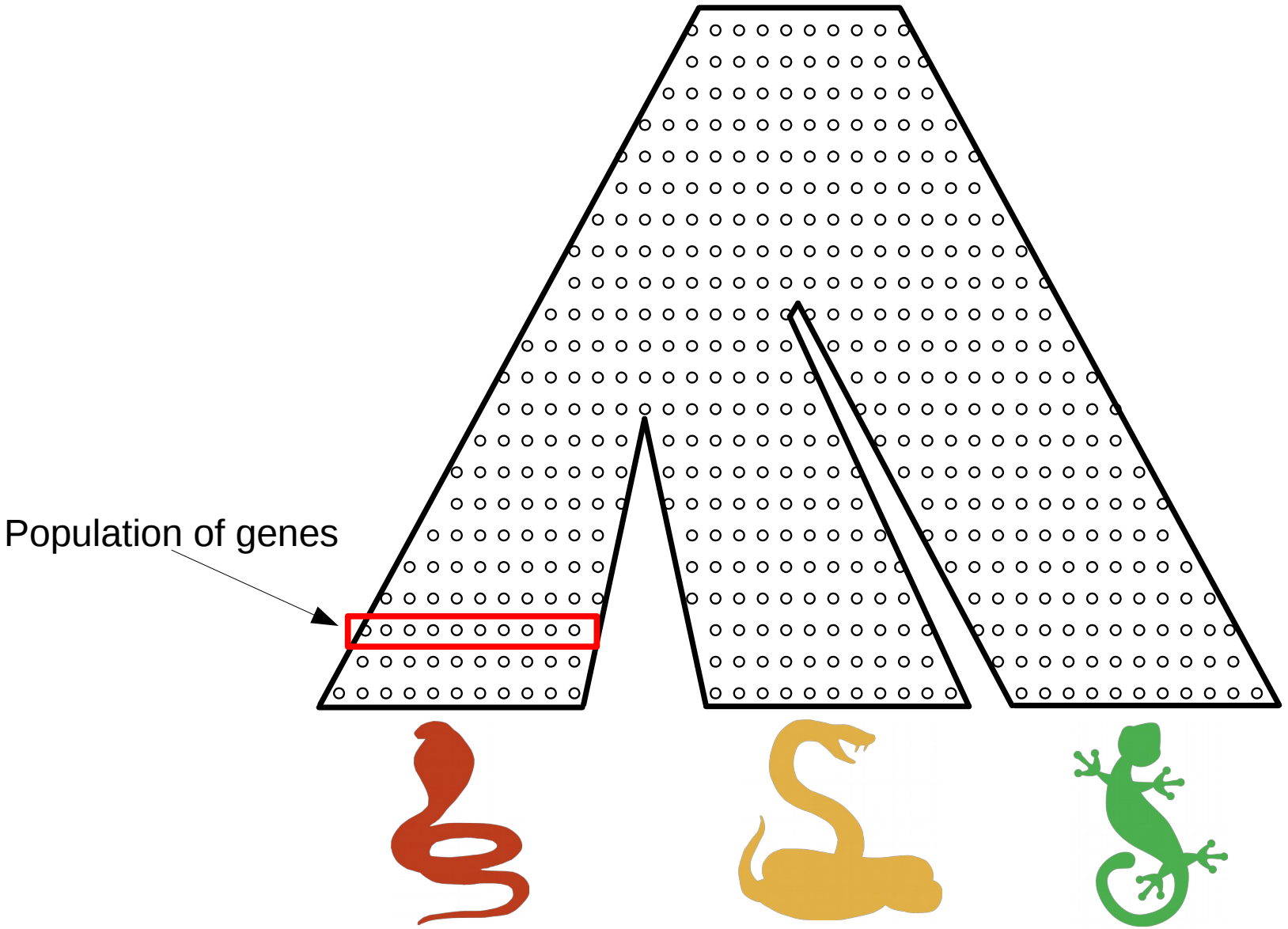
Incomplete lineage sorting



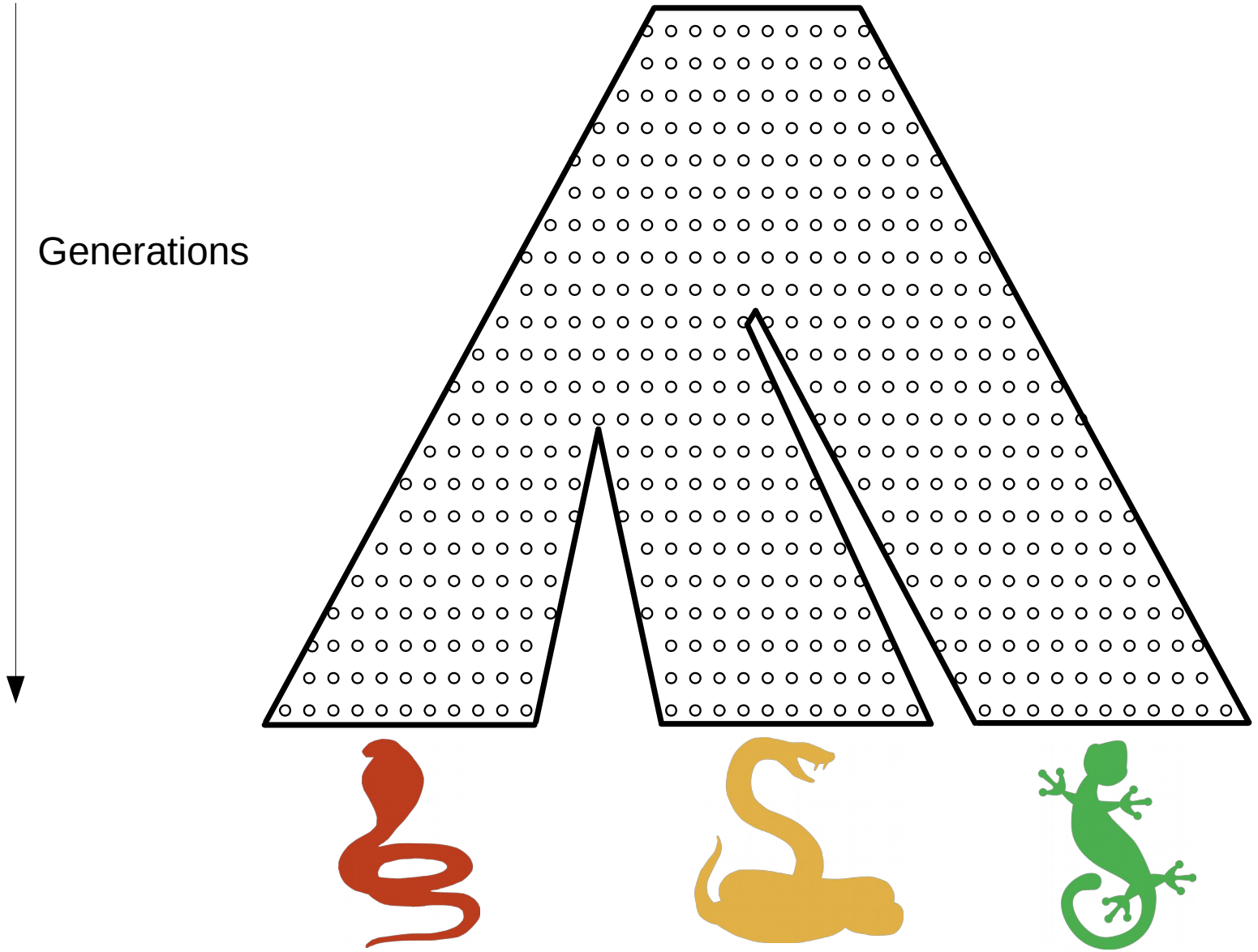
The multi-species coalescent model



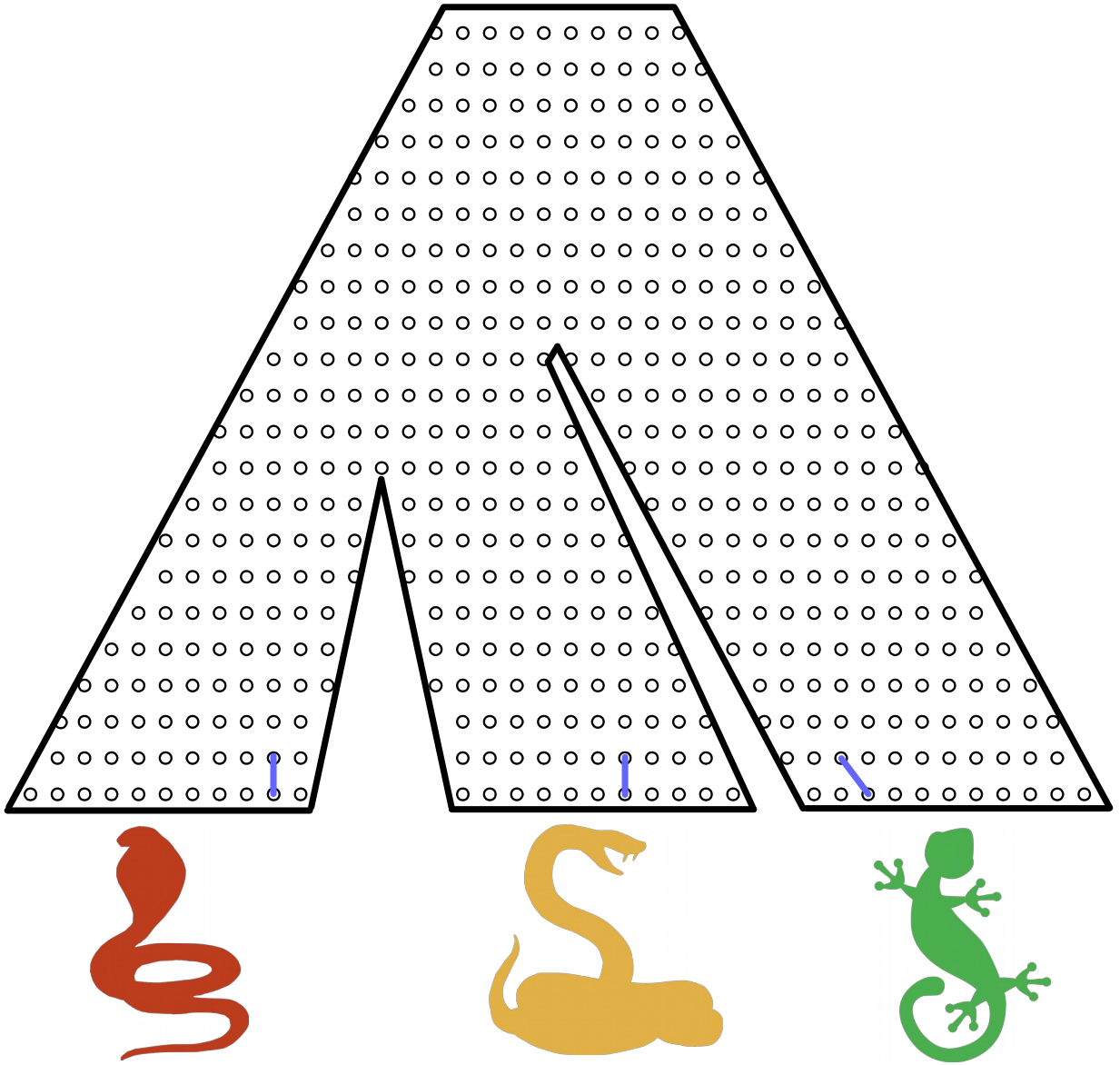
The multi-species coalescent model



The multi-species coalescent model

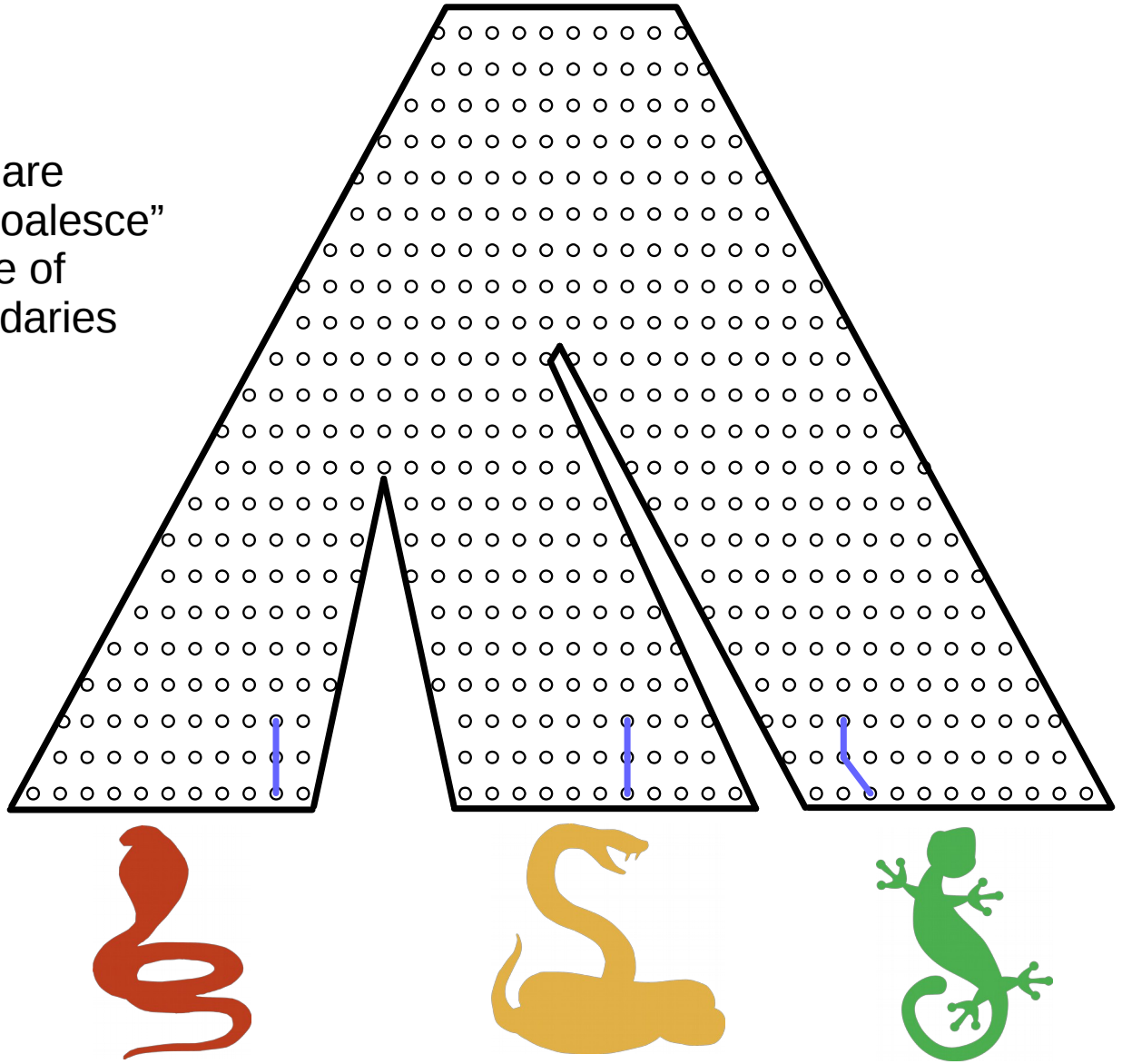


The multi-species coalescent model

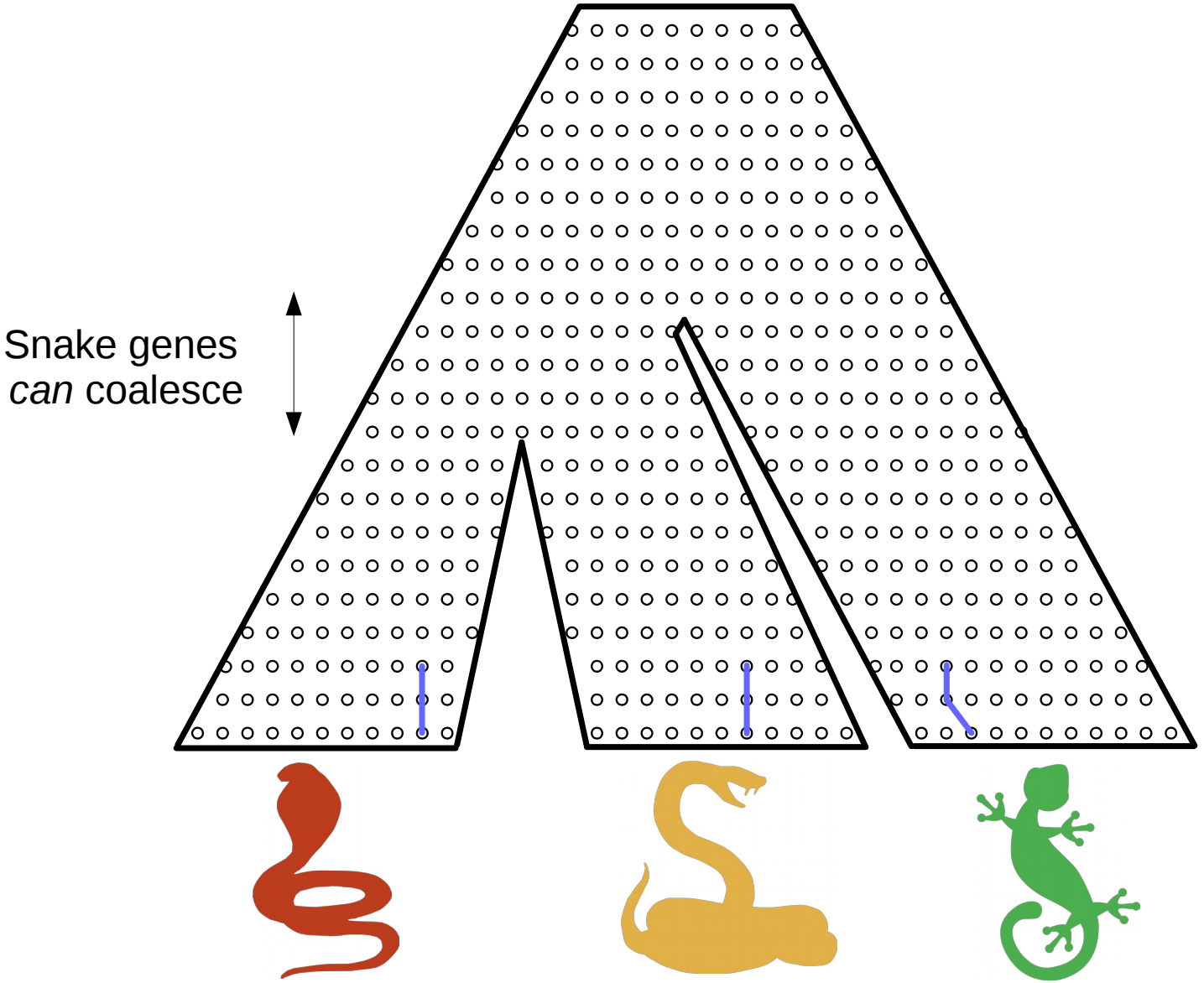


The multi-species coalescent model

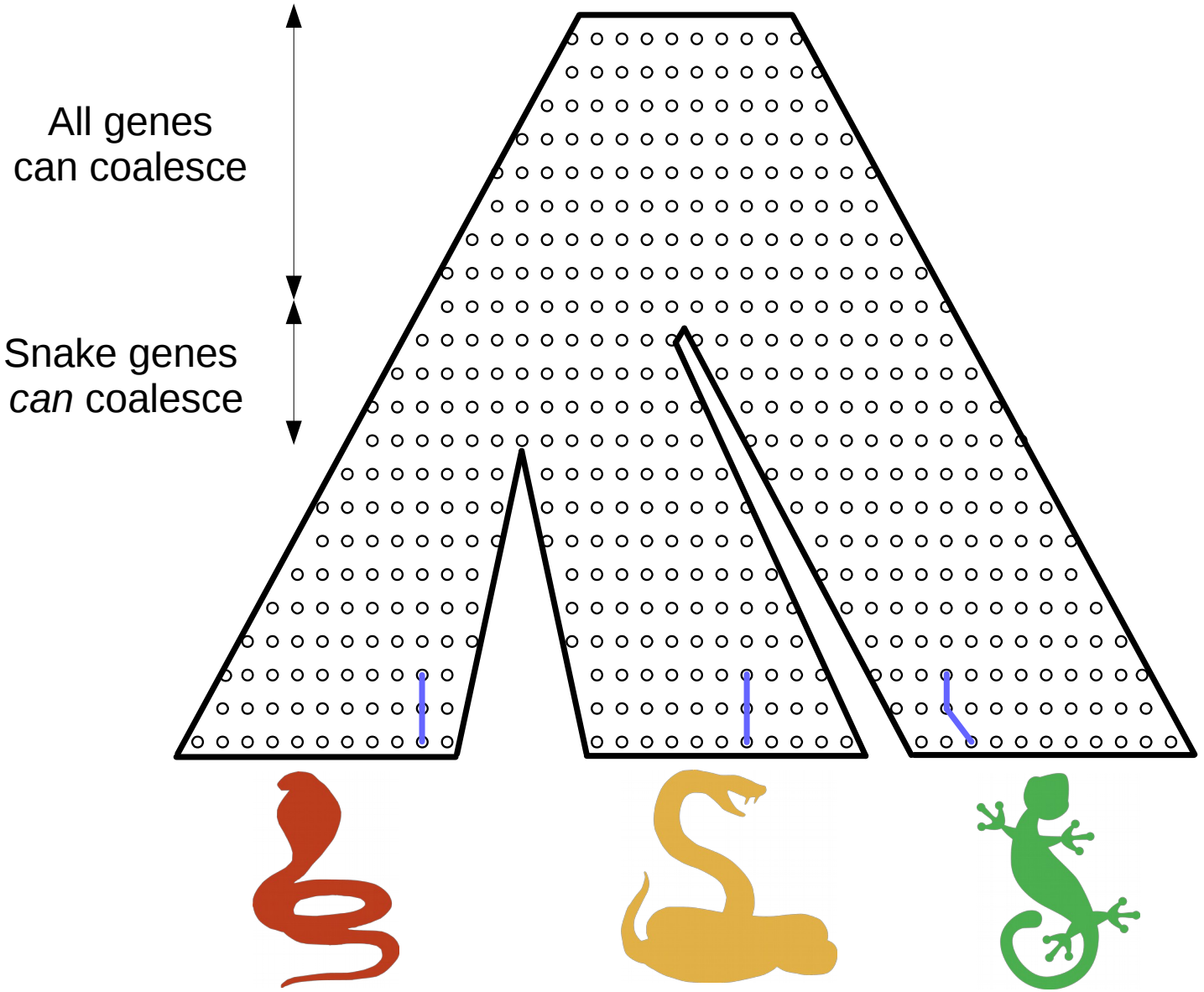
The genes are
“not allowed to coalesce”
yet because of
species boundaries



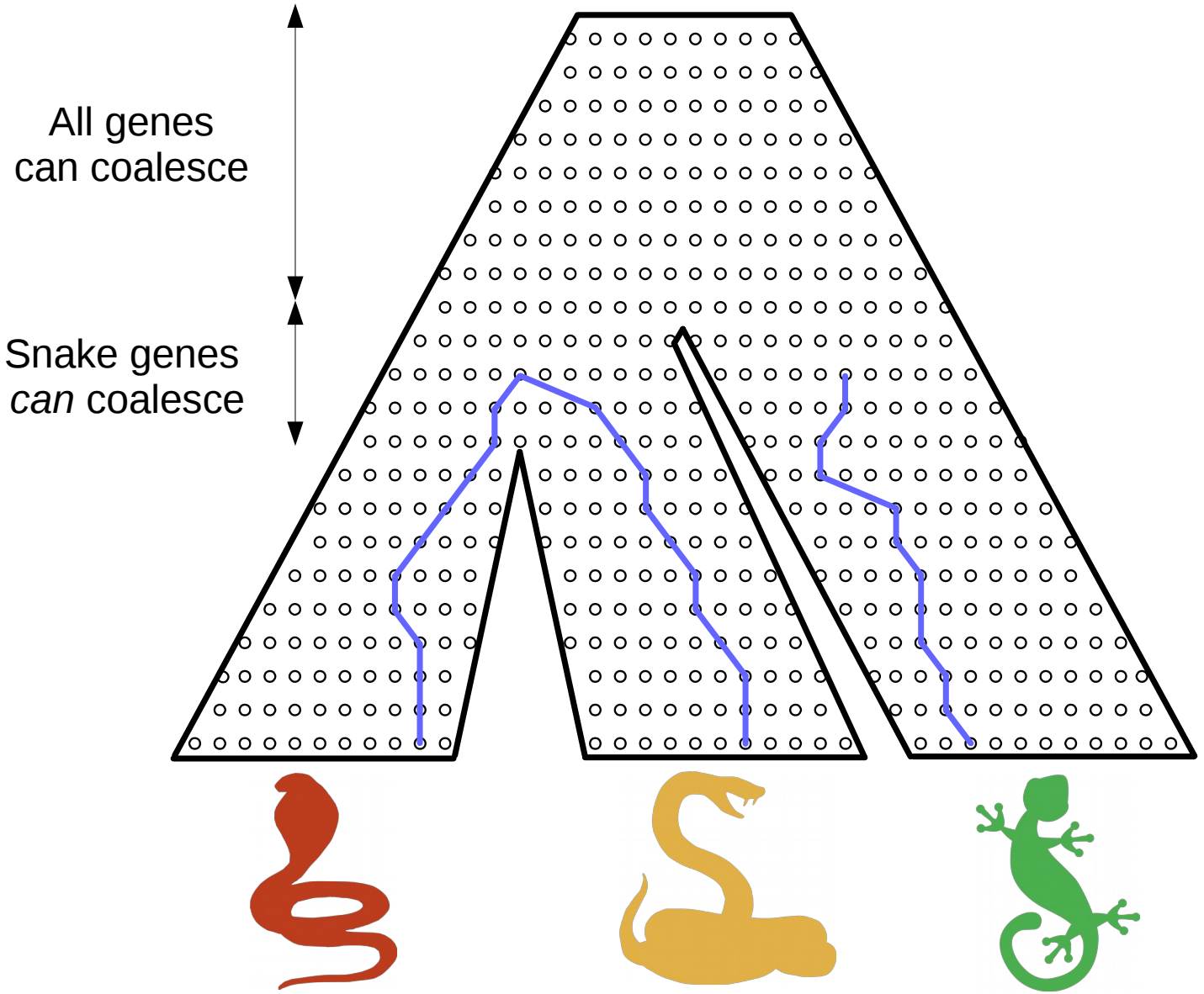
The multi-species coalescent model



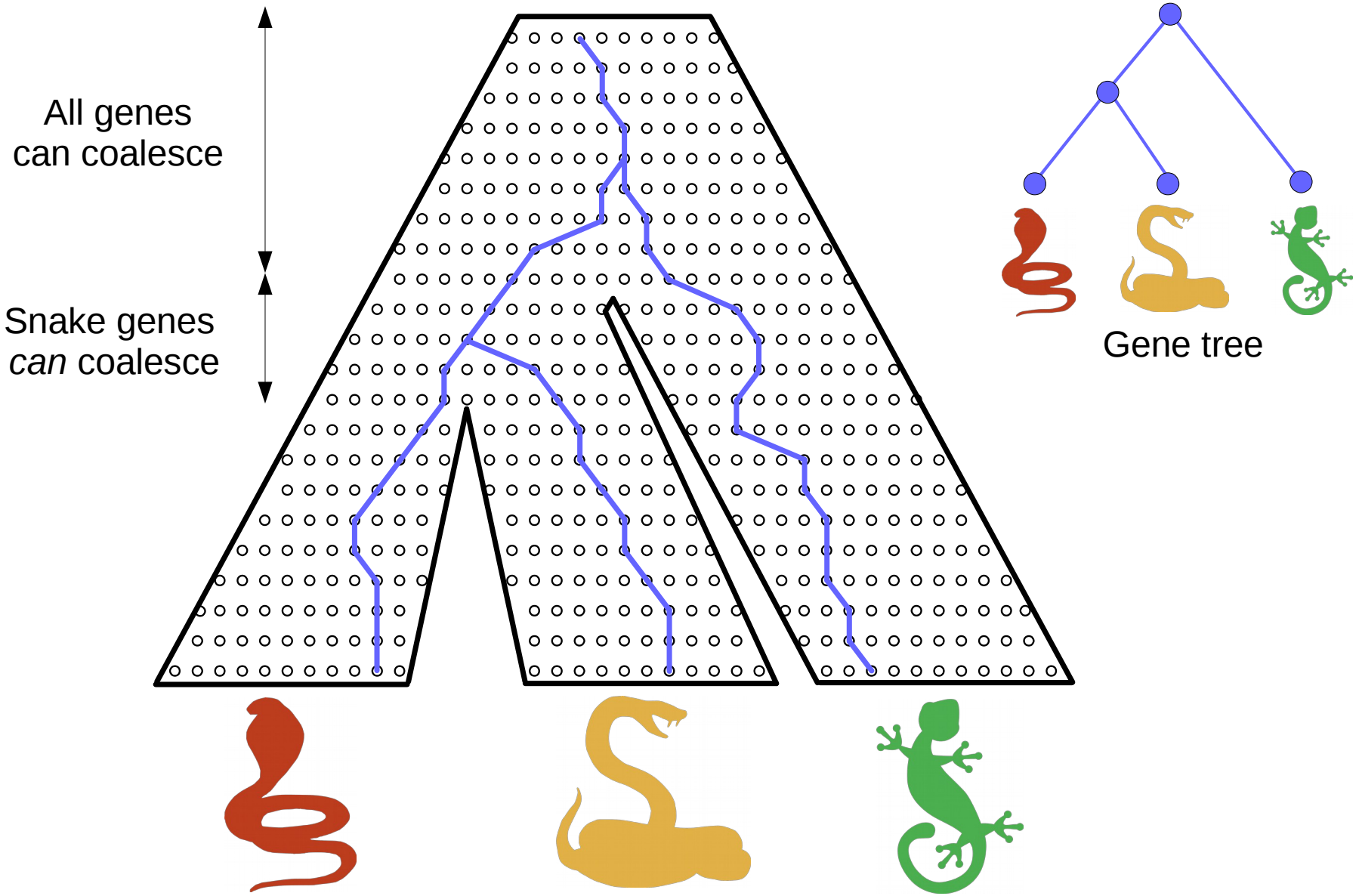
The multi-species coalescent model



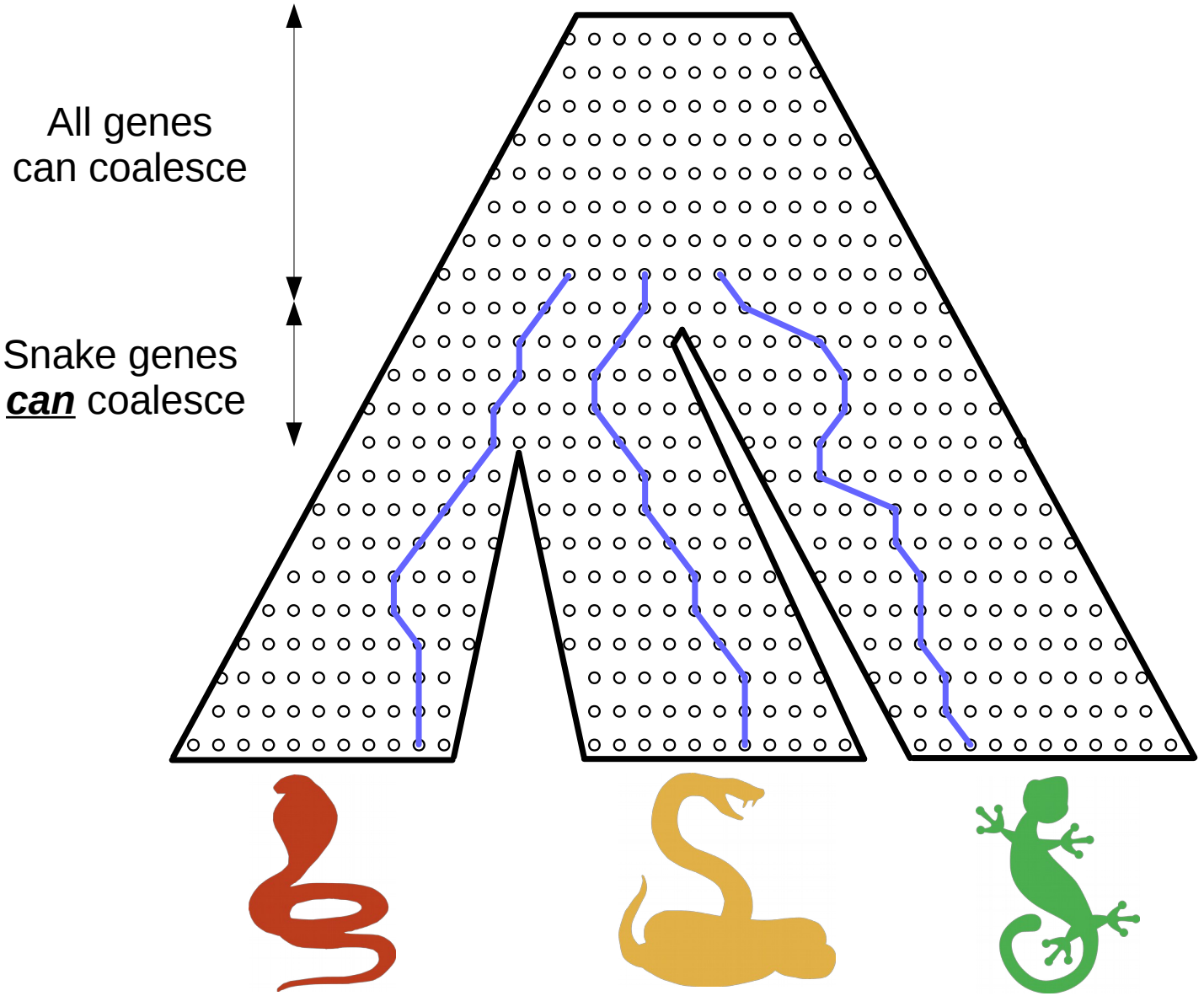
The multi-species coalescent model



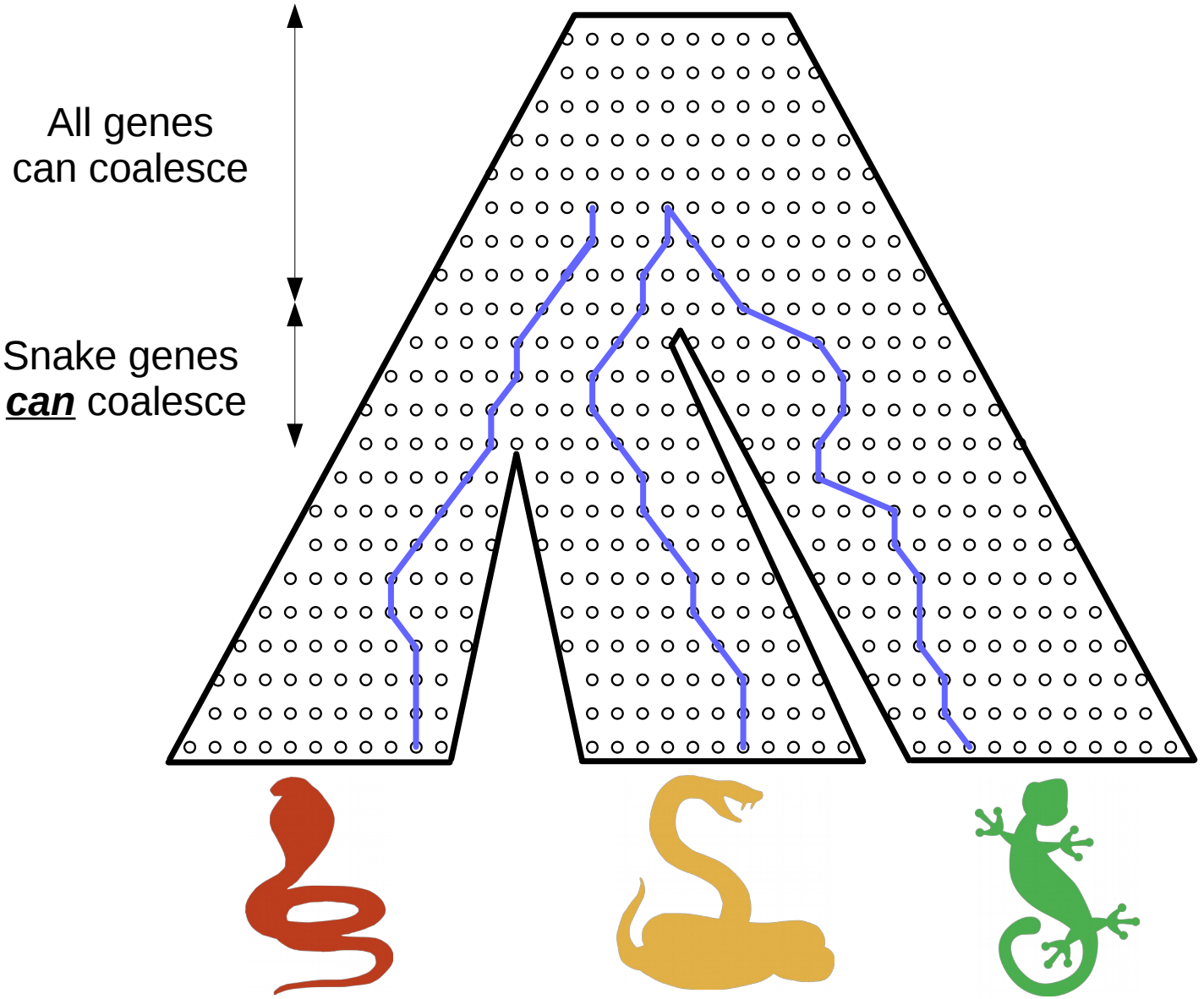
The multi-species coalescent model



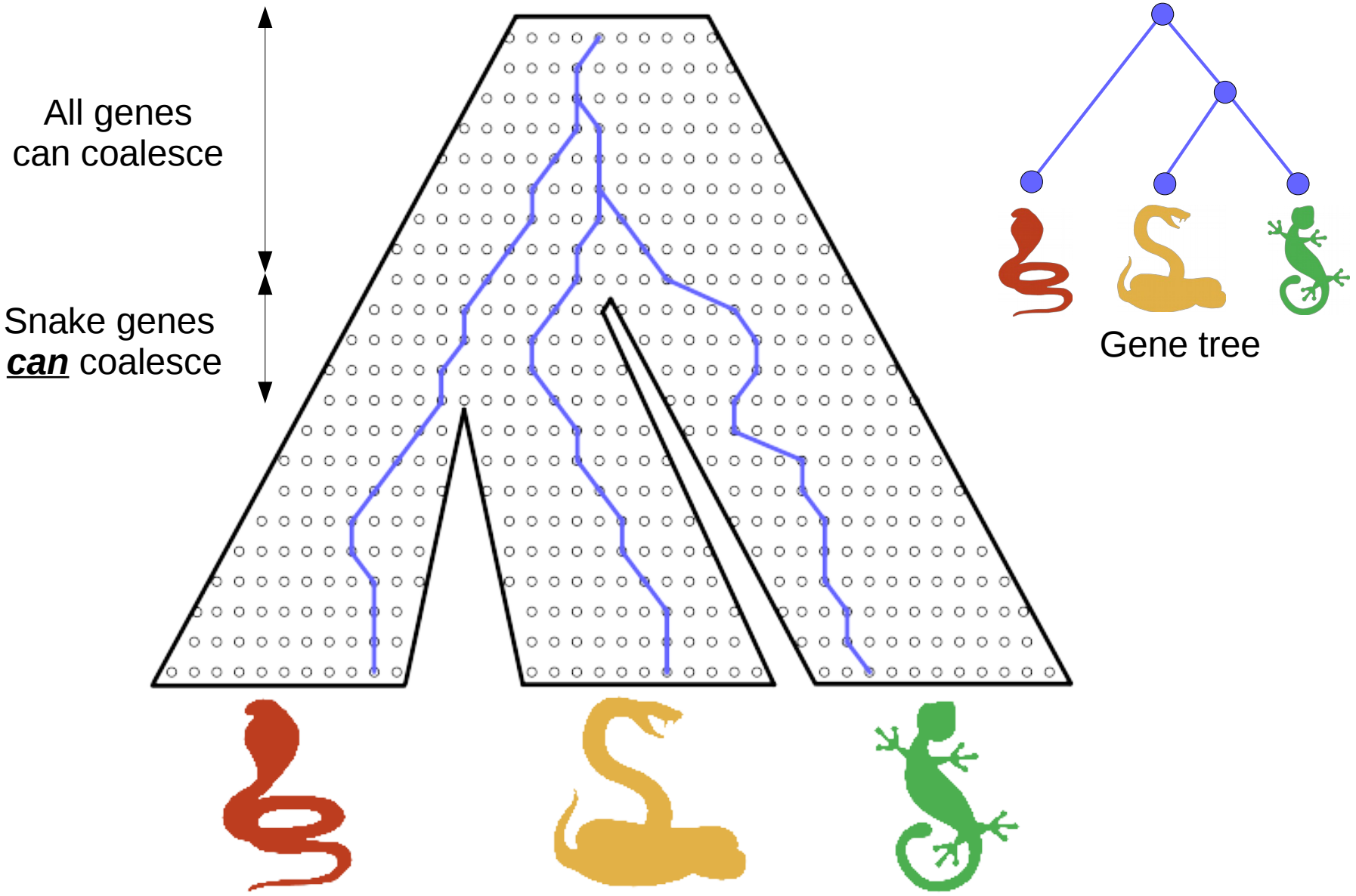
Incomplete lineage sorting



Incomplete lineage sorting



Incomplete lineage sorting



Is ILS a problem?

The following claim looks reasonable:

- *“With enough genes, concatenation will recover the correct topology”*

Is ILS a problem?

The following claim looks reasonable:

- *“With enough genes, concatenation will recover the correct topology”*

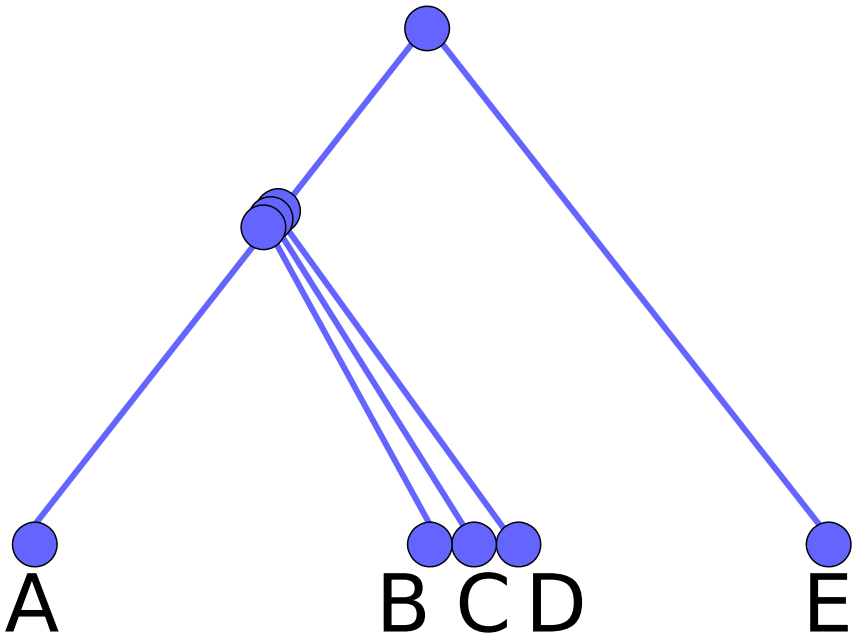
… Always true up to 4 taxa

… **Sometimes wrong from 5 taxa**

[Kubatko & Degnan 2007]

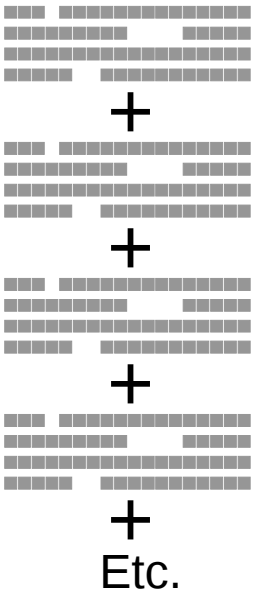
The anomaly zone

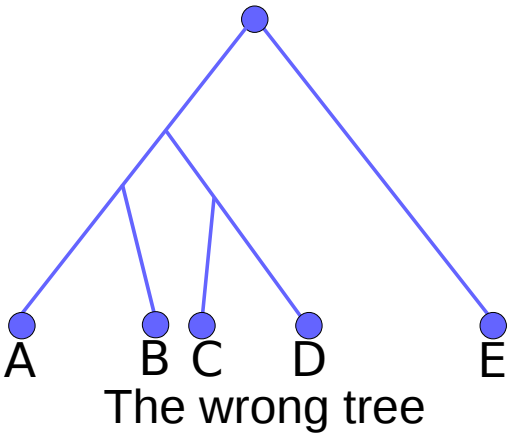
- Set of conditions under which the most frequent gene tree is not the species tree
- Short branches: few generations and high effective population size



Concatenation inconsistency

- Concatenation is inconsistent under the multi-species coalescent
- Under some conditions:

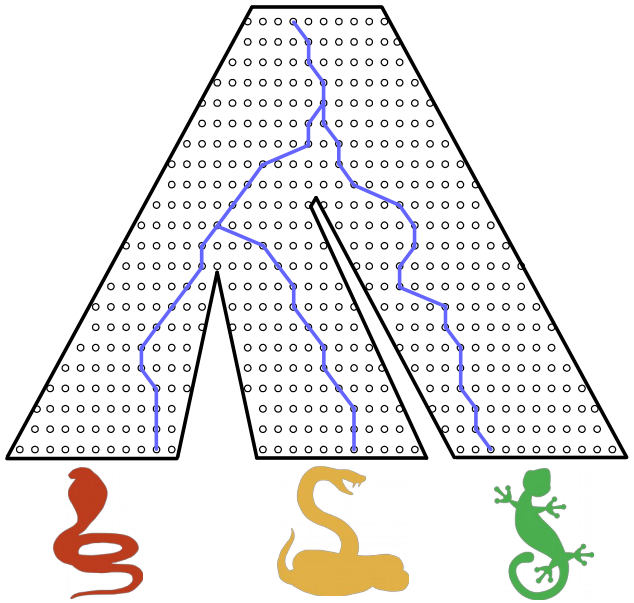


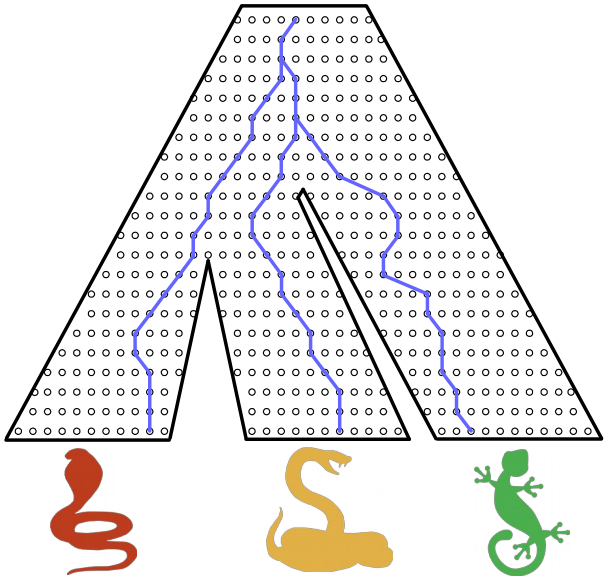


Gene tree methods

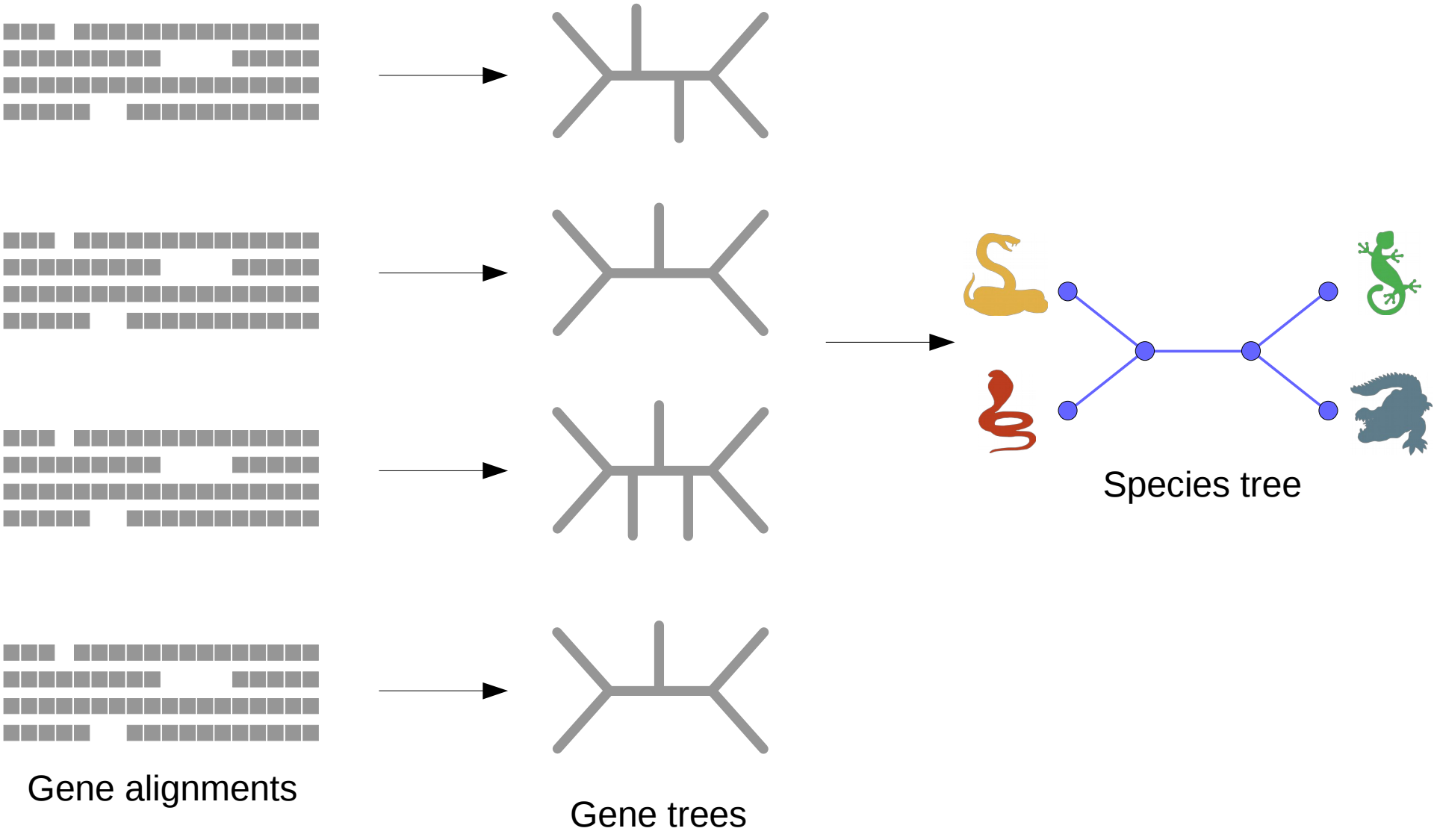
"If genes have different evolutionary histories...

… we should infer each gene tree separately"

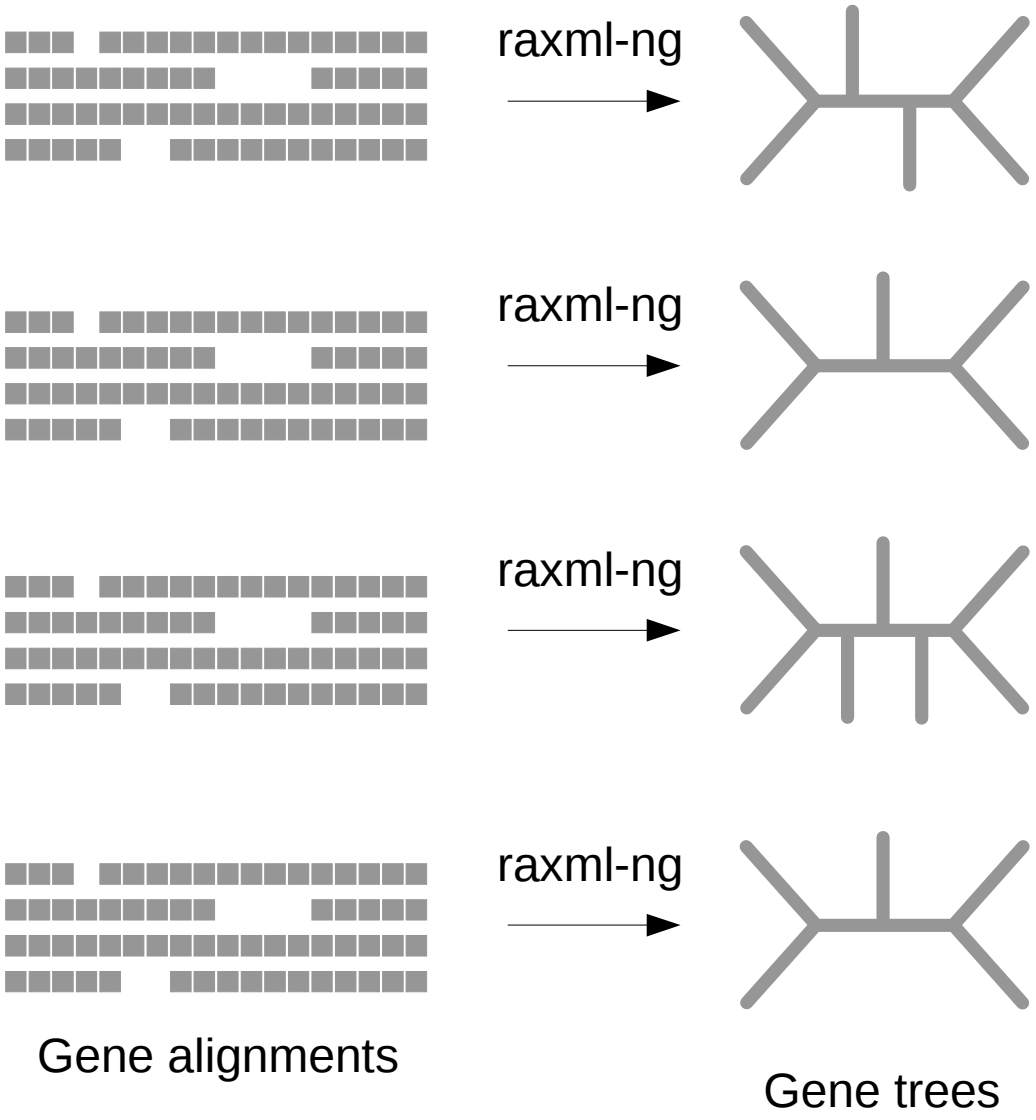




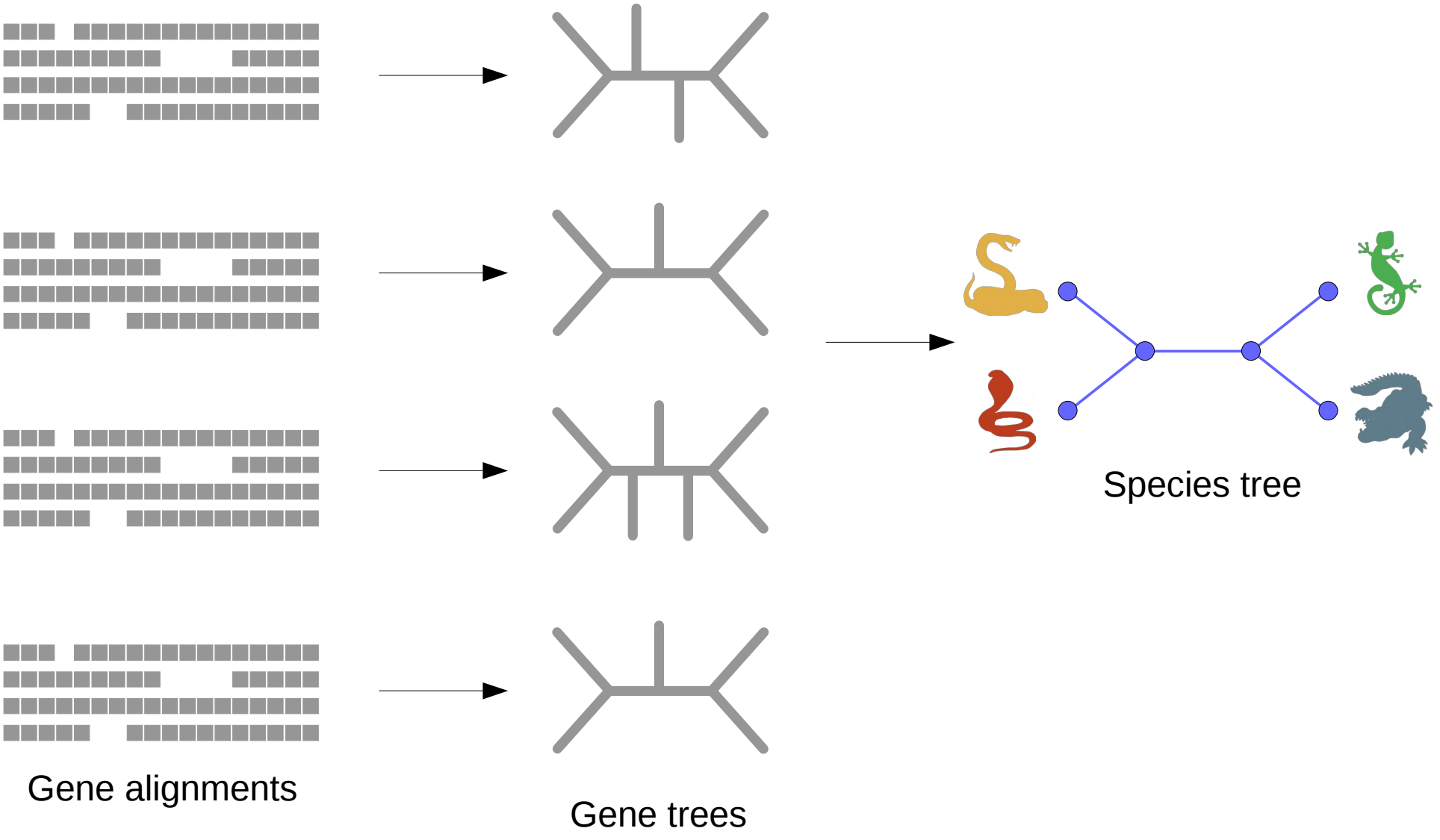
Gene tree methods



ParGenes: gene tree inference [Morel 2019]



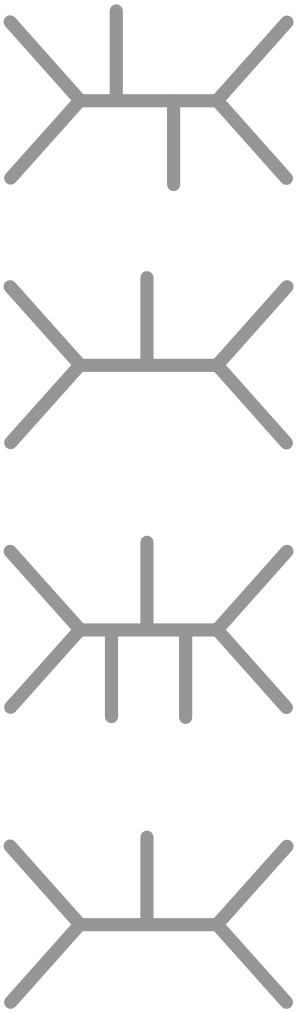
Gene tree methods



Gene tree methods

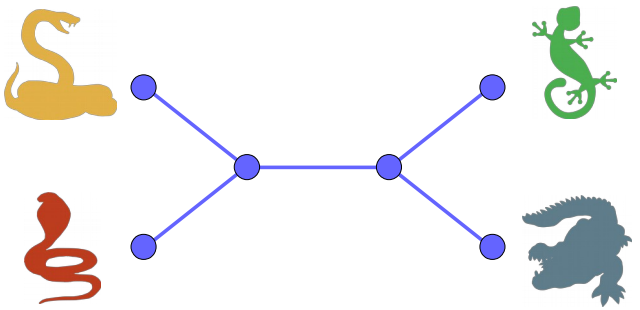
- ASTRAL
- ASTRID
- MP-EST
- STELLS
- etc.

ASTRAL *[Zhang et al. 2017]*



Gene trees

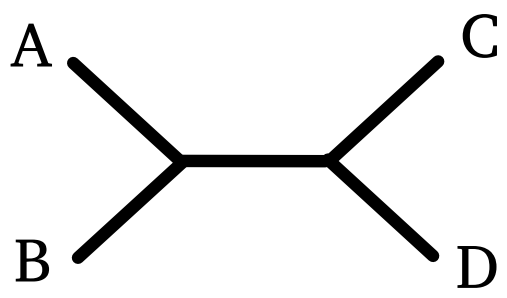
ASTRAL



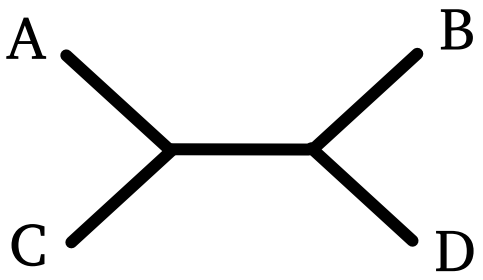
Unrooted
species tree

Quartets

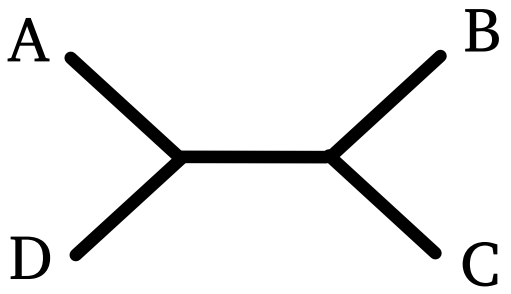
- Unrooted tree with 4 taxa
- 3 possible resolutions



AB | CD



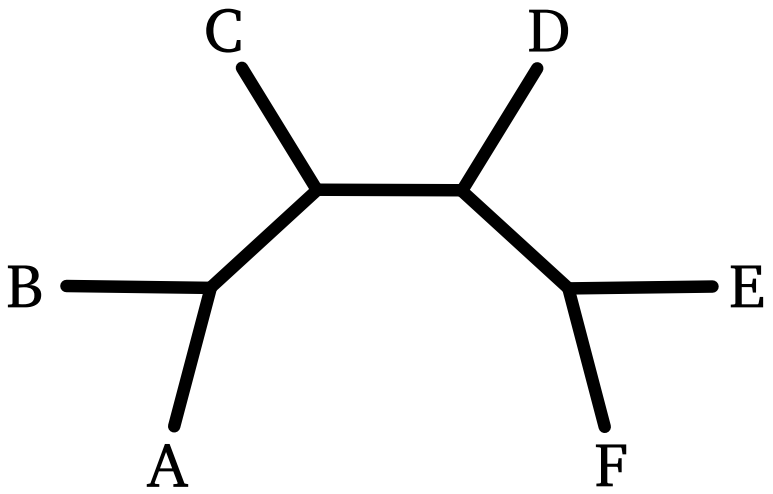
AC | BD



AD | BC

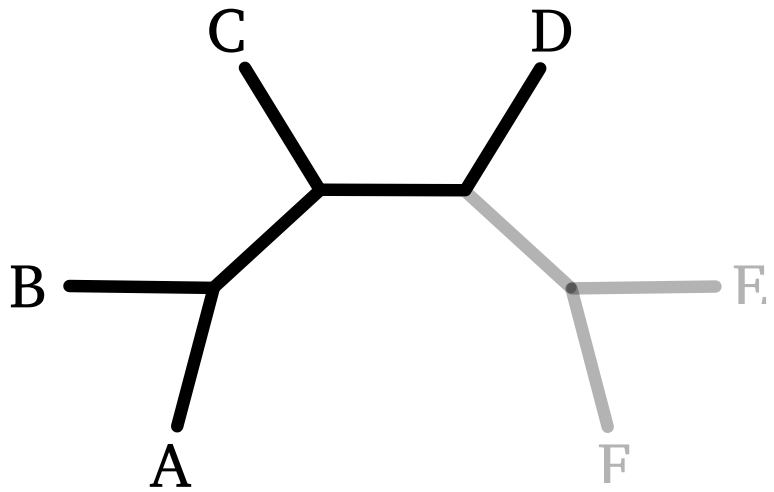
Quartet decomposition

- List all quartets compatible with a tree



Quartet decomposition

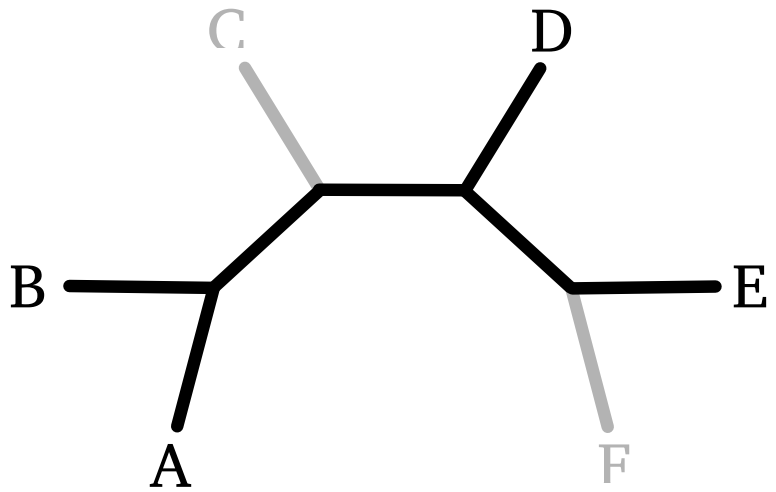
- List all quartets compatible with a tree



AB | CD

Quartet decomposition

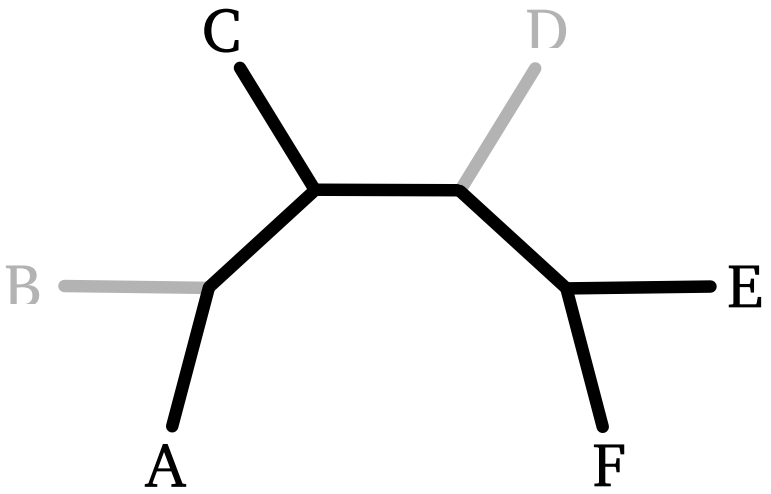
- List all quartets compatible with a tree



AB | CD
AB | DE

Quartet decomposition

- List all quartets compatible with a tree

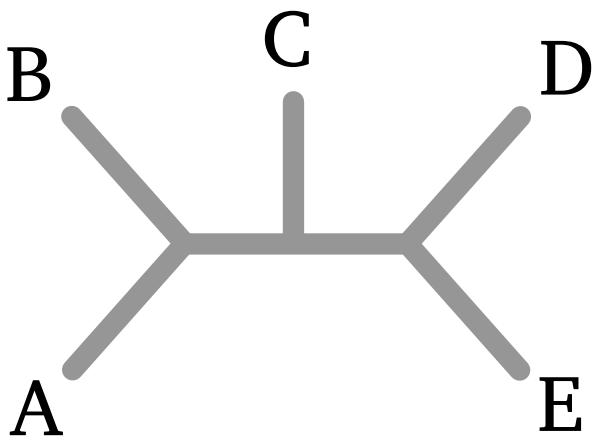


AB | CD
AB | DE
AC | EF
...

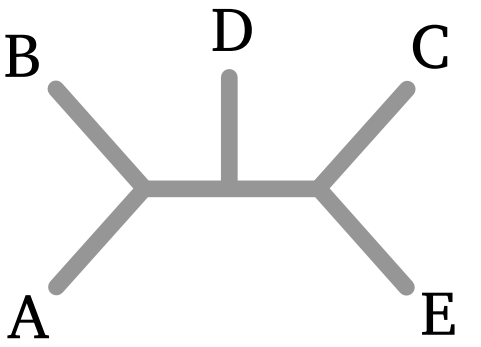
ASTRAL

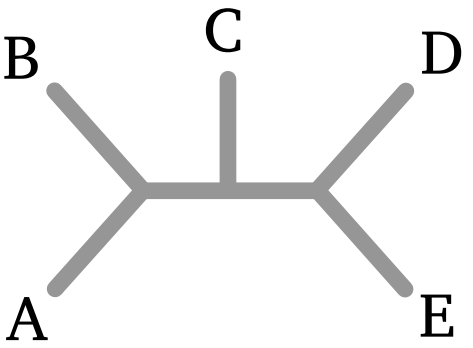
- Quartet score: number of quartets of the gene trees that “agree” with the species tree
- ASTRAL looks for the species tree with the highest quartet score
- Output: unrooted species tree
- Branch lengths: in coalescent units

Quartet score example



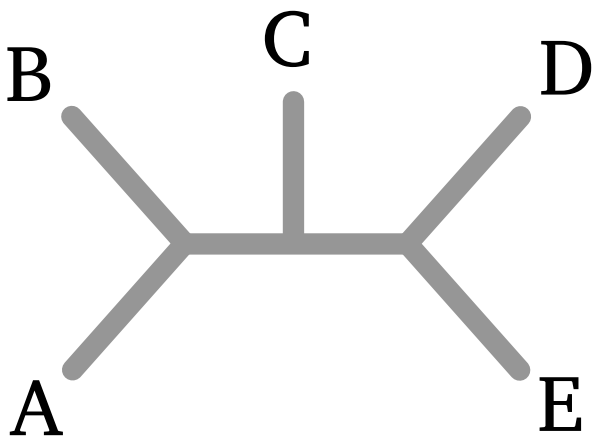
Species tree





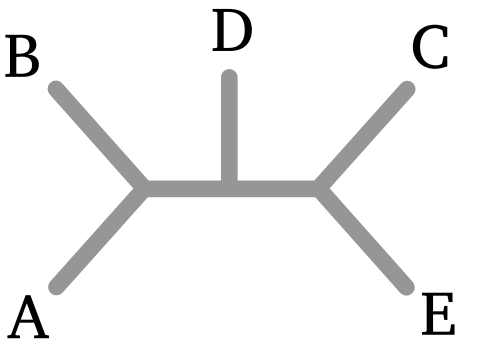
Gene trees

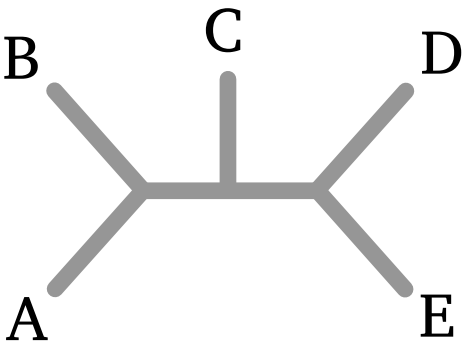
Quartet score example



Species tree

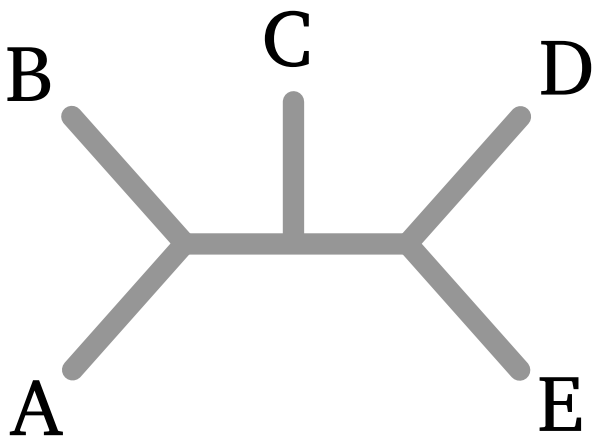
Quartets:
AB|CD
AB|DE
AB|CE
BC|DE
AC|DE





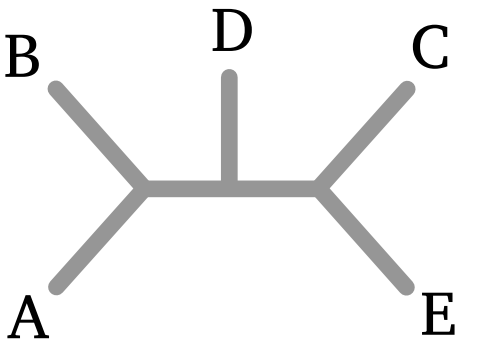
Gene trees

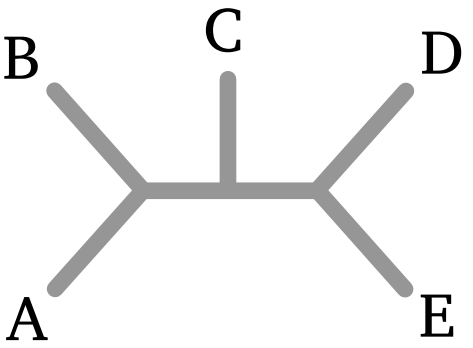
Quartet score example



Species tree

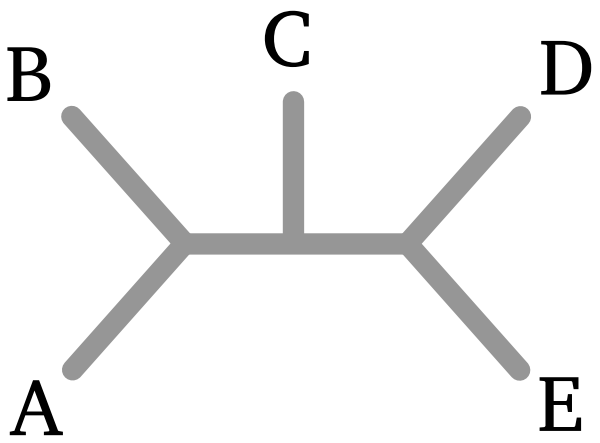
Quartets:	g1	g2
AB|CD	ok	ok
AB|DE	ok	ok
AB|CE	ok	ok
BC|DE		ok
AC|DE		ok





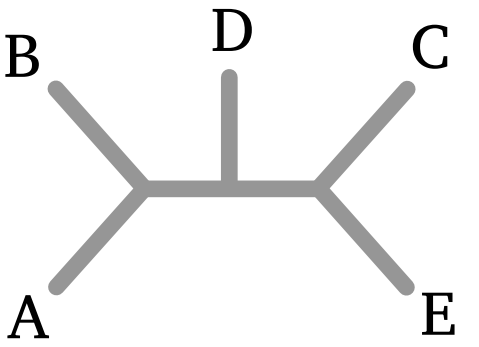
Gene trees

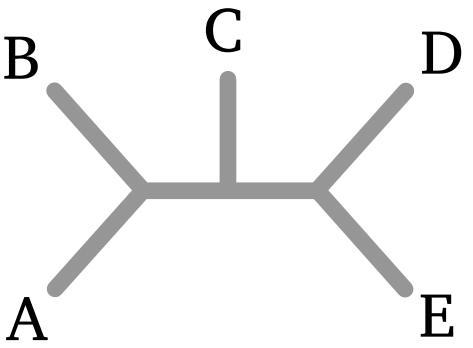
Quartet score example



Species tree

Quartets:	g1	g2
AB|CD	ok	ok
AB|DE	ok	ok
AB|CE	ok	ok
BC|DE		ok
AC|DE		ok
Score =	3 +	5 = 8





Gene trees

ASTRAL: consistent under the MSC

- Intuition: the anomaly zone does not affect quartets!

- ASTRAL is guaranteed to find the true species tree with an infinite number of loci…

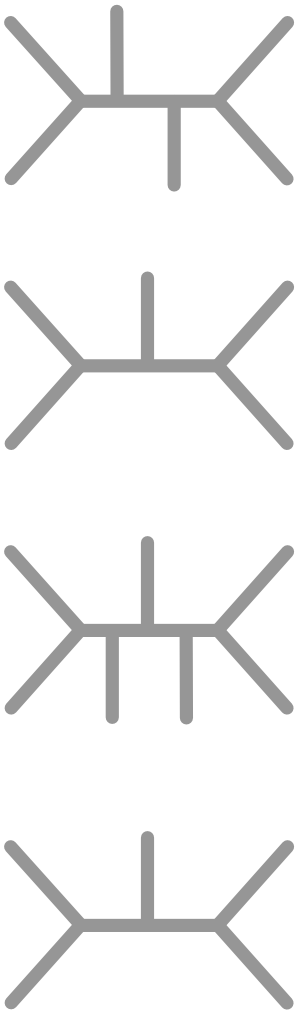
- **… under the assumption that the gene trees are correctly estimated**

Concatenation or ASTRAL?

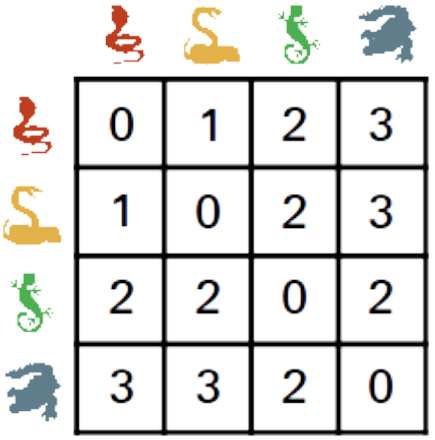
- Concatenation is inconsistent under the MSC
- ASTRAL is sensitive to gene tree error

- → try both and report both trees!

ASTRID [Vachaspati 2015]

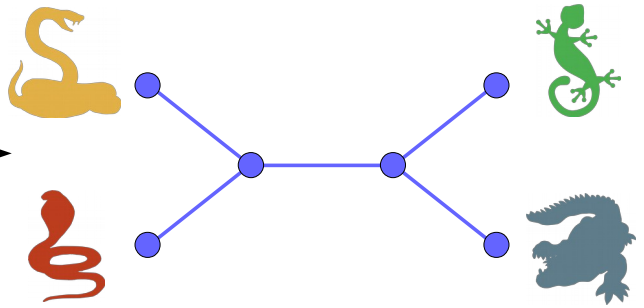


Gene trees



0	1	2	3
1	0	2	3
2	2	0	2
3	3	2	0

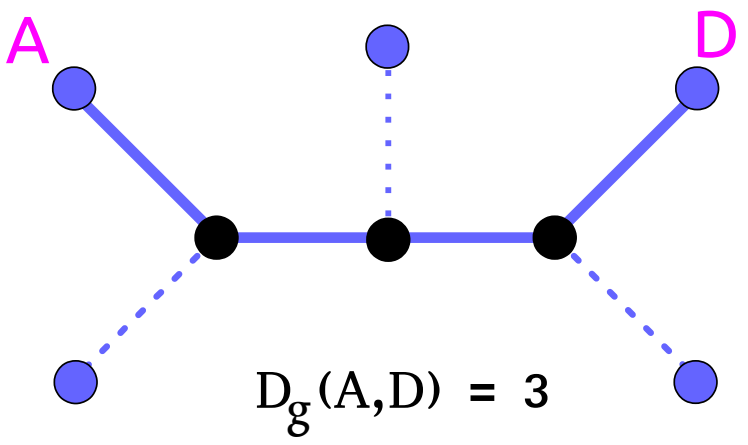
Distance matrix



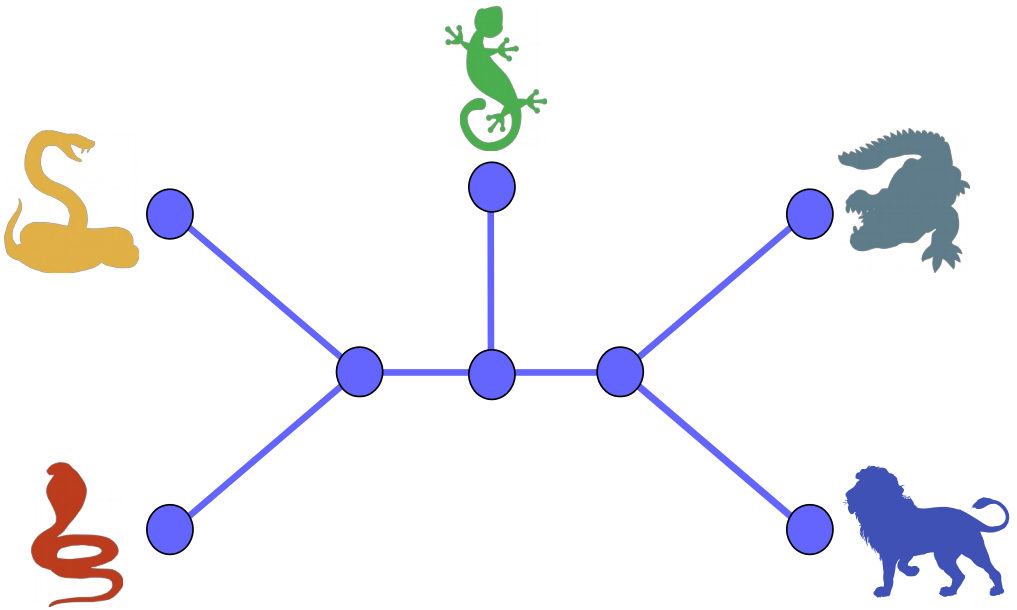
Unrooted species tree

ASTRID

- For each pair of species, for each gene tree, compute the internode distance:

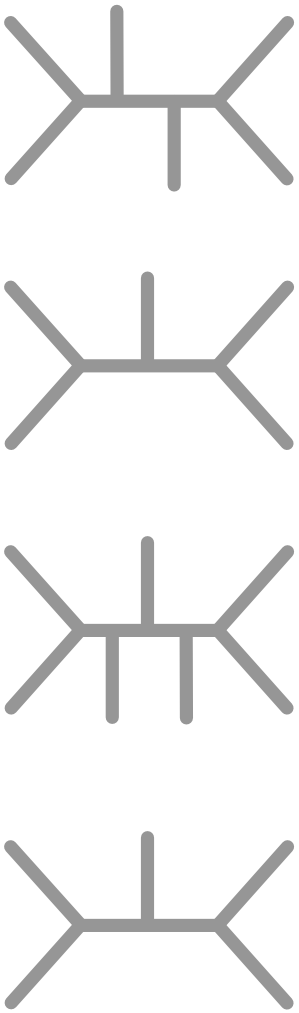


Example

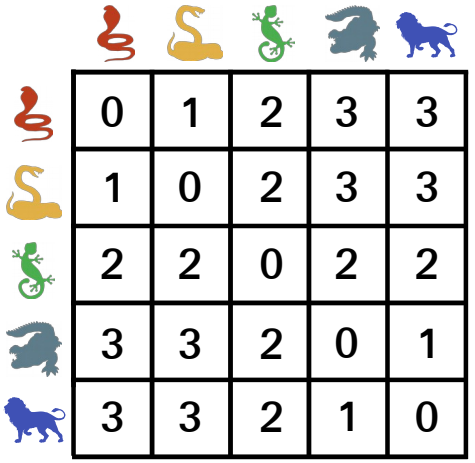


0	1	2	3	3
1	0	2	3	3
2	2	0	2	2
3	3	2	0	1
3	3	2	1	0

ASTRID

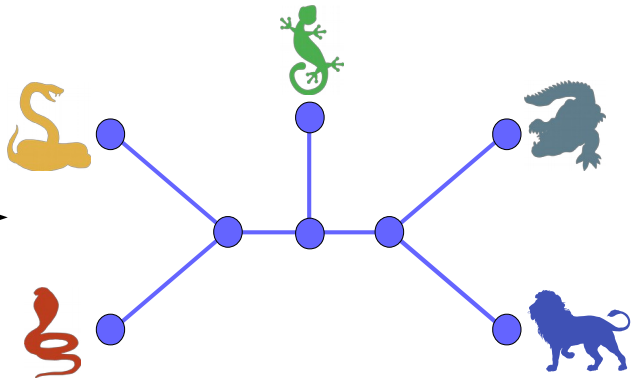


Gene trees



0	1	2	3	3
1	0	2	3	3
2	2	0	2	2
3	3	2	0	1
3	3	2	1	0

Distance matrix



Unrooted species tree

ASTRID consistency

- ASTRID is guaranteed to find the true species tree with an infinite number of loci…

- **… under the assumption that the gene trees are correctly estimated**

ASTRID consistency

- ASTRID is guaranteed to find the true species tree with an infinite number of loci…

- **… under the assumption that the gene trees are correctly estimated**
- **… and under the assumption that the gene trees are complete**

Missing data

- Family with missing data = family that does not cover all species
- Can be the result of:
  - Gene loss
  - Unsampled genes
  - Errors in gene family clustering
  - Filtering step in the pipeline

Effects of missing data

- Missing data can affect species tree reconstruction by:
  - Reducing the amount of information in the dataset
  - Misleading some methods (introduction of a bias)

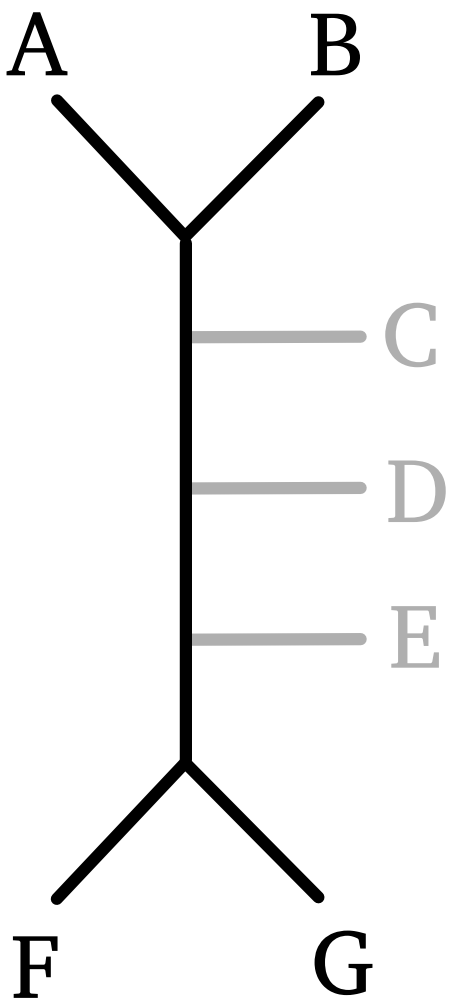
ASTRID and missing data bias

- ASTRID counts the number of nodes between each pair of genes
- If some genes are missing, ASTRID underestimates distances
- This might bias species tree inference

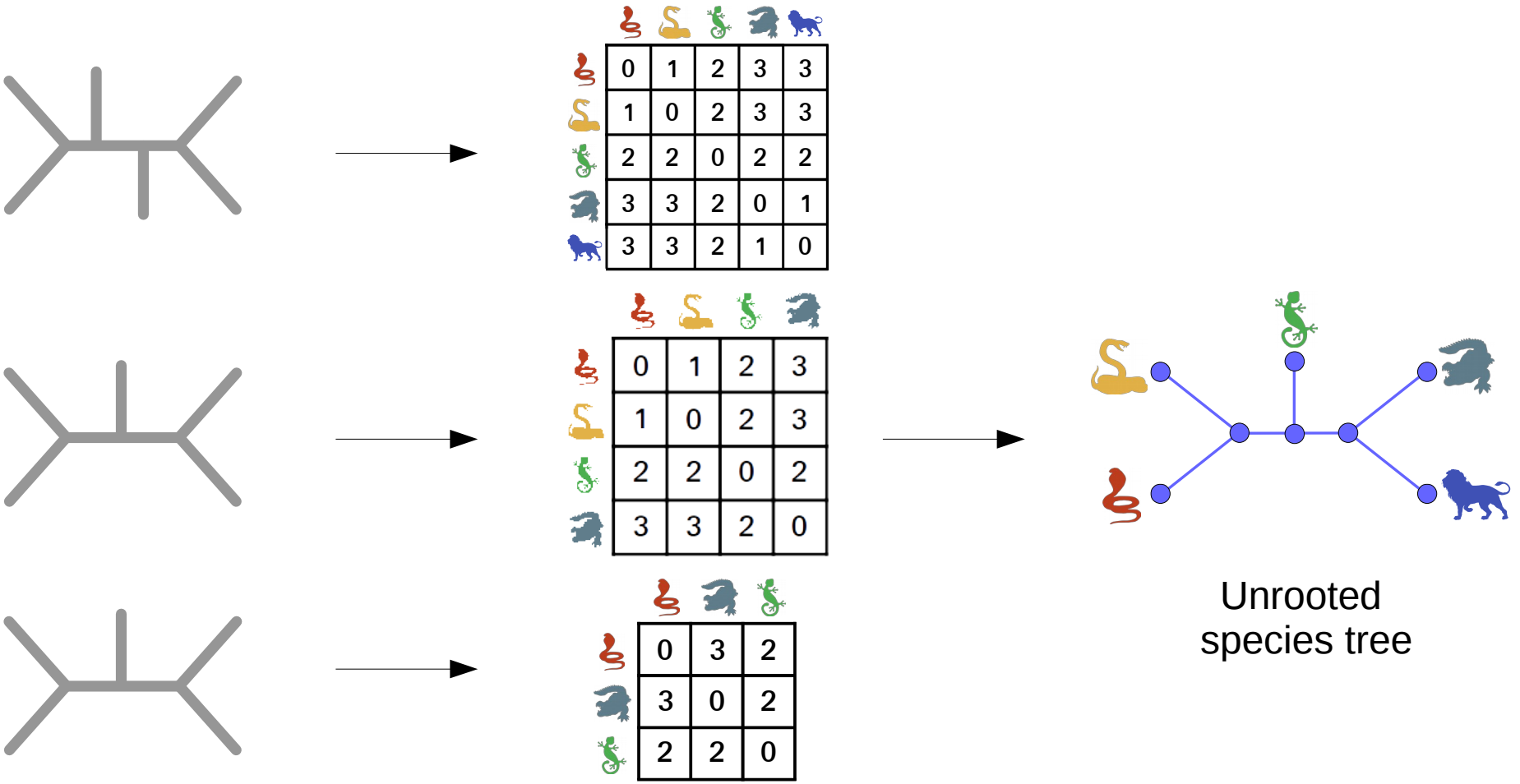
ASTRID and missing data bias

In this example, the genes C, D, and E were not sampled. This affects the distance between A and G:

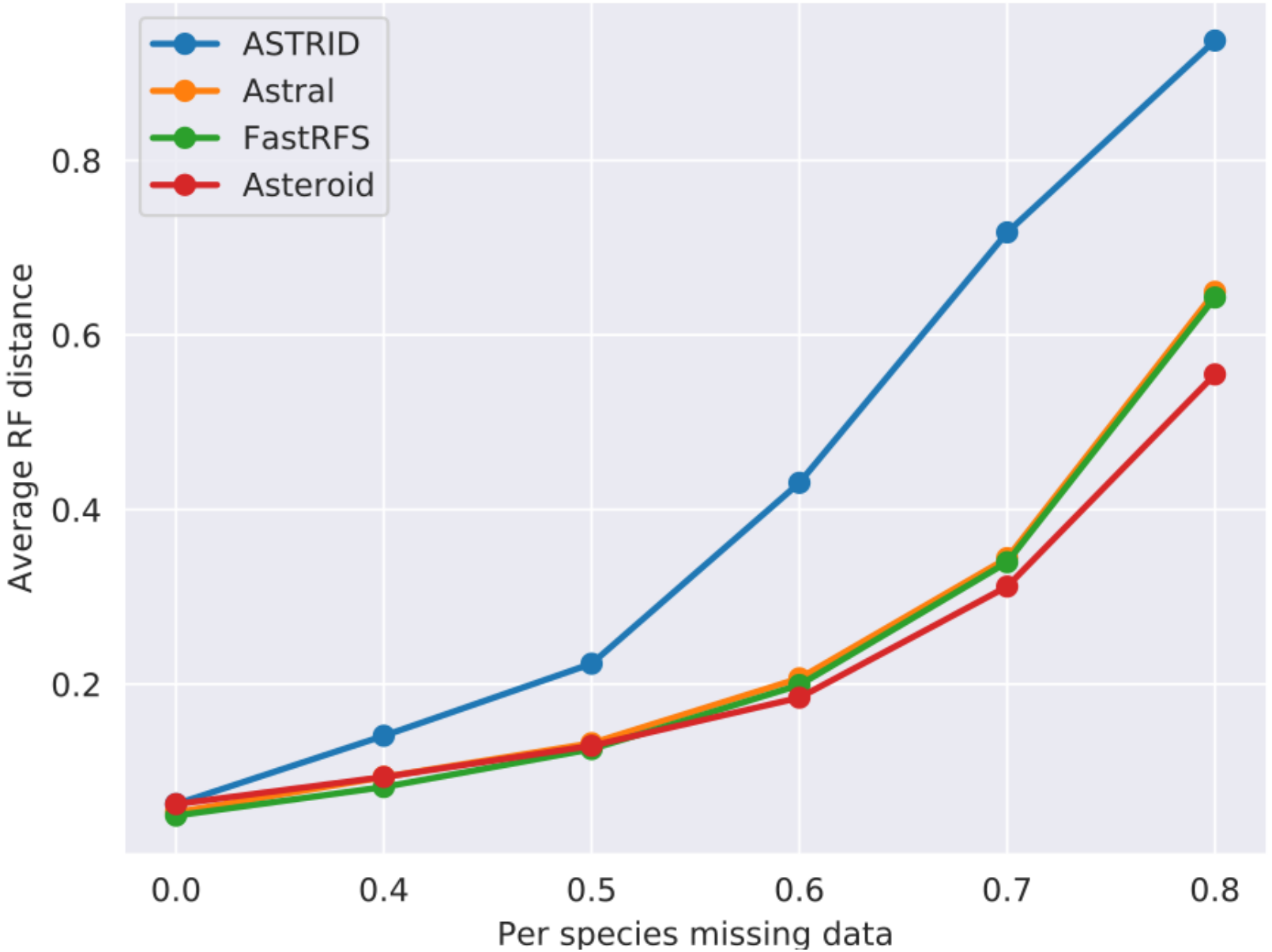
- Real D(A,G) = 5
- Observed D(A,G) = 2



Asteroid: robust to extreme levels of missing data [Morel 2023]



Accuracy for extreme levels of missing data



Single-copy methods, summary

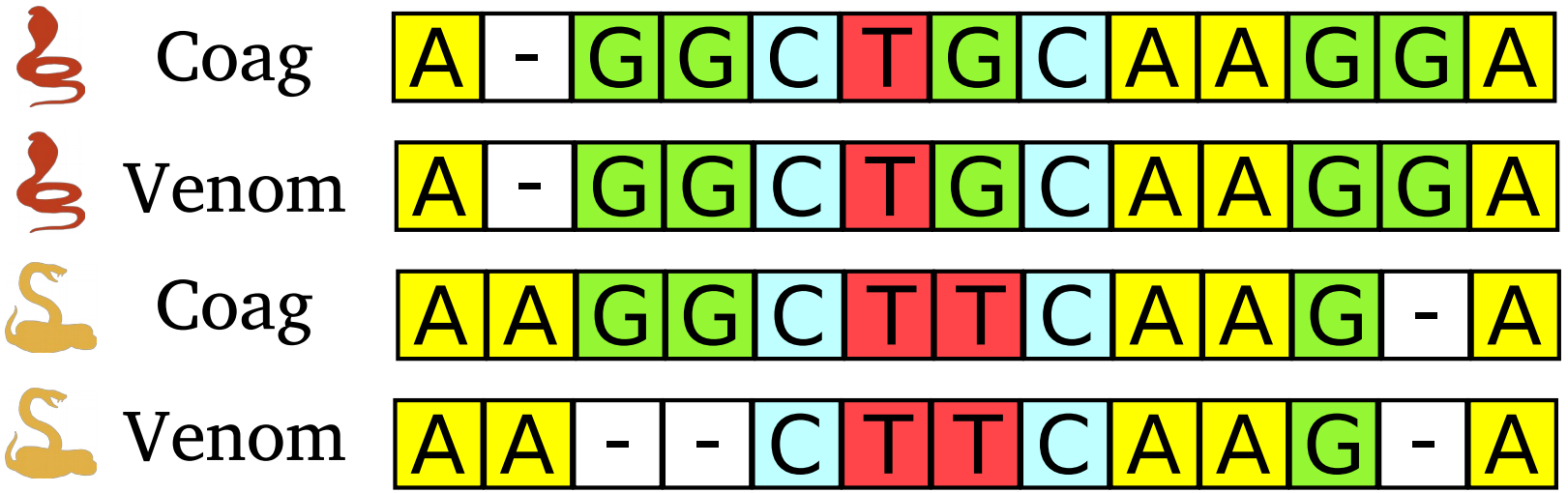
- Concatenation:
  - Distance methods
  - Parsimony methods
  - Probabilistic methods
- Gene tree methods:
  - ASTRAL (quartet method)
  - ASTRID (distance method)
  - Asteroid (distance method)
  - … many other excellent methods!

Gene tree species tree co-estimation

- Some method co-estimate the gene trees and the species tree (for instance StarBeast2)
- Those methods are less sensitive to gene tree reconstruction errors
- But they are also slower…

Multi-copy gene families

Multi-copy gene families: the same species can have several genes in the same family



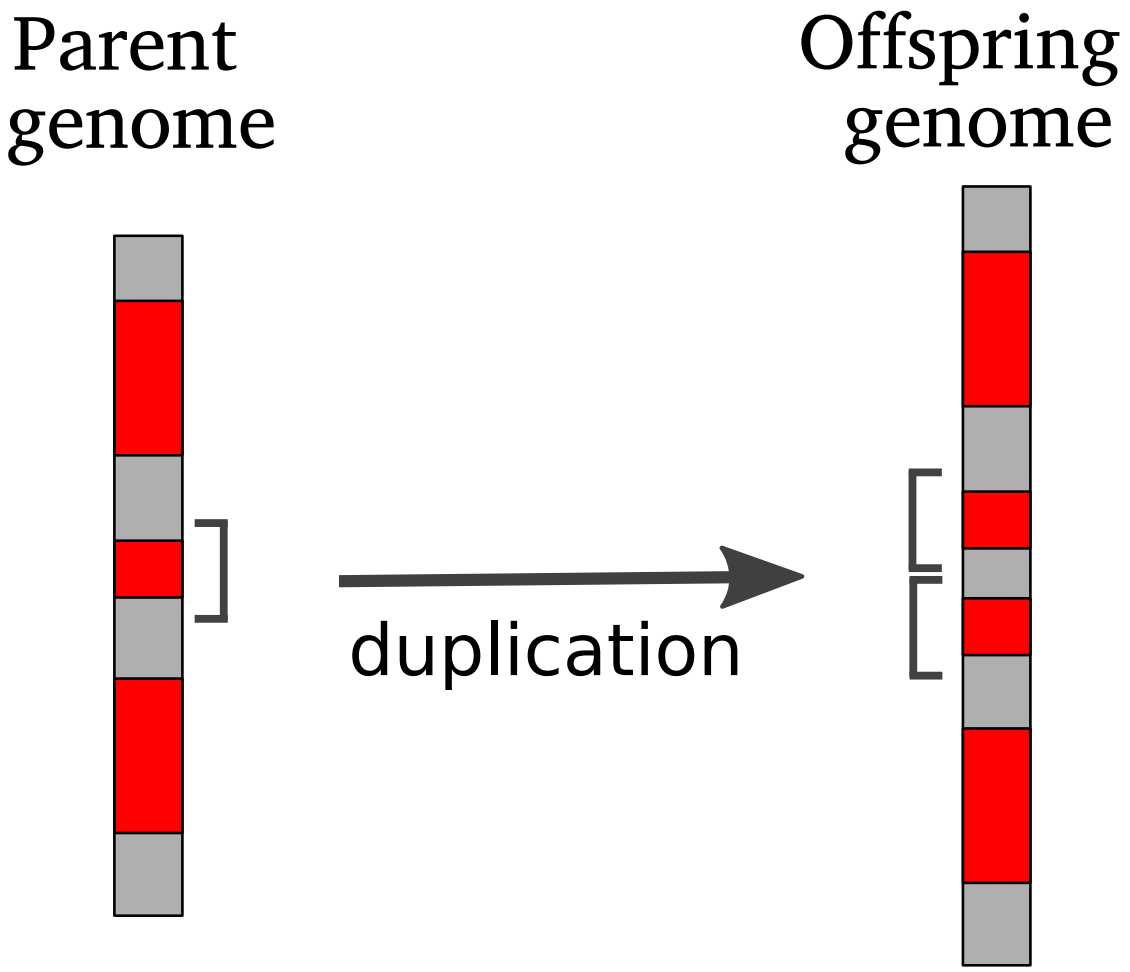
Multi-copy gene families

Causes:

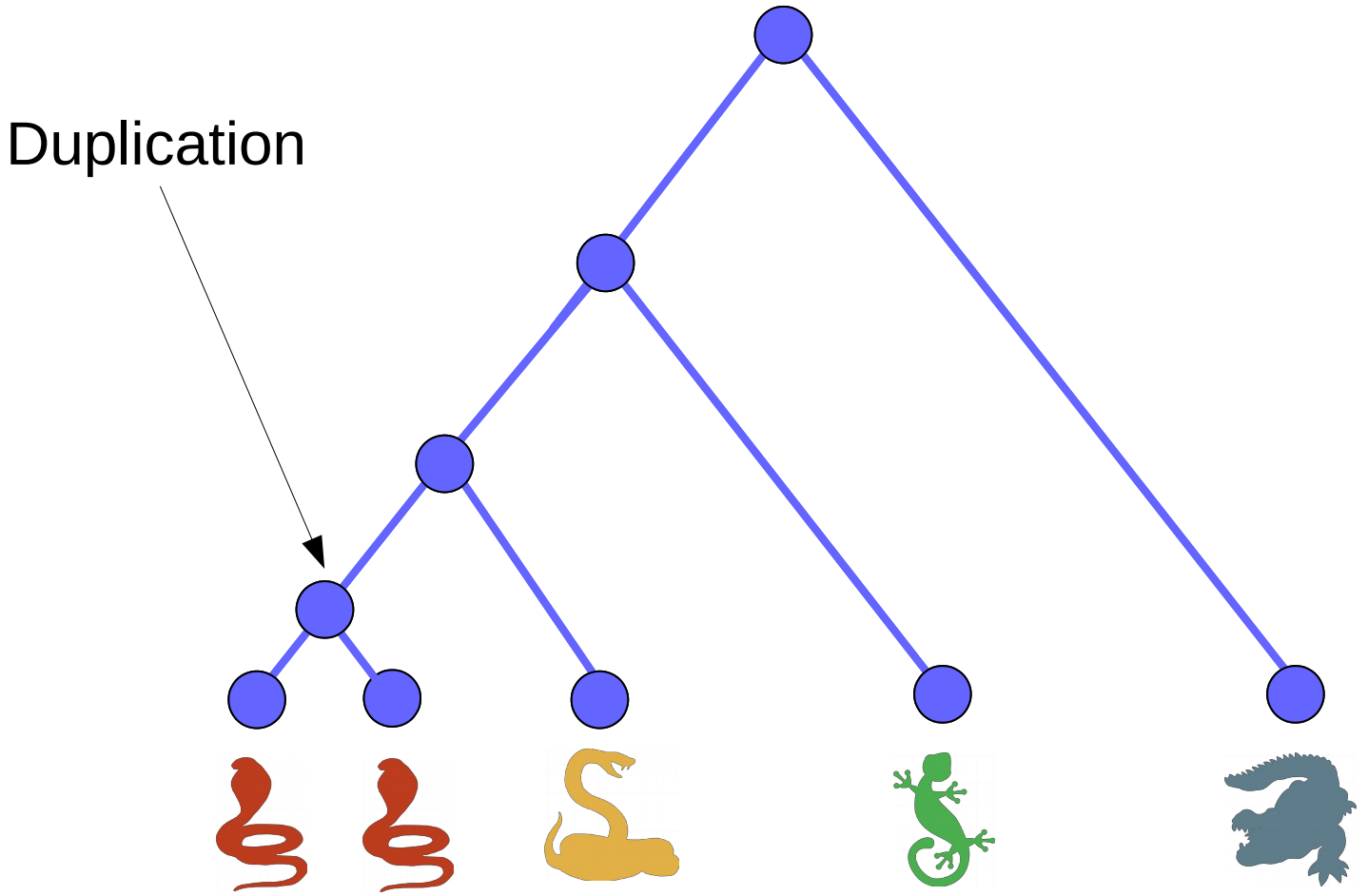
- Gene duplications
- Gene losses
- Horizontal gene transfers

DTL events

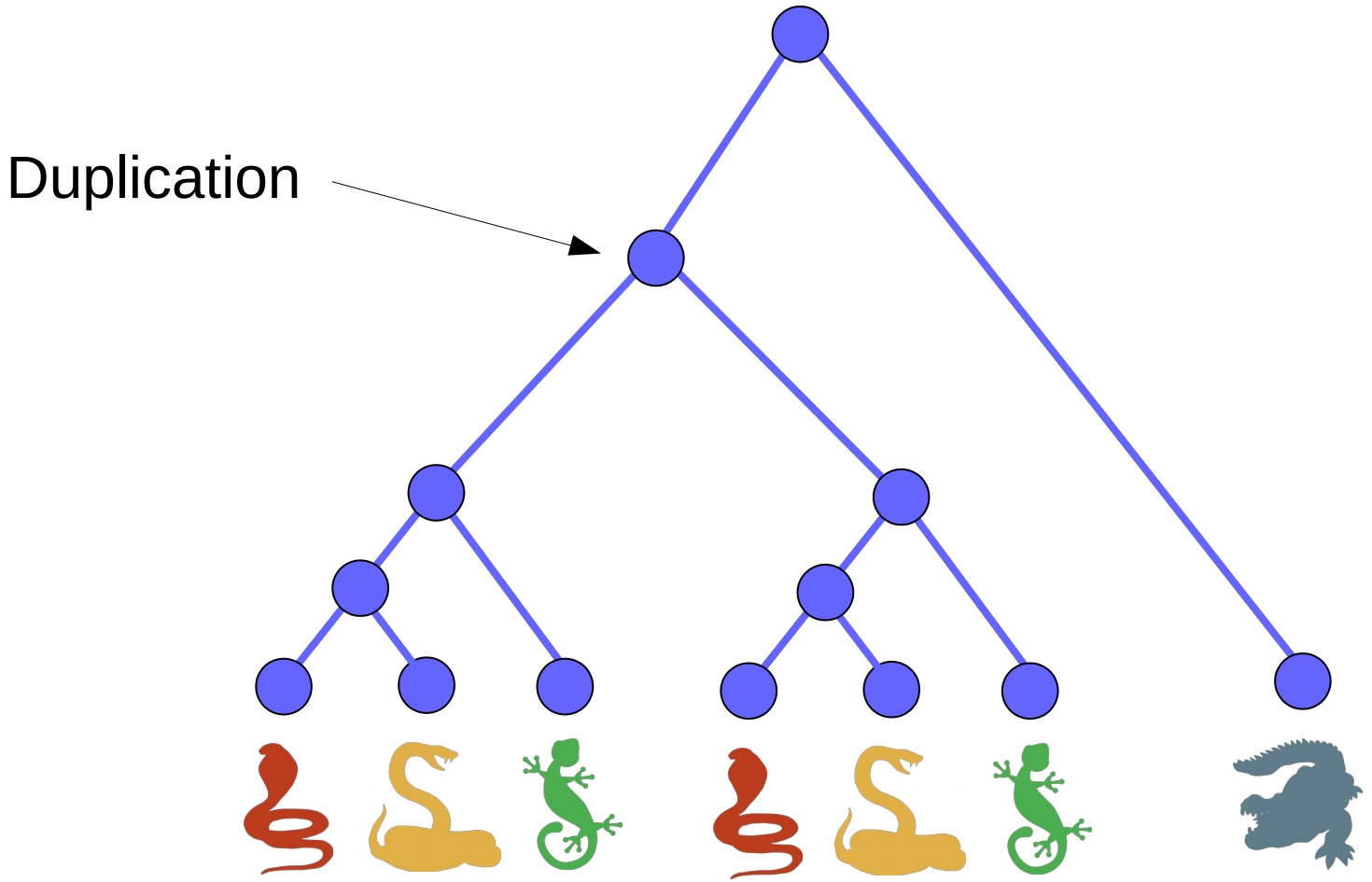
Gene duplication



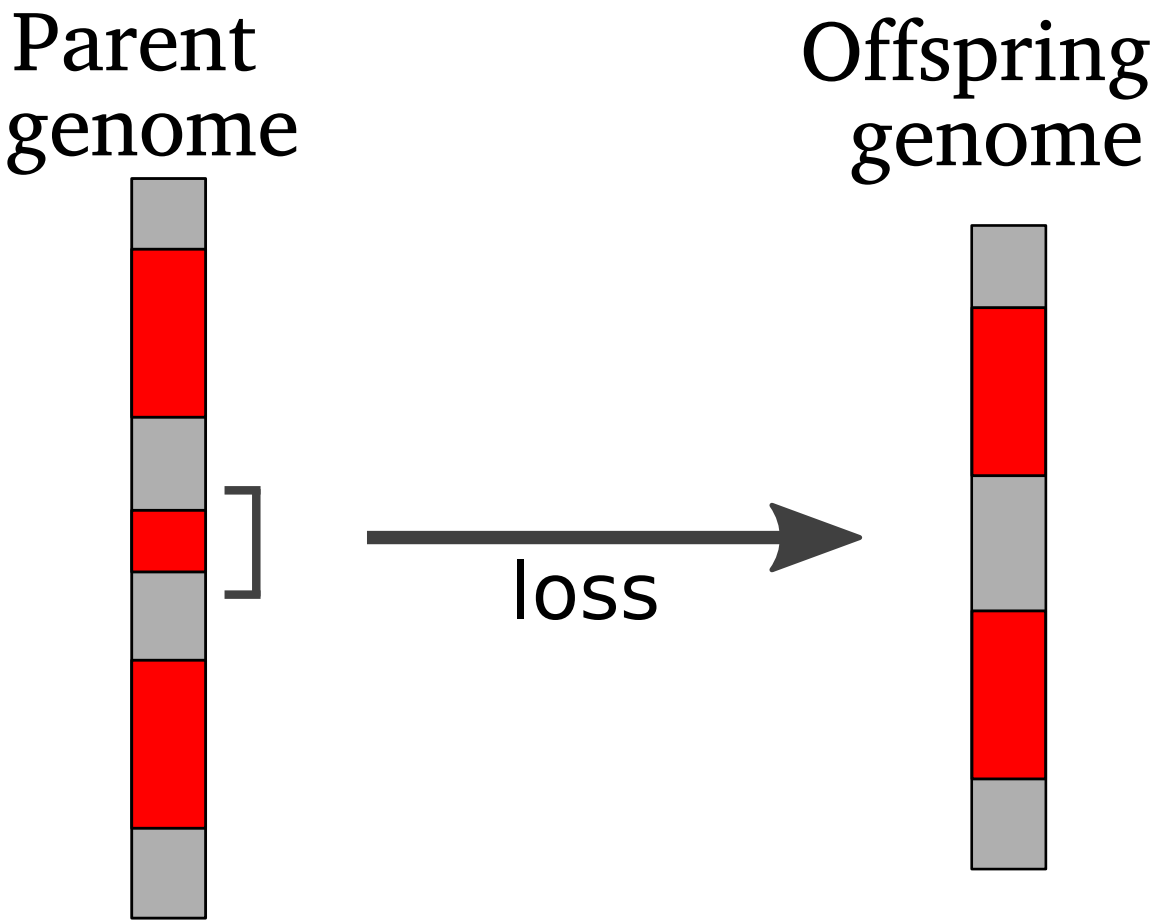
Gene tree and duplication



Gene tree and duplication



Gene loss

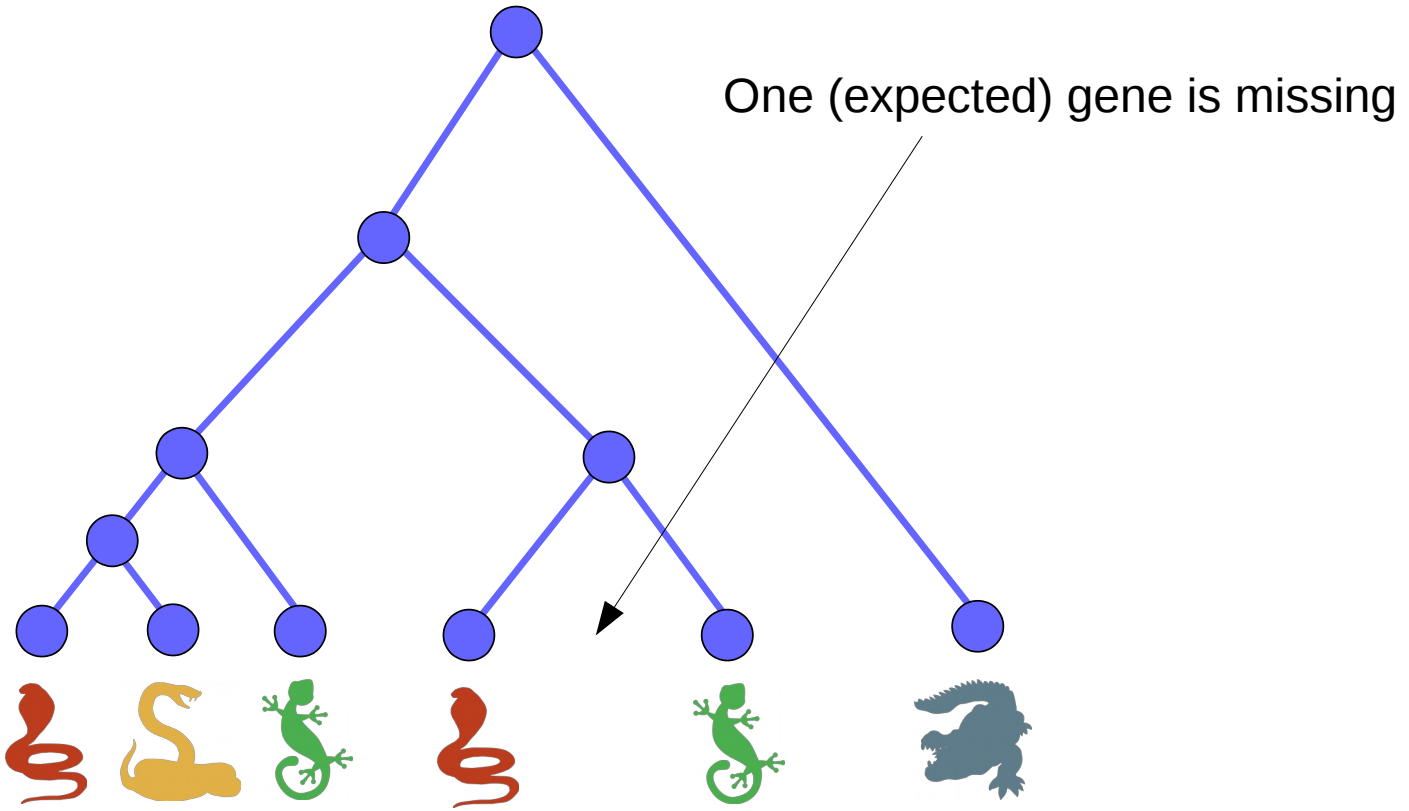


Gene tree and gene loss

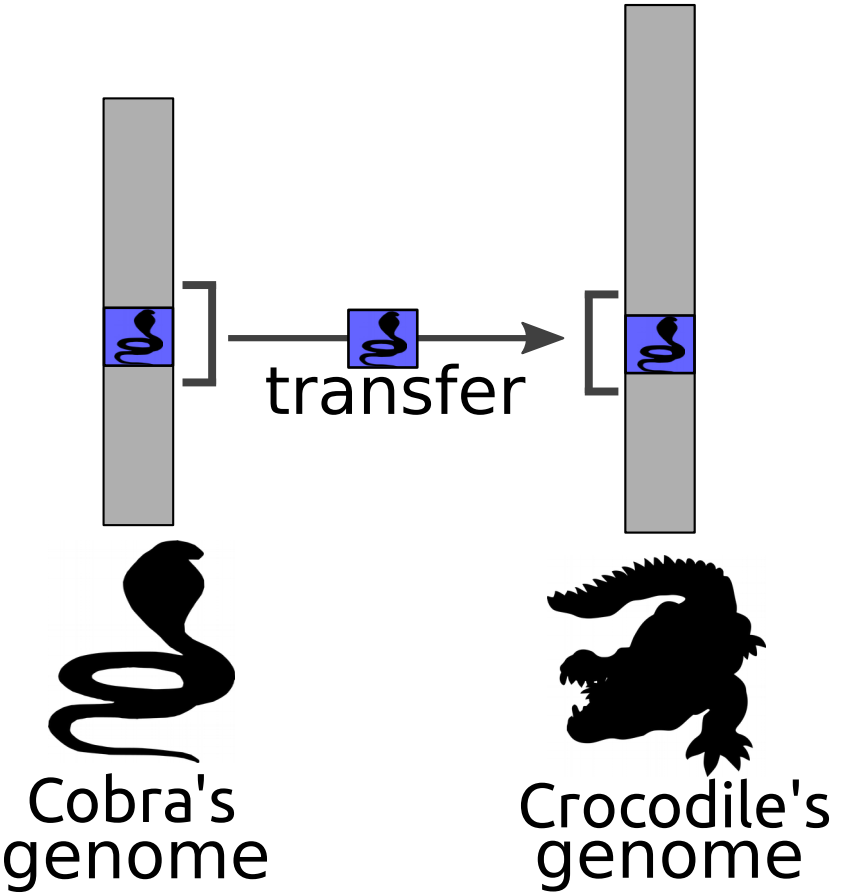
- Obviously, we do not observe lost genes

Gene tree and gene loss

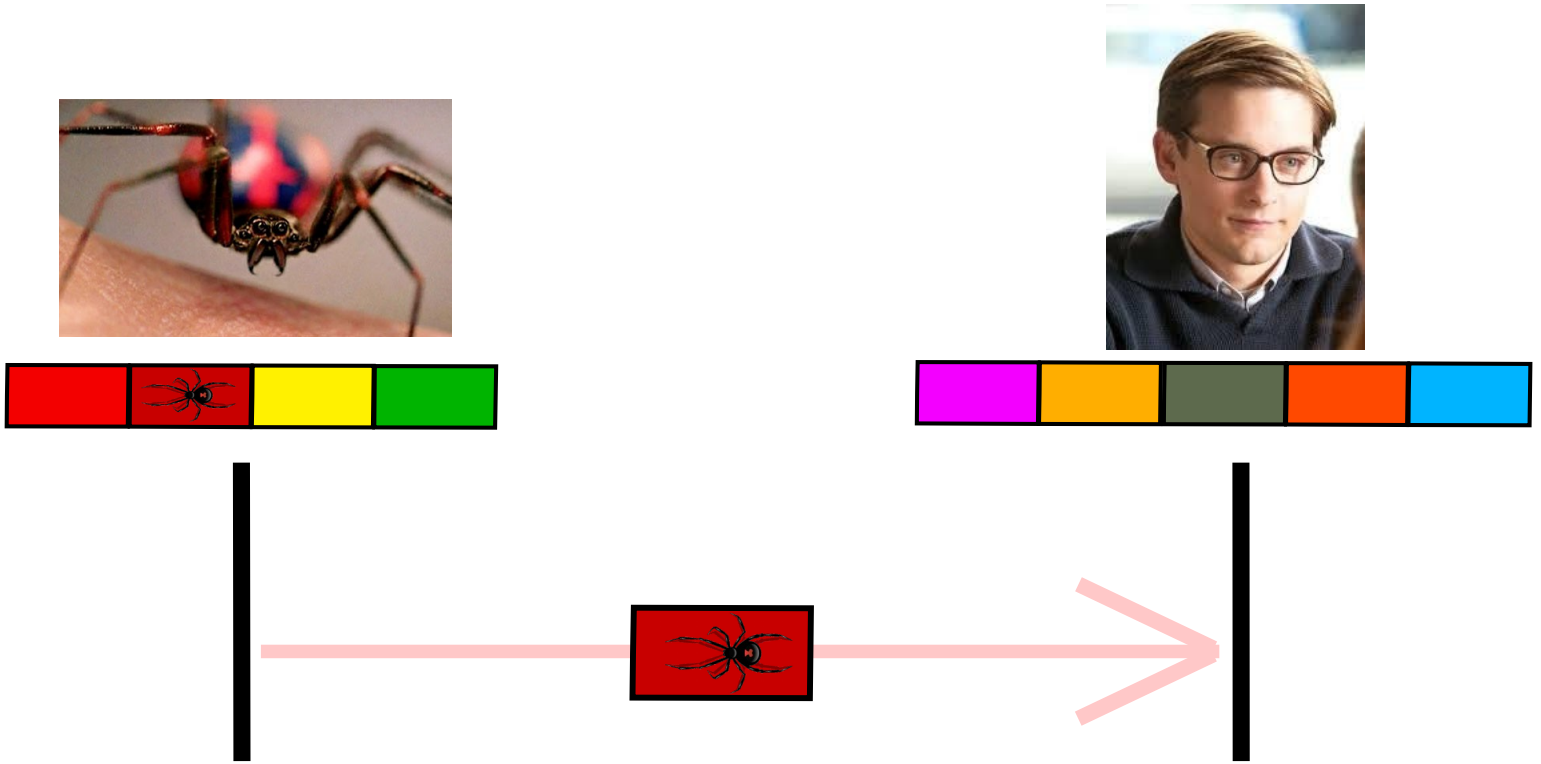
- Obviously, we do not observe lost genes
- But we can infer loss events



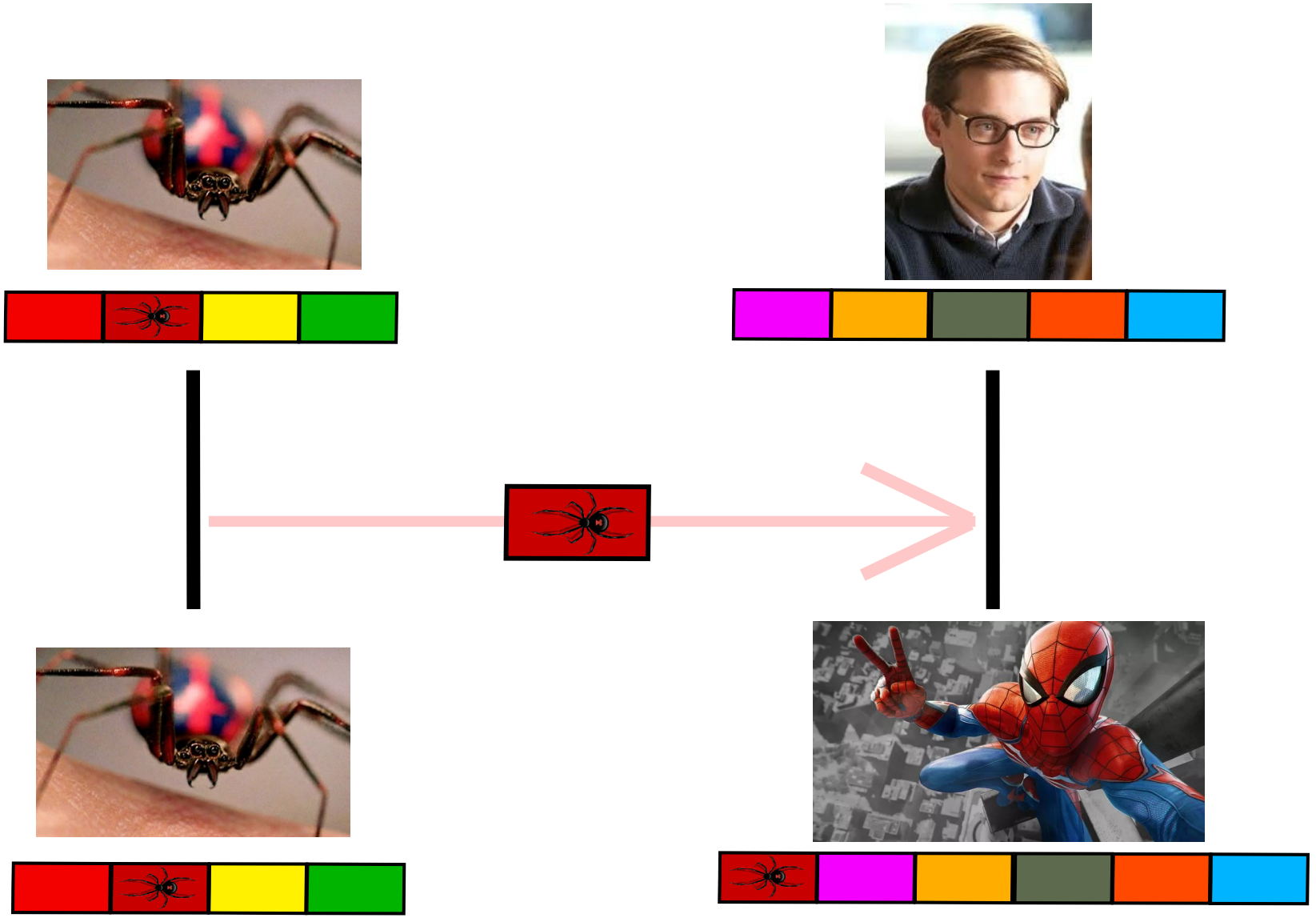
Horizontal gene transfer



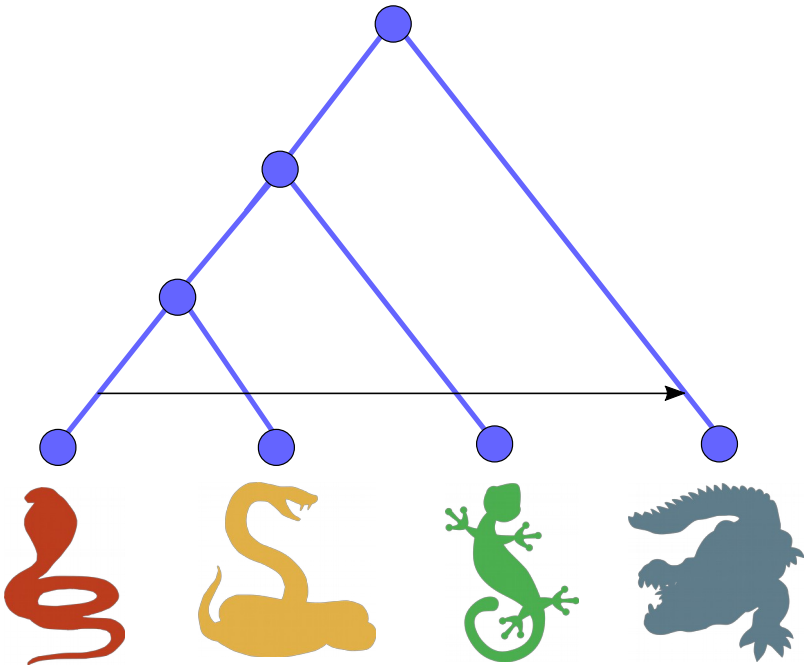
The most famous transfer



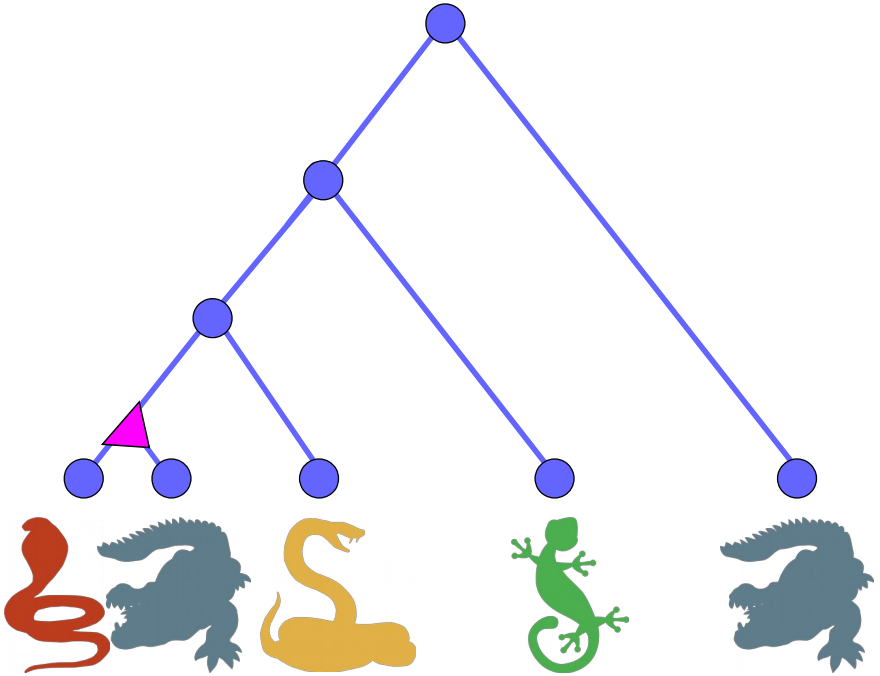
The most famous transfer



Gene tree and HGT



A HGT in the species tree



Resulting gene tree

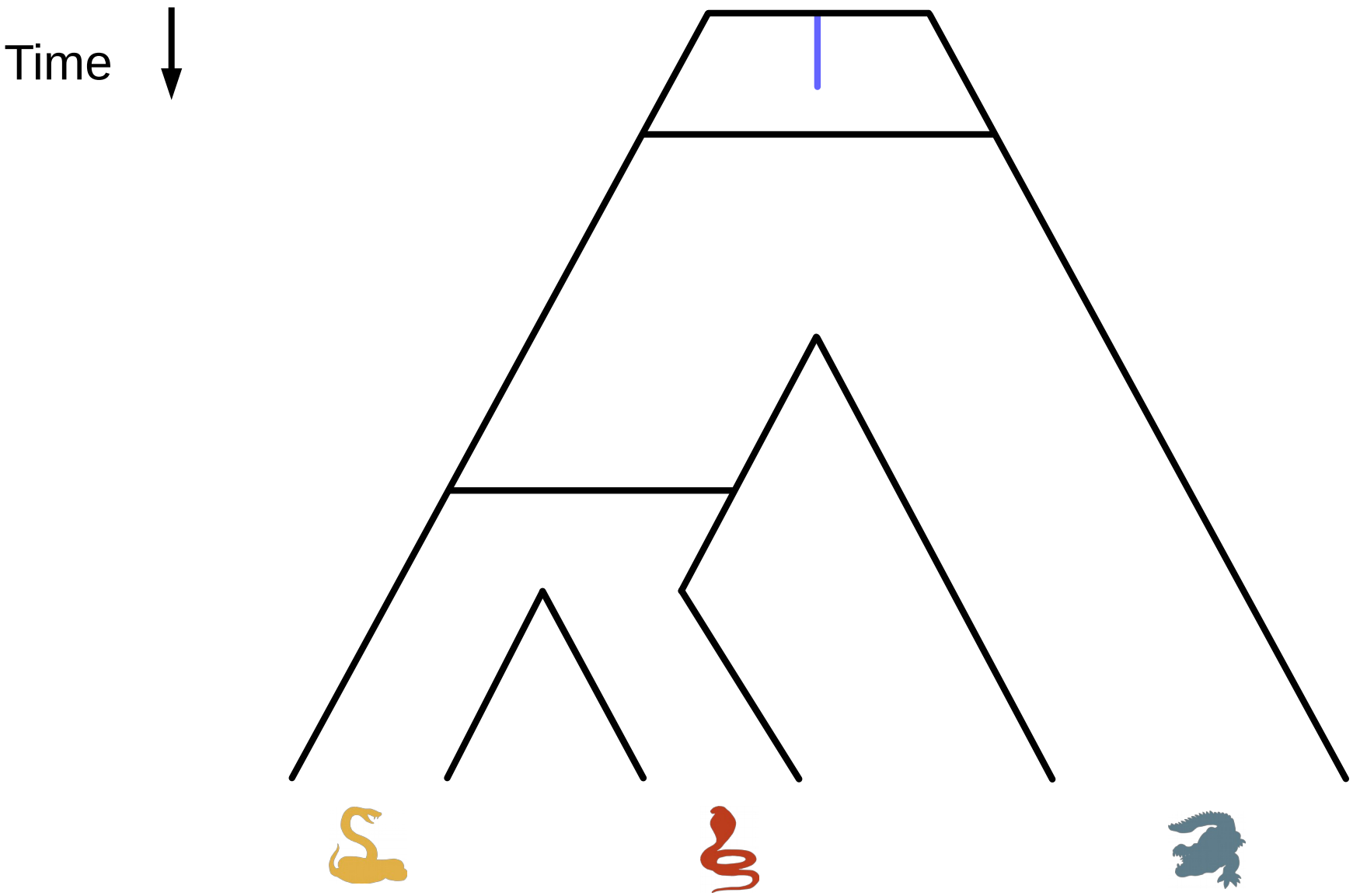
Horizontal gene transfers

- Convenient mechanism to quickly acquire “super powers” from other species

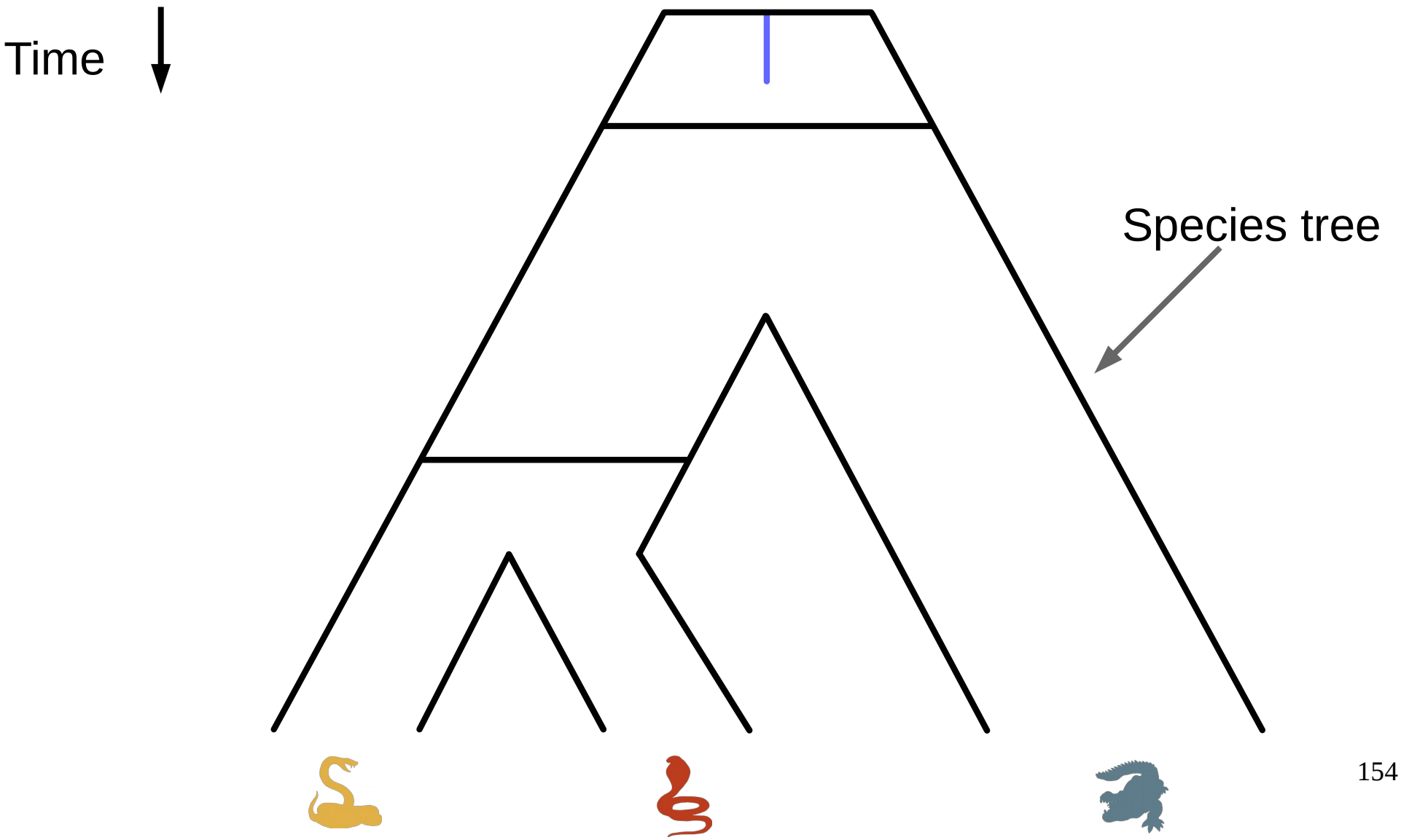
- … But very annoying for species tree inference!

  Cause drastic conflicts between gene trees and the species tree.

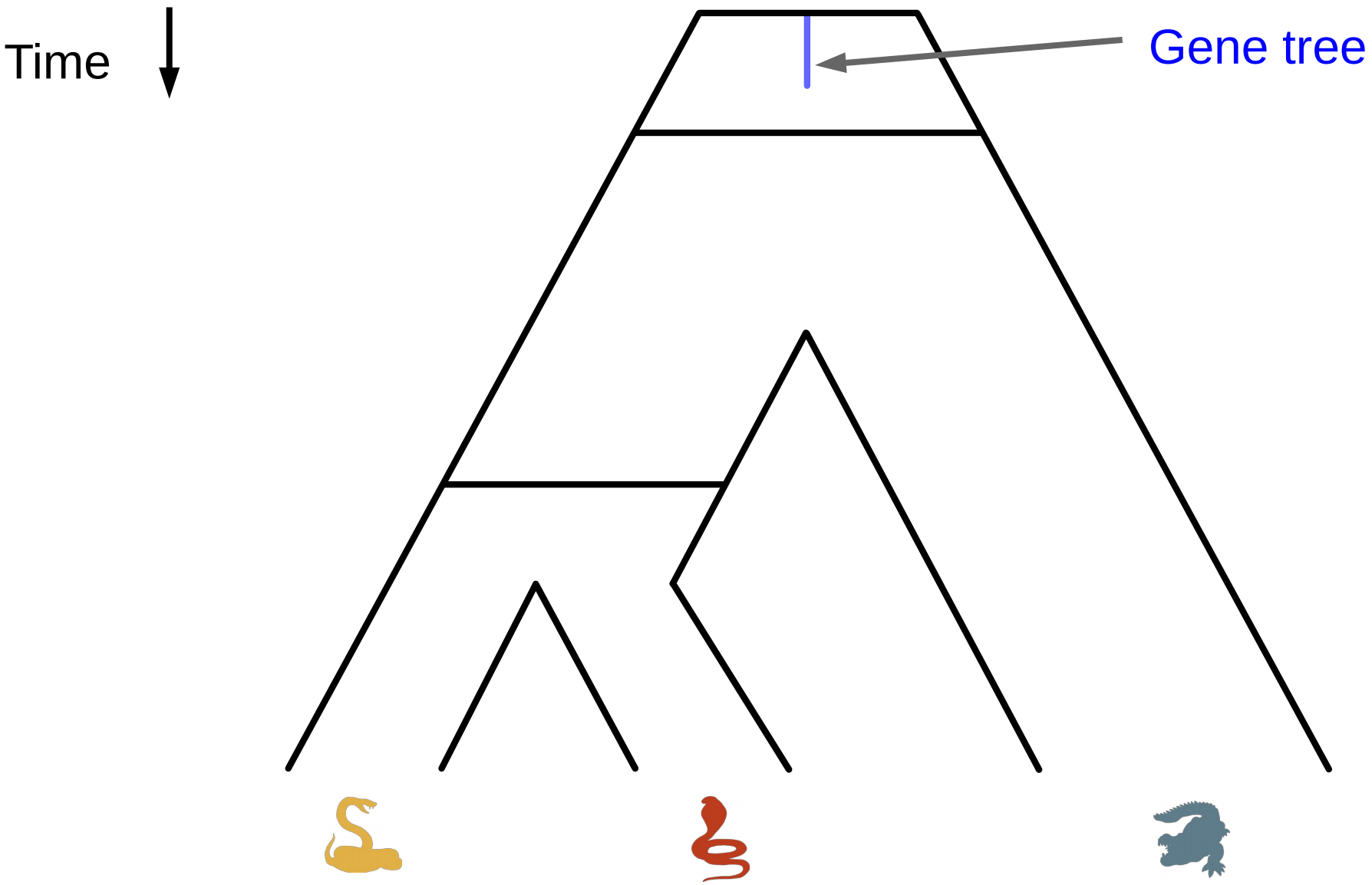
Gene evolution



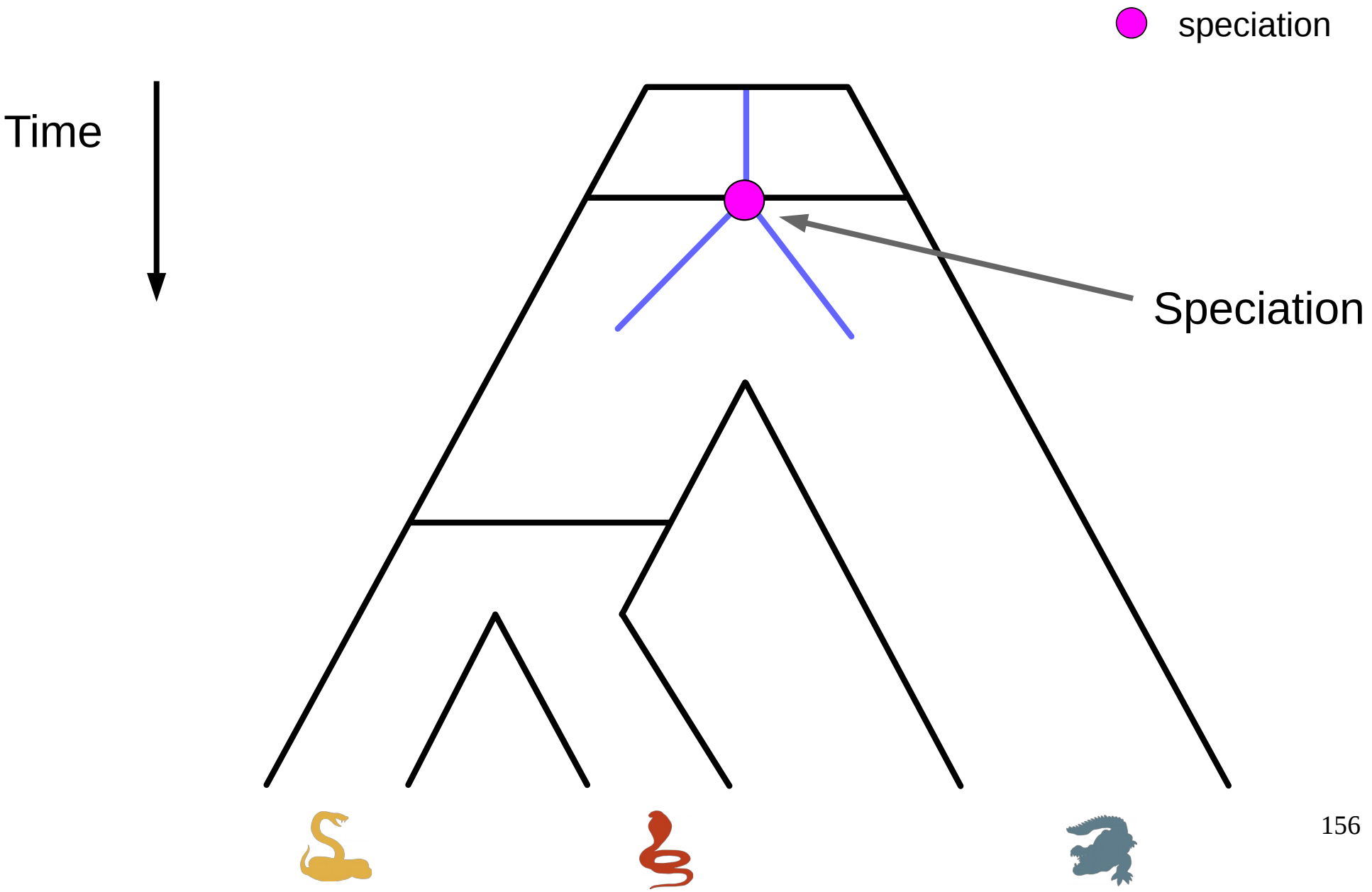
Gene evolution



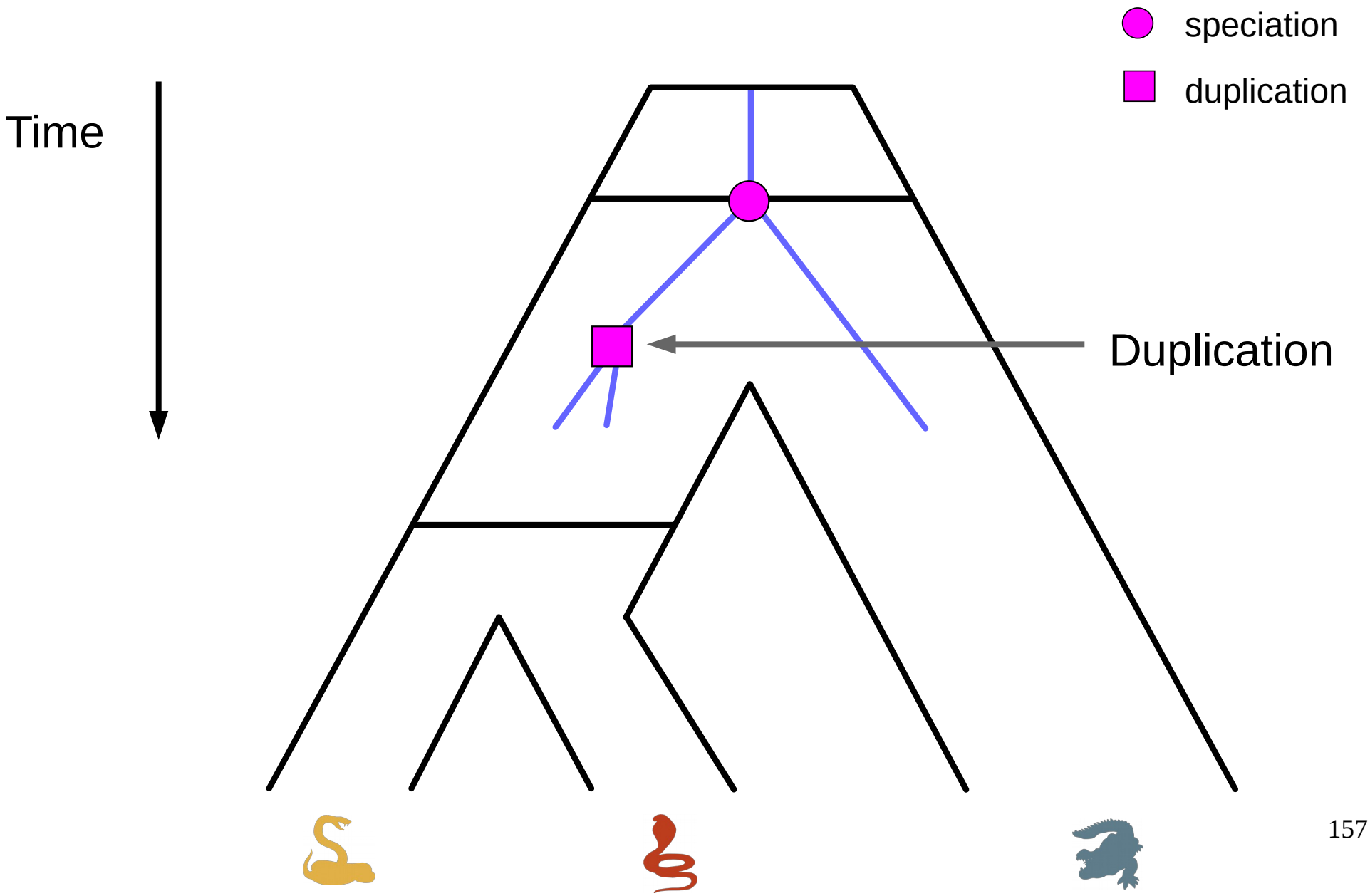
Gene evolution



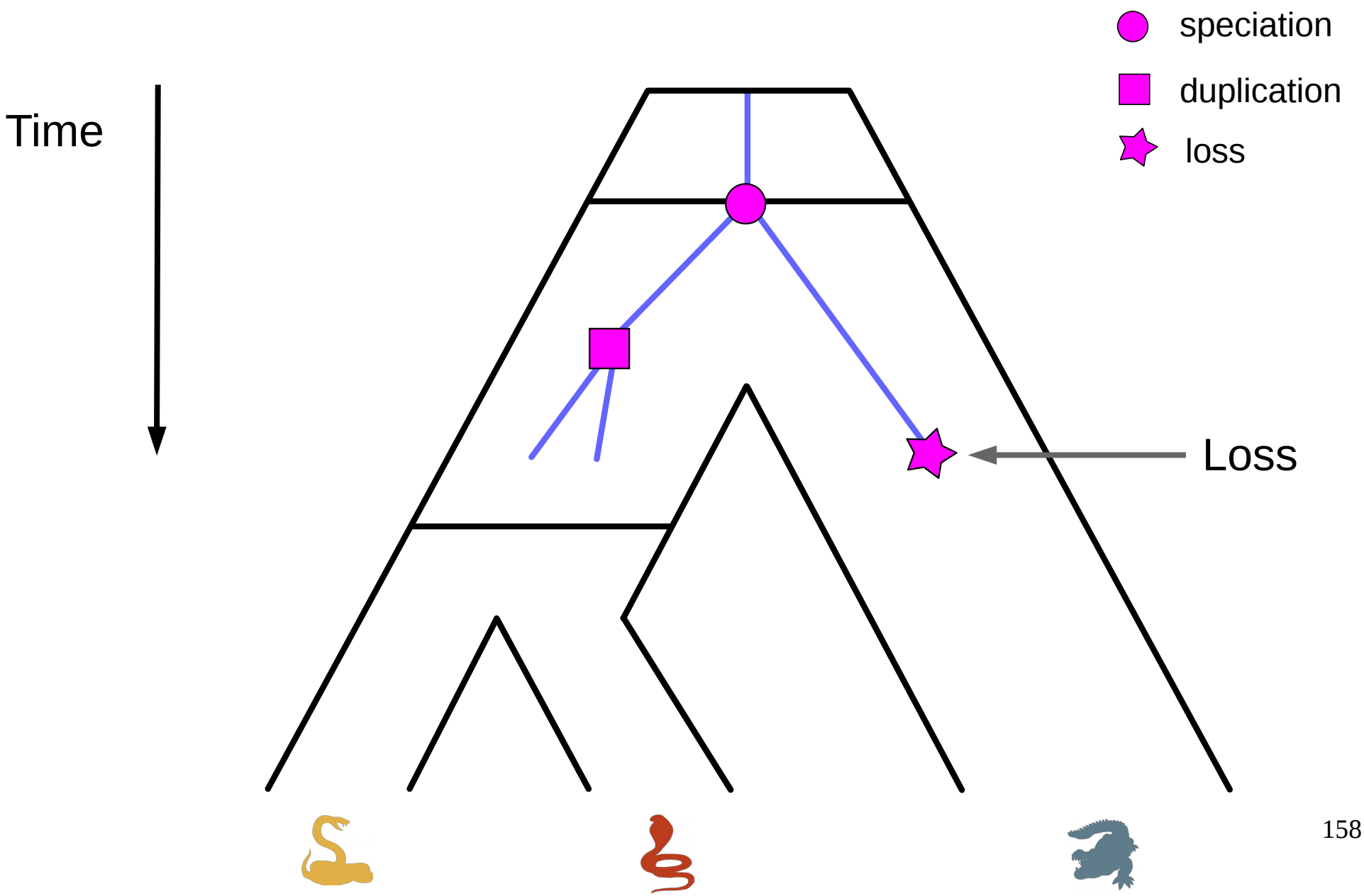
Gene evolution



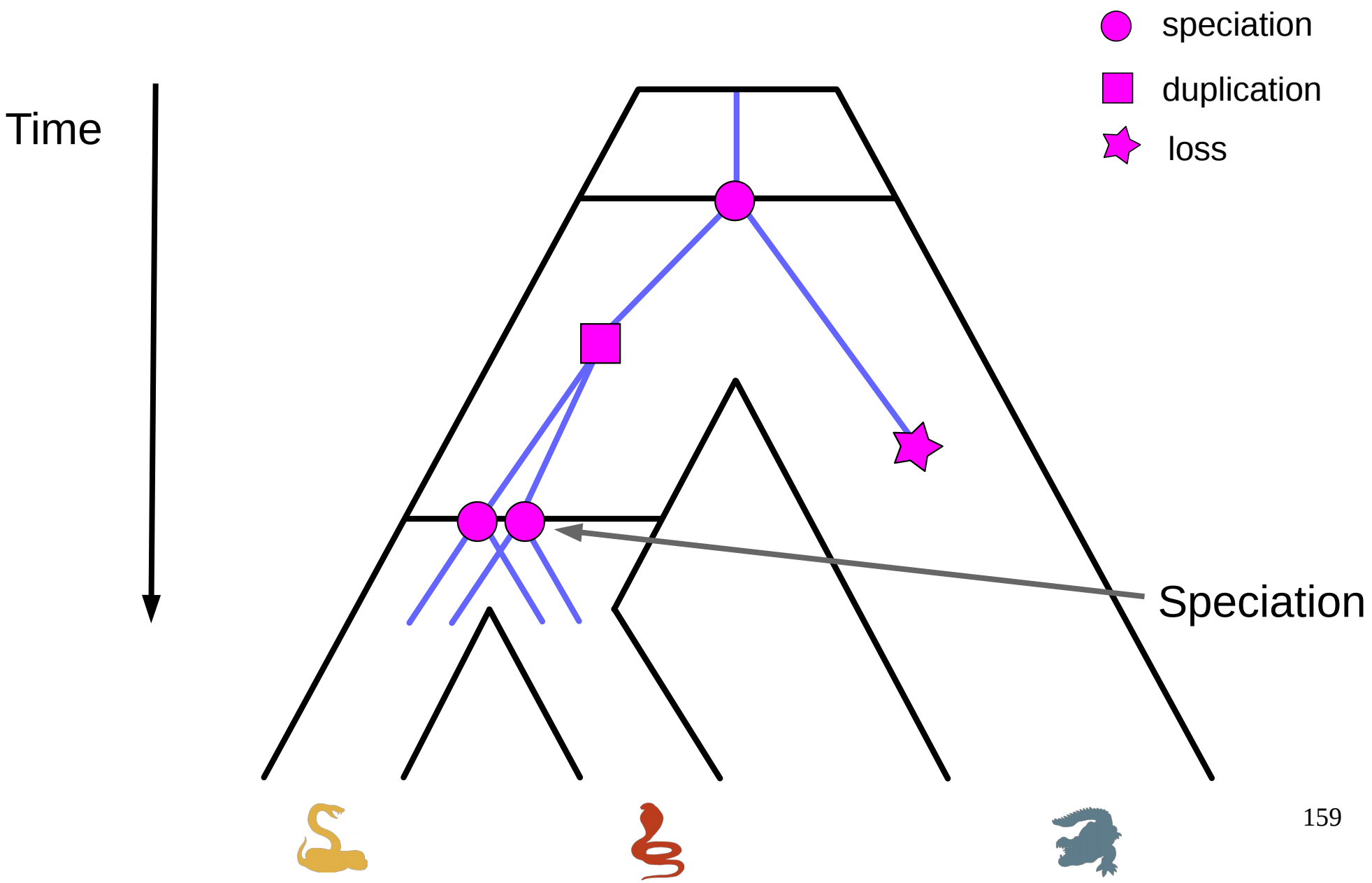
Gene evolution



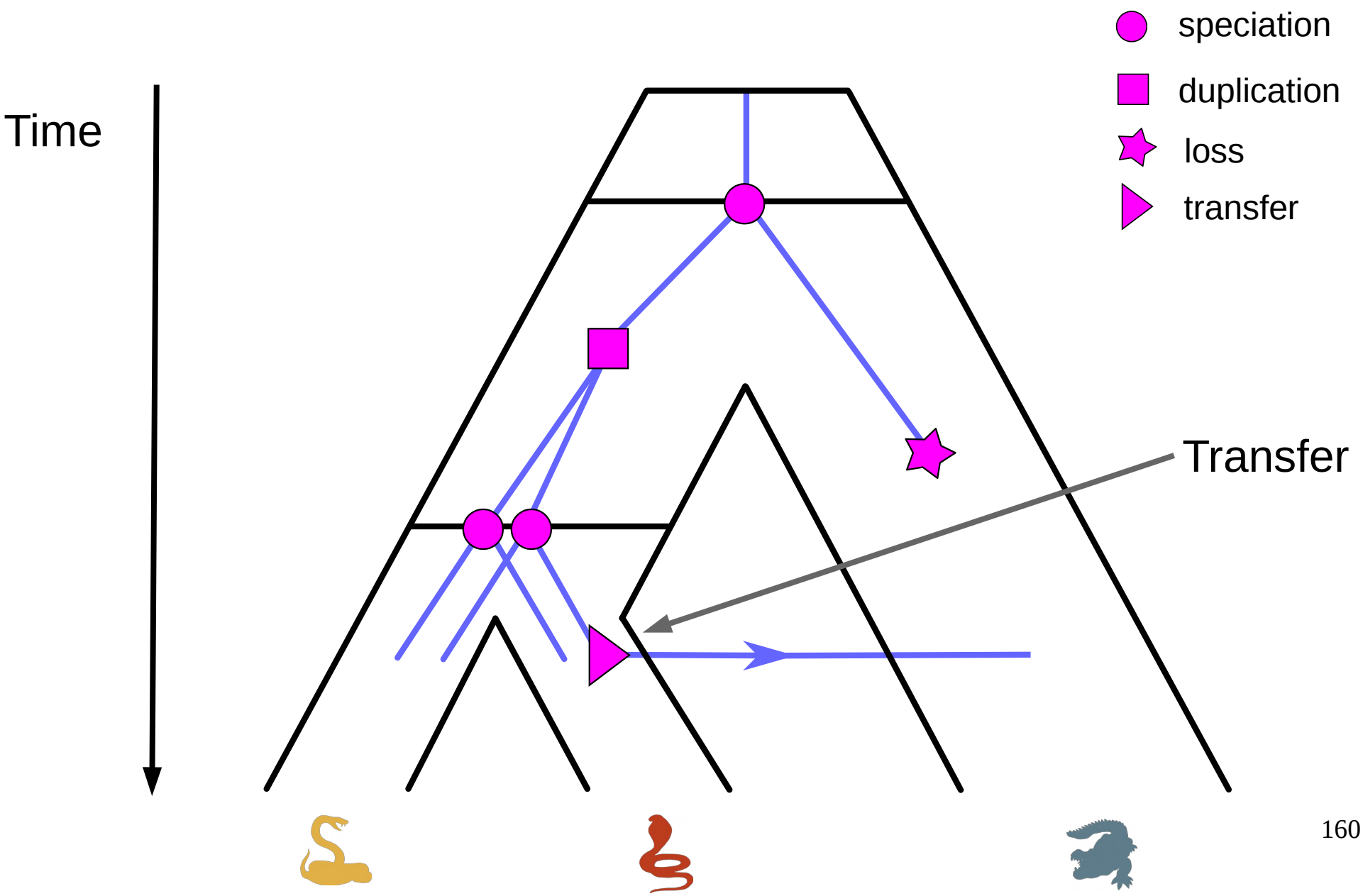
Gene evolution



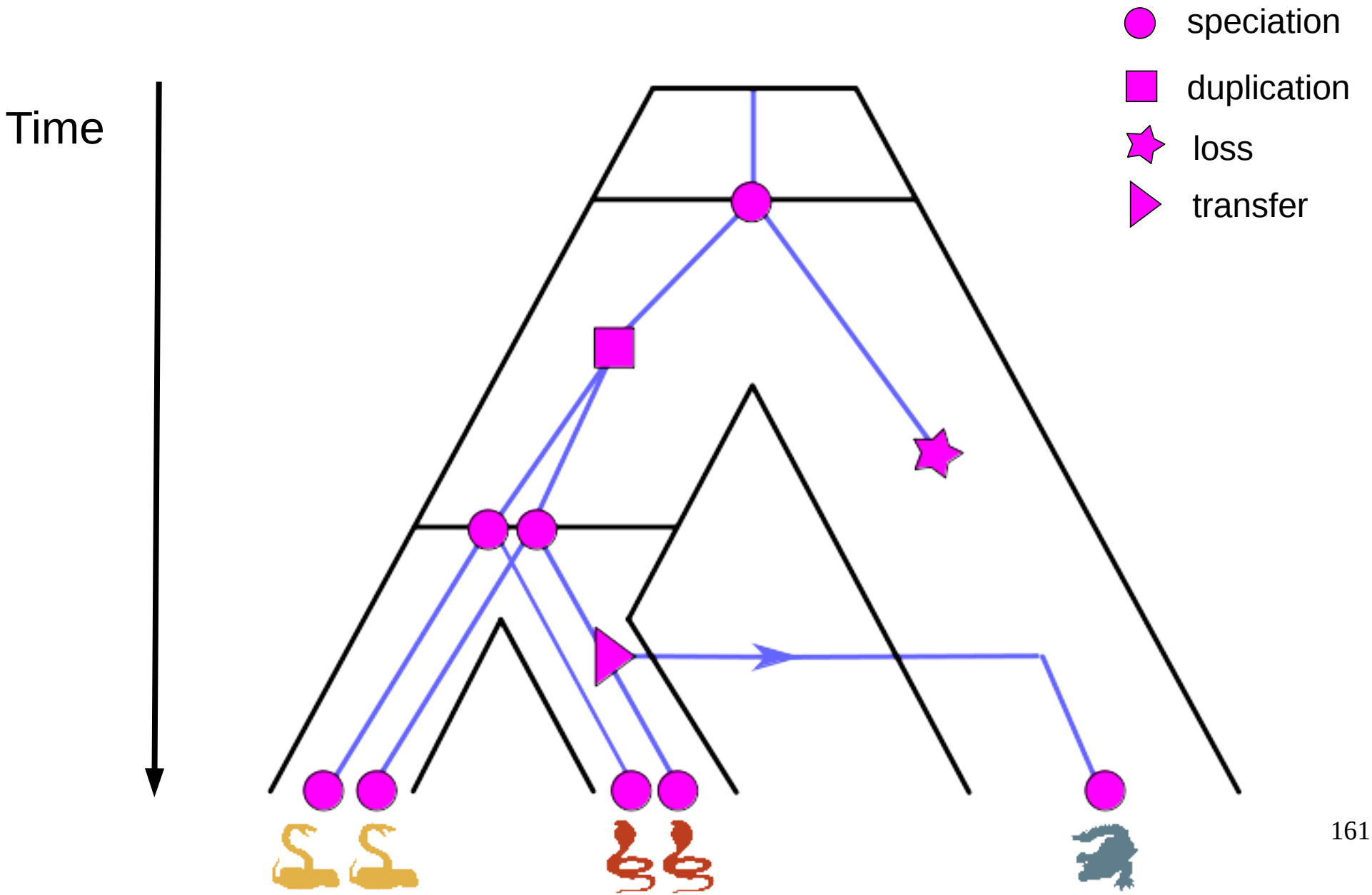
Gene evolution



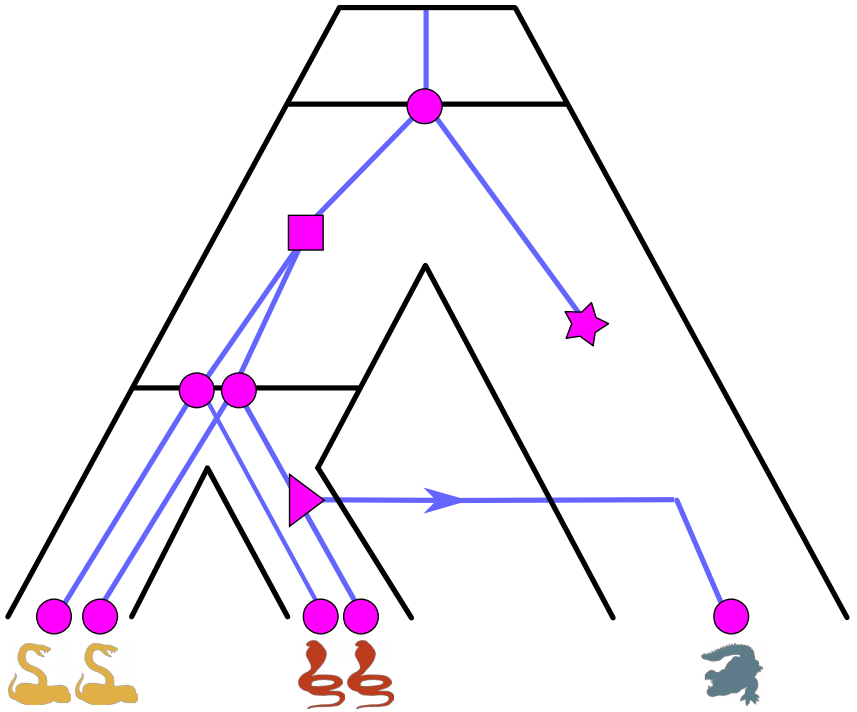
Gene evolution



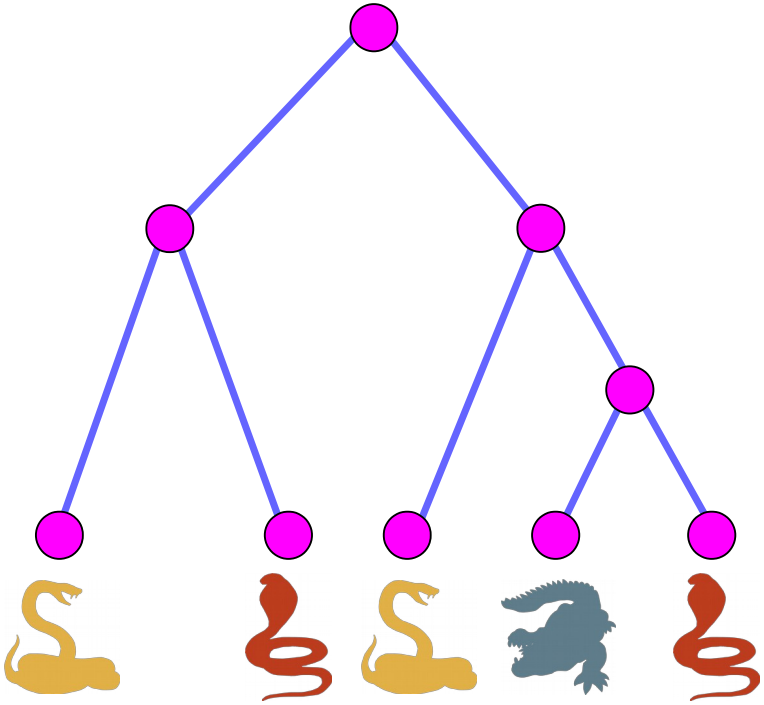
Gene evolution



Gene tree



Reconciliation scenario

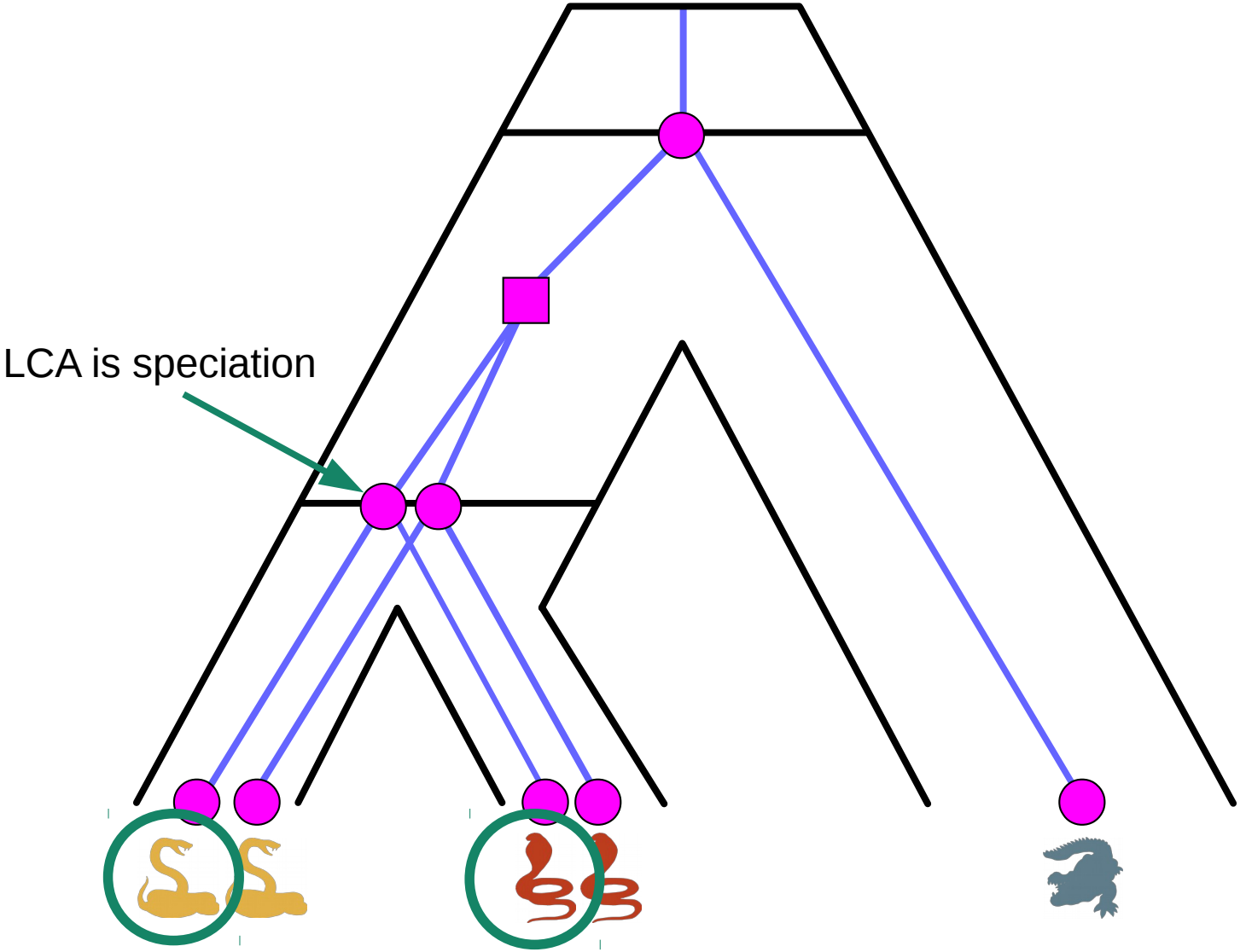


Resulting gene tree

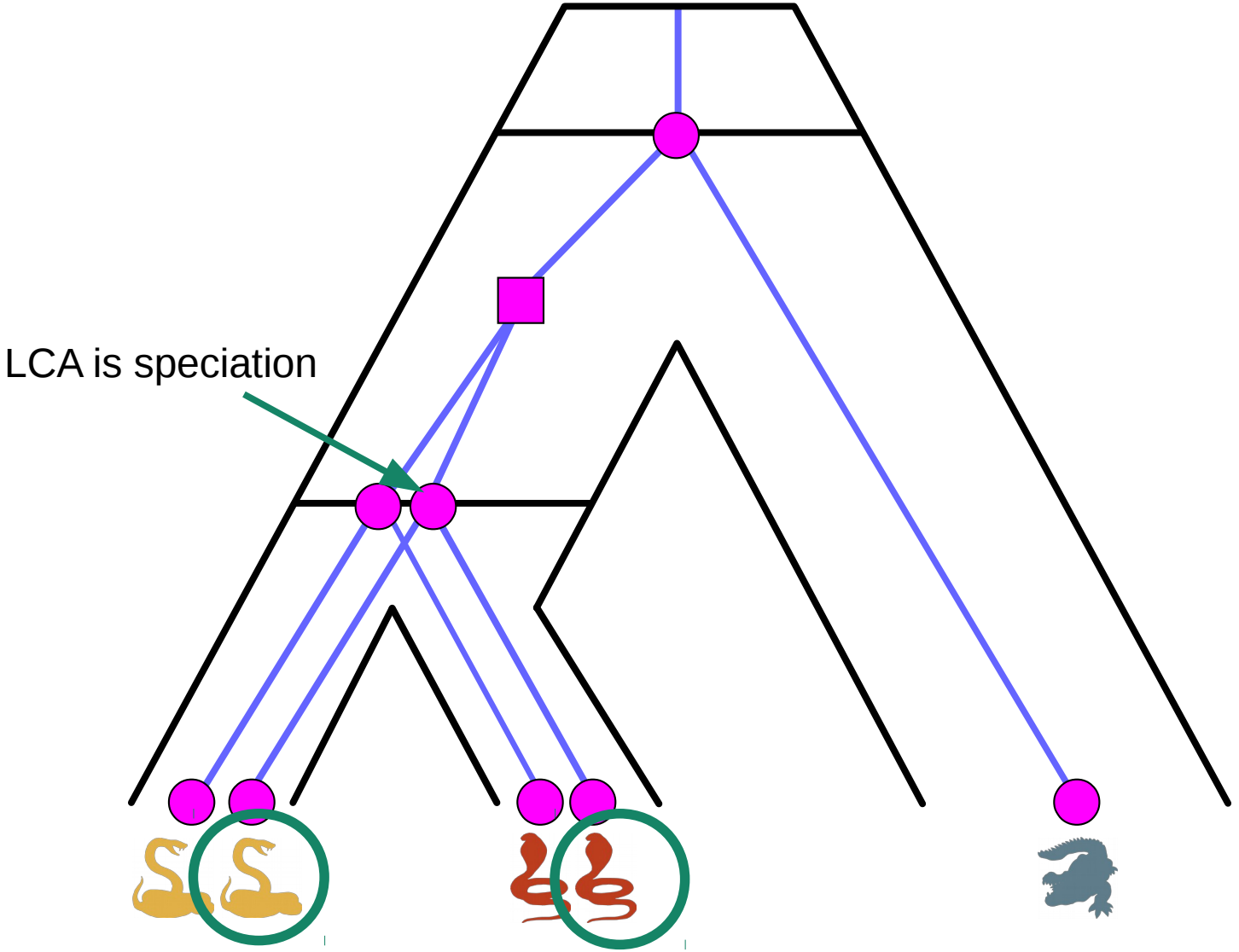
Ortholog and paralog genes

- LCA = last common ancestor
- Two genes are ortholog if their LCA is a speciation event
- Two genes are paralog if their LCA is a duplication event
- Two genes are xenolog if their LCA is a HGT event

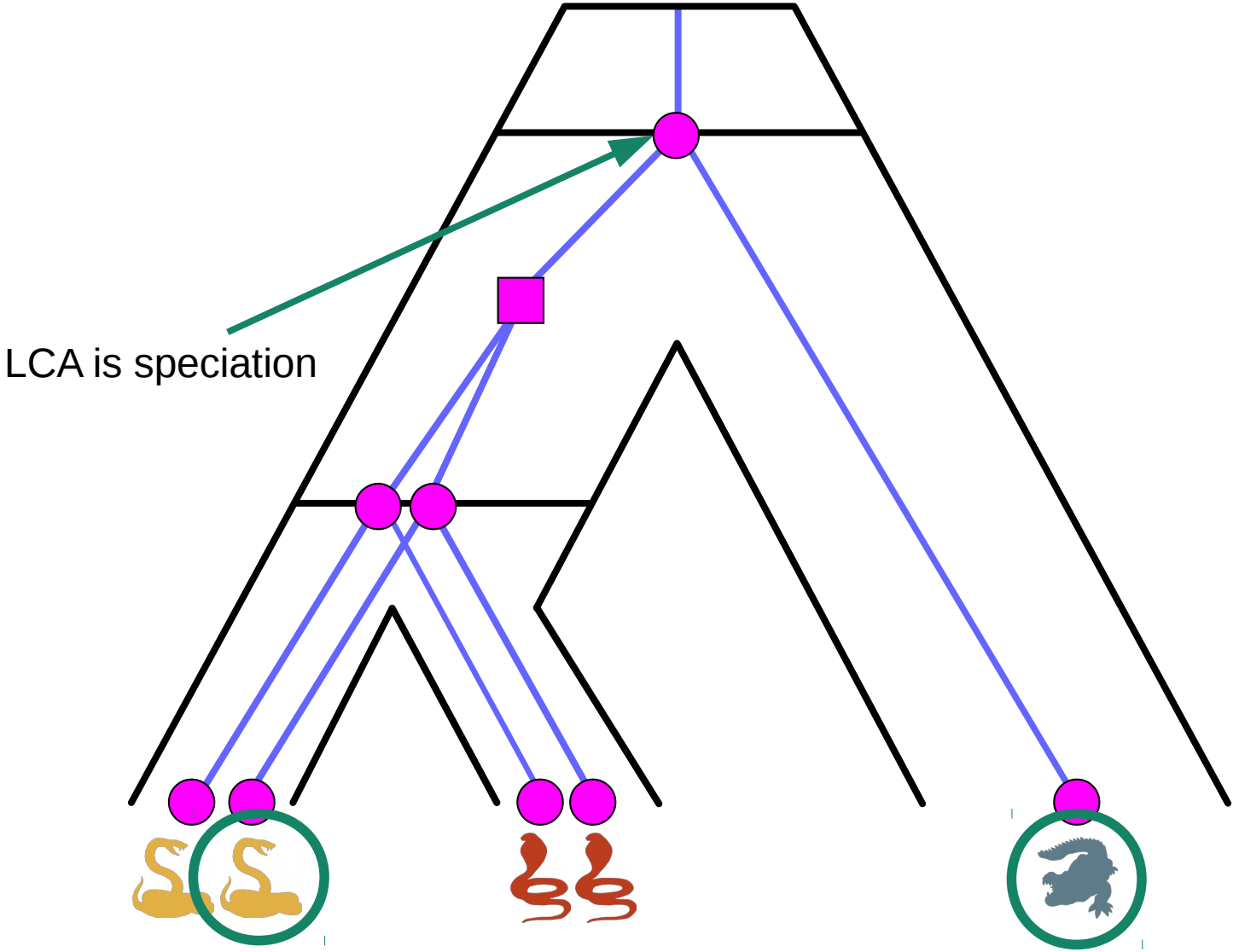
Example: orthologs



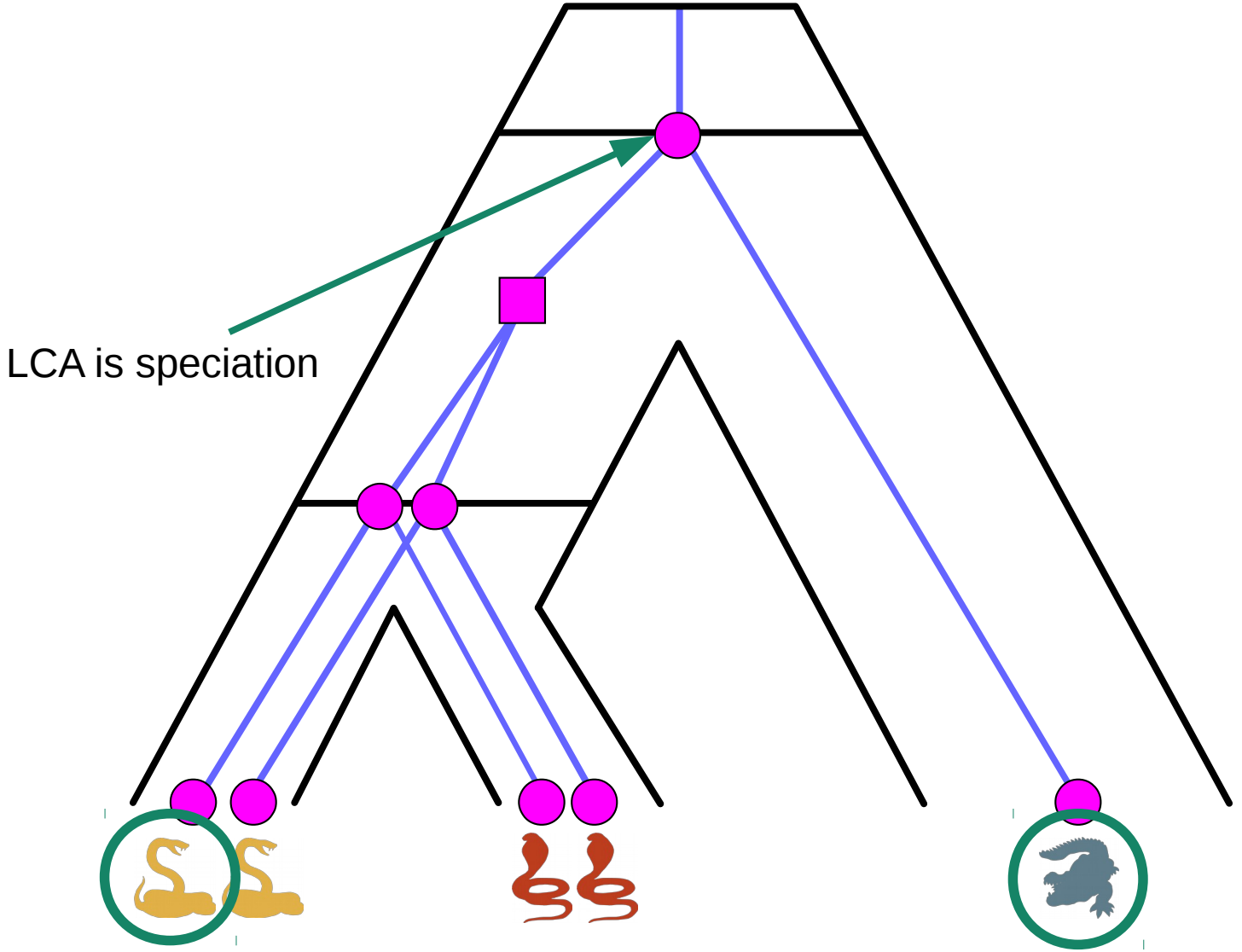
Example: orthologs



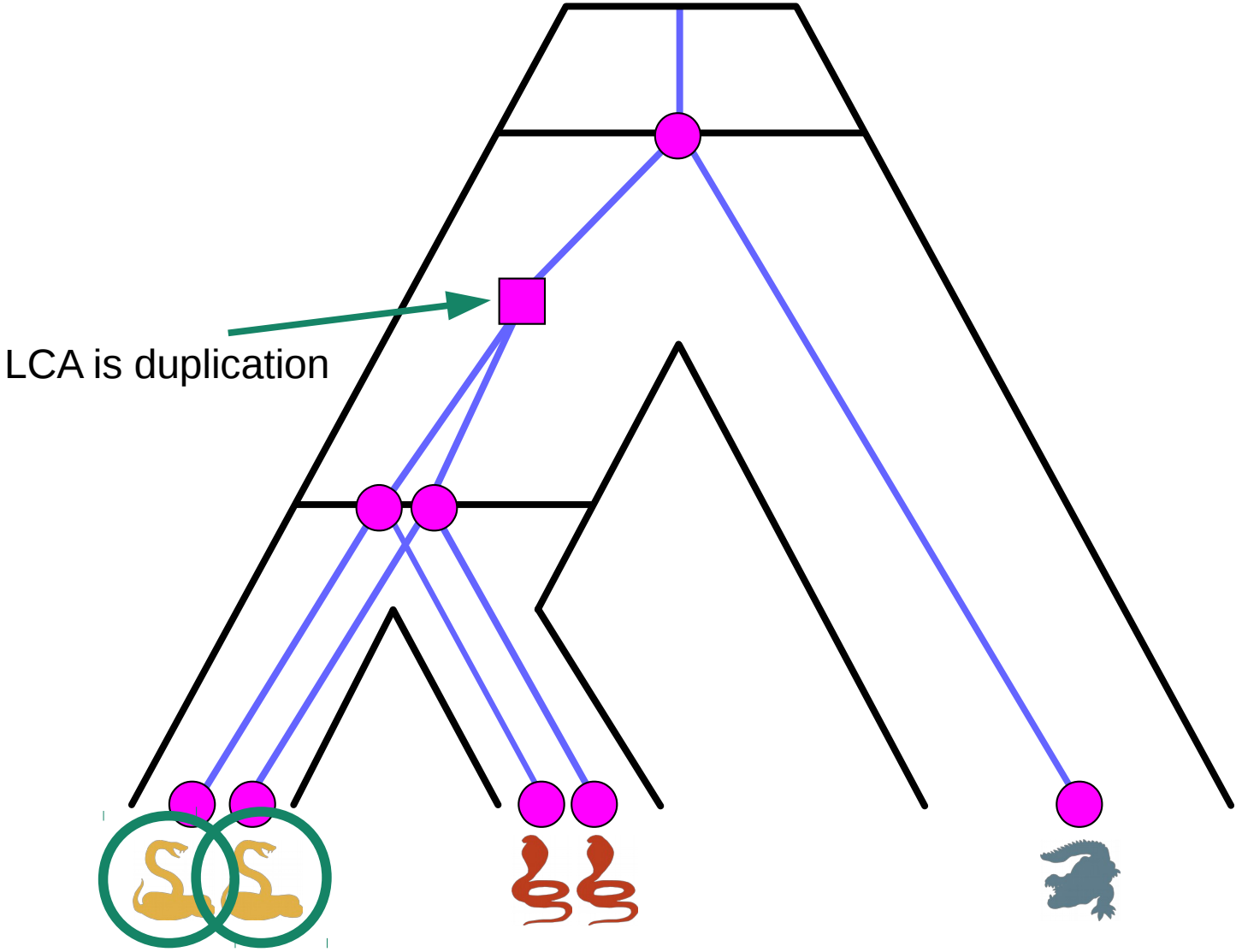
Example: orthologs



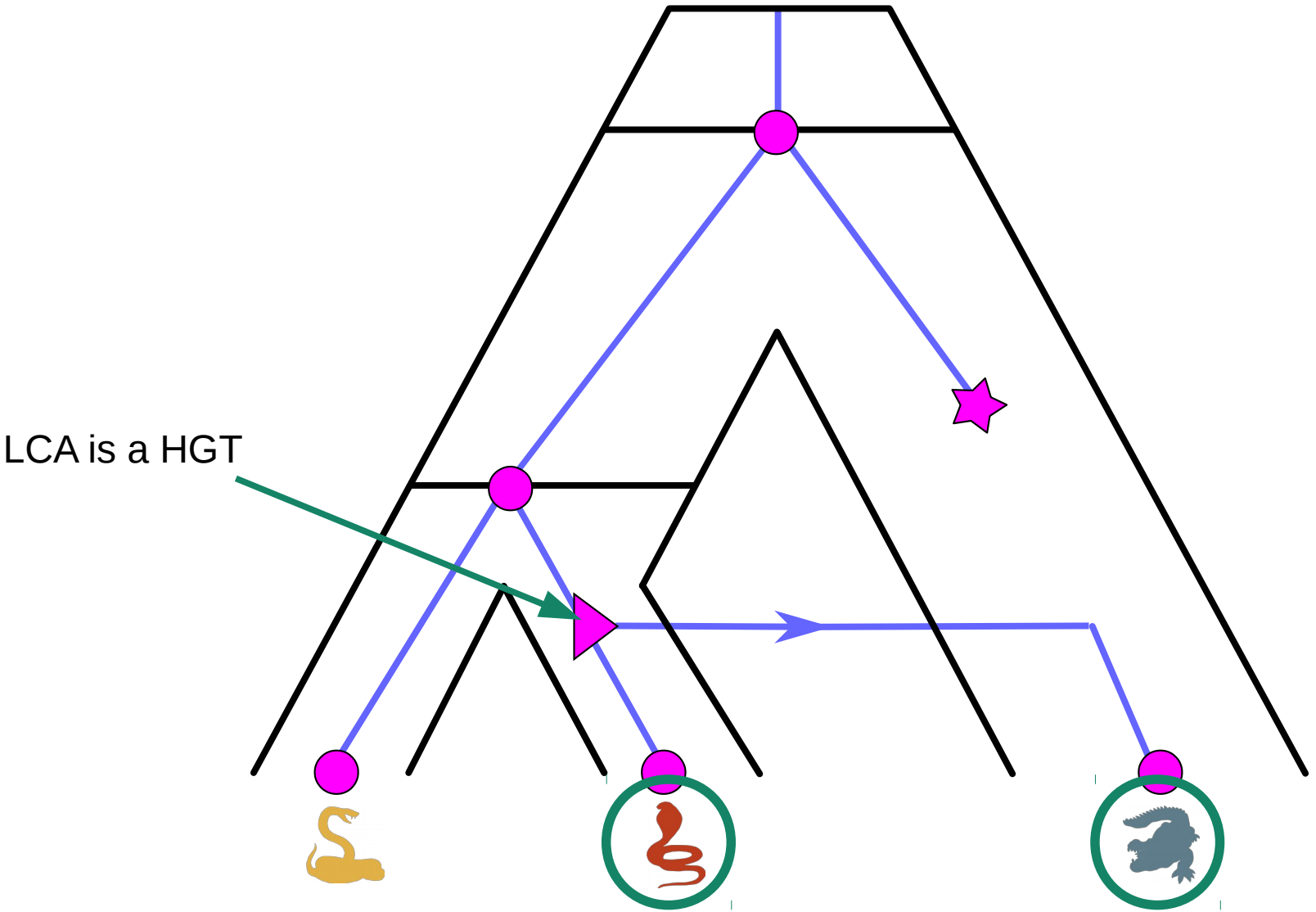
Example: orthologs



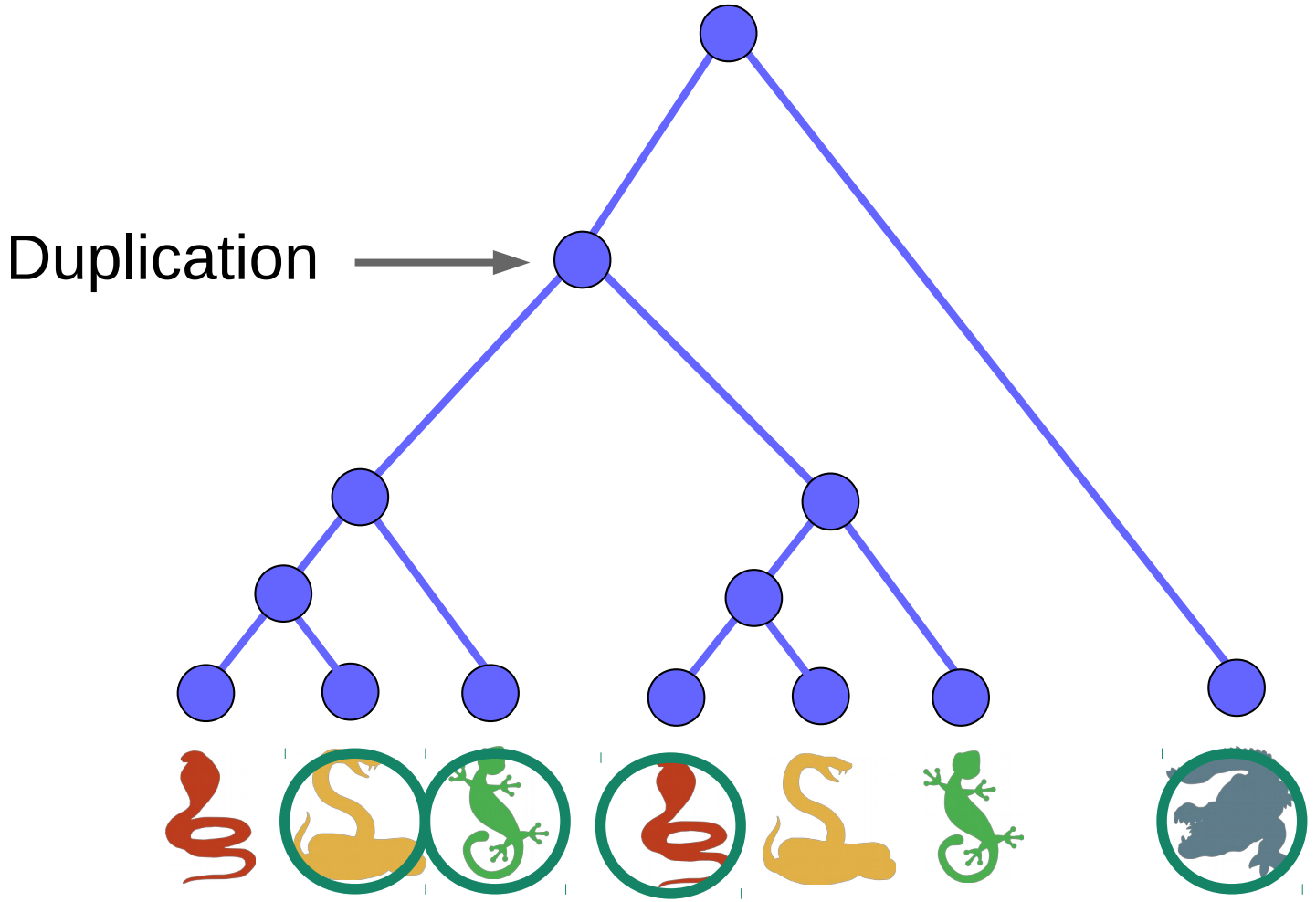
Example: paralogs



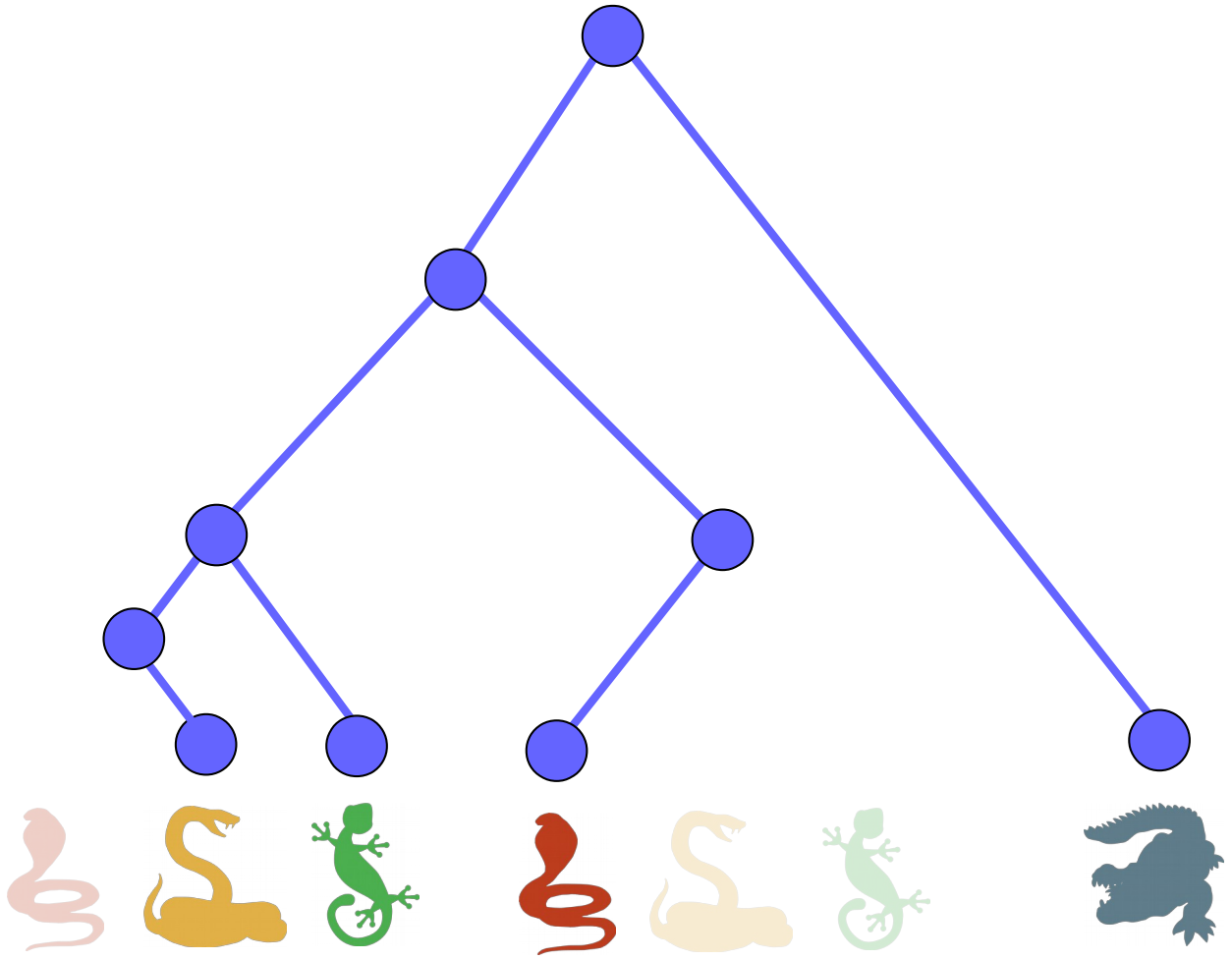
Example: xenologs



What’s wrong with paralog genes?



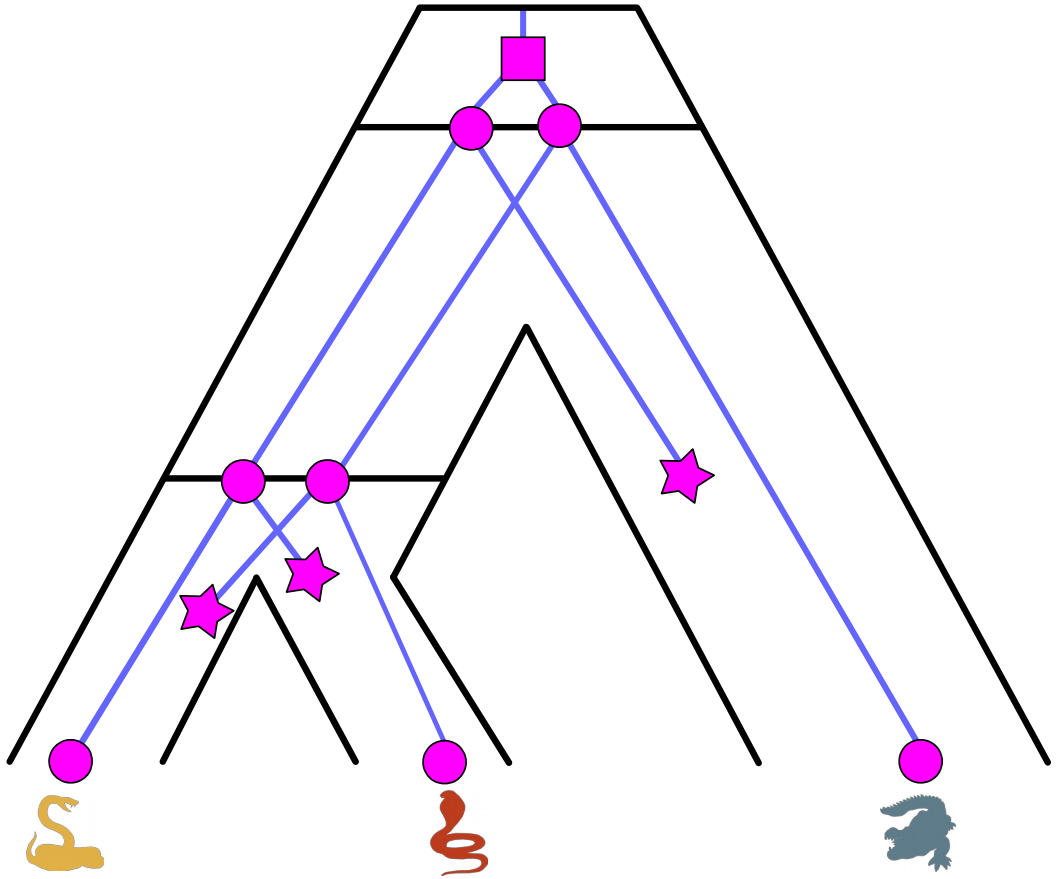
What’s wrong with paralog genes?



Conflict with the real species tree

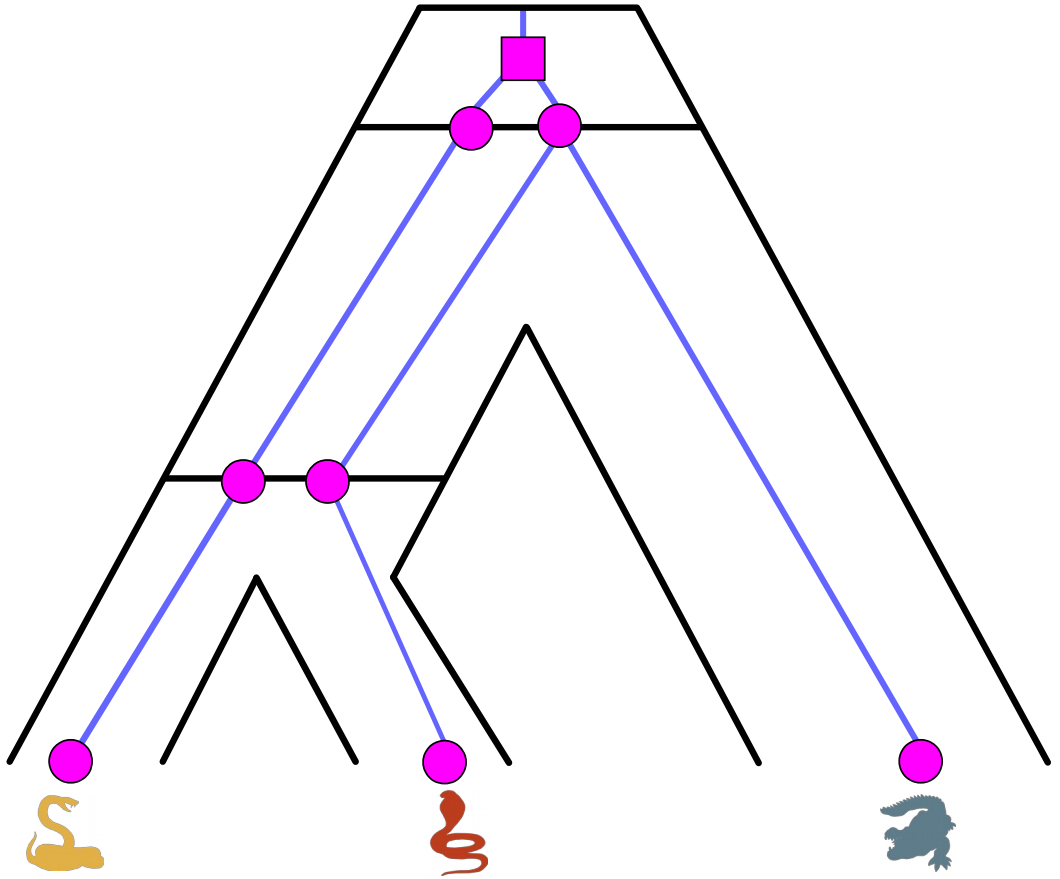
Hidden paralogy

- Single copy family with paralog genes



Hidden paralogy

- Single copy family with paralog genes



Paralog genes affect tree reconstruction

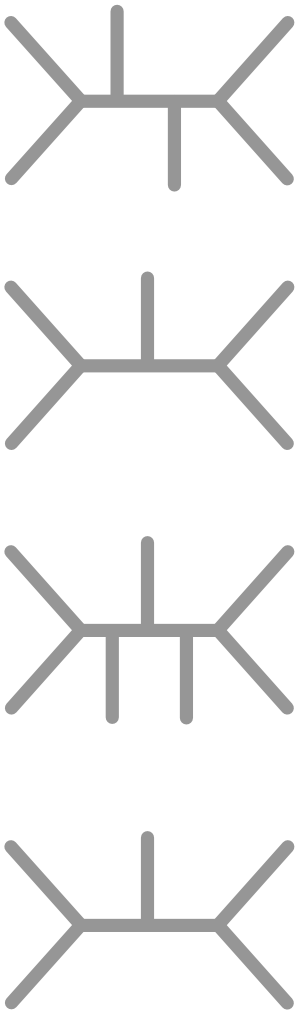
- DTL events generate gene-species tree conflicts and are problematic for species tree inference

- Single-copy methods (concatenation, ASTRAL etc.) require ortholog genes

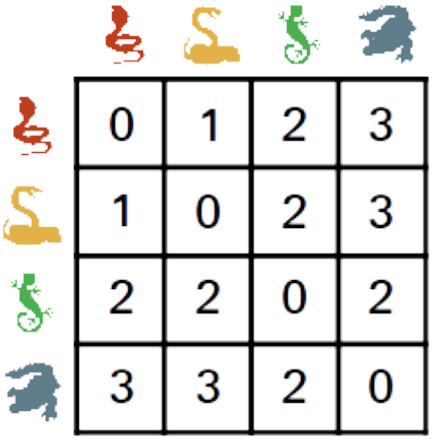
Multi-copy methods

- Distance: Astrid/Asteroid
- Quartet method: AstralPro
- Parsimony method: DupTree
- Maximum likelihood method: SpeciesRax

Astrid / Asteroid

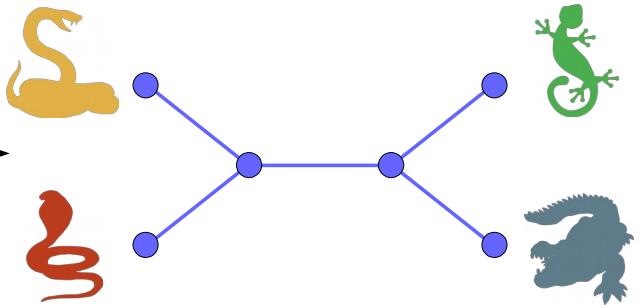


Gene trees



0	1	2	3
1	0	2	3
2	2	0	2
3	3	2	0

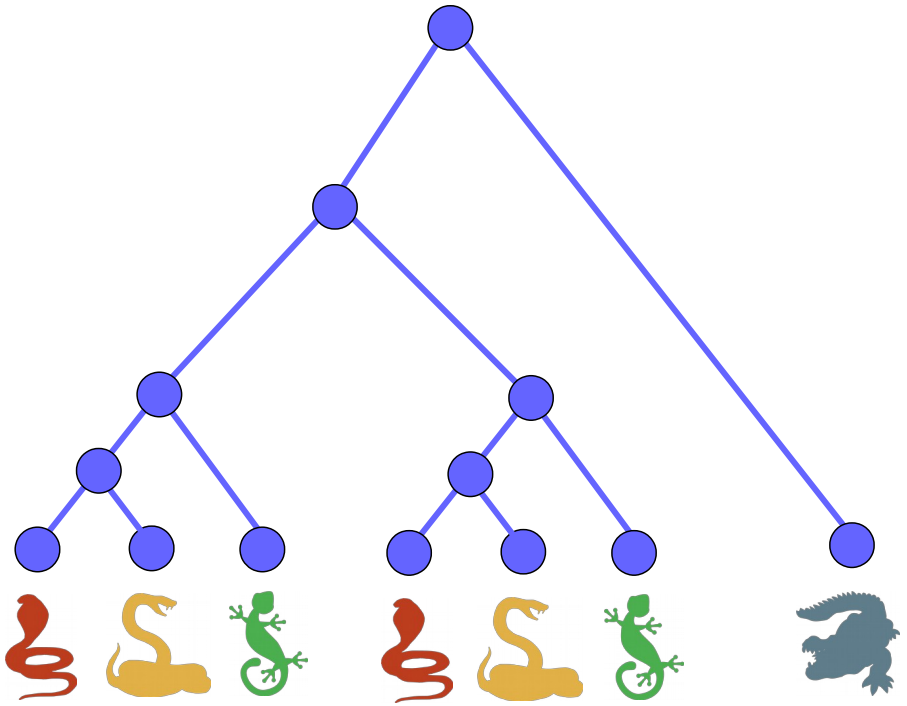
Distance matrix



Unrooted species tree

Astrid / Asteroid

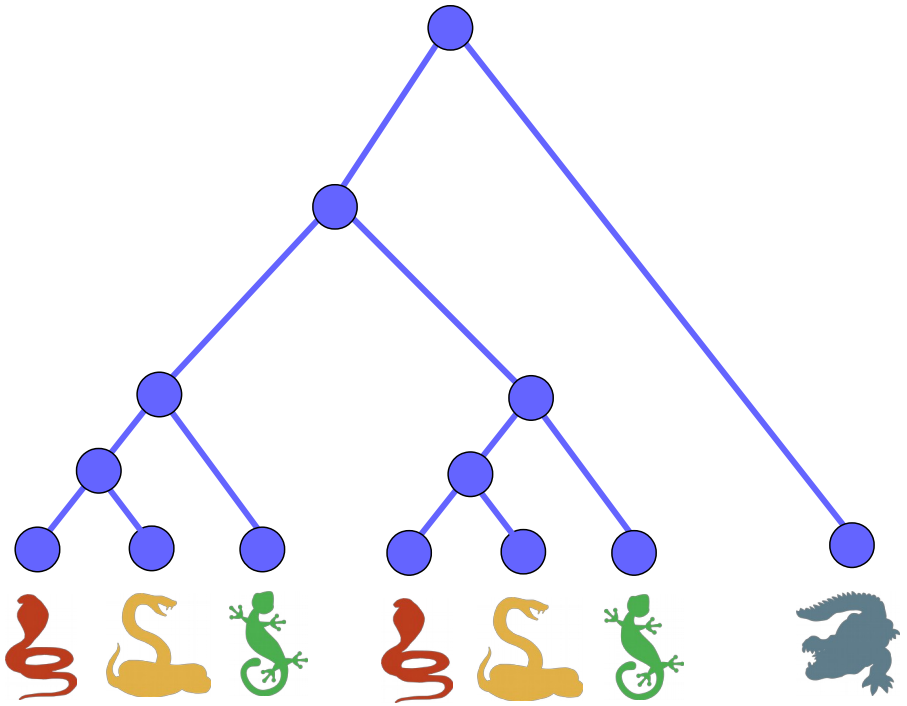
Challenge: how to compute the internode distance between two species?



D(,) = ??

Astrid / Asteroid

Challenge: how to compute the internode distance between two species?

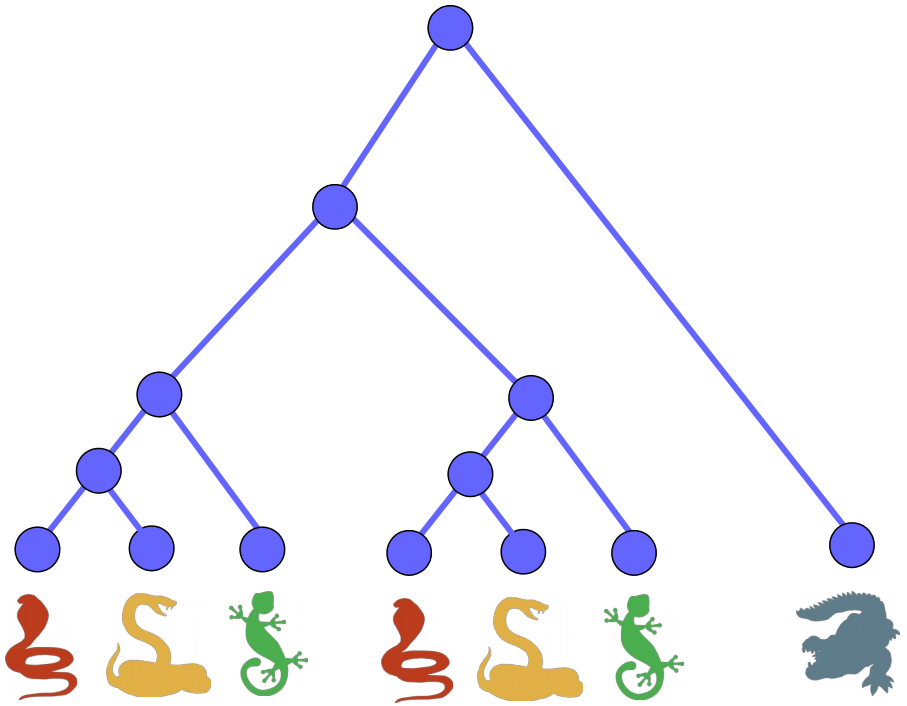


USTAR: take the average of all distances

D(,)
= (1 + 1 + 5 + 5) / 4
= 3

Astrid / Asteroid

Challenge: how to compute the internode distance between two species?



MiniNJ: take the min of all distances

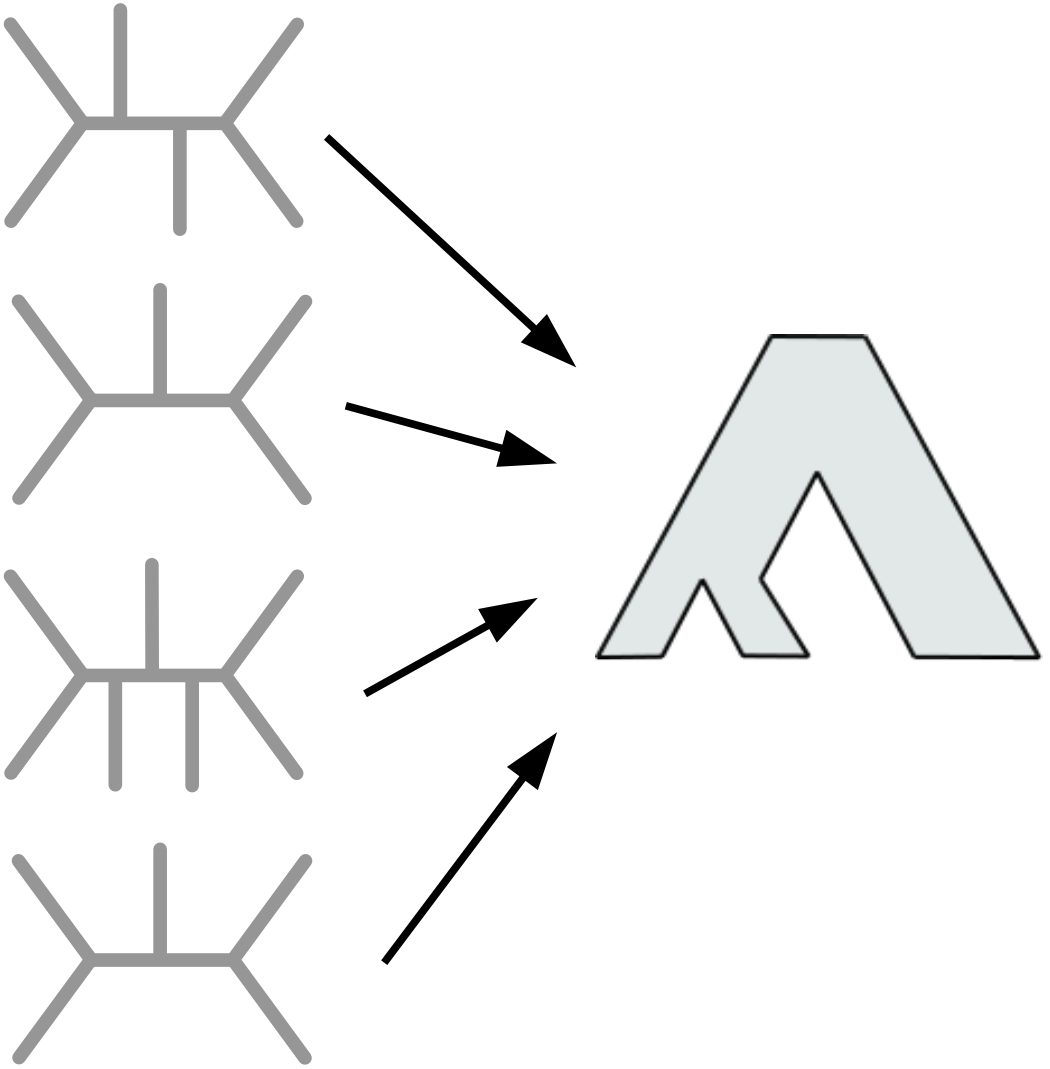
D(,)
= min(1 + 1 + 5 + 5)
= 1

Astral-Pro

- Astral: count the compatible quartets
- Astral-Pro: count the quartets that result from speciation events (and not the ones that result from duplication events)

SpeciesRax: maximum likelihood species tree inference

Maximize the reconciliation likelihood



Models of evolution

- We need to define probabilistic models to describe evolutionary processes
- Once we have defined a model, we can use it to compute likelihood scores and estimate parameters
- Those models are a trade-off between the complexity of the real process and the simplicity that will allow us to implement fast algorithms

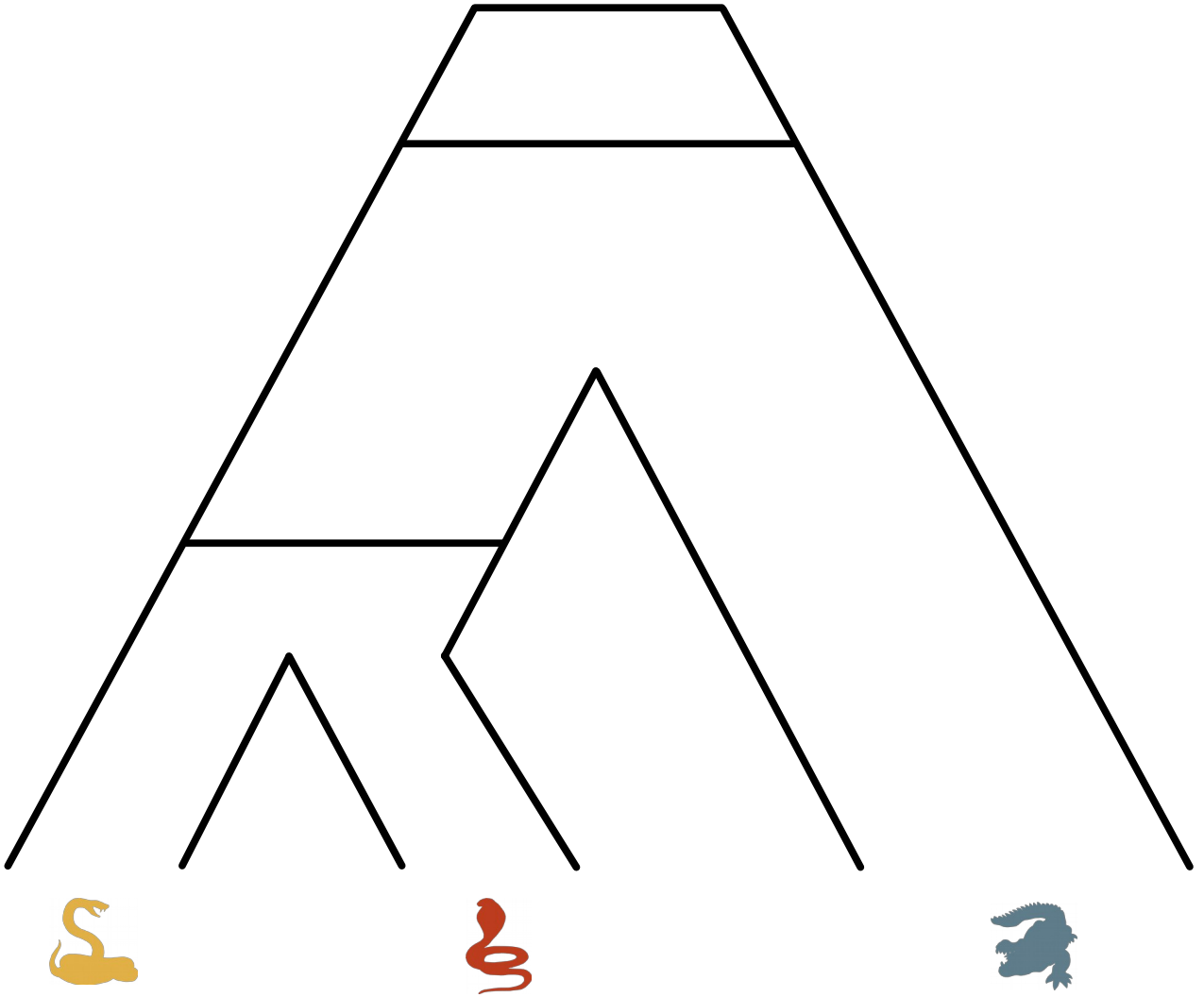
The UndatedDTL model

- The UndatedDTL model describes how a gene tree evolves in a species tree
- It is parametrized by its event probabilities: pD, pL, pT, pS
- pD + pL + pT + pS = 1

The UndatedDTL model

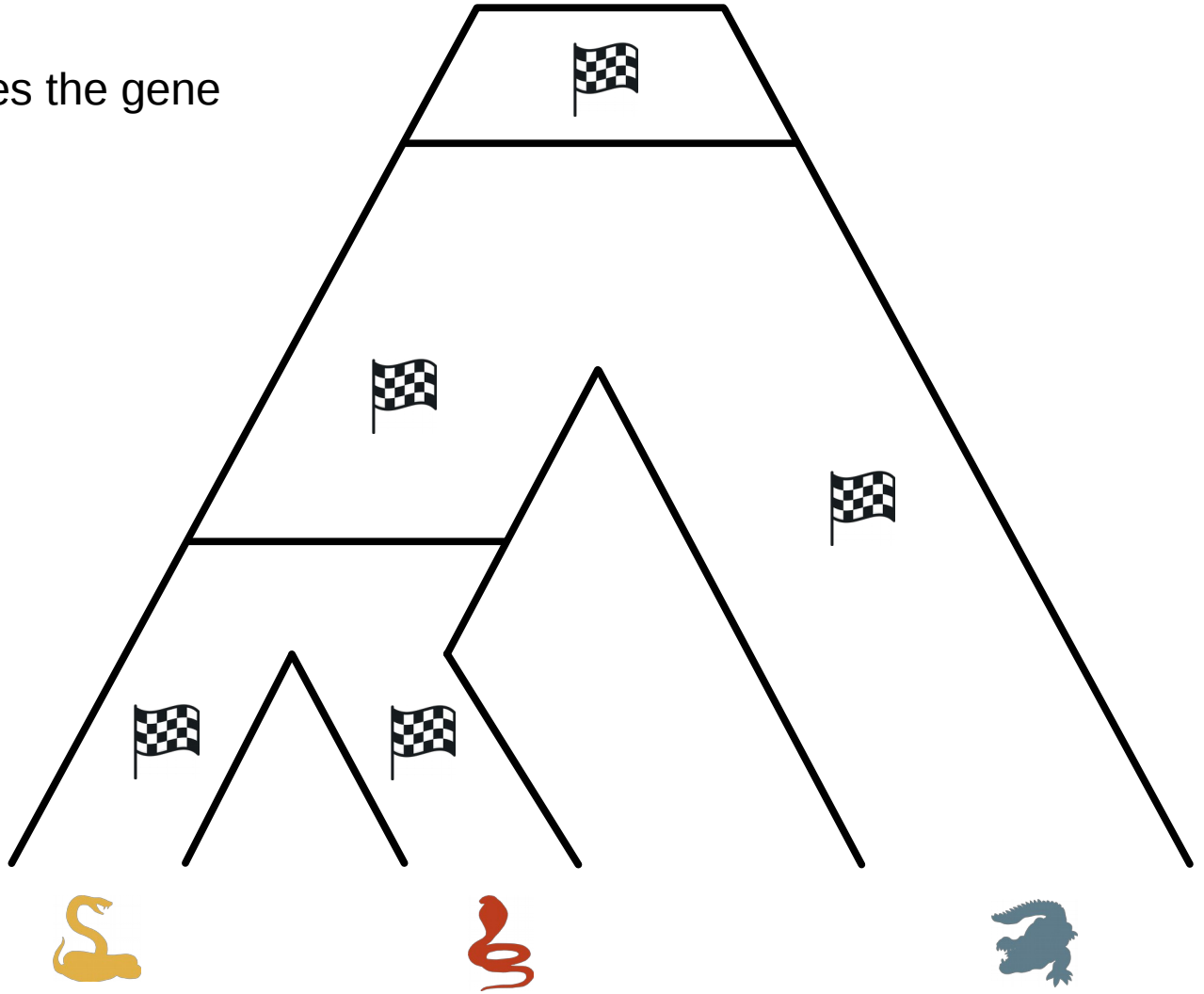
- The UndatedDTL model describes how a gene tree evolves in a species tree
- It is parametrized by its event probabilities: pD, pL, pT, pS
- pD + pL + pT + pS = 1
- Let us assume:
  - pS = 0.5
  - pD = 0.2
  - pL = 0.2
  - pT = 0.1

The UndatedDTL model

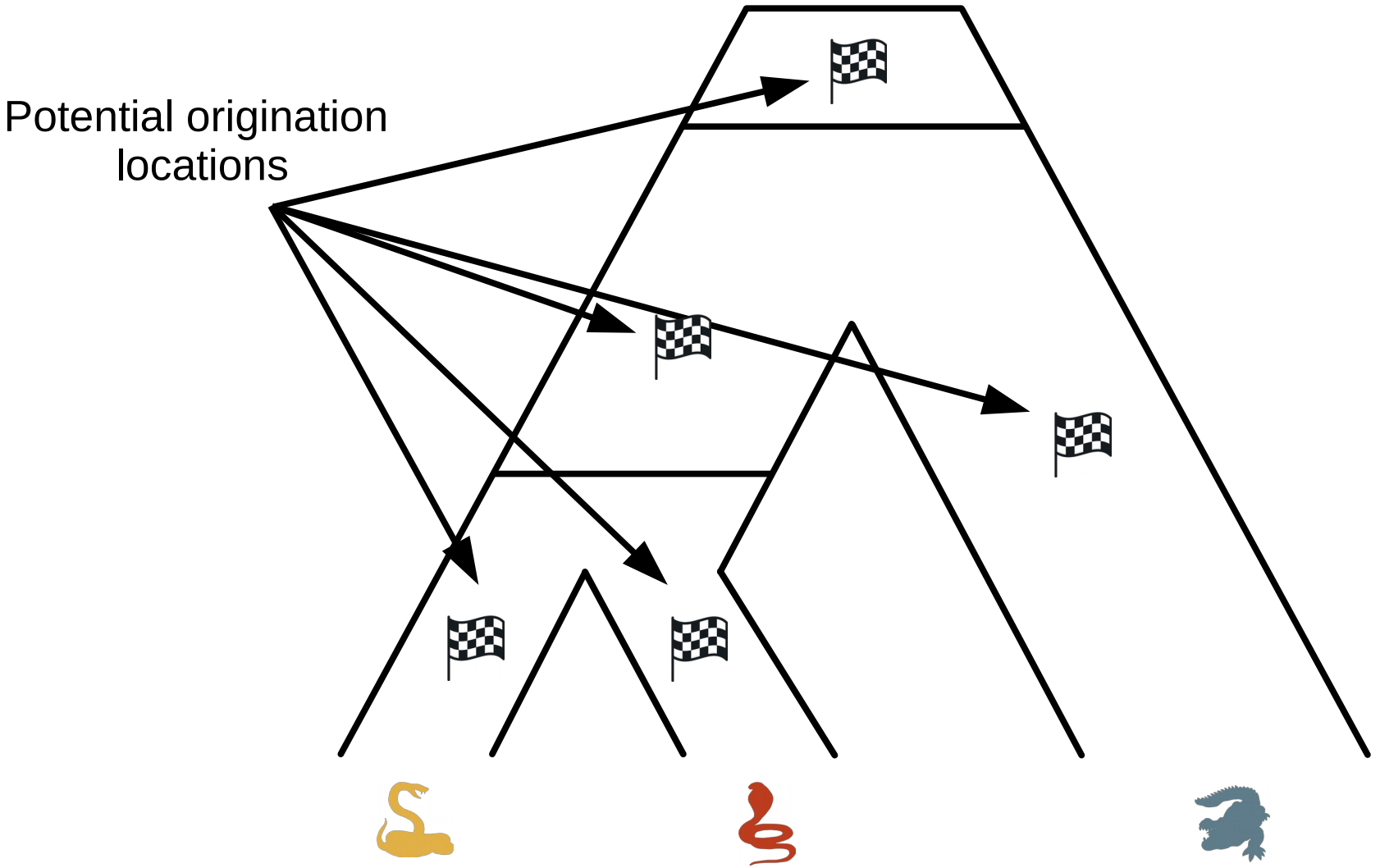


Origination

Where does the gene tree start?

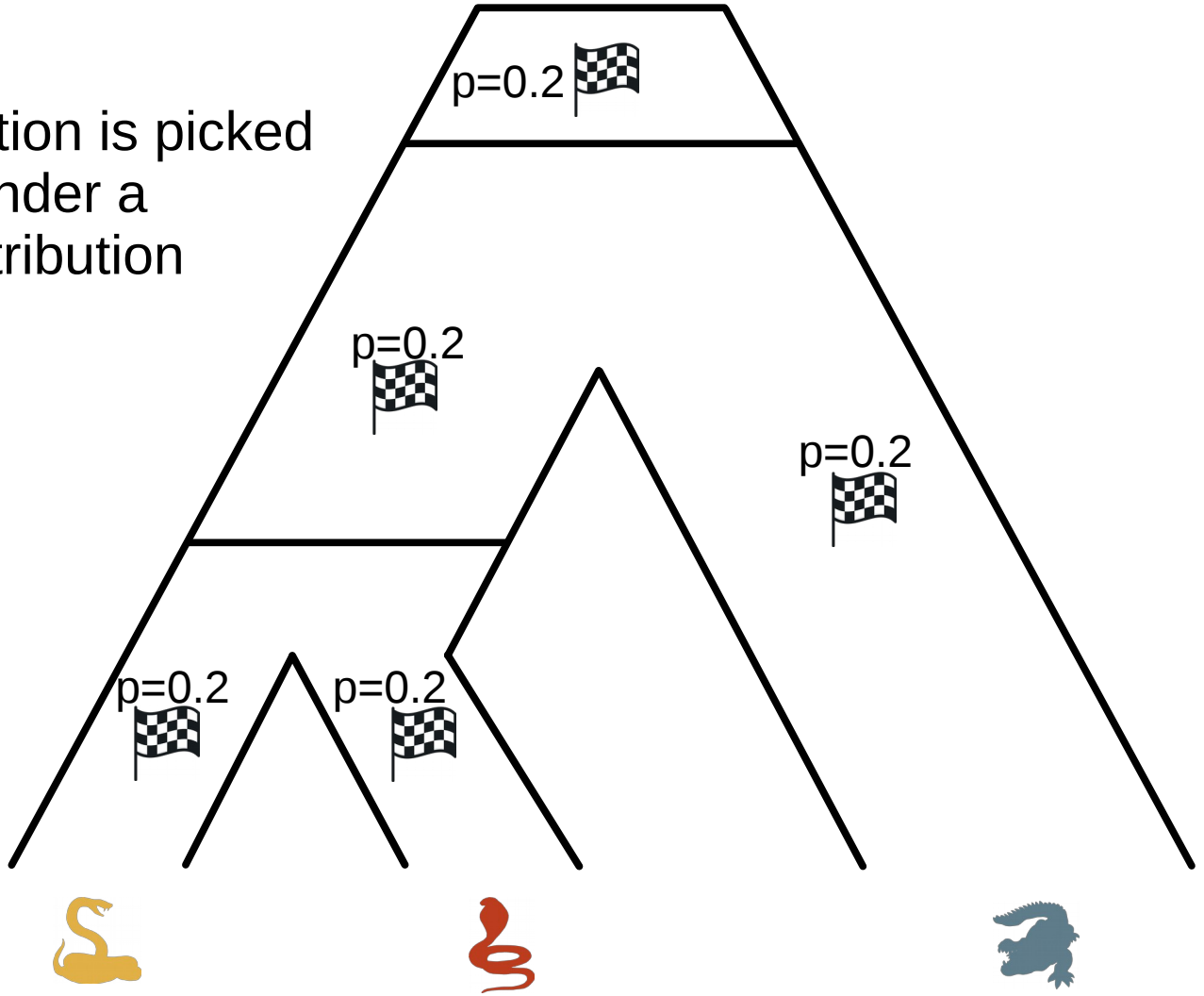


Origination



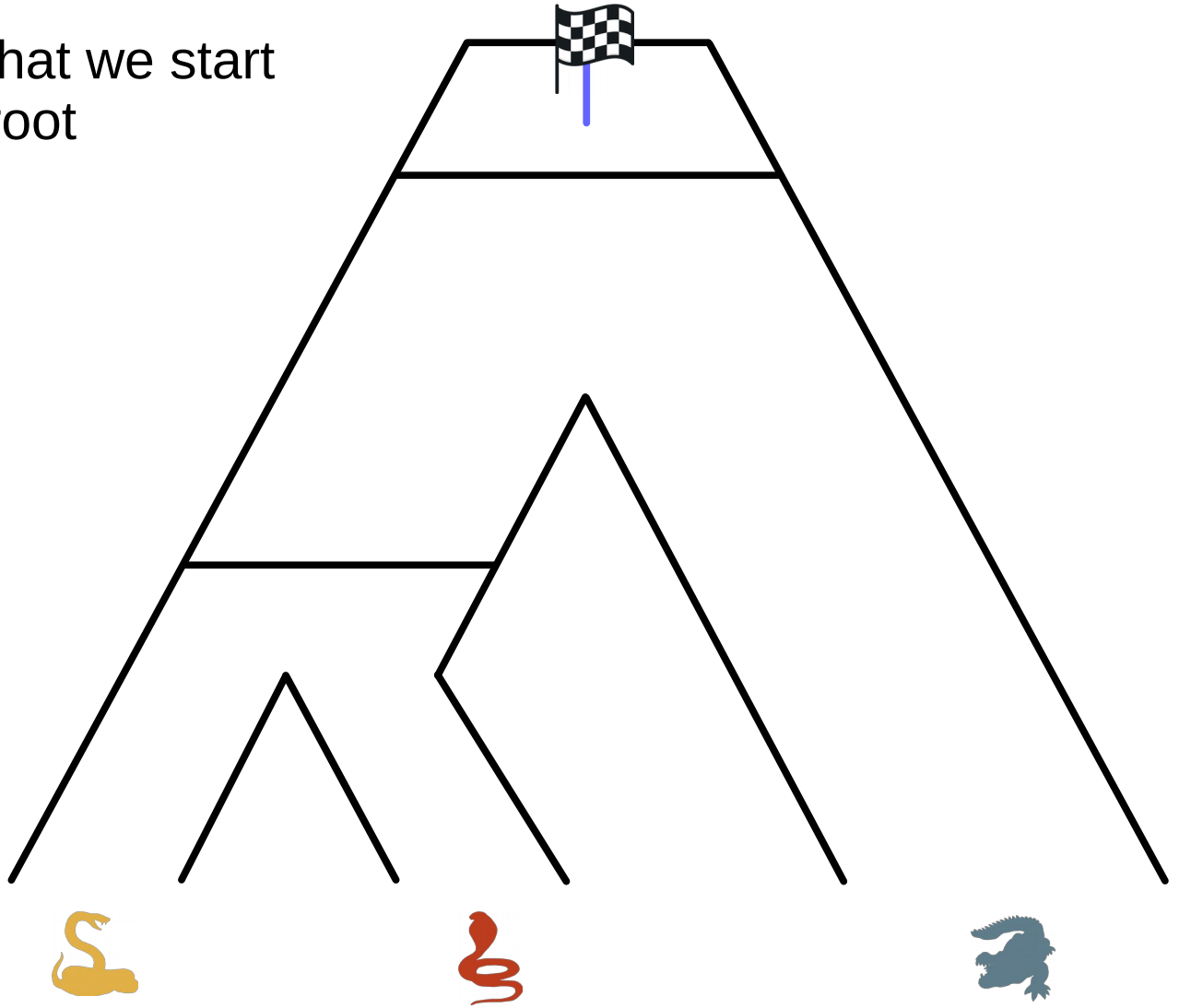
Origination

The origination is picked randomly under a uniform distribution



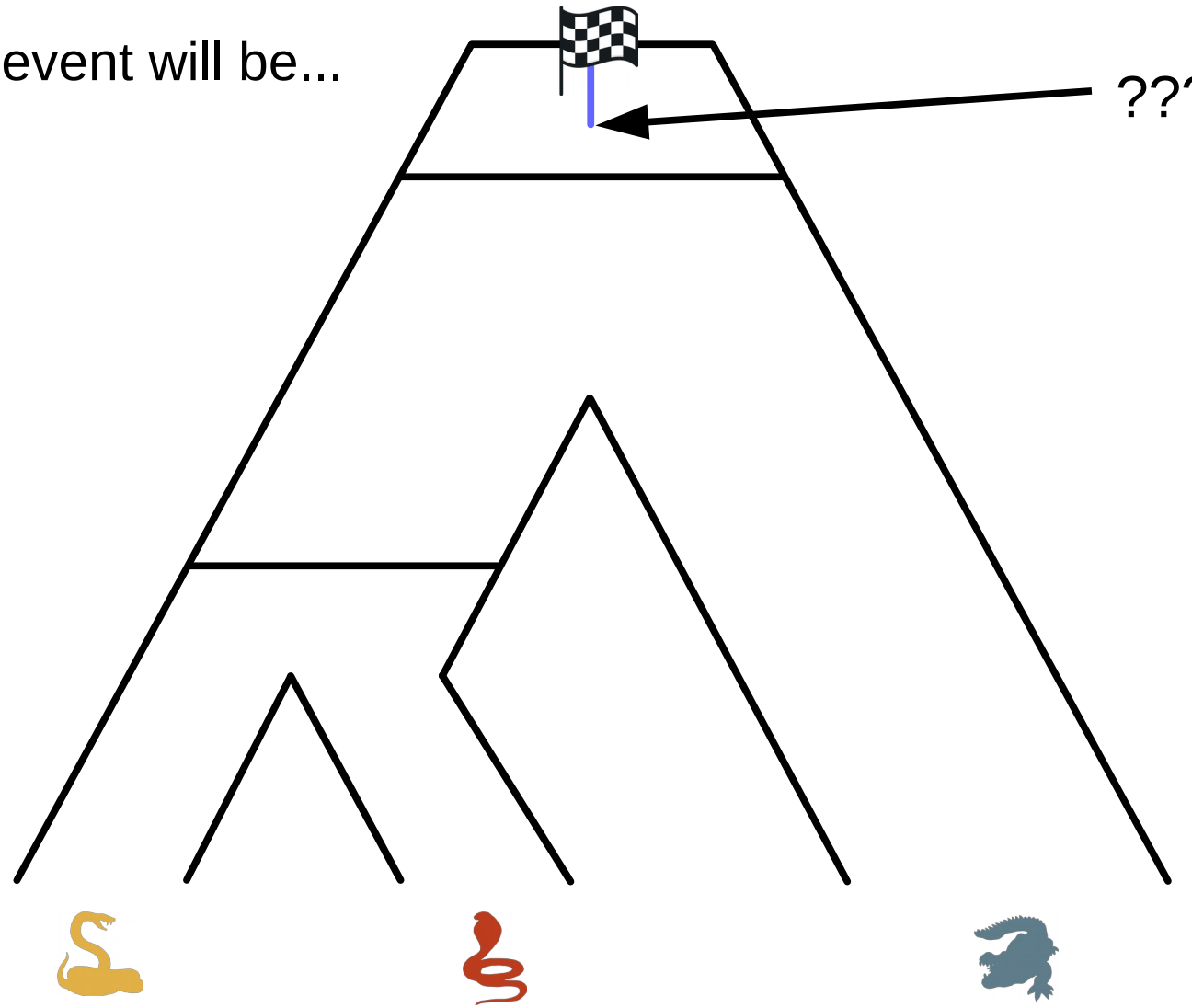
Simulating under the UndatedDTL

Assume that we start from the root

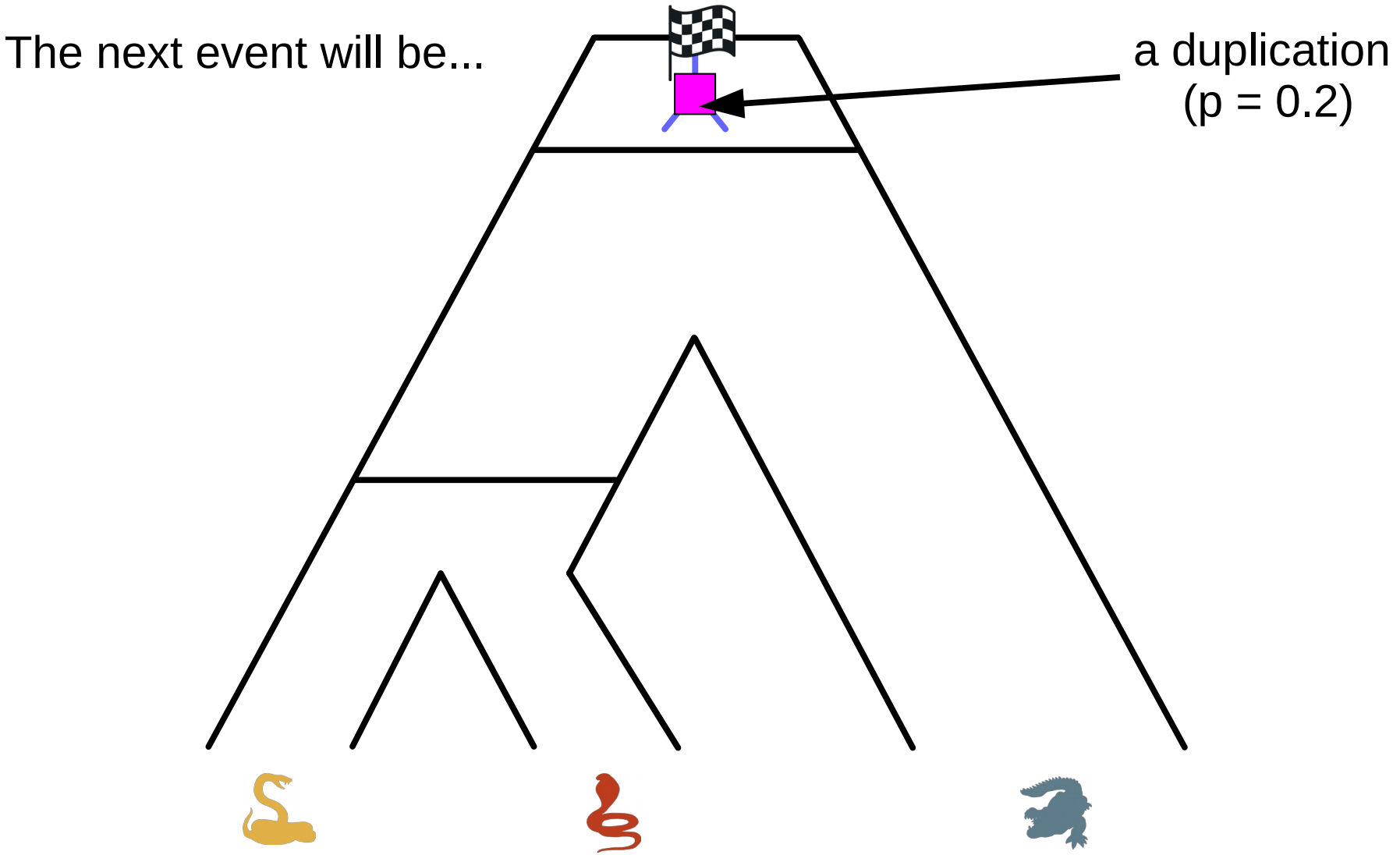


Simulating under the UndatedDTL

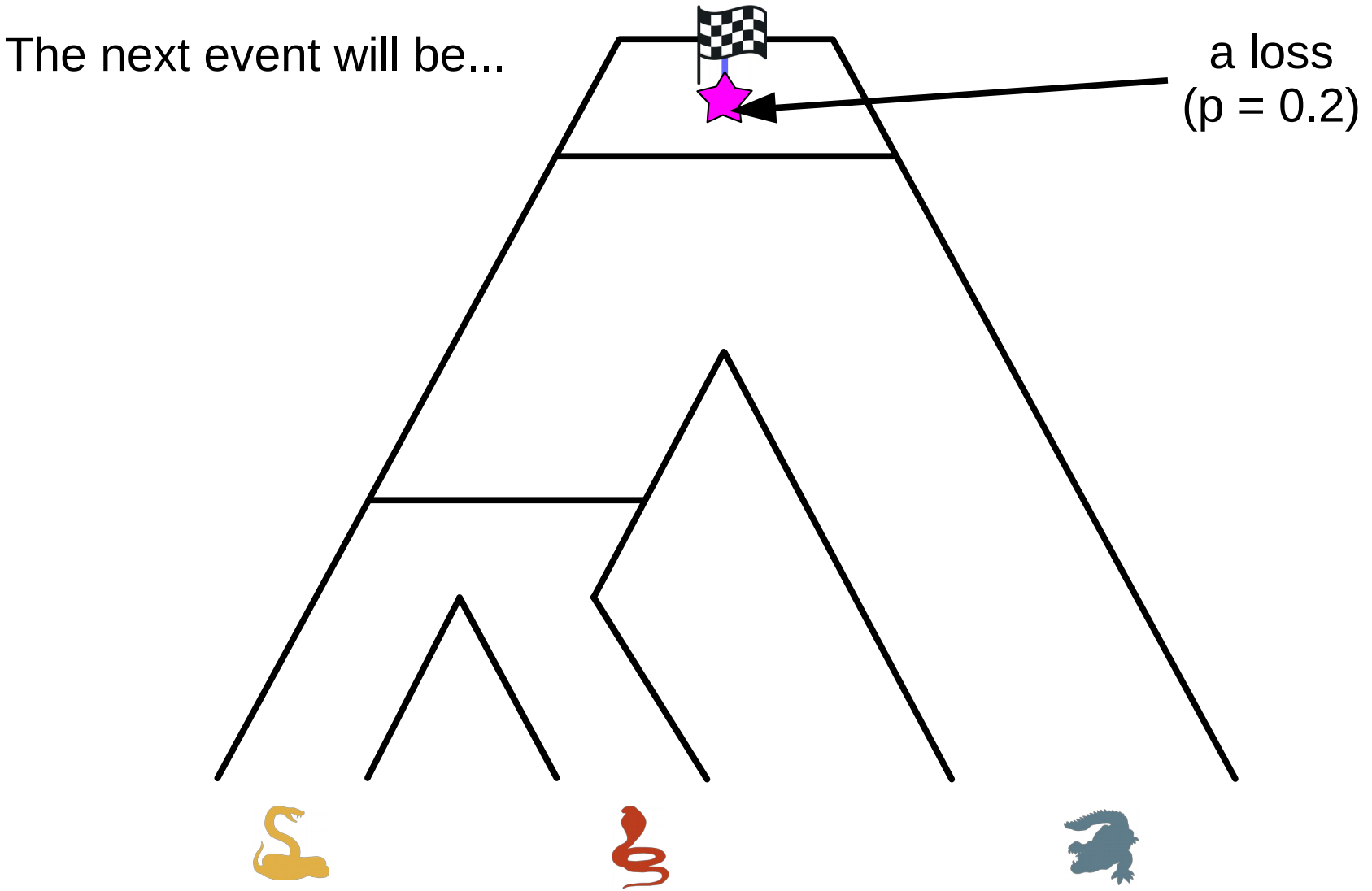
The next event will be...



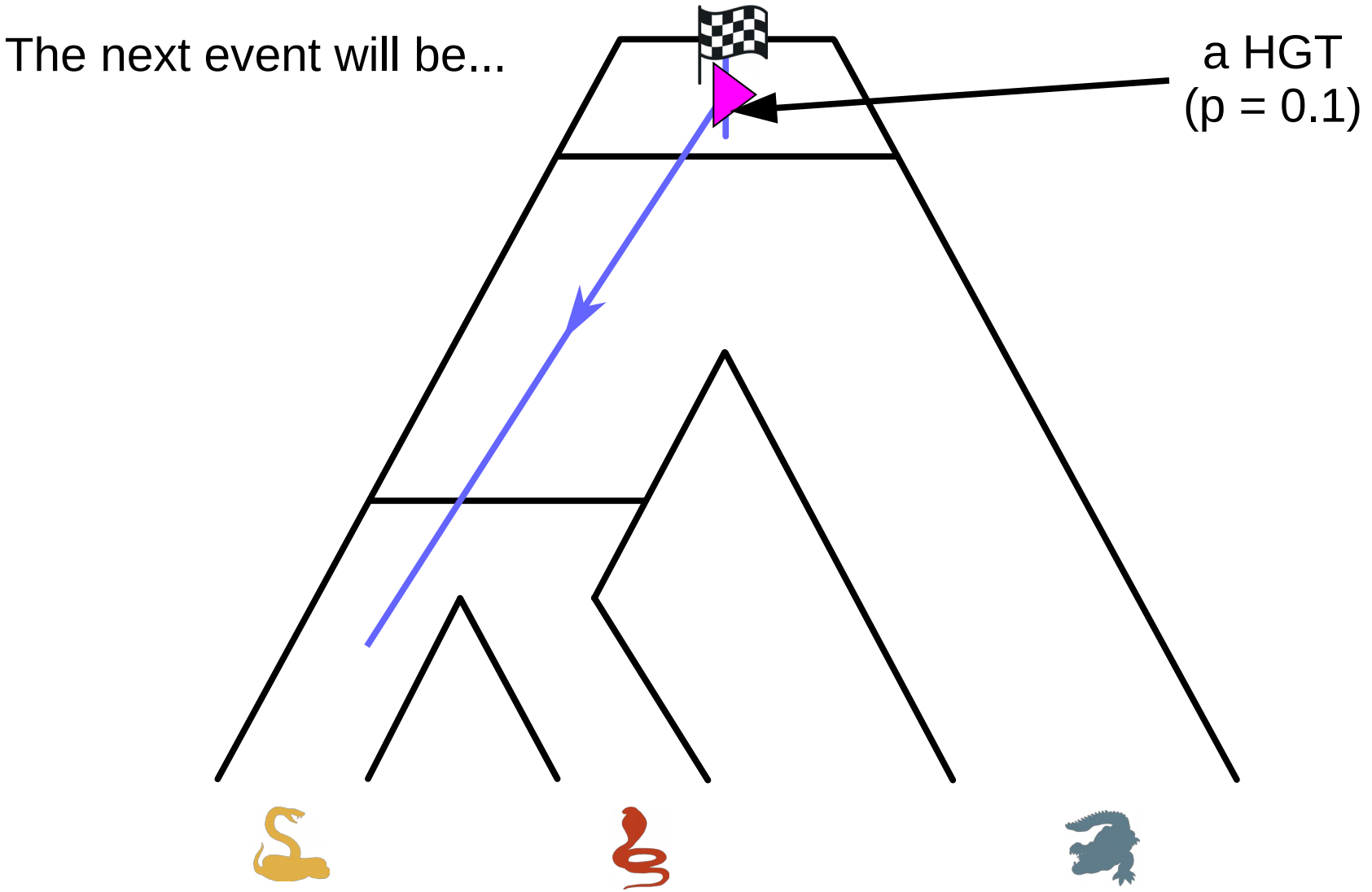
Simulating under the UndatedDTL



Simulating under the UndatedDTL



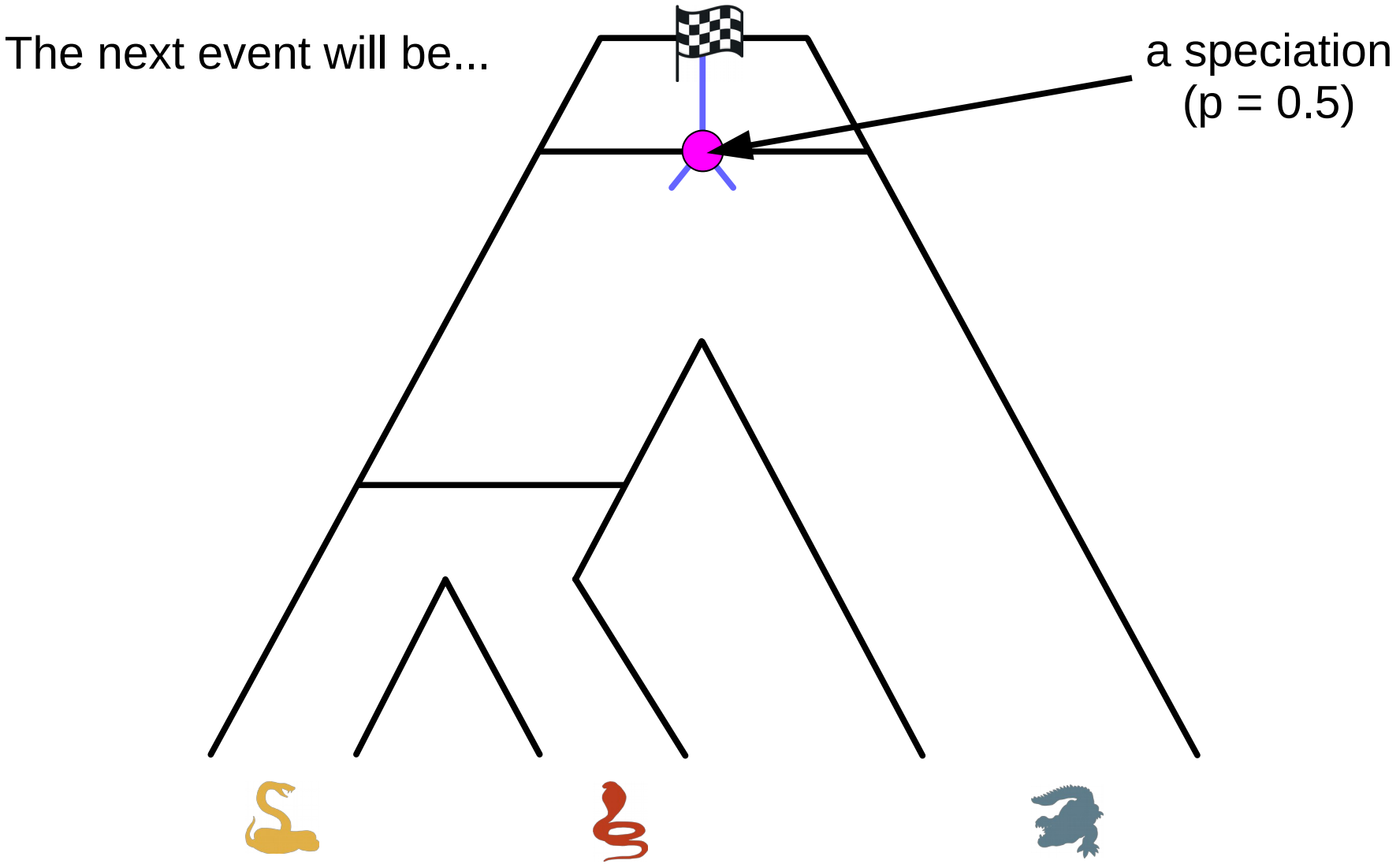
Simulating under the UndatedDTL



Simulating under the UndatedDTL

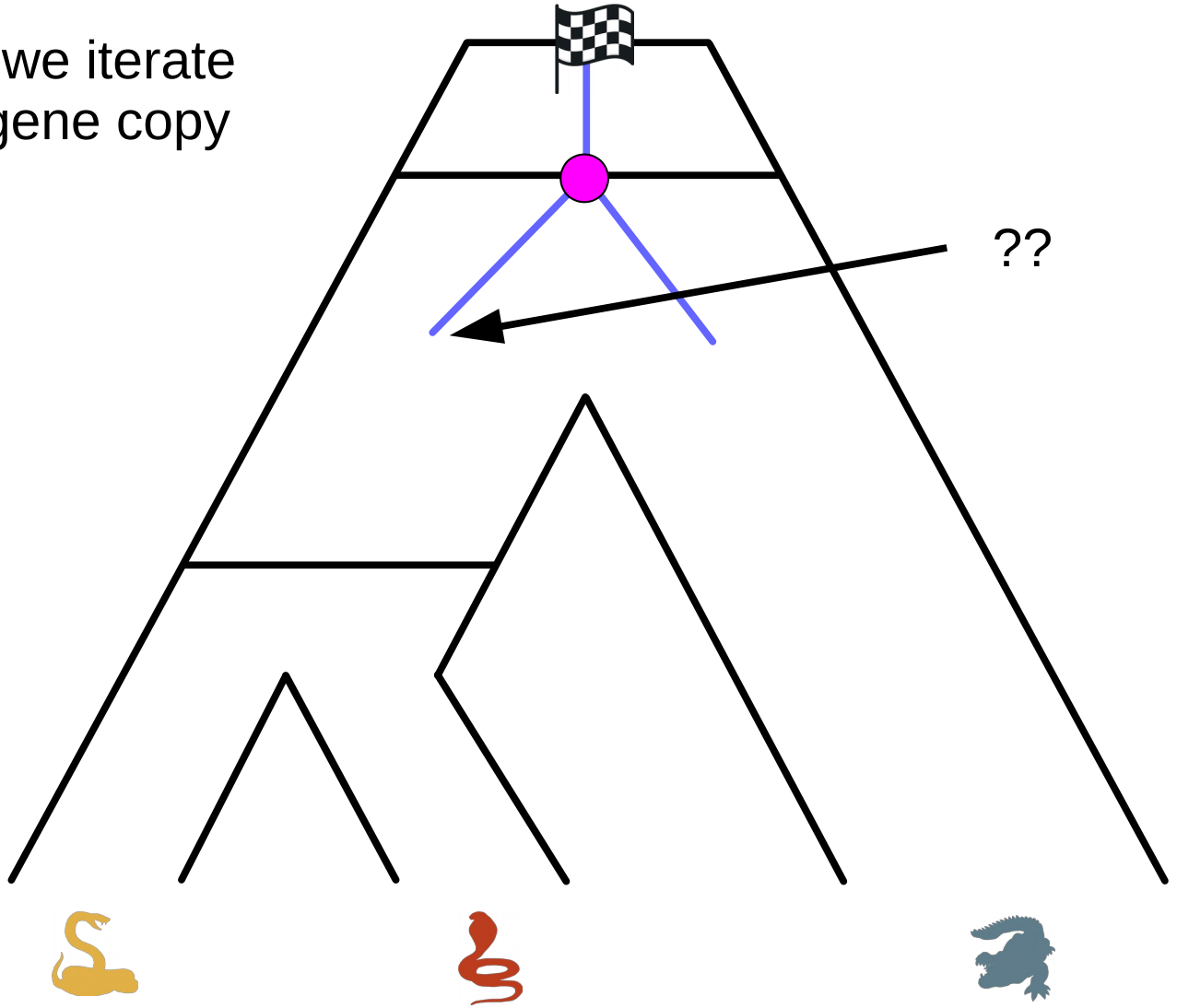
Simulating under the UndatedDTL

Simulating under the UndatedDTL



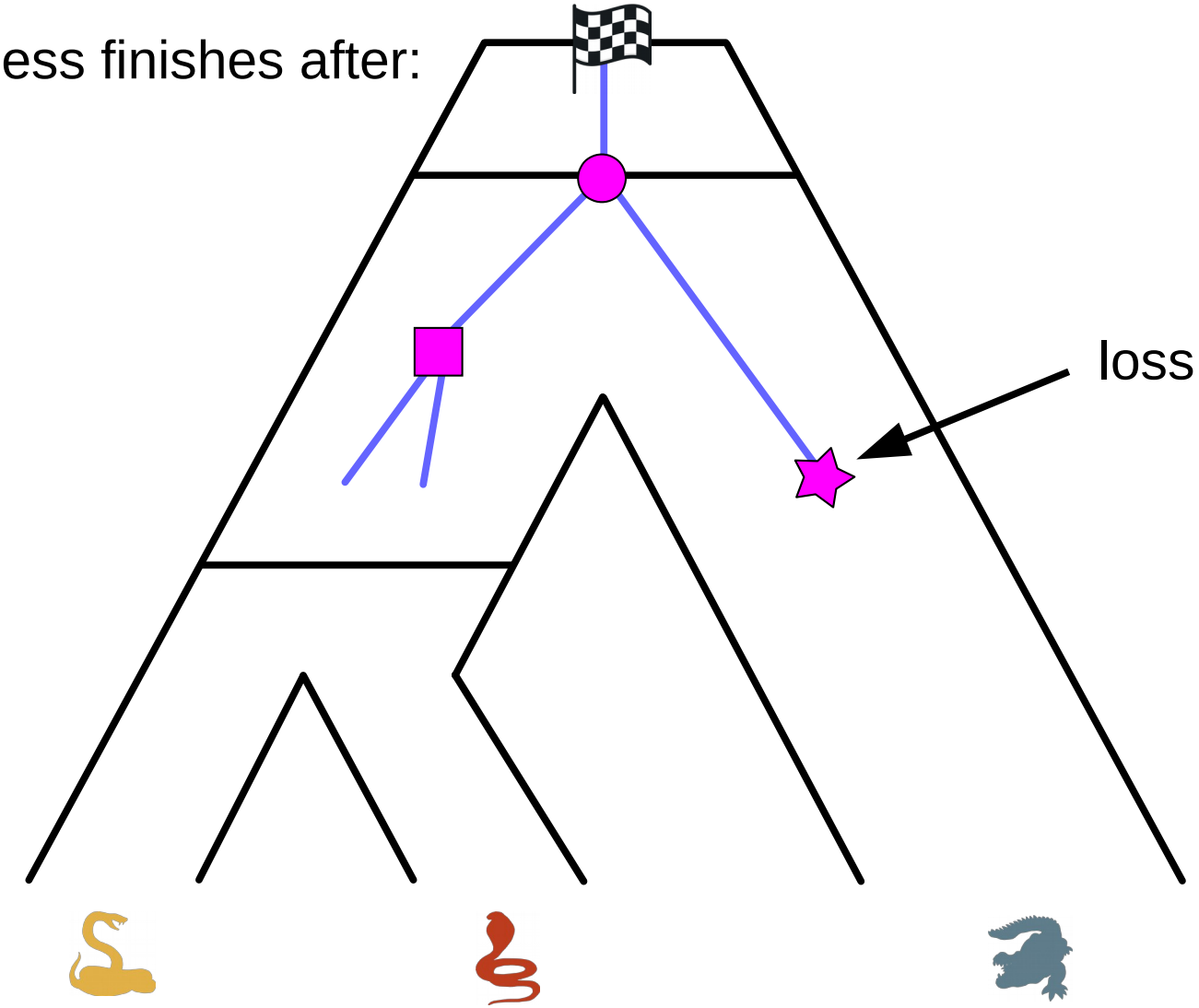
Simulating under the UndatedDTL

And then we iterate
for each gene copy



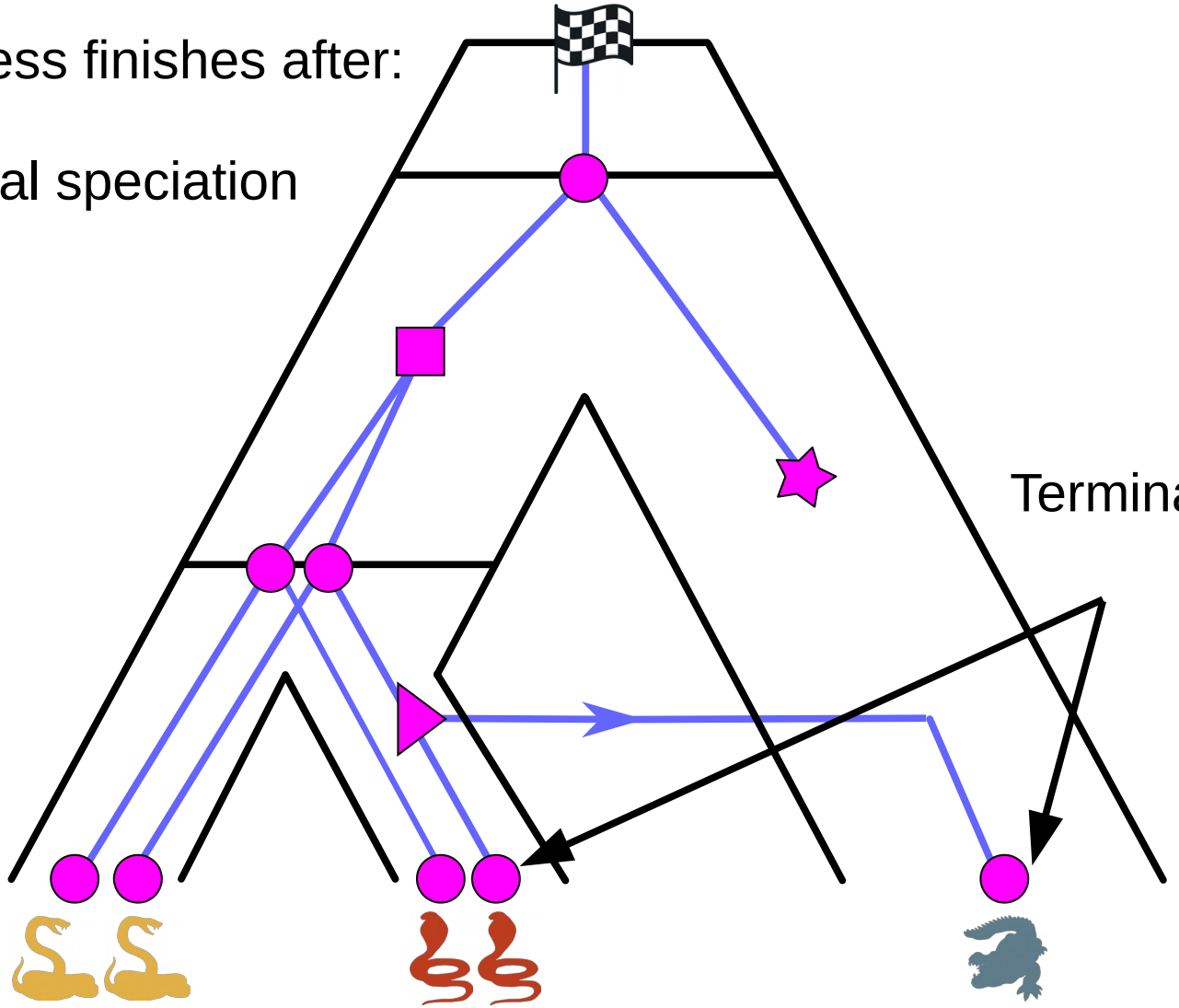
Simulating under the UndatedDTL

The process finishes after:
- a loss



Simulating under the UndatedDTL

The process finishes after:
- a loss
- a terminal speciation



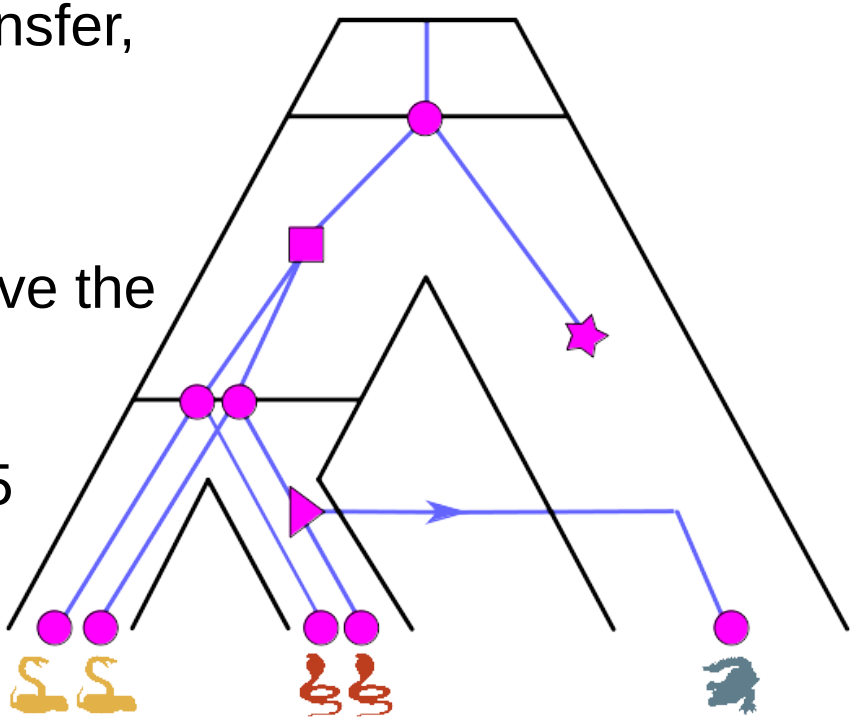
Terminal speciations

Probability of a reconciliation scenario

3 speciations, 1 duplication, 1 loss, 1 transfer,
5 terminal speciations

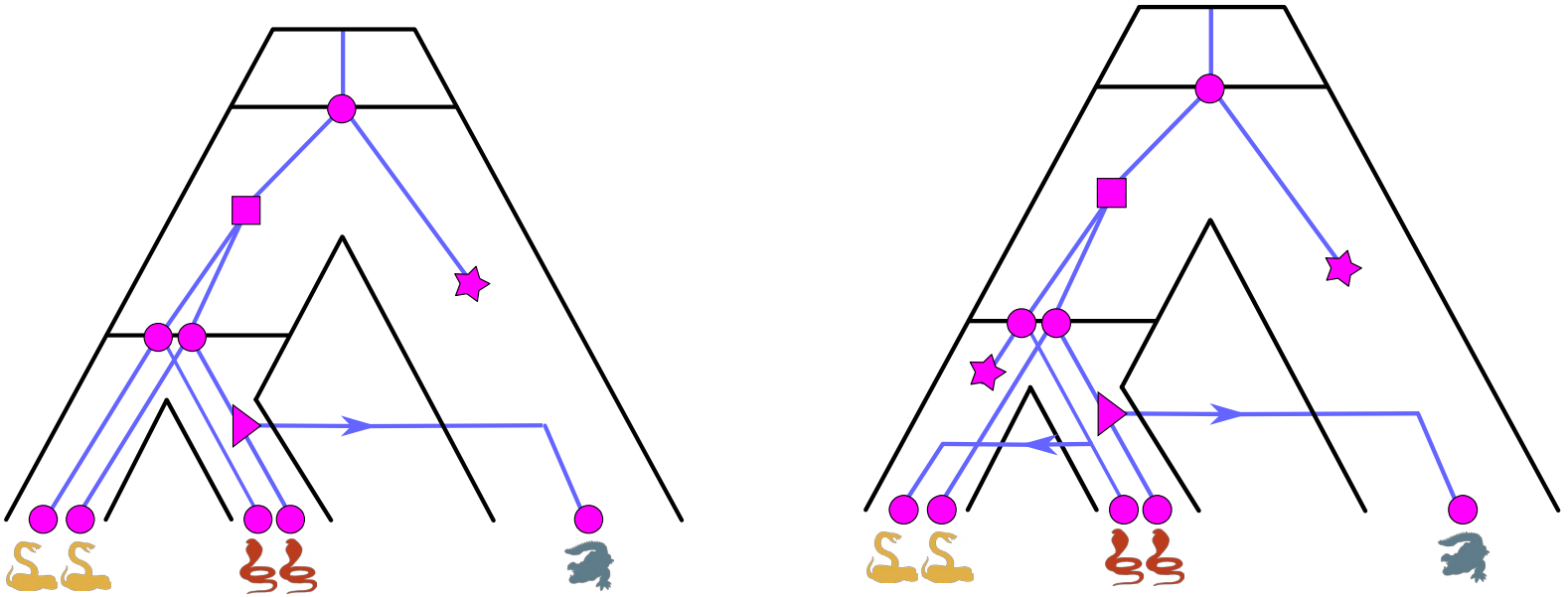
N = the number of species nodes
n = the number of species that can receive the transferred gene

→ N^-1 * pS^3 * pD * pL * (pT / n) * pS^5



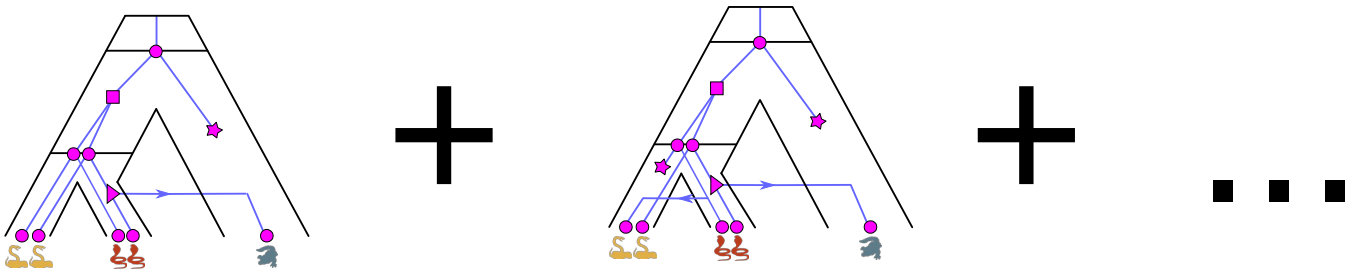
Gene tree

Several distinct senarios can produce the same gene tree



Probability of a gene tree given a species tree

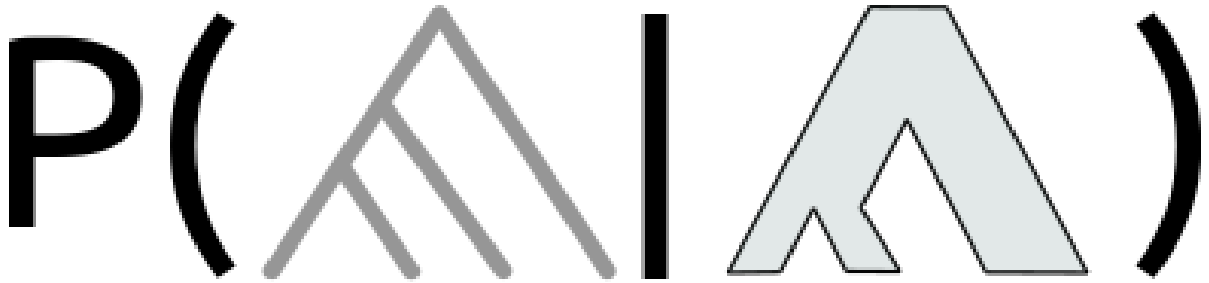
Sum of the probabilities of all the reconciliation scenarios that induce this gene tree



The reconciliation likelihood under the UndatedDTL model

- Reconciliation likelihood of the species tree:

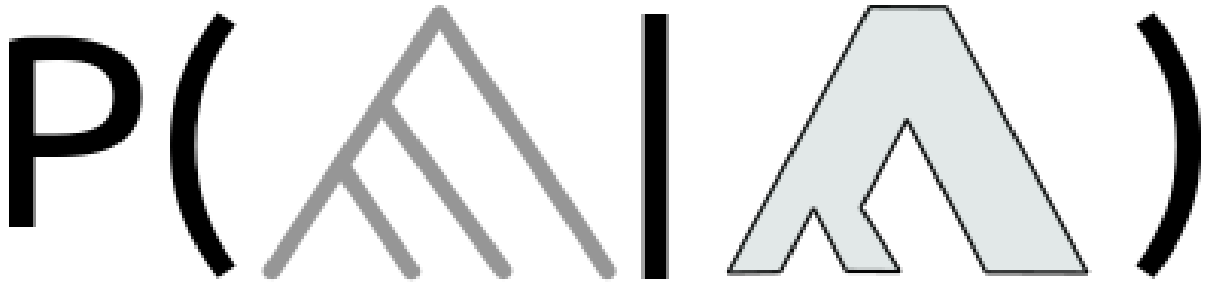
  probability of the gene tree given the species tree



The reconciliation likelihood under the UndatedDTL model

- Reconciliation likelihood of the species tree:

  probability of the gene tree given the species tree

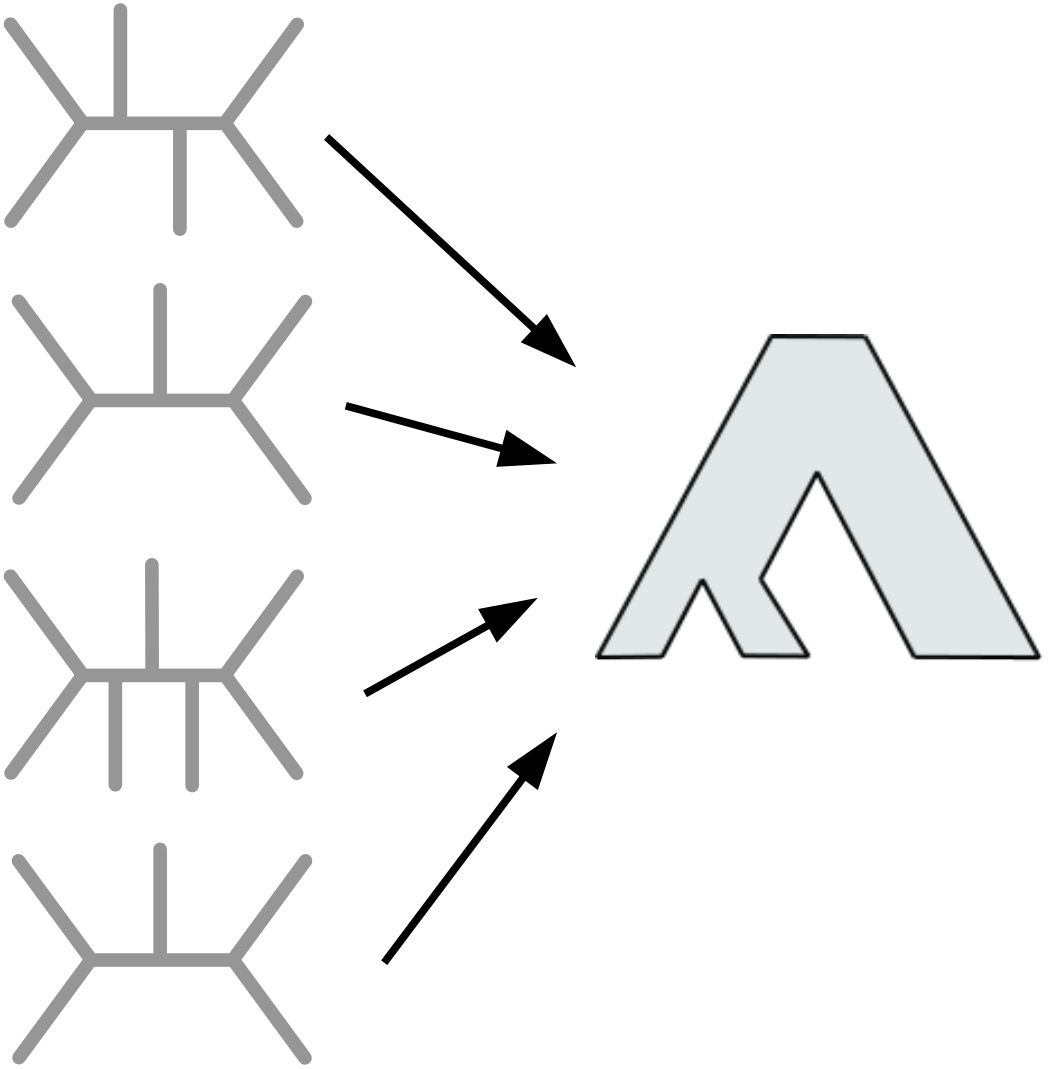


- If there are several gene trees, the likelihood of the species tree is the product over all gene trees

SpeciesRax: maximum likelihood species tree inference

Maximize the reconciliation likelihood

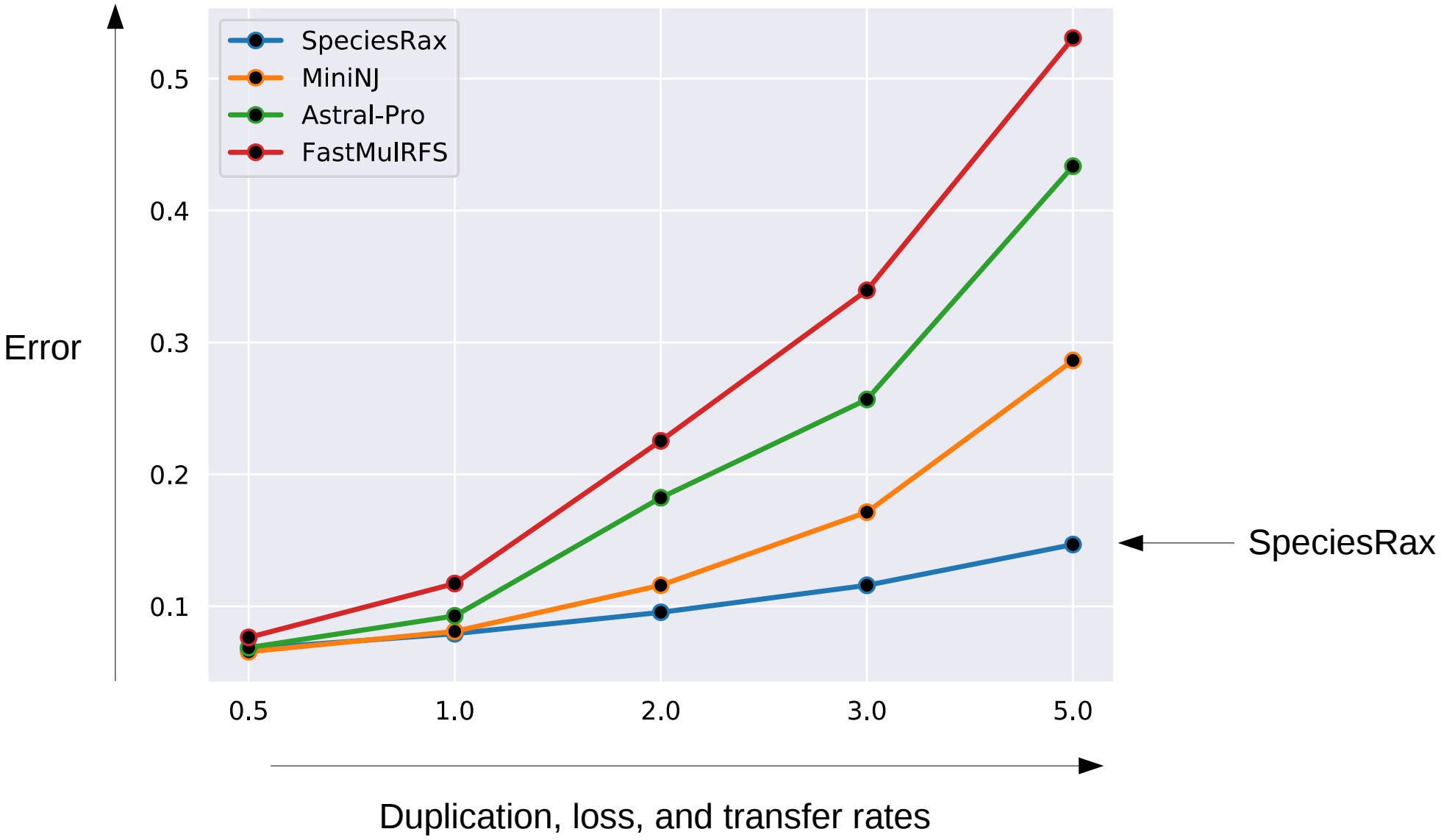
with a tree search algorithm



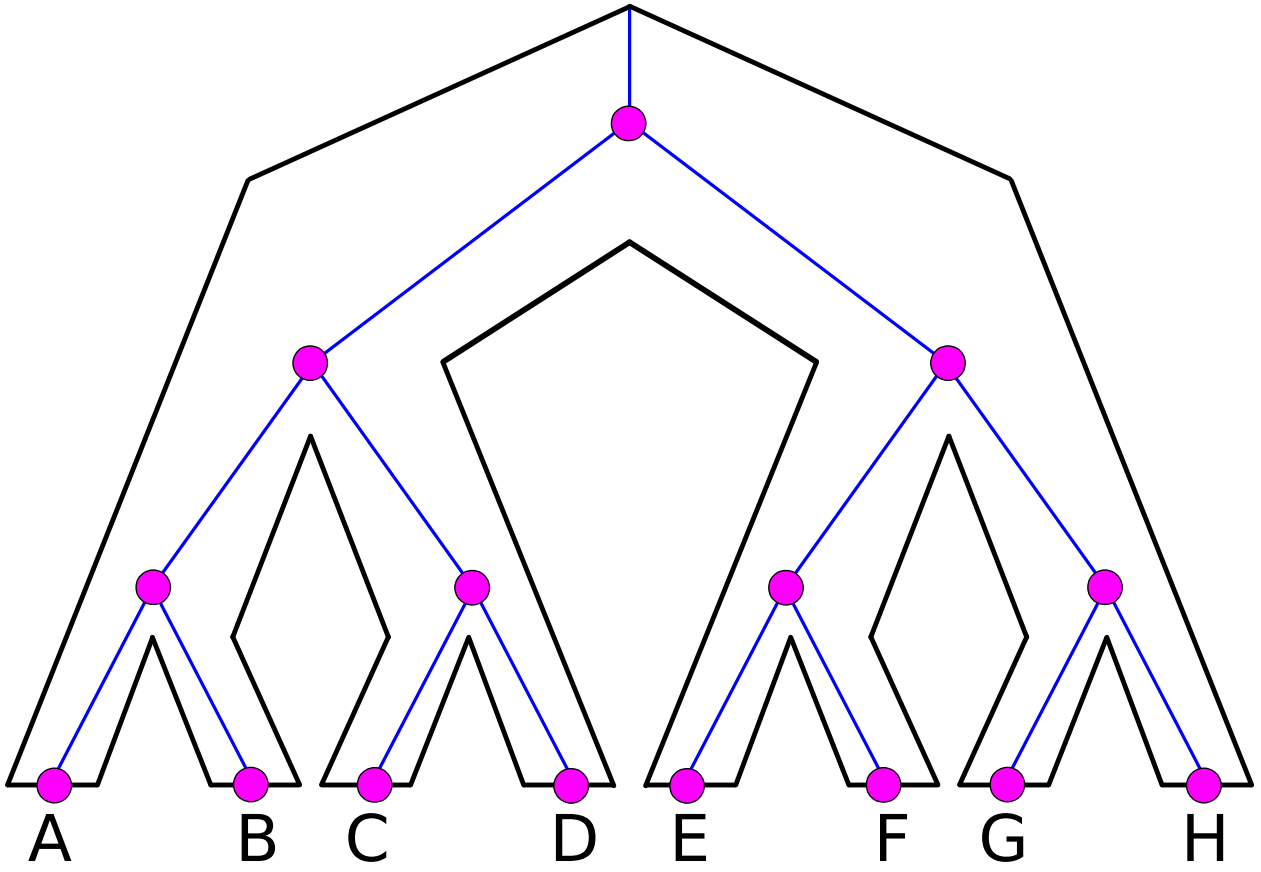
SpeciesRax infers a rooted species tree

- The reconciliation likelihood depends on the root position of the species tree
- SpeciesRax infers a **rooted** tree

Accuracy on simulated datasets

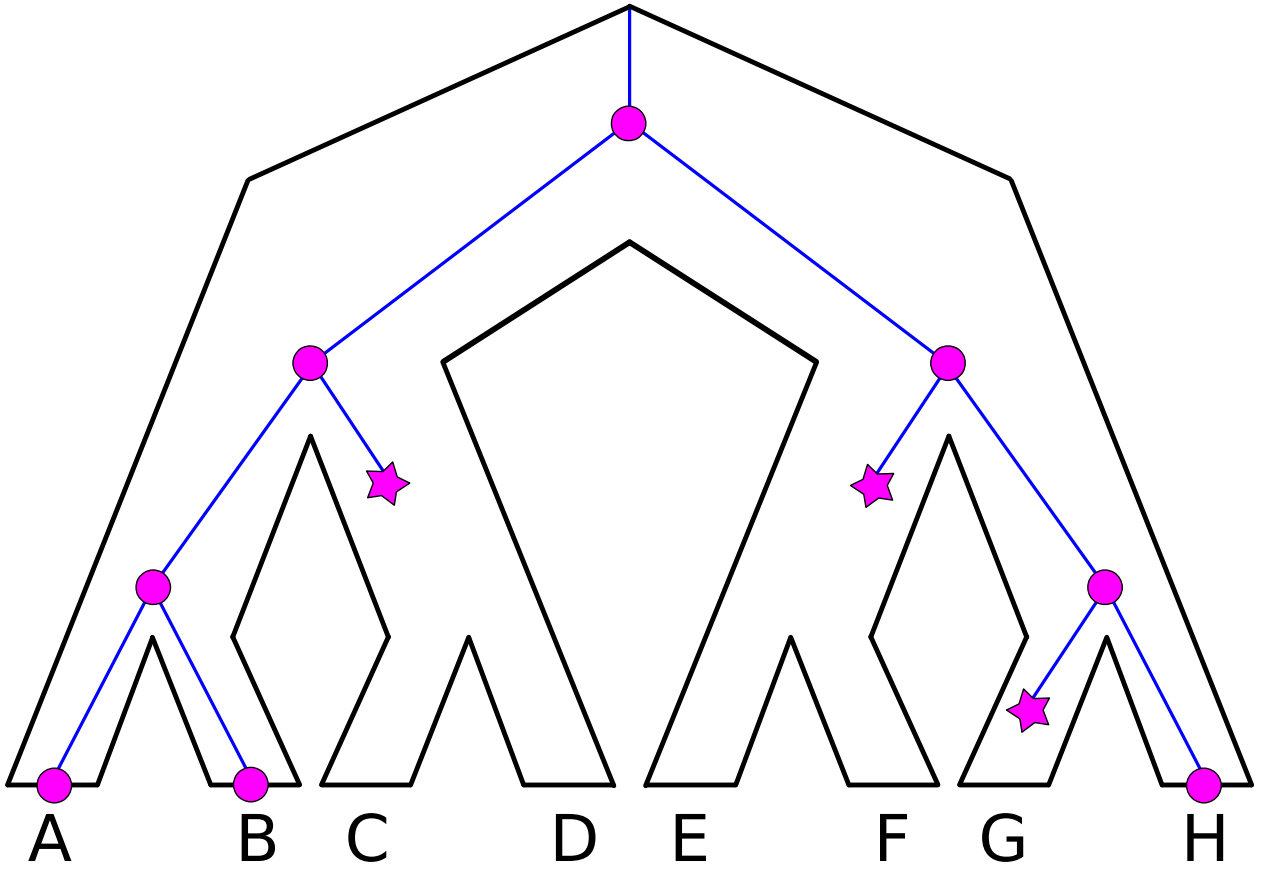


Missing data bias



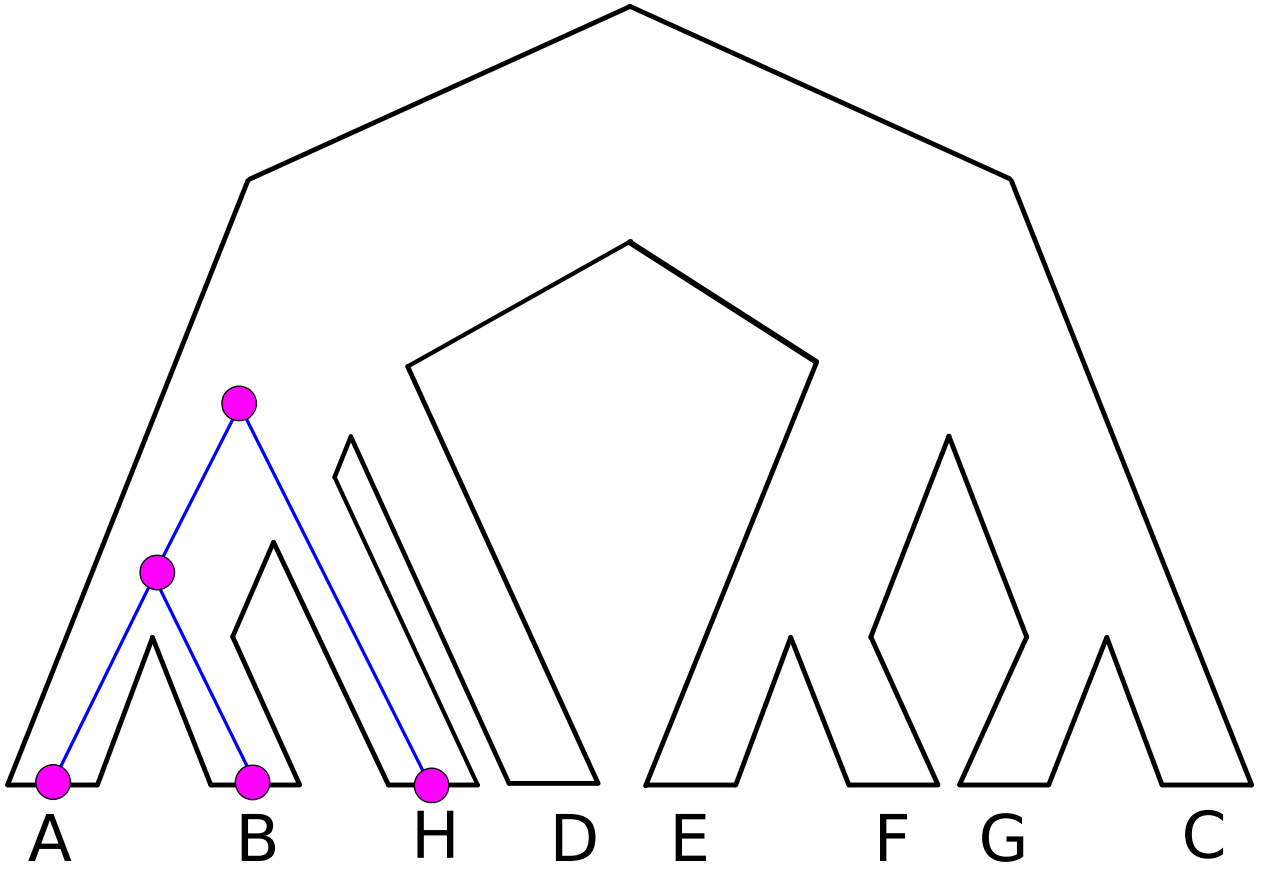
Reconciled gene tree without missing data

Missing data bias



Reconciled gene tree if this species C, D, E, F, and G were not correctly sampled

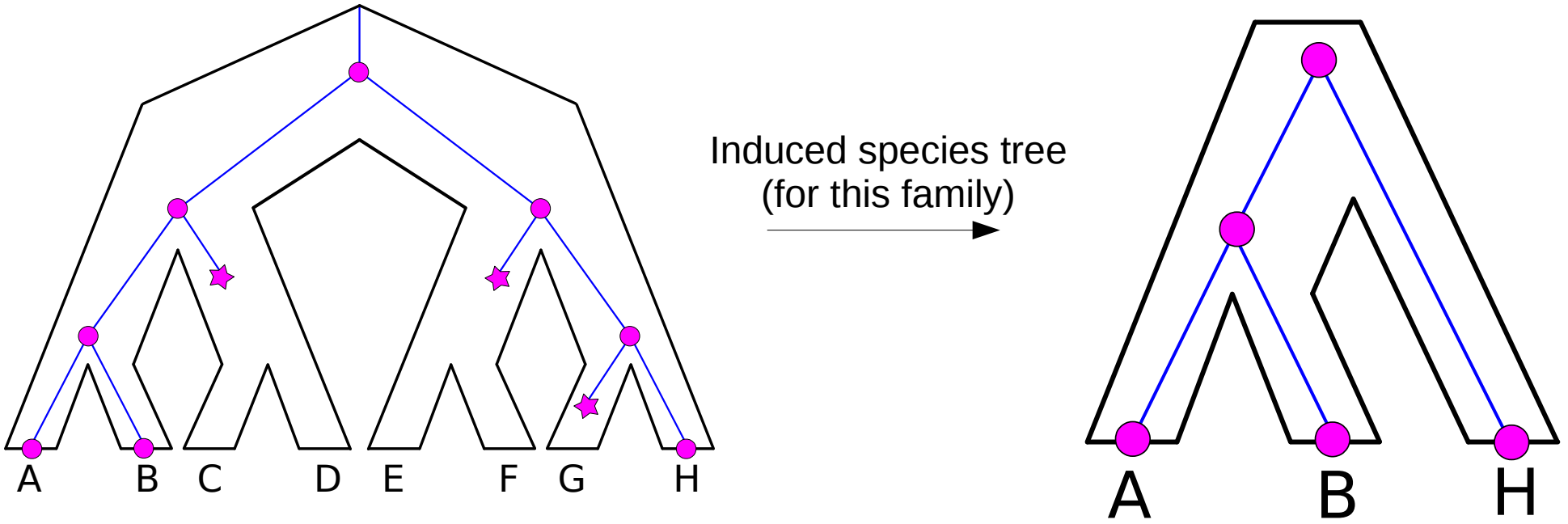
Missing data bias



Alternative (but wrong!) topology with a better likelihiood

Prune mode in SpeciesRax

- With `--prune-species-tree`, SpeciesRax computes the likelihood of the species induced by the covered species:



When to use the prune mode?

- Do you expect the missing genes to be the result of:
  - Real gene losses? → use the default mode
  - Missing data (e.g. bad sampling) → use the prune mode

Time for a break!

Rooting a species tree

- Which strategies do you know to root a species tree?

Rooting a species tree

- Which strategies do you know to root a species tree?
  - Outgroup rooting (add an external species)
  - Midpoint/MAD/MinVar rooting
  - Use reversible models
  - Use the DTL events