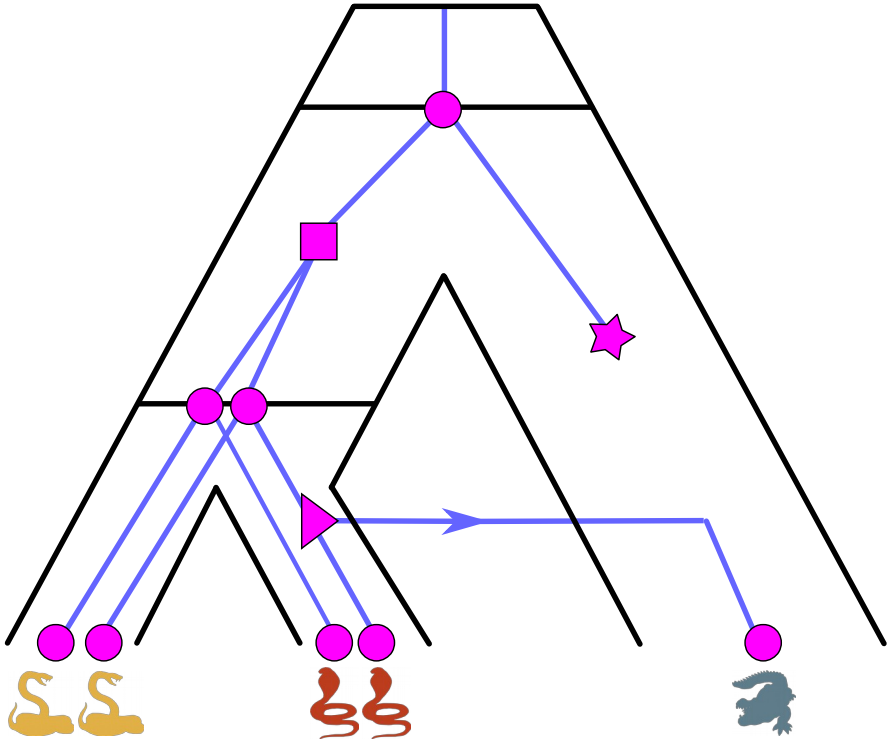
Gene tree and species tree reconciliation

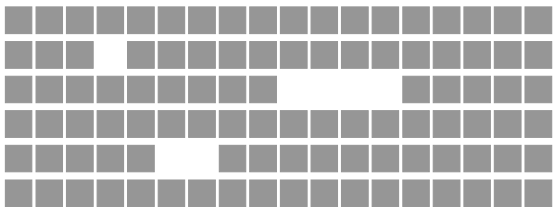
Goal

- Infer the reconciliation scenario:
  - Gene tree topology
  - Mapping between the gene tree and the species tree
  - DTL events

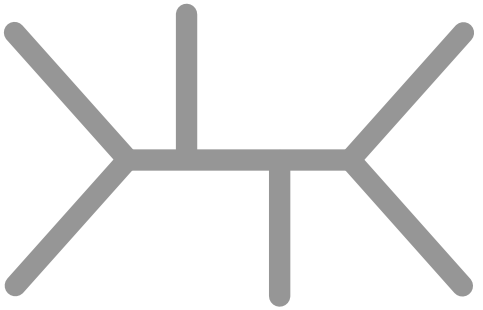


From the sequences to the gene trees

Use any tree inference method

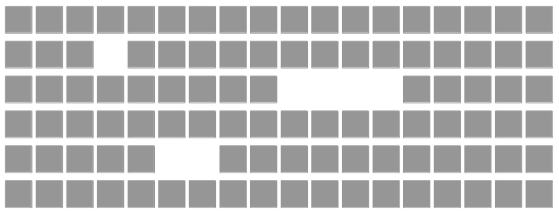


- RAxML-NG
- IQTREE-2



From the sequences to the gene trees

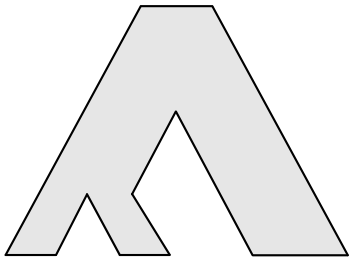
Let us assume that we can perfectly estimate the gene trees.



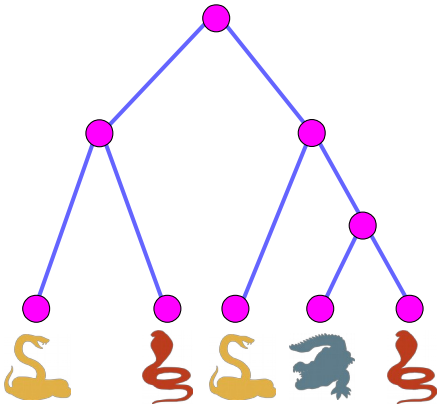
Miraculous method

True gene tree

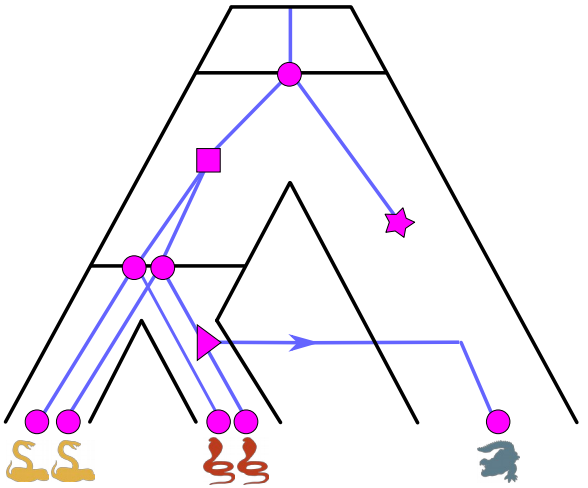
Gene tree reconciliation



Species tree

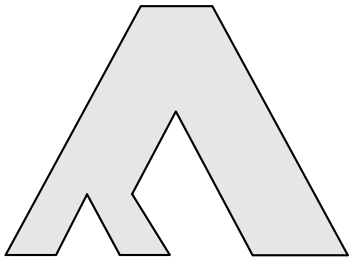


Gene tree

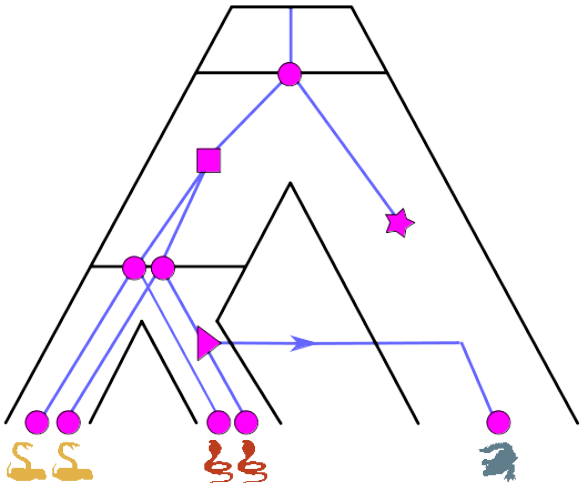


Reconciliation scenario

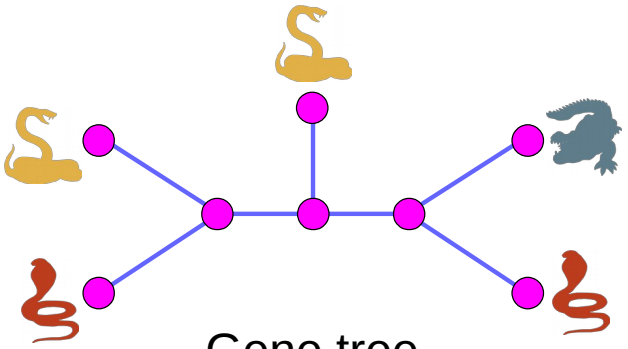
Gene tree reconciliation



Species tree

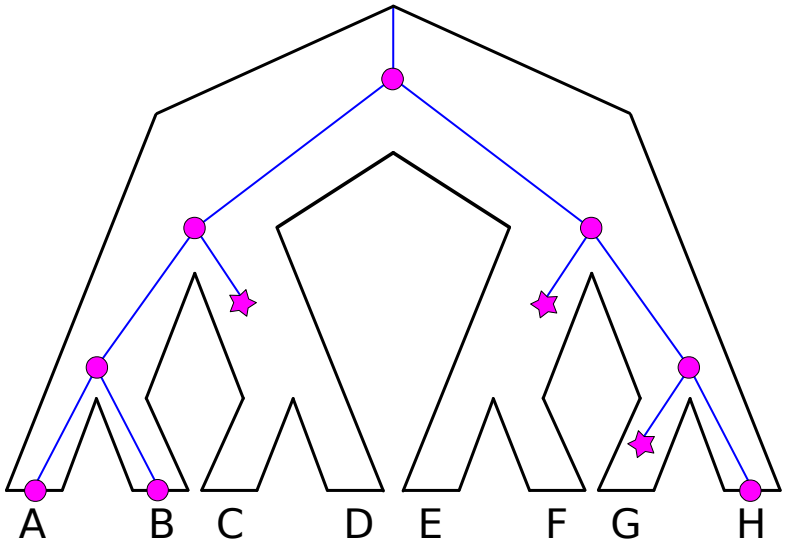


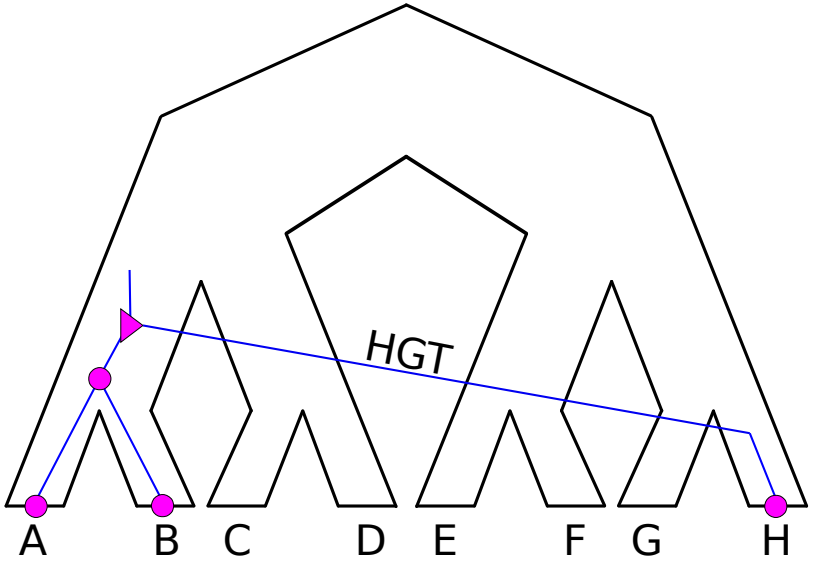
Reconciliation scenario



Gene tree

Many compatible reconciliation scenarios



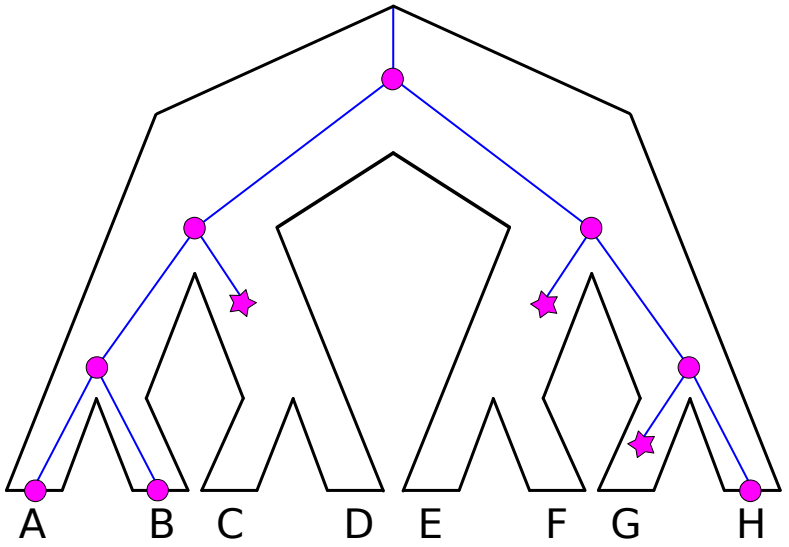


Both scenarios would produce the same gene tree.
Which one is the most plausible?

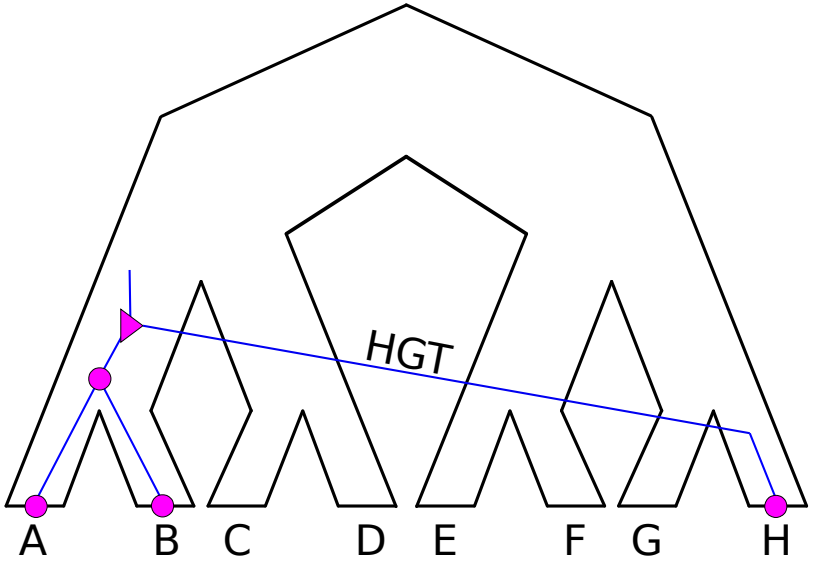
Parsimony: count the number of events

- Event costs (example):
  - D → 1
  - L → 1
  - T → 2

Parsimony: count the number of events

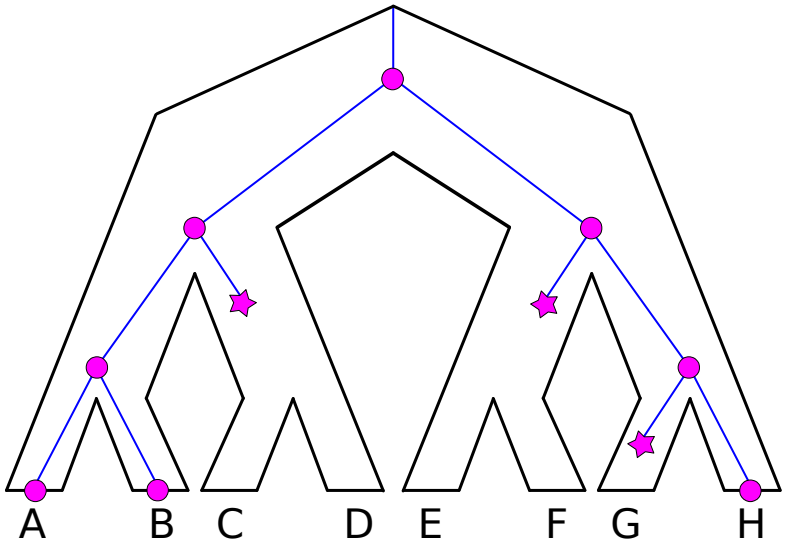


Score = 3

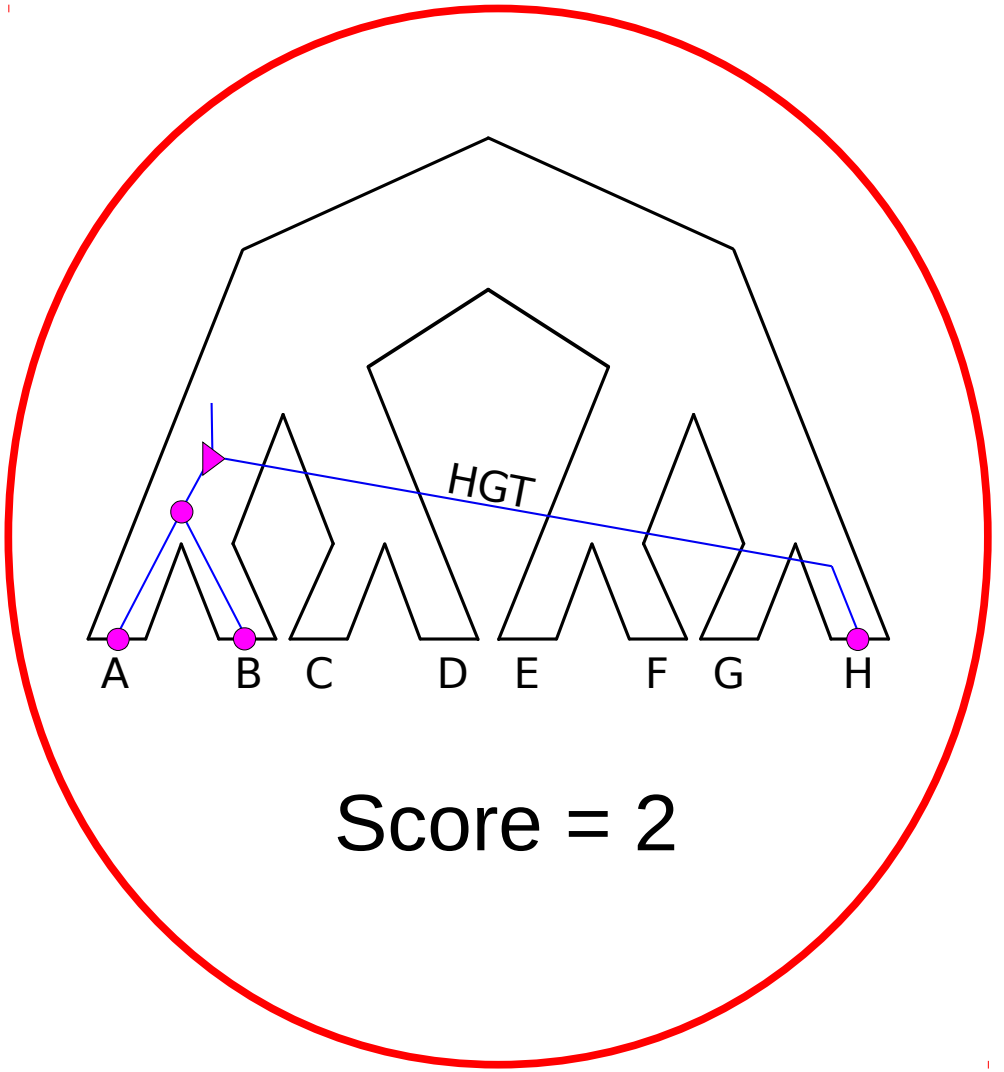


Score = 2

Parsimony: count the number of events



Score = 3

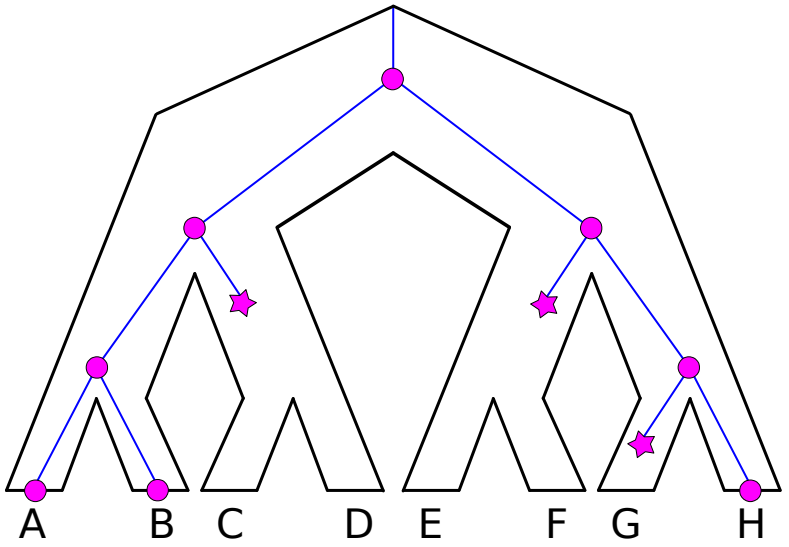


Score = 2

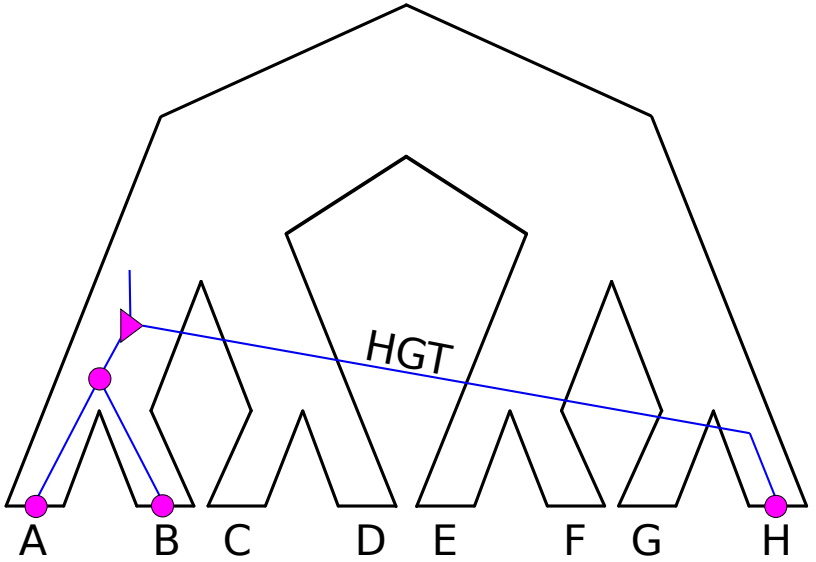
Parsimony: count the number of events

- Event costs (example):
  - D → 1
  - L → 1
  - T → 10

Parsimony: count the number of events

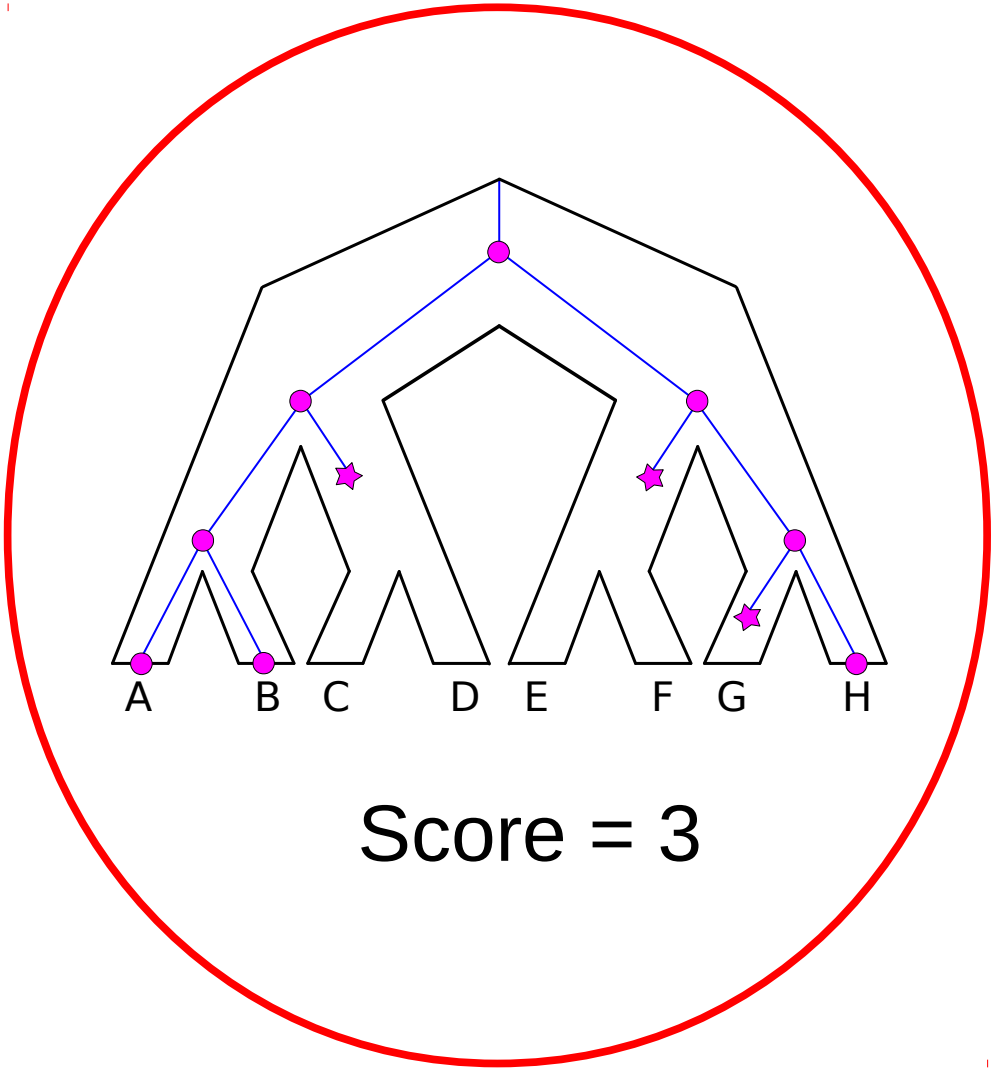


Score = 3

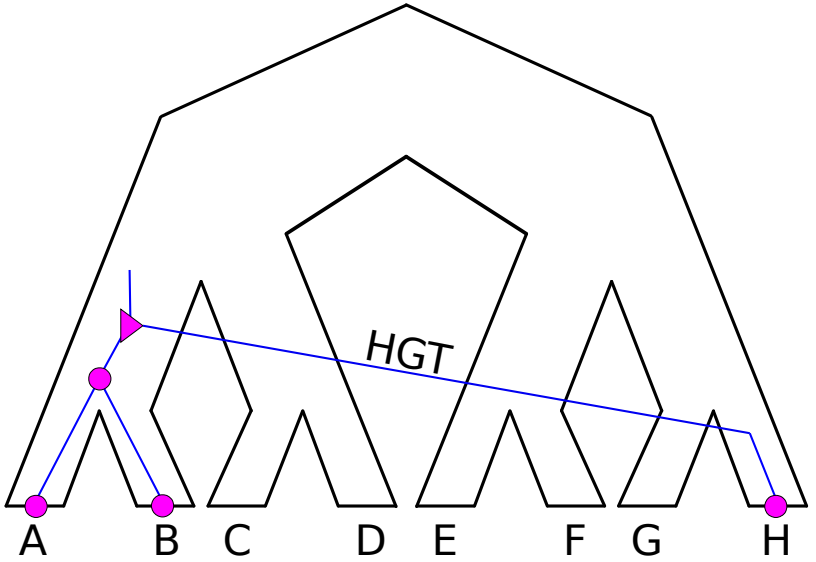


Score = 10

Parsimony: count the number of events



Score = 3

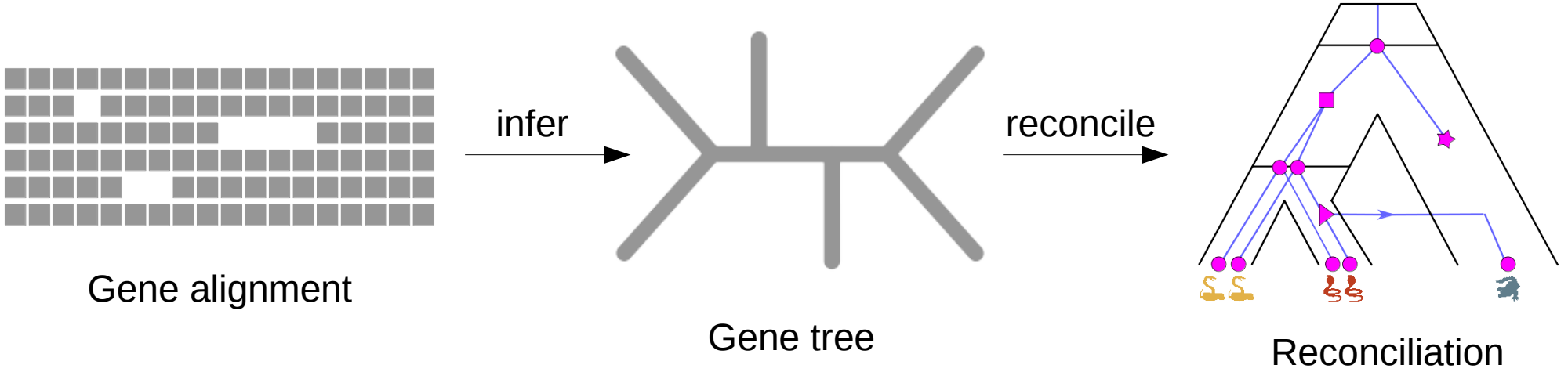


Score = 10

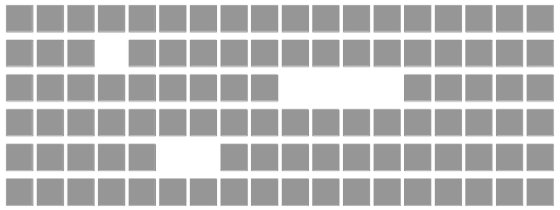
Parsimony: limitation

- Arbitrary costs!
- Ideally, the costs should be automatically estimated from the data
- Parsimony methods cannot estimate parameters

Main obstacle to reconciliation



Main obstacle to reconciliation

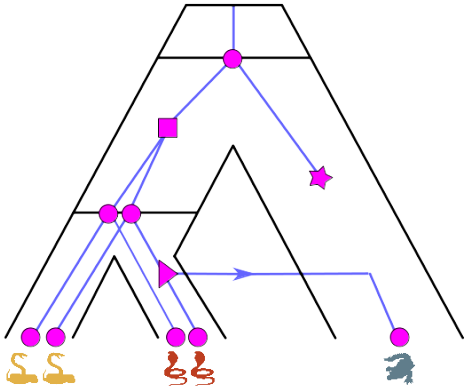


Gene alignment

infer

Estimated gene tree

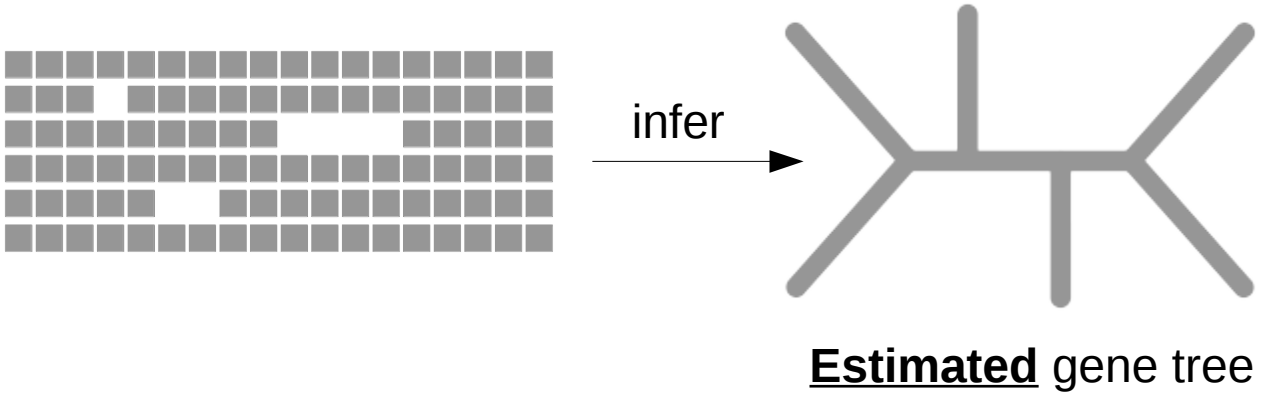
reconcile



Reconciliation

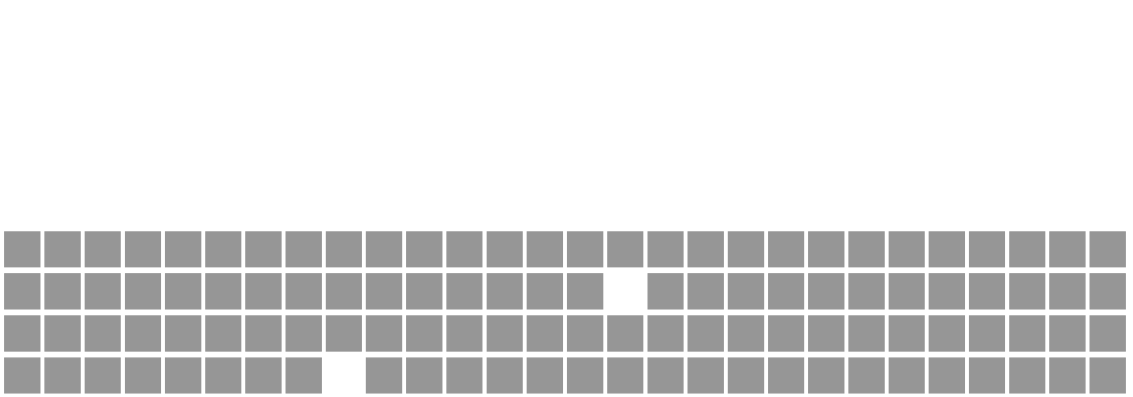
Gene tree error

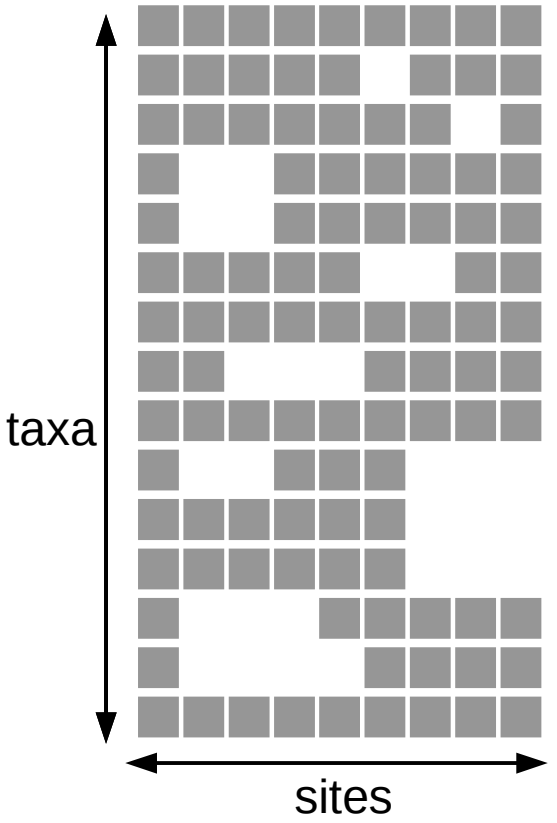
- Another source of conflict: gene tree error
- Tree inference is difficult!



Easy and difficult alignments

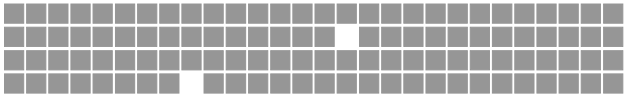
Which alignment is “easy” and which one is “difficult”? Why?



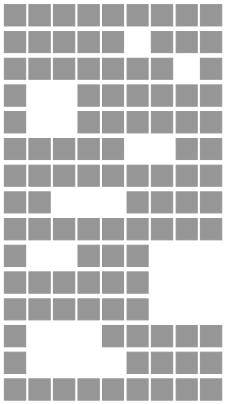


Alignment difficulty

- Easy alignments:
  - Many sites, few taxa, few gaps

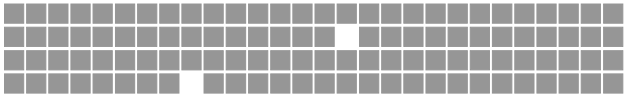


- Difficult alignments:
  - Few sites, many taxa, many gaps

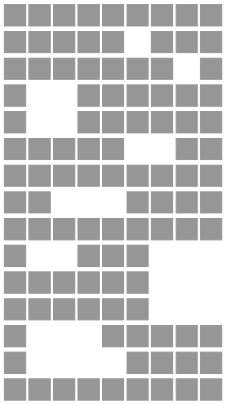


Alignment difficulty

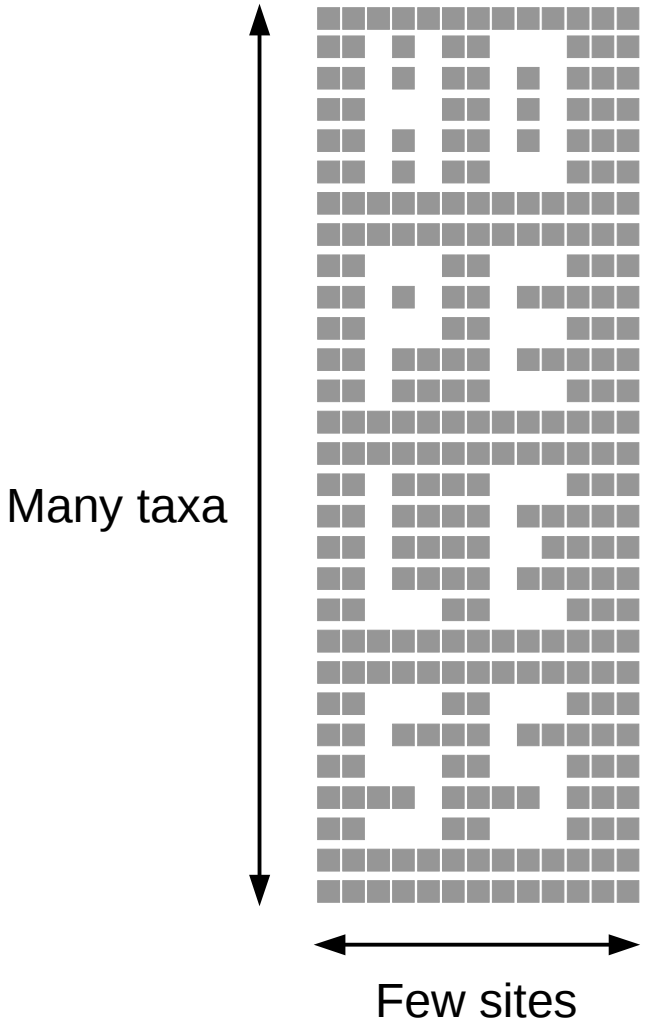
- Easy alignments:
  - Many sites, few taxa, few gaps
  - Medium distance between taxa



- Difficult alignments:
  - Few sites, many taxa, many gaps
  - Extremely short or long distances between taxa



Gene alignments are (often) difficult



- RAxML-NG
- IQTREE-2

→ Unreliable tree!

Quiz

What happens when trying to reconcile **wrong gene trees** with a species tree?

Quizz

What happens when trying to reconcile **wrong gene trees** with a species tree?

→ more conflicts need to be reconciled

→ overestimation of DTL events

(on average)

Quiz

What happens when trying to reconcile gene trees with a **wrong species tree**?

Quizz

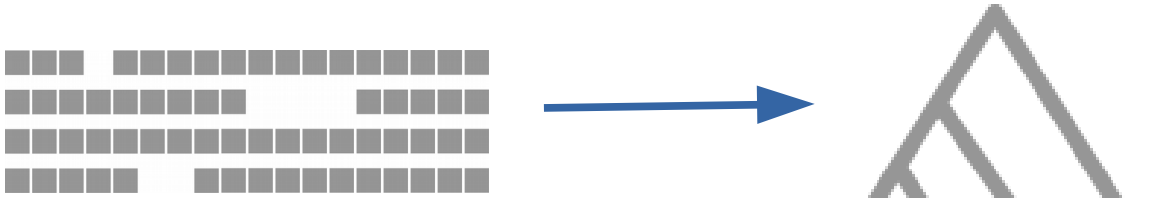
What happens when trying to reconcile gene trees with a **wrong species tree**?

→ more conflicts need to be reconciled

→ overestimation of DTL events, in particular around the wrong parts of the species tree

Species tree aware methods

- Main problem: limited amount of data

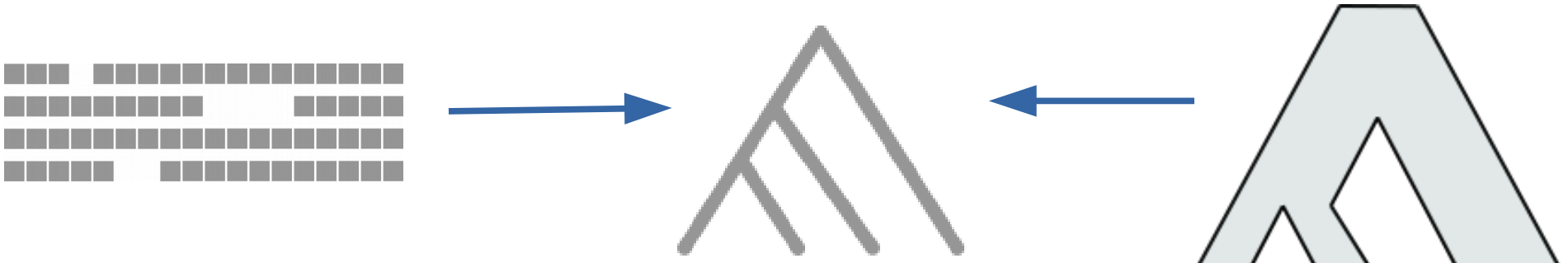


Gene sequence alignment

Gene tree

Species tree aware methods

- Main problem: limited amount of data
- Solution: use the species tree!

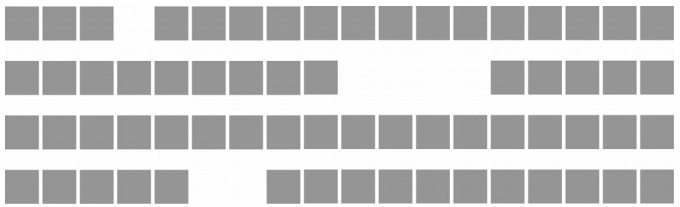


Species tree aware methods

Today, we will cover:

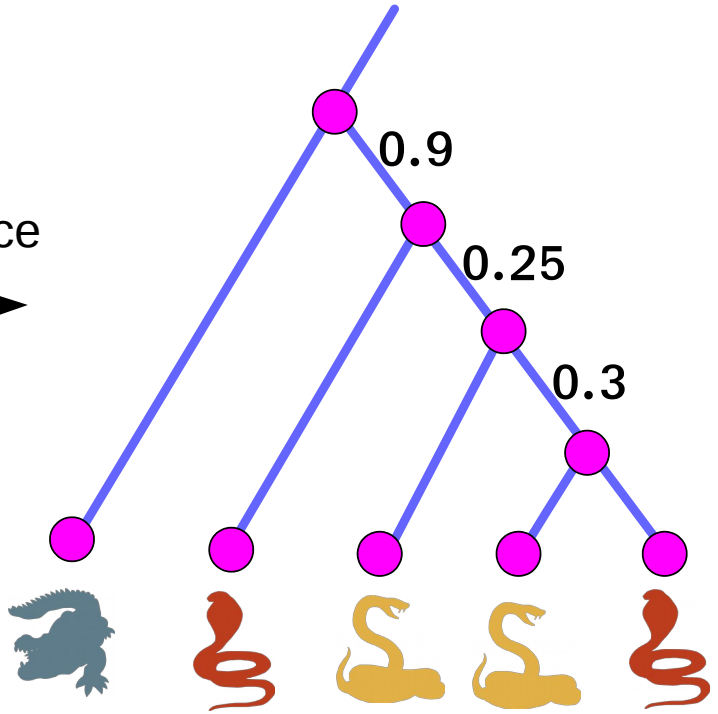
- Parsimony (Treerecs, Notung etc.)
- Maximum likelihood (GeneRax, Phyldog)
- Gene tree distributions (ALE, AleRax)

Parsimony methods



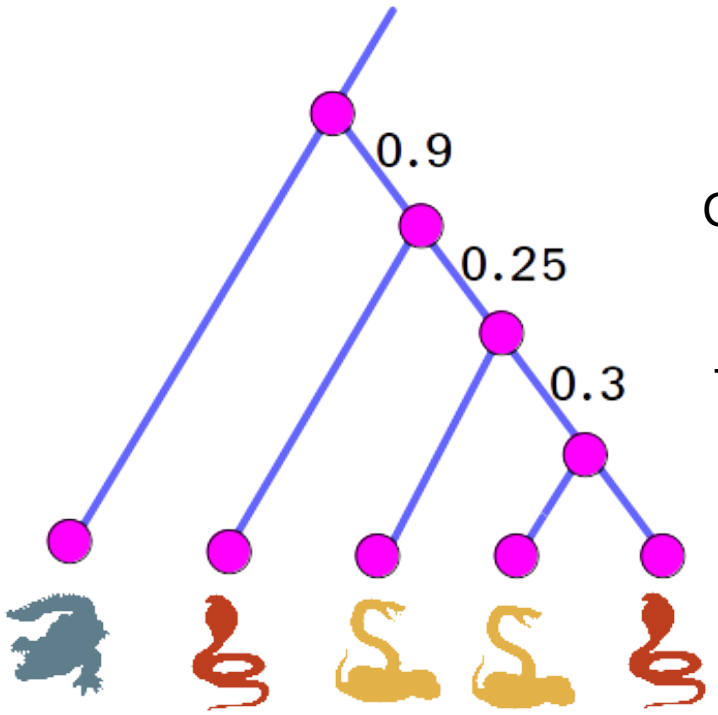
Sequences

Tree inference



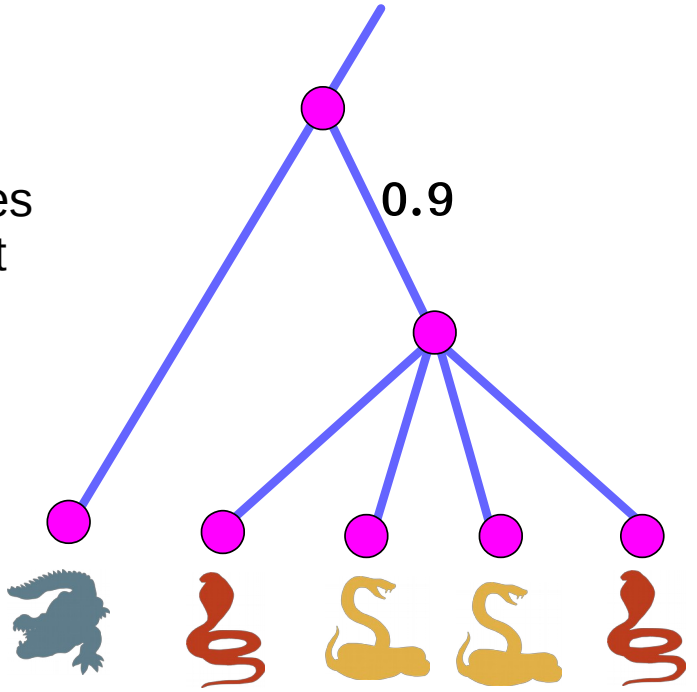
Gene tree with confidence values

Parsimony methods



Collapse branches with low support

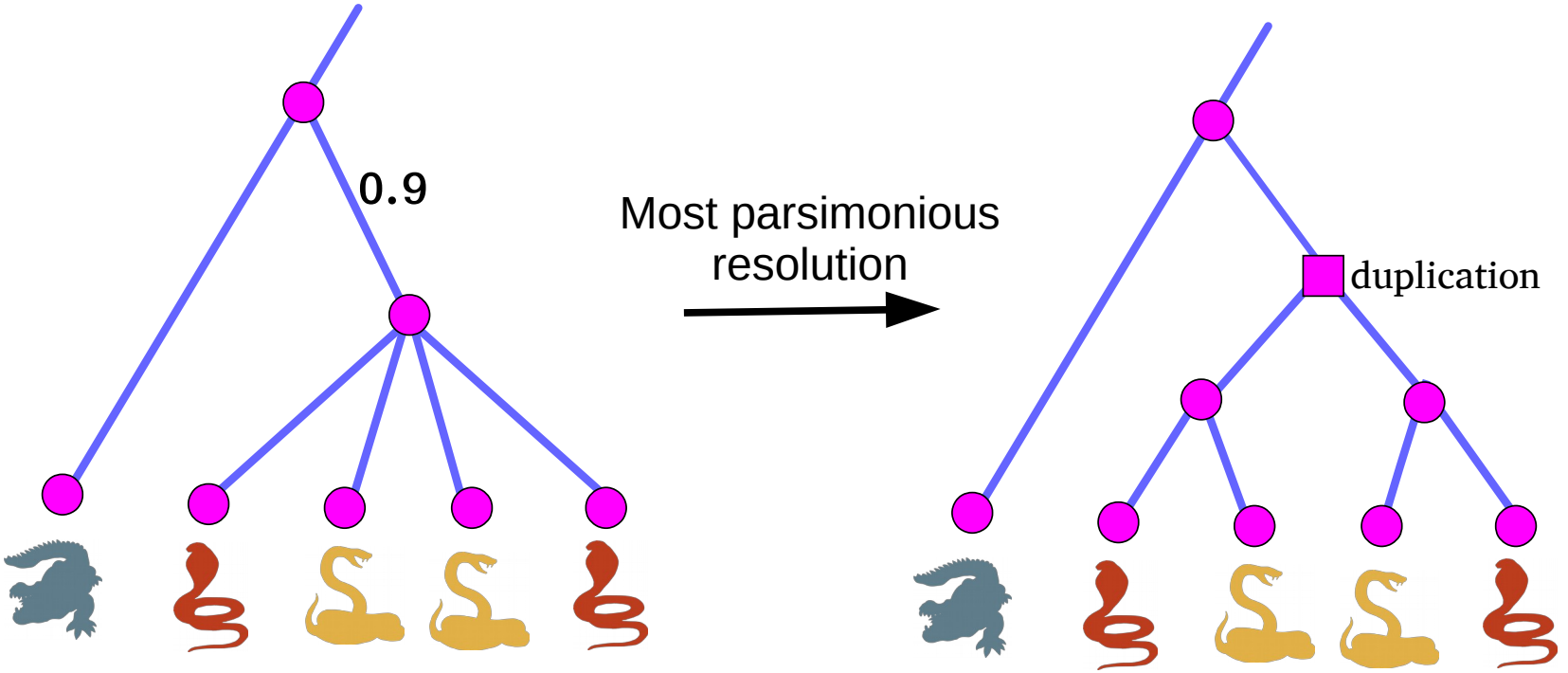
Threshold = 0.5



Gene tree with confidence values

Gene tree with polytomies

Parsimony methods



Gene tree with polytomies

Corrected gene tree + reconciliation

Limitations of non-probabilistic methods

- Confidence values:
  - expensive to compute
  - difficult to interpret (they are not probabilities!!)
- How to set the confidence threshold?

  → arbitrary choice from the user
- How to set the DTL parsimony costs?

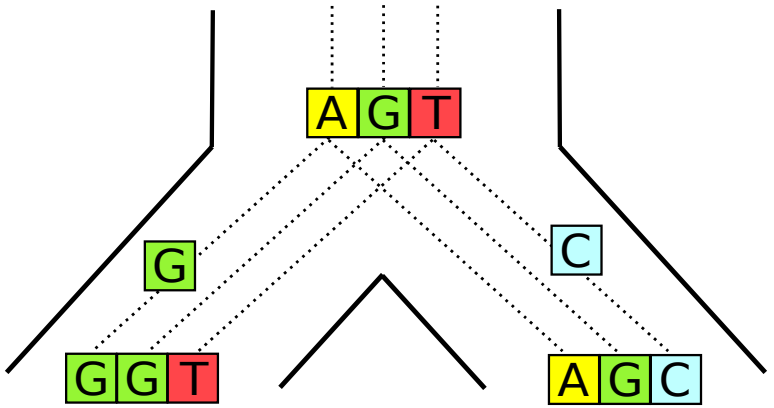
  → arbitrary choice from the user
- Do not handle HGT (as far as I know)

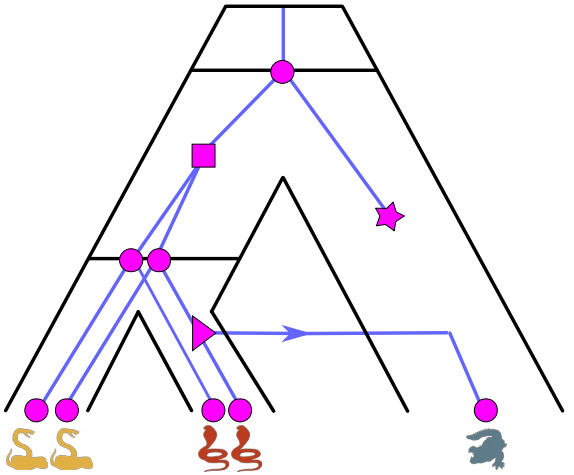
Arbitrary choices

- Threshold values:
  - High values → favor signal from the sequences
  - Low values → favor signal from the species tree

Solution: probabilistic approaches

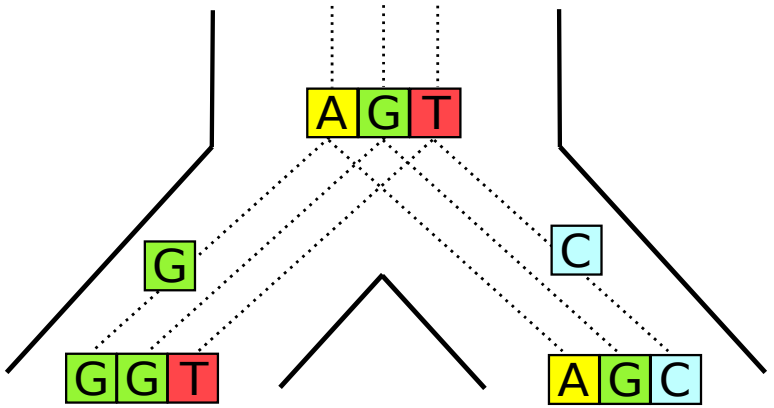
- Remember, we can describe:
  - How sequences evolve in the gene tree
  - How genes evolve in the species tree

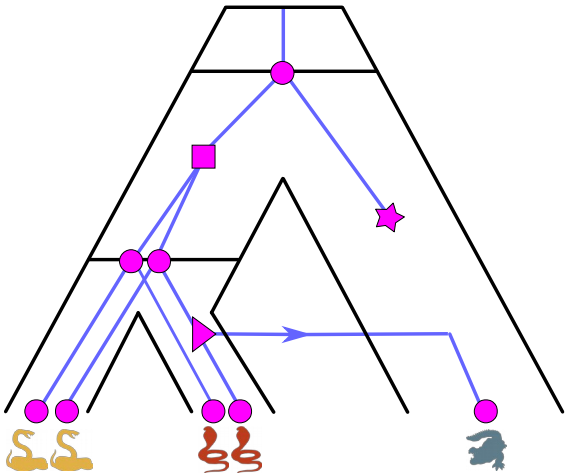




Solution: probabilistic approaches

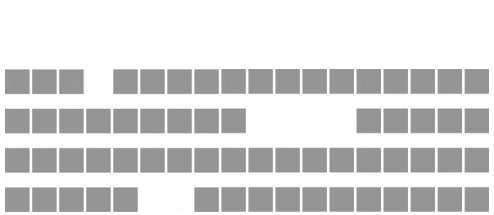
- We can define and compute:
  - The phylogenetic likelihood P(|)
  - The reconciliation likelihood P(|)

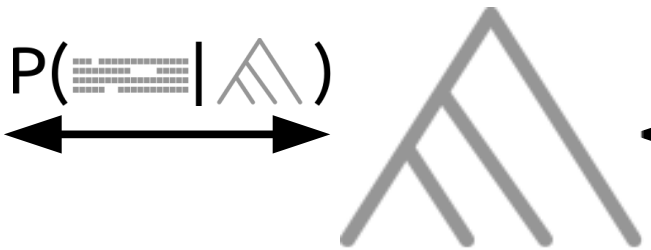




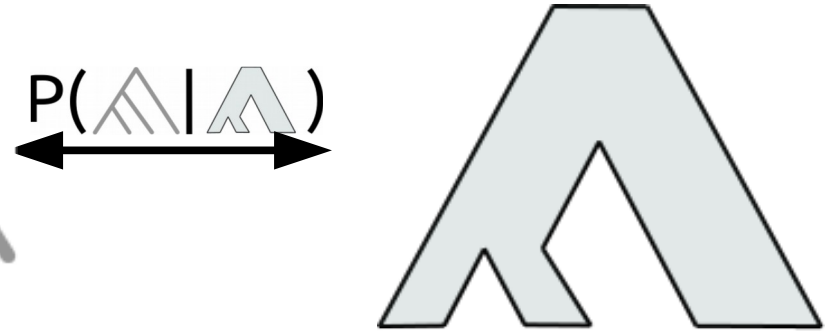
Models of sequence and gene evolution

Phylogenetic likelihood





Reconciliation likelihood

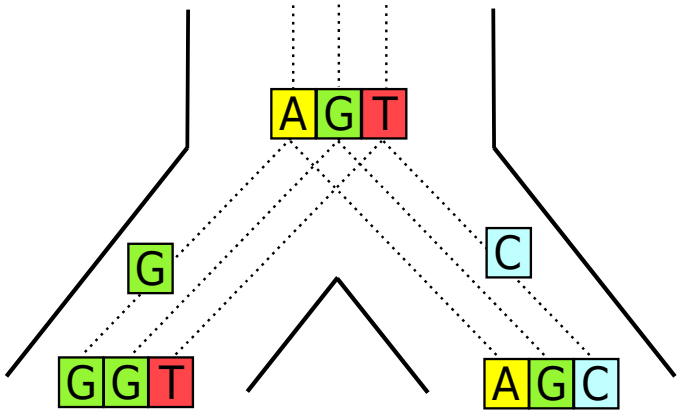


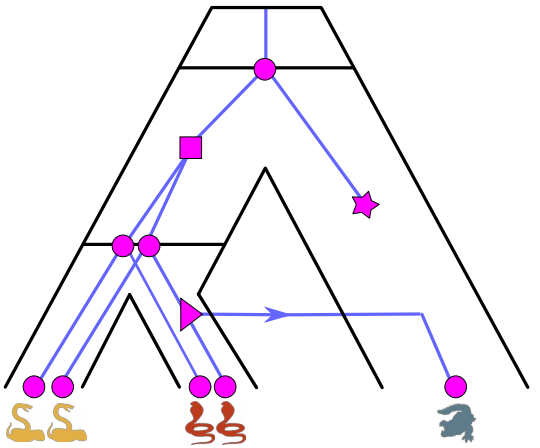
GeneRax

Find the gene tree that maximizes the joint likelihood:

Phylogenetic likelihood

Reconciliation likelihood



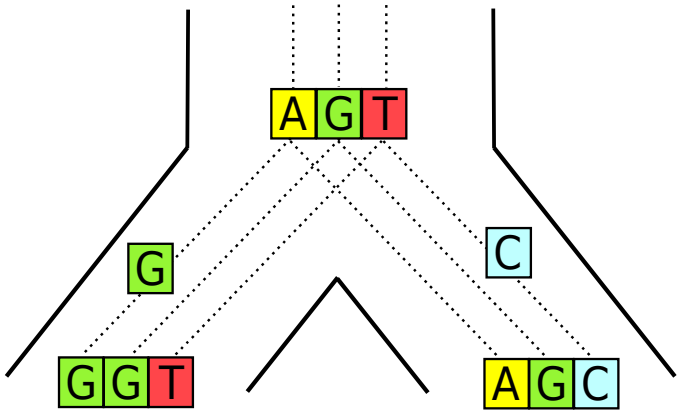


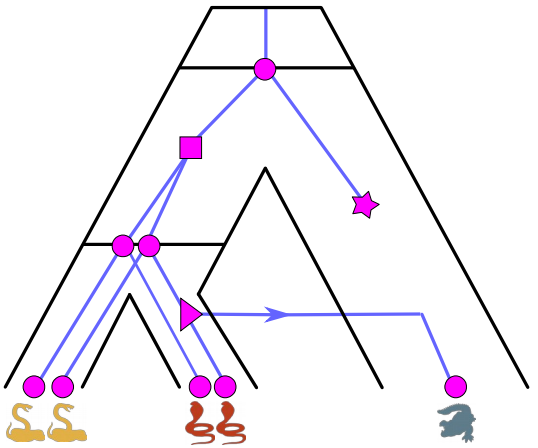
GeneRax

Find the gene tree that maximizes the joint likelihood:

Phylogenetic likelihood

Reconciliation likelihood





GeneRax

Find the gene tree that maximizes the joint likelihood:

The balance between the signals from the sequence and from the species tree is “automatic”

GeneRax parameters

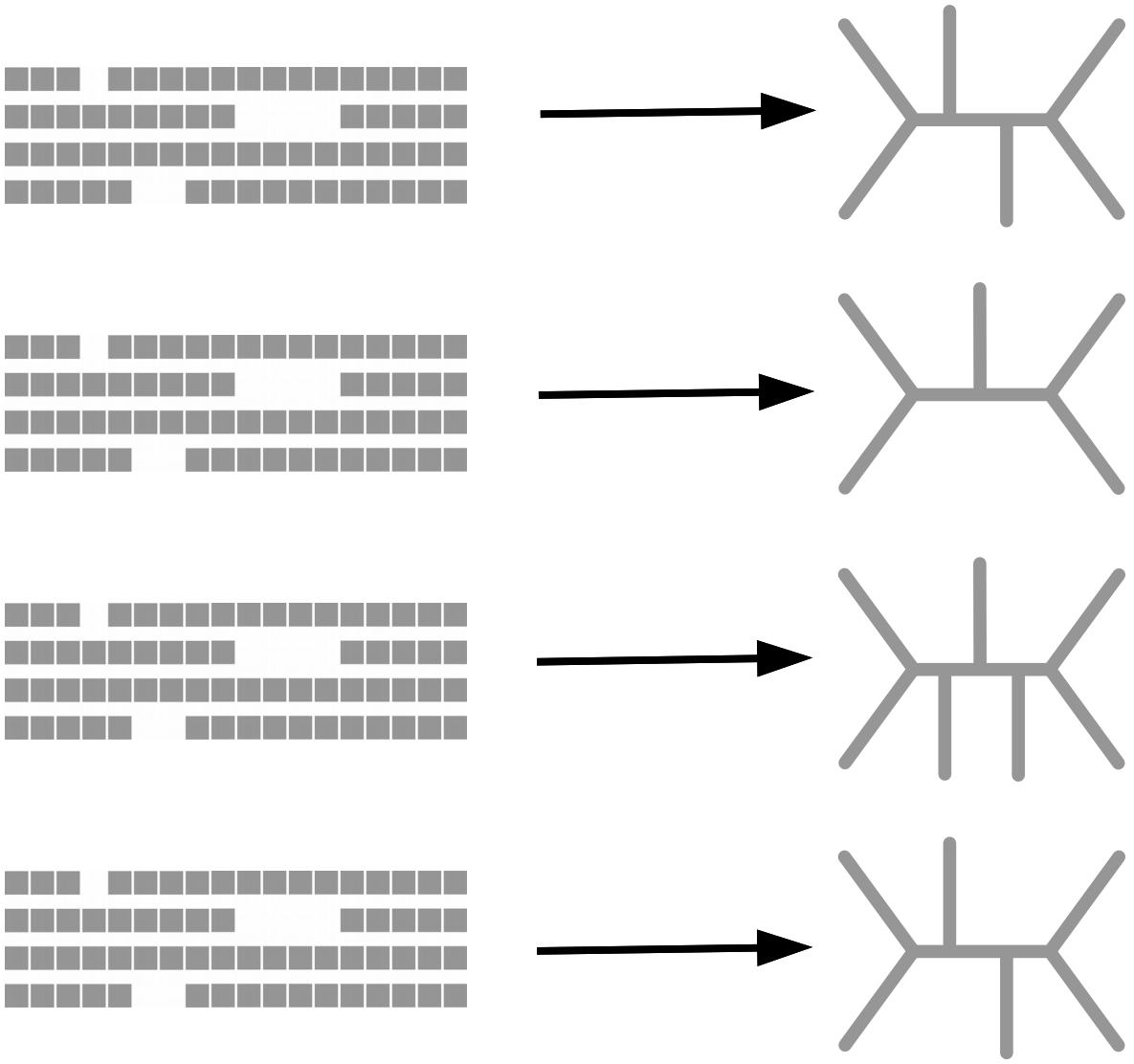
- GeneRax co-estimates:
  - the gene tree rooted topology
  - its branch lengths
  - the DTL probabilities
  - the substitution model parameters
  - the reconciliation with the species tree

  with maximum likelihood

Search algorithm

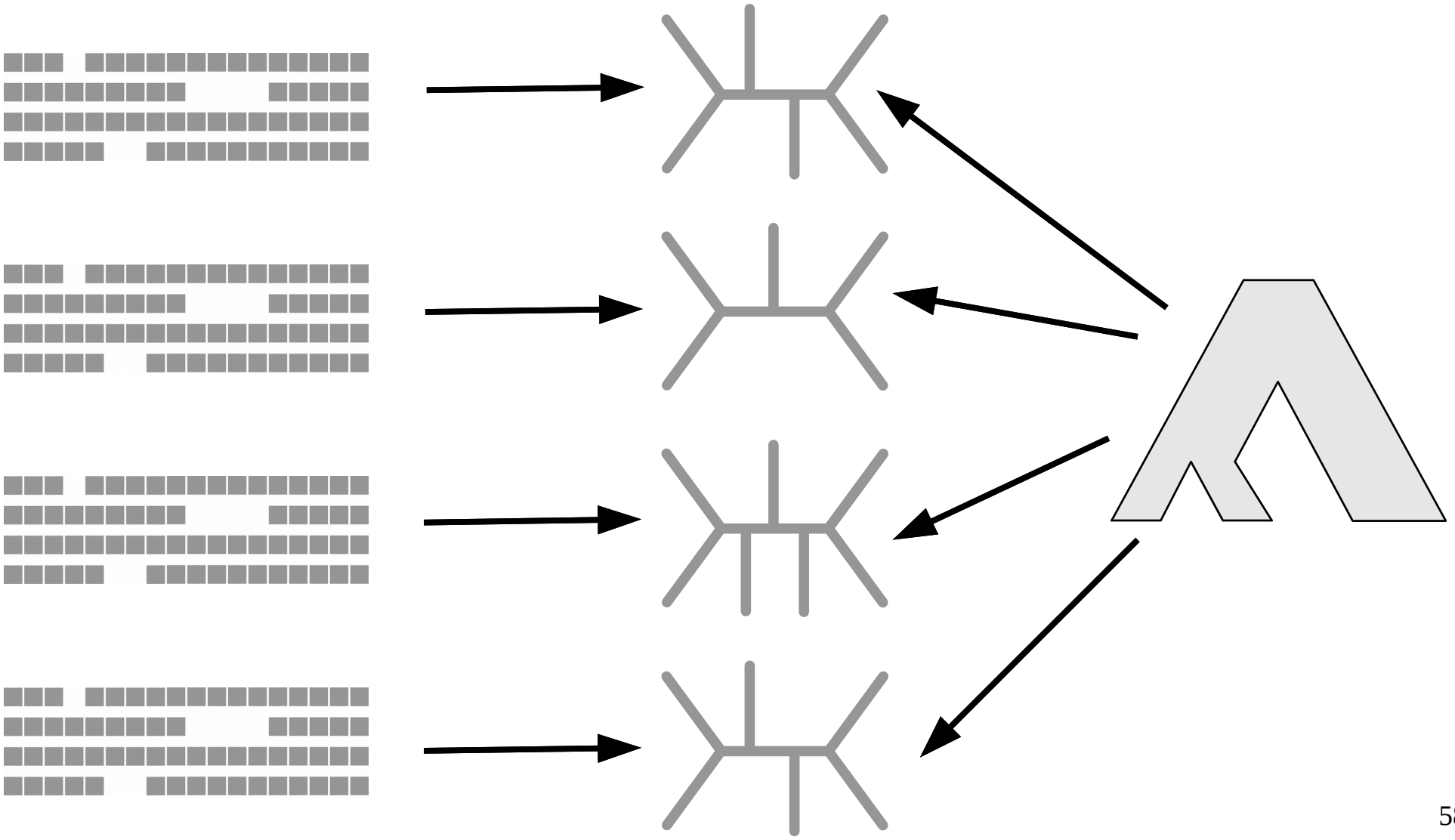
- Compute an initial starting gene tree from the sequences
- Iterate several times:
  - Optimize the DTL probabilities
  - Optimize the substitution model parameters
  - Tree search on the gene tree
- And then:
  - Infer the reconciliation

GeneRax simultaneously processes multiple families



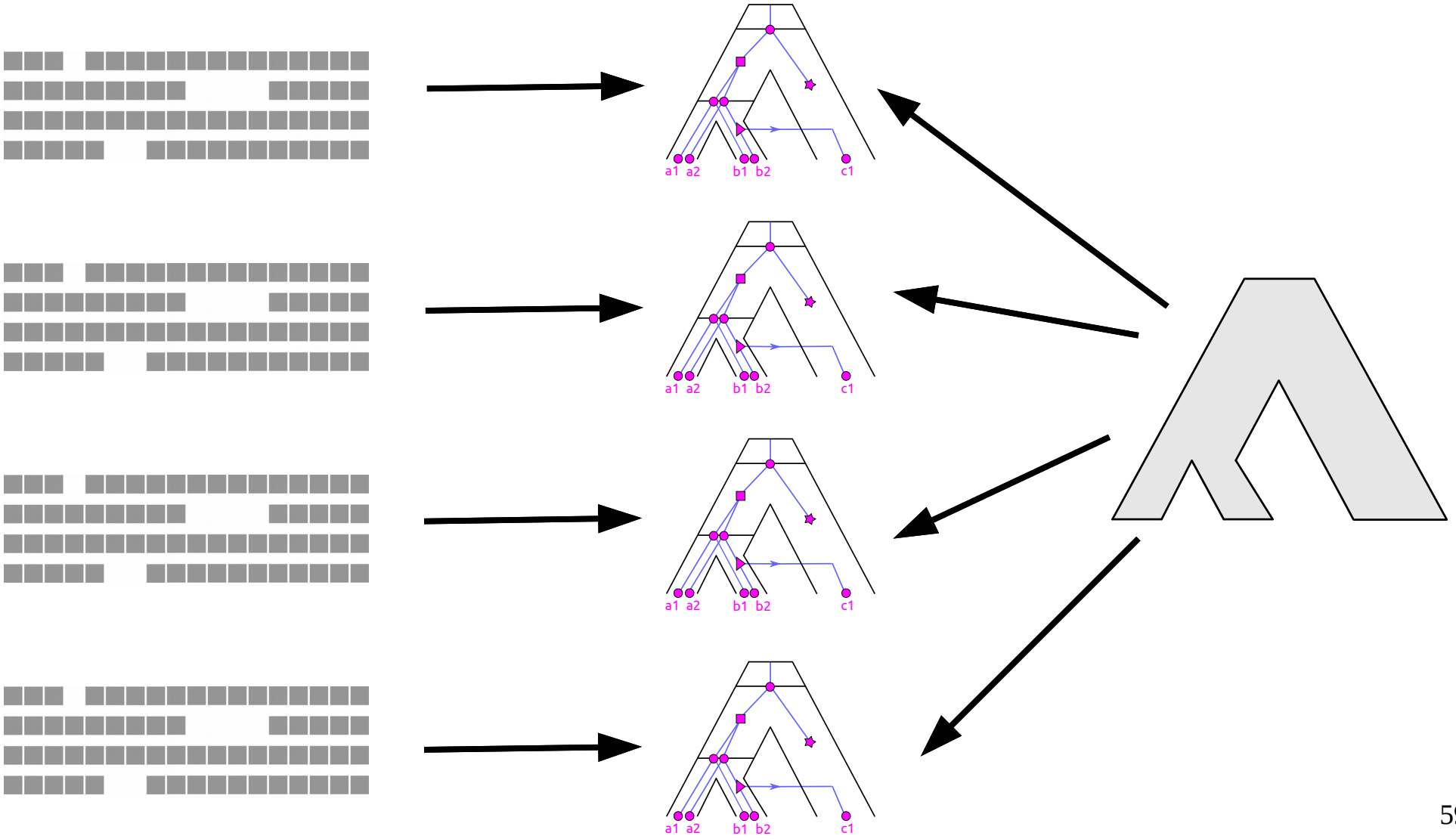
First estimation of the gene trees

GeneRax simultaneously processes multiple families



Gene tree correction with the species tree

GeneRax simultaneously processes multiple families



Reconcile the corrected gene trees

Starting gene tree

- Users can specify an input starting gene tree (typically inferred with RAxML or IQTREE)
- You can:
  - Enable gene tree optimization (default). The input gene tree is just the starting topology for the gene tree search.
  - Disable gene tree optimization (--strategy EVAL) and reconcile the input gene tree

DTL probabilities estimation

- DTL probabilities are estimated from a fixed gene tree and a fixed species tree with maximum likelihood

DTL probabilities estimation

- DTL probabilities are estimated from a fixed gene tree and a fixed species tree with maximum likelihood

Example:

D=0.1 L=0.1 T=0.5 S=0.3 L = -105

D=0.2 L=0.2 T=0.1 S=0.5 L = -99

DTL probabilities estimation

- DTL probabilities are estimated from a fixed gene tree and a fixed species tree with maximum likelihood

Example:

~~D=0.1 L=0.1 T=0.5 S=0.3 L = -105~~

D=0.2 L=0.2 T=0.1 S=0.5 L = -99

DTL probabilities estimation

- DTL probabilities are estimated from a fixed gene tree and a fixed species tree with maximum likelihood

- In practice, we optimize them with gradient descent

DTL probabilities parameters

- Global probabilities: all families and all species have the same DTL probabilities
- Per-family probabilities: each family has different DTL probabilities
- Per-species probabilities: each species has different DTL probabilities

- GeneRax outputs those DTL probabilities

Substitution model parameters

- GeneRax only implements reversible models
- It estimates the substitution model parameters from the gene tree and the sequences (same code as RaxML-NG…)
- Each gene family has its own model parameters

GeneRax and gene tree branch lengths

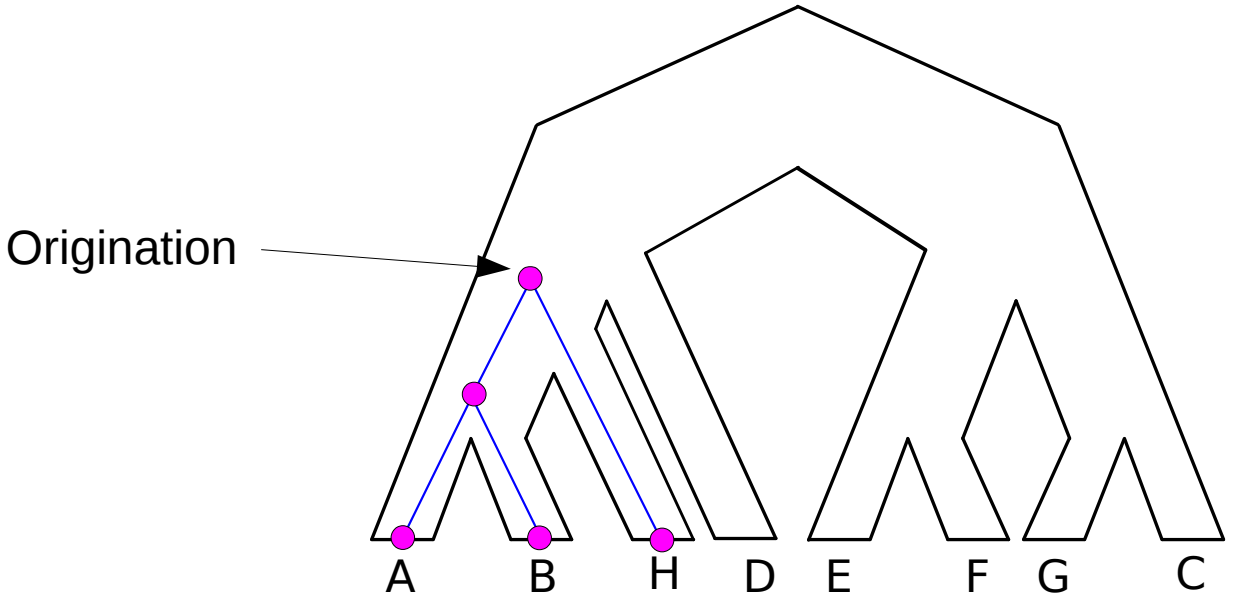
- Unit: expected average number of mutations per site
- Same unit as RAxML or IQTree
- Depend on:
  - The unrooted gene tree
  - The MSA
- Do **not** depend on:
  - The species tree
  - The DTL events

Root position

- GeneRax returns rooted gene trees
- The root is inferred from the unrooted gene tree topology and the species tree
- The root is not inferred from the sequences (we do not support non-reversible substitution models)

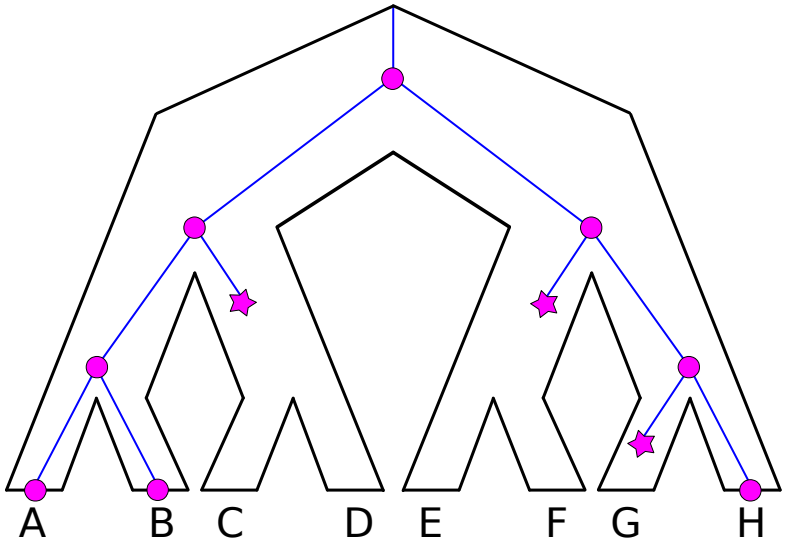
Origination

- Origination: position of the root of the gene tree in the species tree
- Can be anywhere in the species tree
- We pick the maximum likelihood origination point

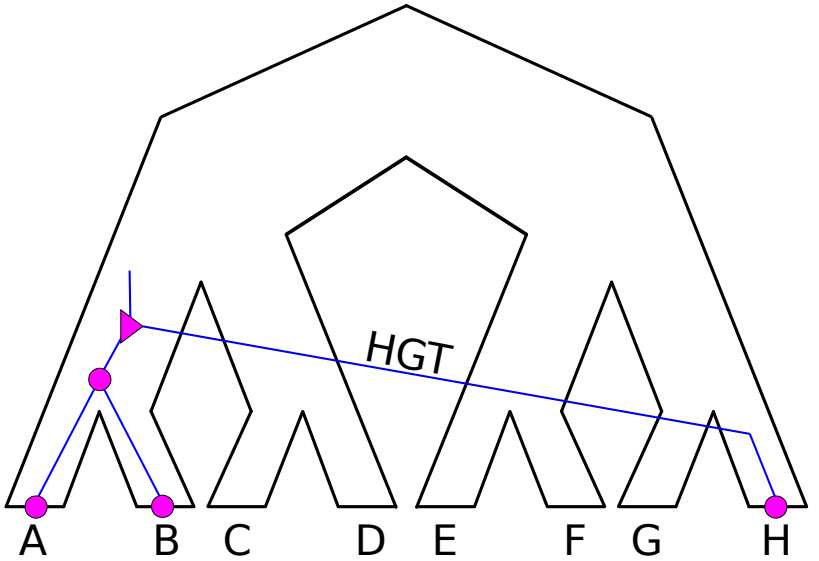


Reconciliation

- There is an infinite amount of compatible scenarios
- GeneRax estimates the scenario with the highest likelihood



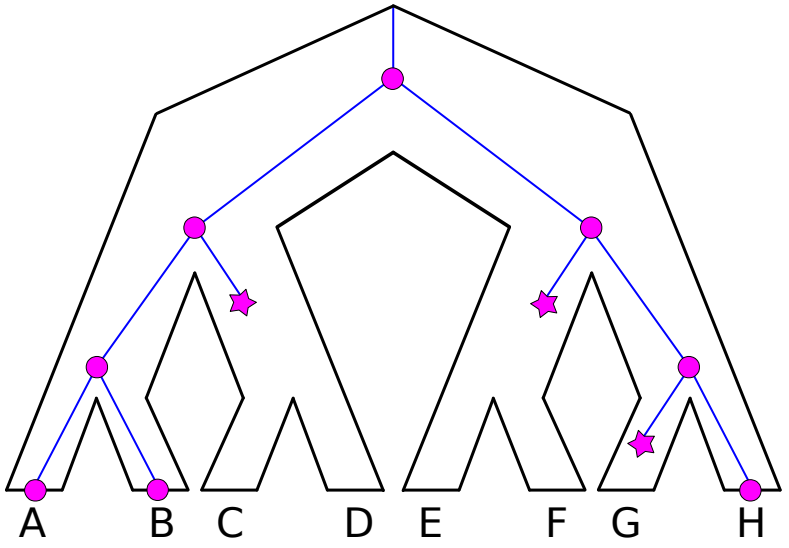
log(L)=-10



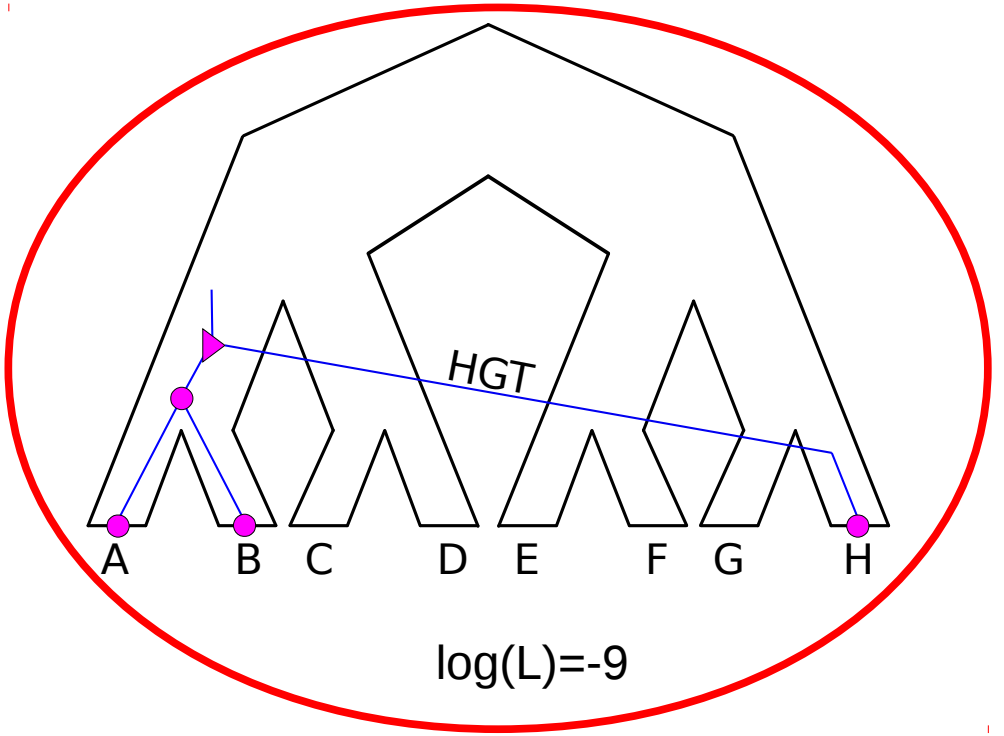
log(L)=-9

Reconciliation

- There is an infinite amount of compatible scenarios
- GeneRax estimates the scenario with the highest likelihood



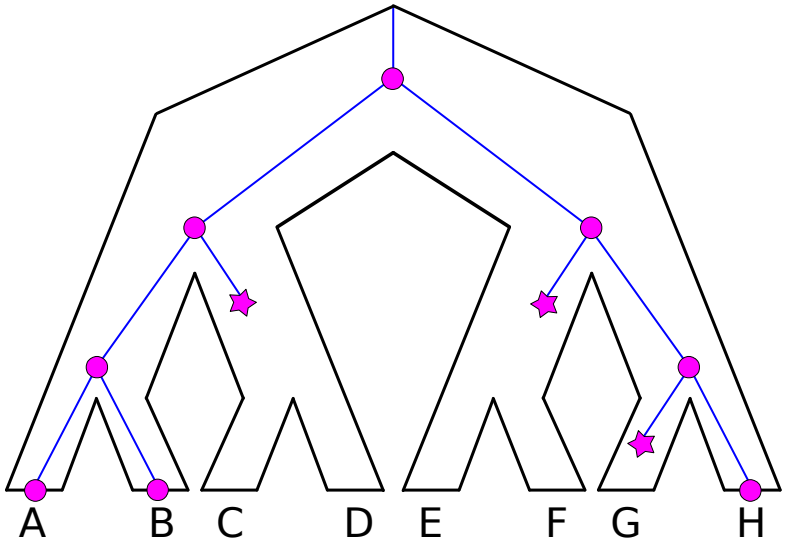
log(L)=-10



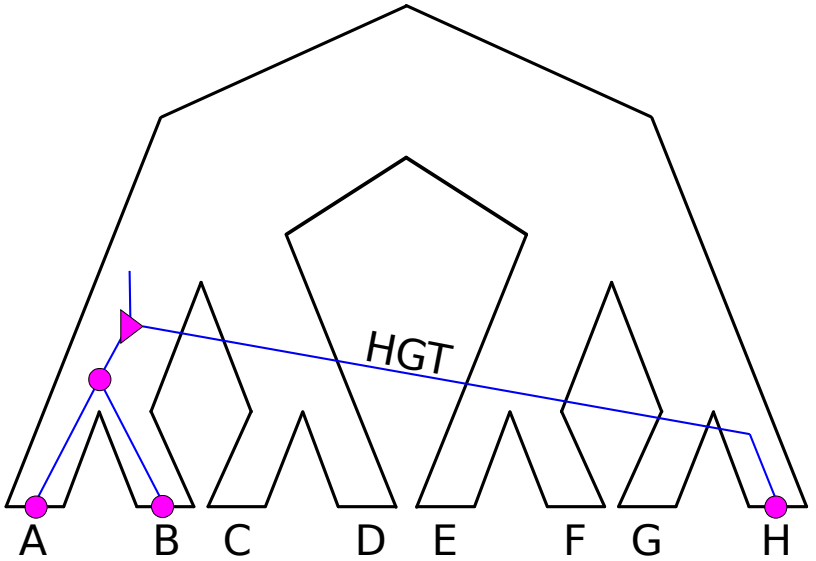
log(L)=-9

Sampling reconciliations

- There might be many scenarios with similar likelihoods
- GeneRax can **sample** scenarios (scenarios with a high likelihood will appear more often)



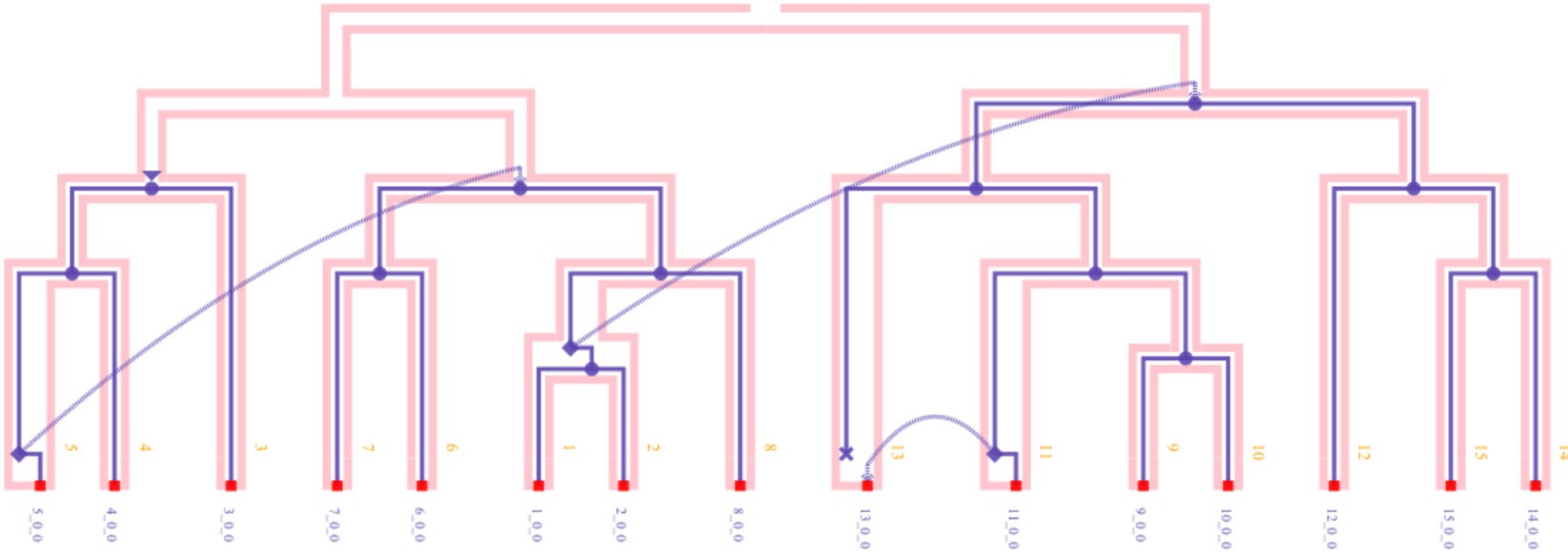
Sample 1



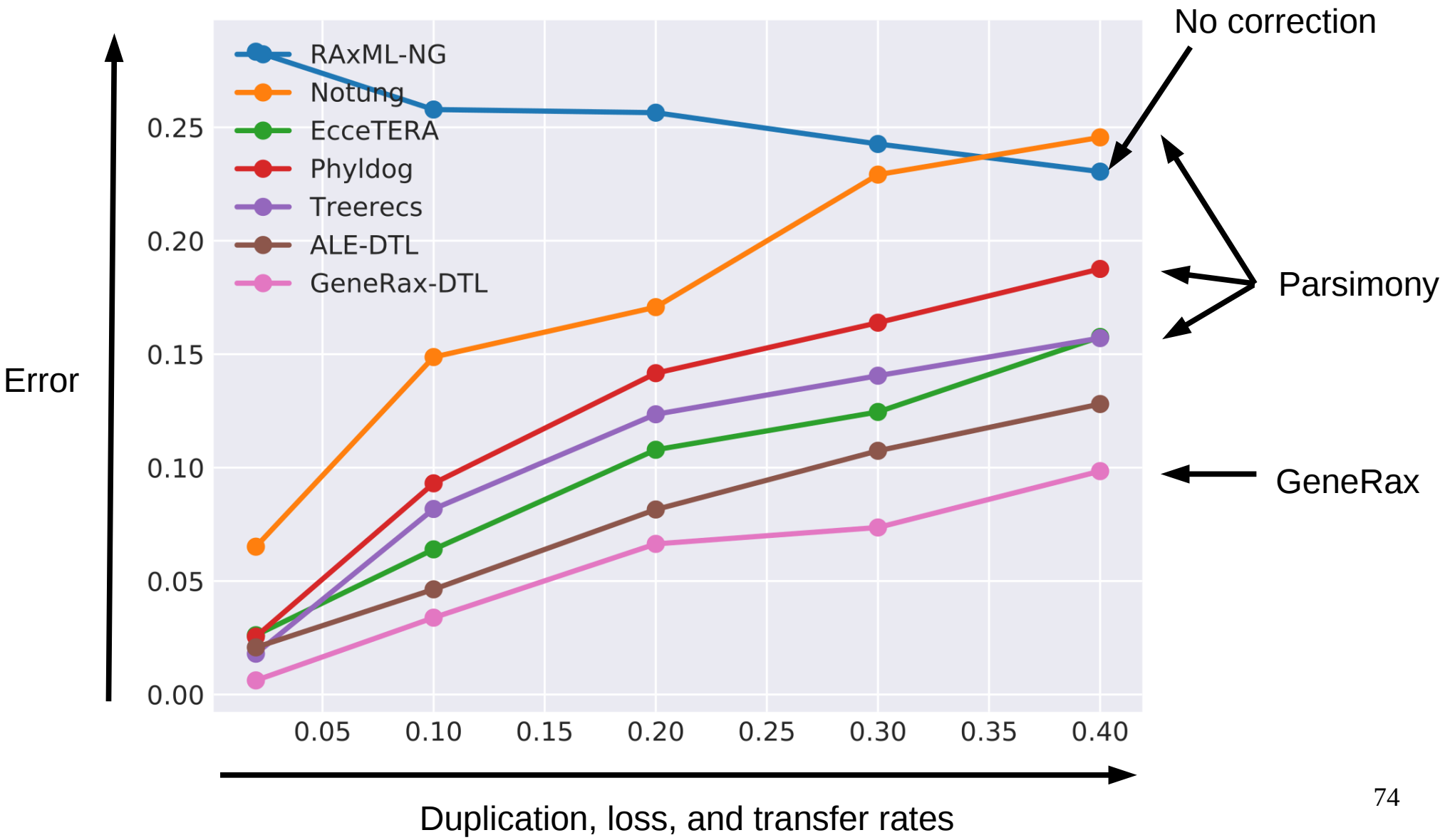
Sample 2

Visualizing the results: ThirdKind

- Developed in another team
- Webserver and software



Accuracy


RAxML-NG
Notung
ExceTERA
Phyldog
Treerecs
ALE-DTL
GeneRax-DTL
0.25
0.20
0.15
0.10
0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Error
Duplication, loss, and transfer rates
No correction
Parsimony
GeneRax

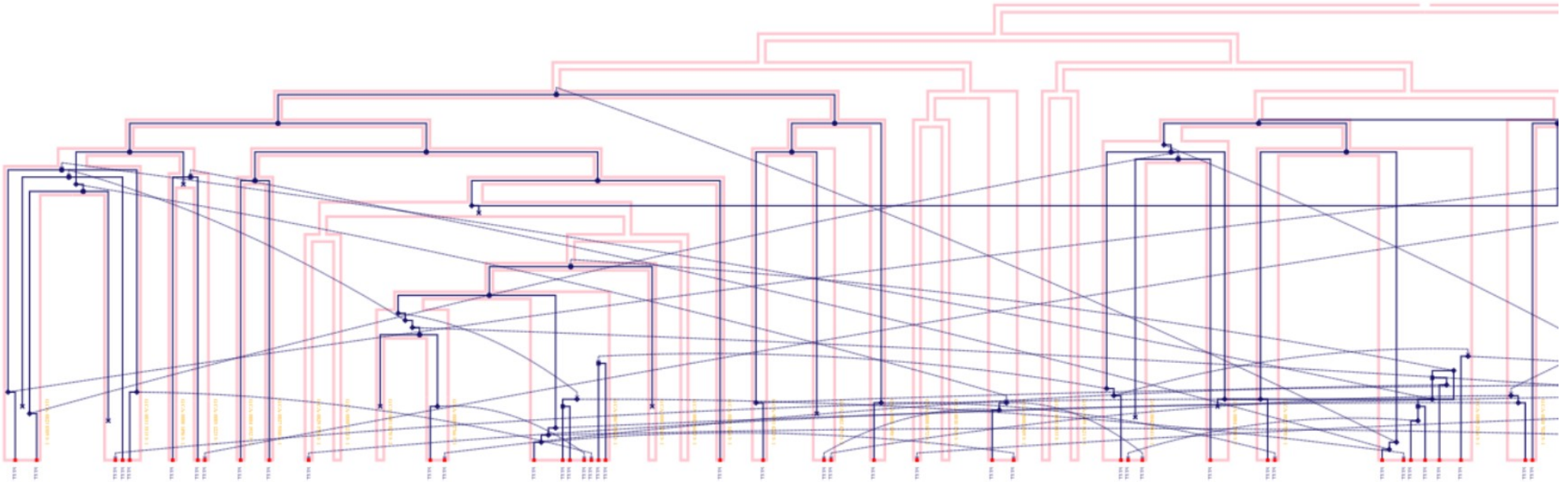
What could go wrong?

- Wrong input species tree
- Wrong gene alignment
- Other sources of conflict (e.g. ILS)
- Wrong gene families
- Missing data

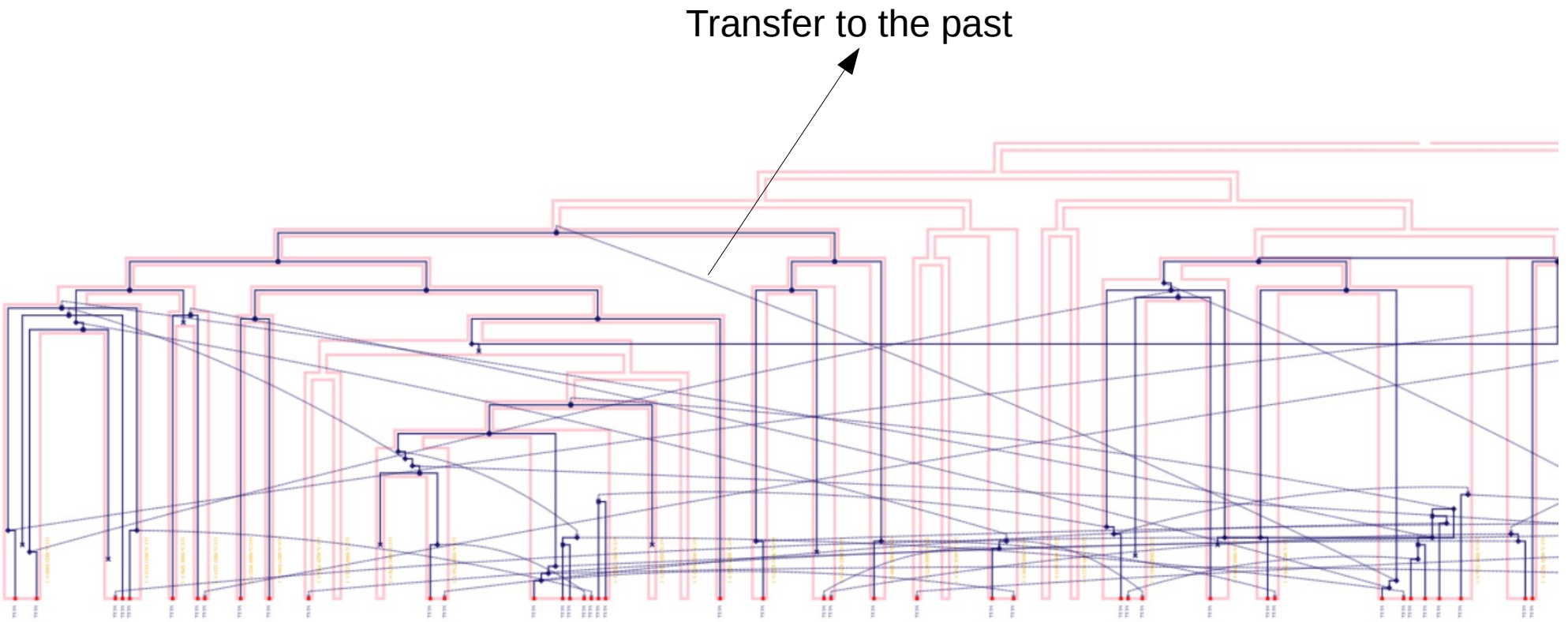
Difficult datasets

- Low signal from the sequences
- Many transfers or very conflicting gene tree topologies
- Unclear origination position

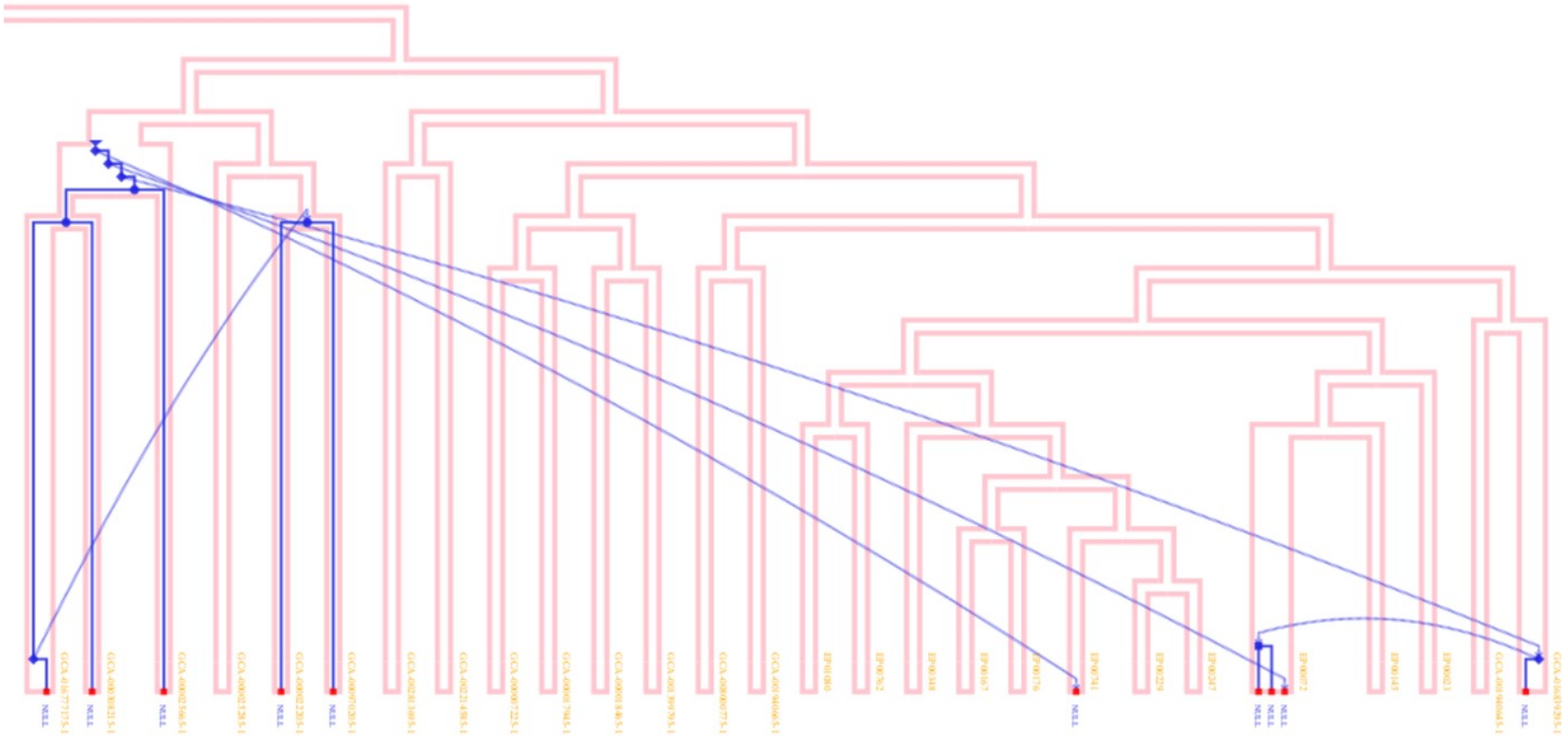
Difficult datasets
Too many transfers



Difficult datasets
Too many transfers



Difficult dataset



Be aware of the limitations

- Gene tree reconciliation is a difficult problem
- GeneRax will return the reconciled gene tree with the highest likelihood…
- … but it can be wrong, especially for difficult datasets
- Do not blindly trust the results

This is also true for any other phylogenetic problem :)

Working with large datasets

Large datasets: thousands of families, hundreds of species

Challenging!

- Runtime limitations
- Memory limitations

Large datasets: **bad** practices

- Analyze the whole dataset
- Wait for two weeks to get the results (and massively burn CO2)
- Check the results
- Realize that you messed up something
- Fix and iterate

(Don’t do this!)

Large datasets: good practices

- Start small: run your analysis on a subset of your dataset
  - Check that the program does what you expect
  - Roughly estimate the time required to analyze the whole dataset
  - Decide if you can afford to run this analysis

My dataset is too large

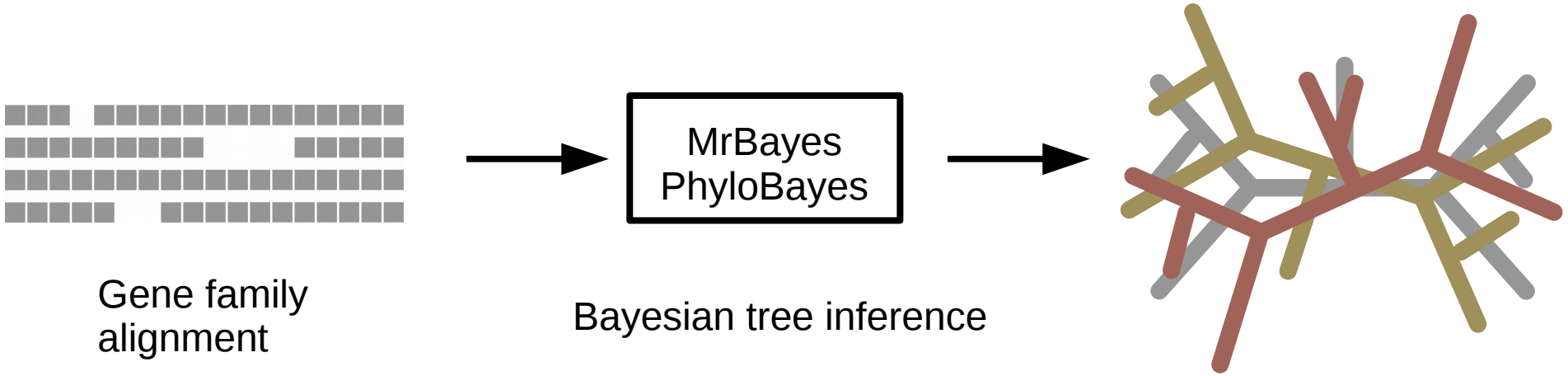
- Run it on a cluster
- Subsample:
  - Reduce the number of species
  - Filter out some families, either:
    - Randomly
    - The largest ones

Limitations of GeneRax

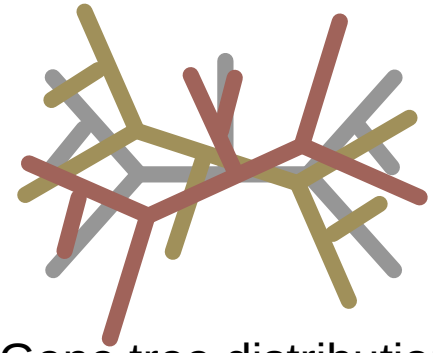
- GeneRax does not provide support values
- GeneRax only supports a predefined set of substitution models (the ones implemented in RAxML-NG)

- Alternative solution: gene tree distribution methods

Gene tree distributions

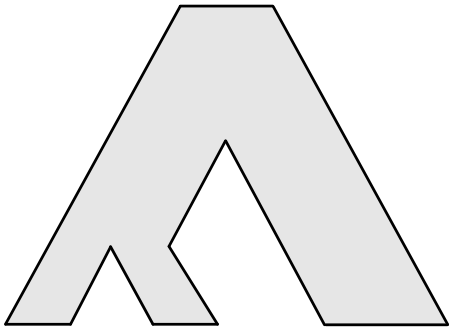


ALE

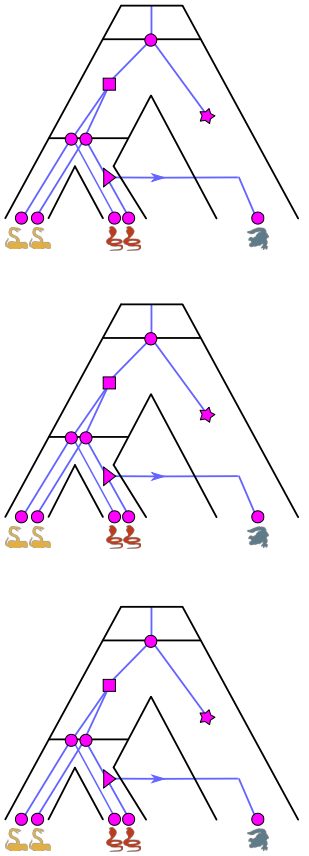


Gene tree distribution

Sample under
the UndatedDTL



Species tree



Reconciled gene
tree distribution

Advantages of ALE

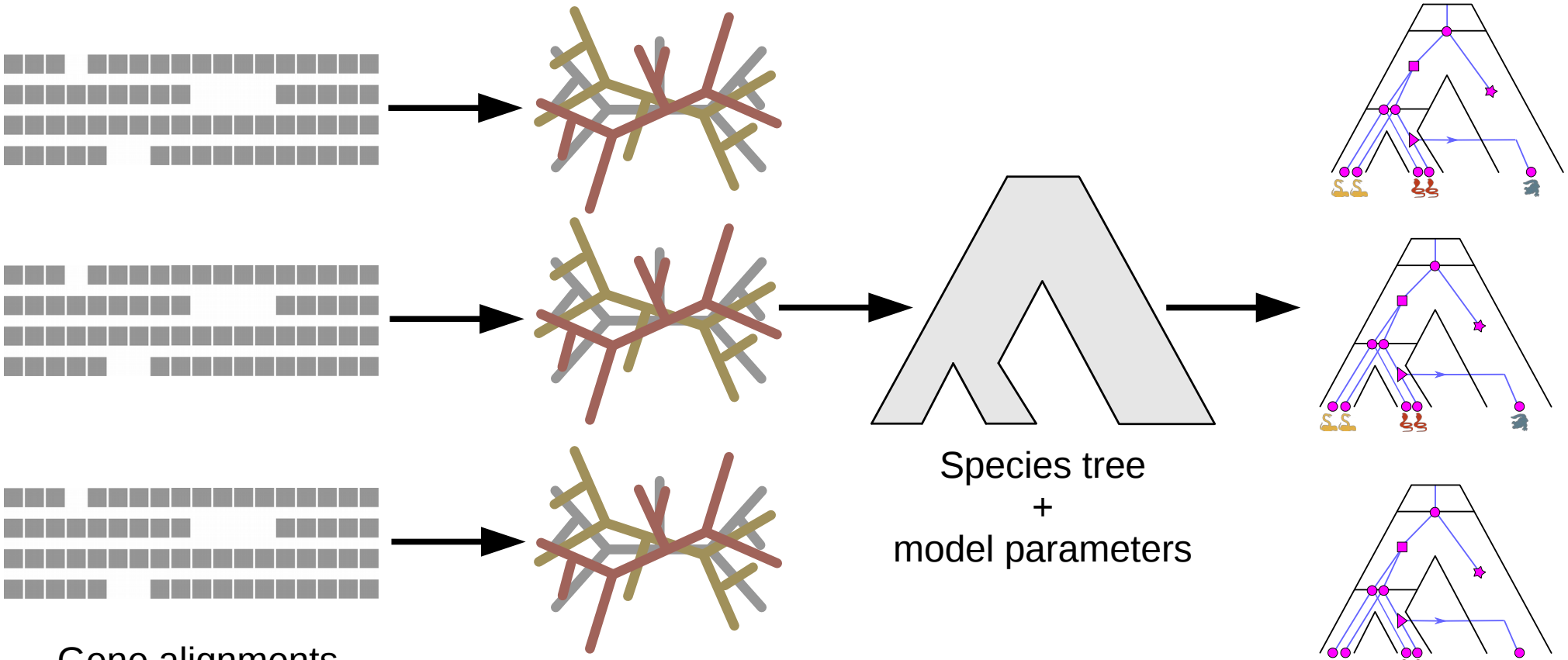
- You can use any substitution model (as long as there is a bayesian inference tool that supports it)
- ALE returns a distribution of reconciled gene trees → measure of uncertainty

AleRax

- AleRax is a reimplementation of ALE
- Not published yet
- Adds more features

- Hopefully, this is the future of reconciliation :-)

AleRax


Gene alignments
Gene tree distributions
Species tree
+
model parameters
Gene events

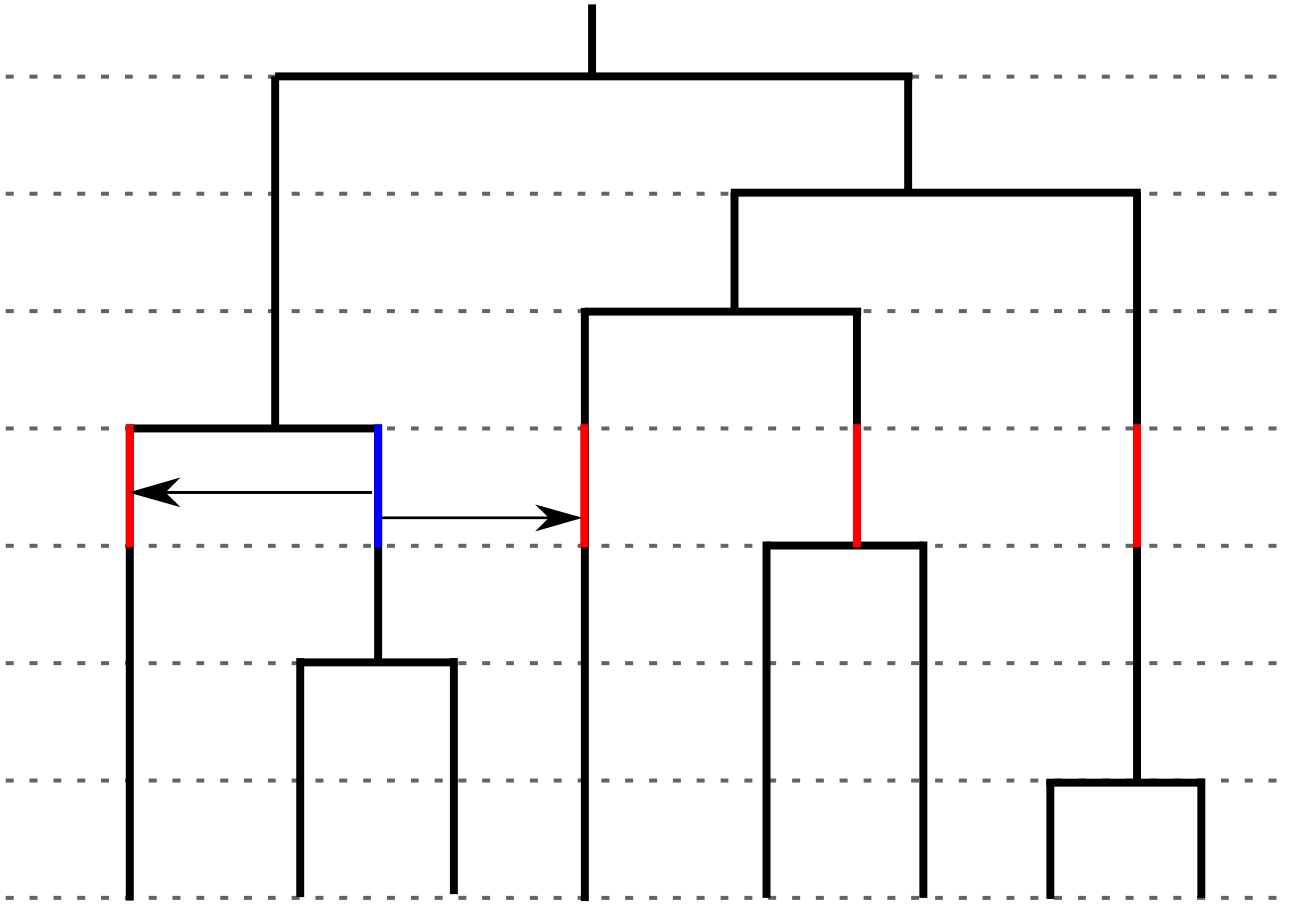
Challenges

- Transfer constraints
- Highways of transfers
- Incomplete lineage sorting

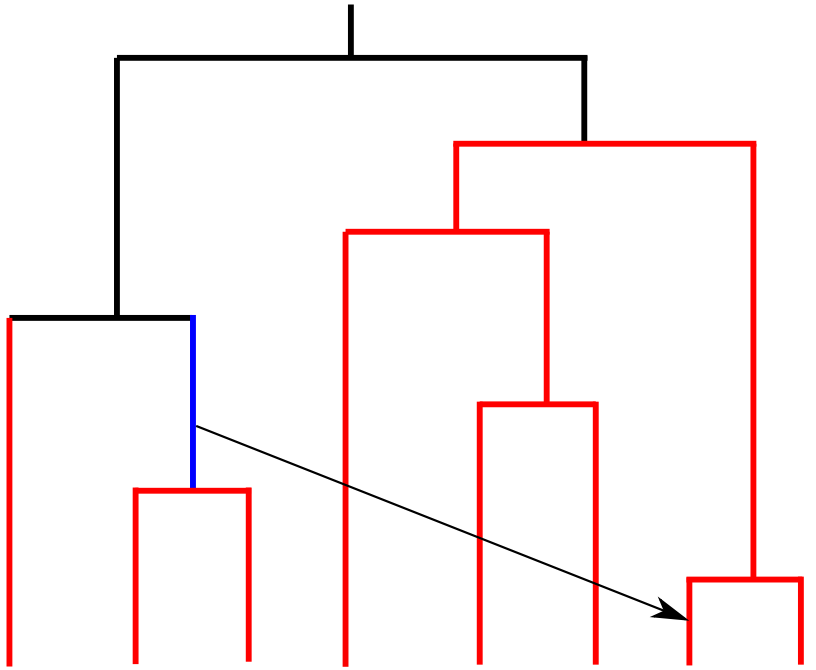
Challenges

- **Transfer constraints**
- Highways of transfers
- Incomplete lineage sorting

Transfers between contemporary species

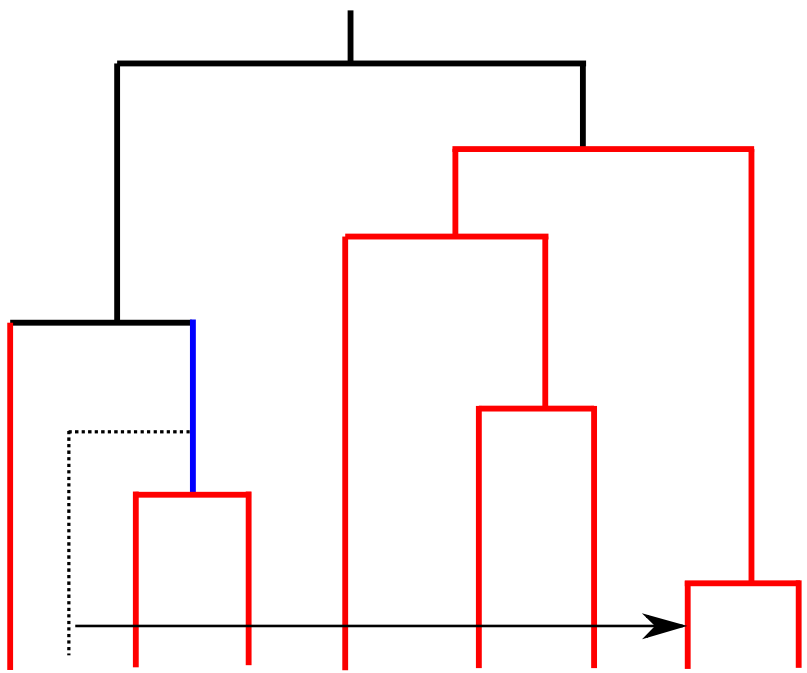


Transfers to the future

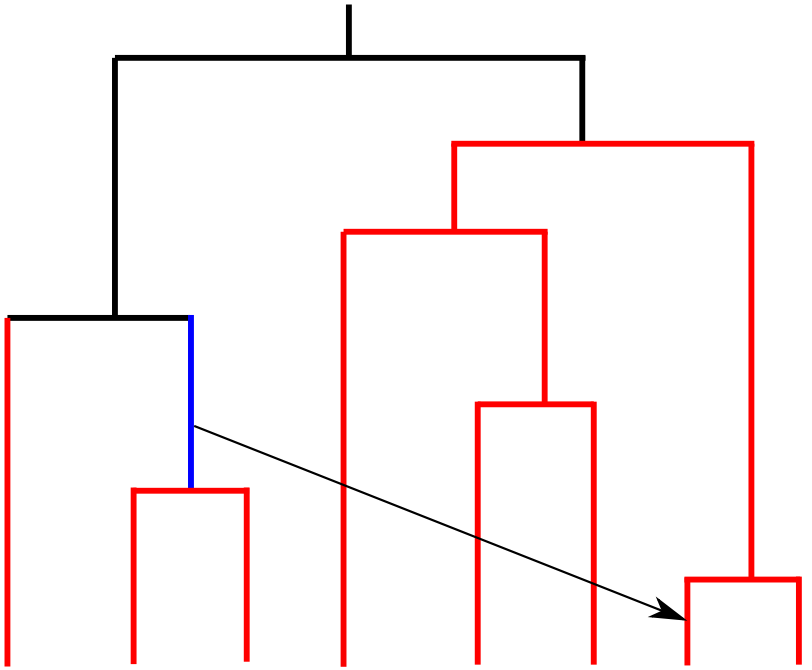


Is this possible?

Transfers to the future (via extinct species)

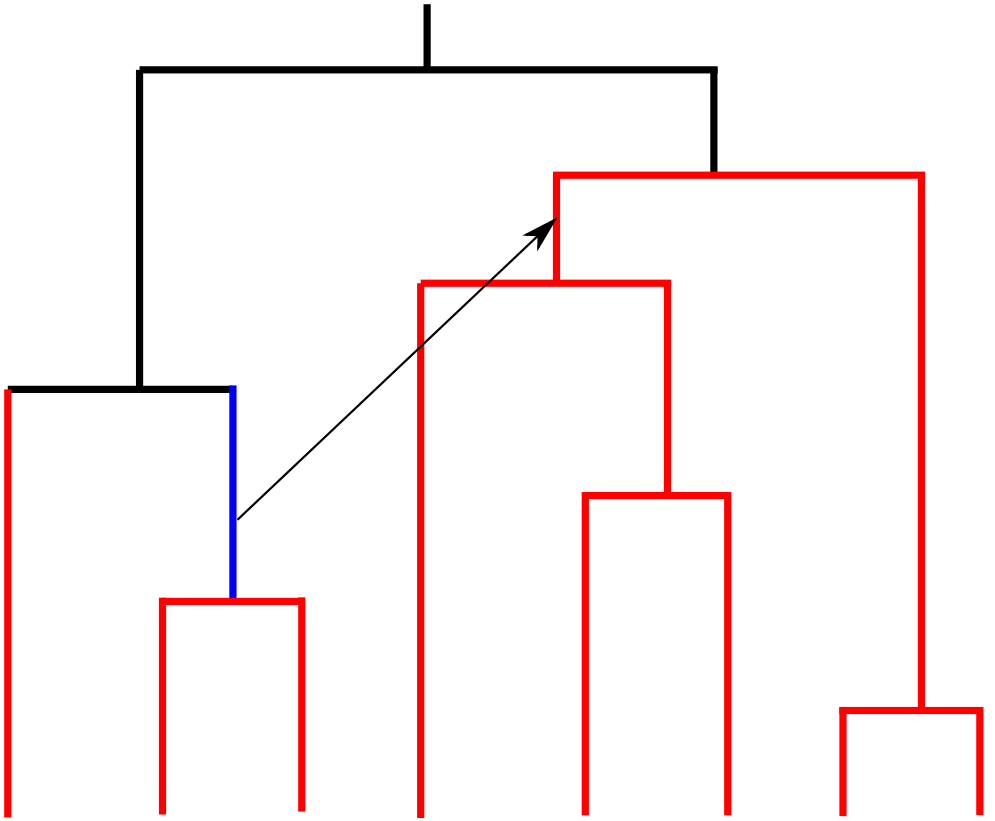


Real transfer



Inferred transfer

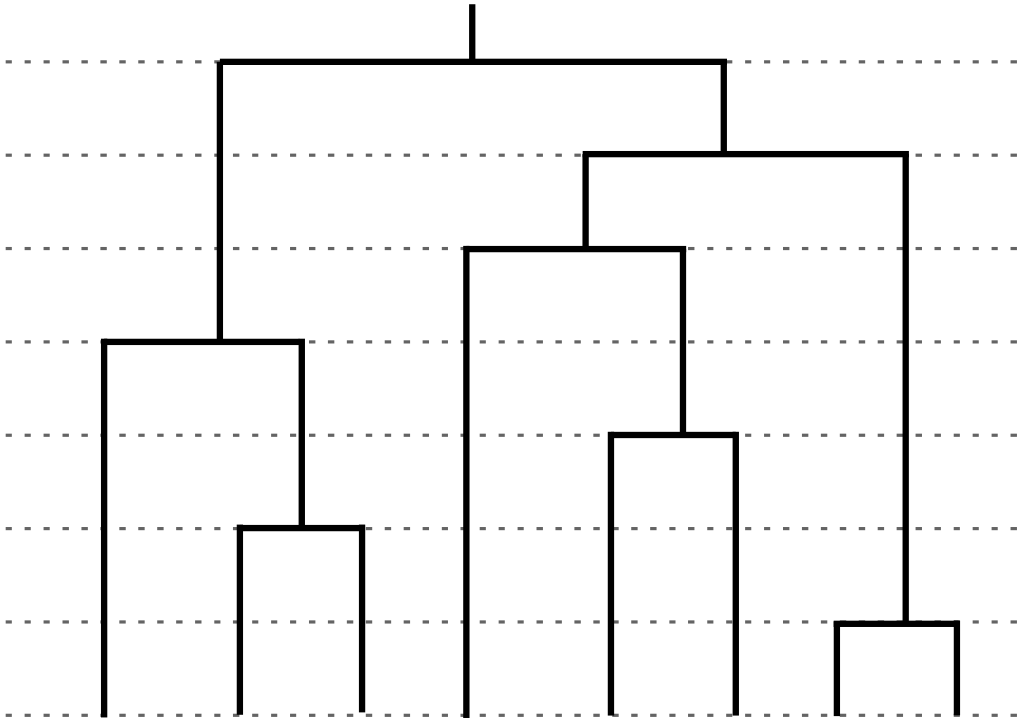
Transfers to the past



Impossible!!

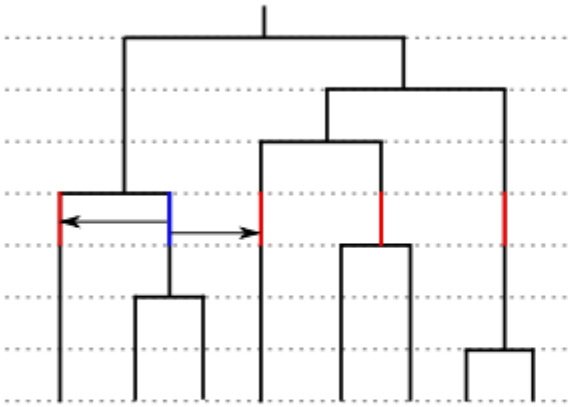
Relative dated species trees

- Models the order of the speciation events
- Does not model the branch lenghts
- Enough to forbid impossible transfers



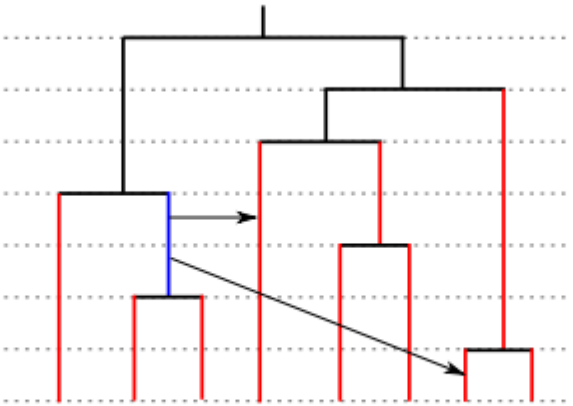
Existing models

ALE



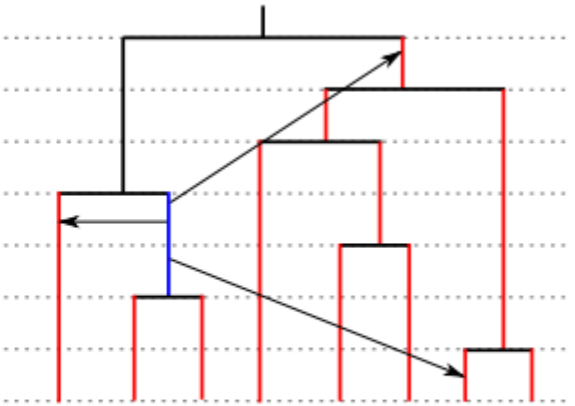
Dated

AleRax (and Ale?)



RelDated

ALE, AleRax, GeneRax



Undated

Use cases

Transfer constraints can be used to:

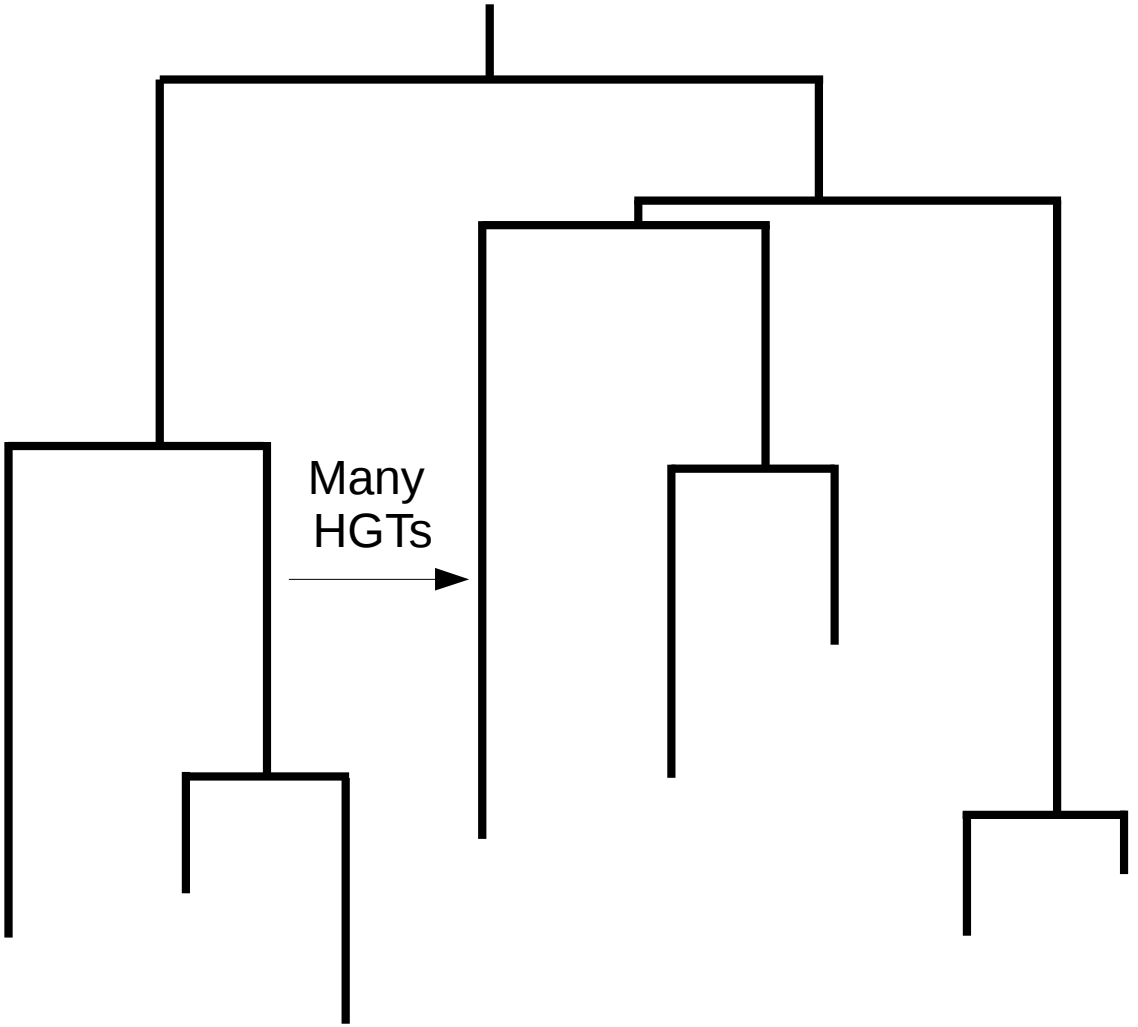
- Infer more realistic reconciliations
- Infer the relative order of speciation events
- Improve species tree rooting

Challenges

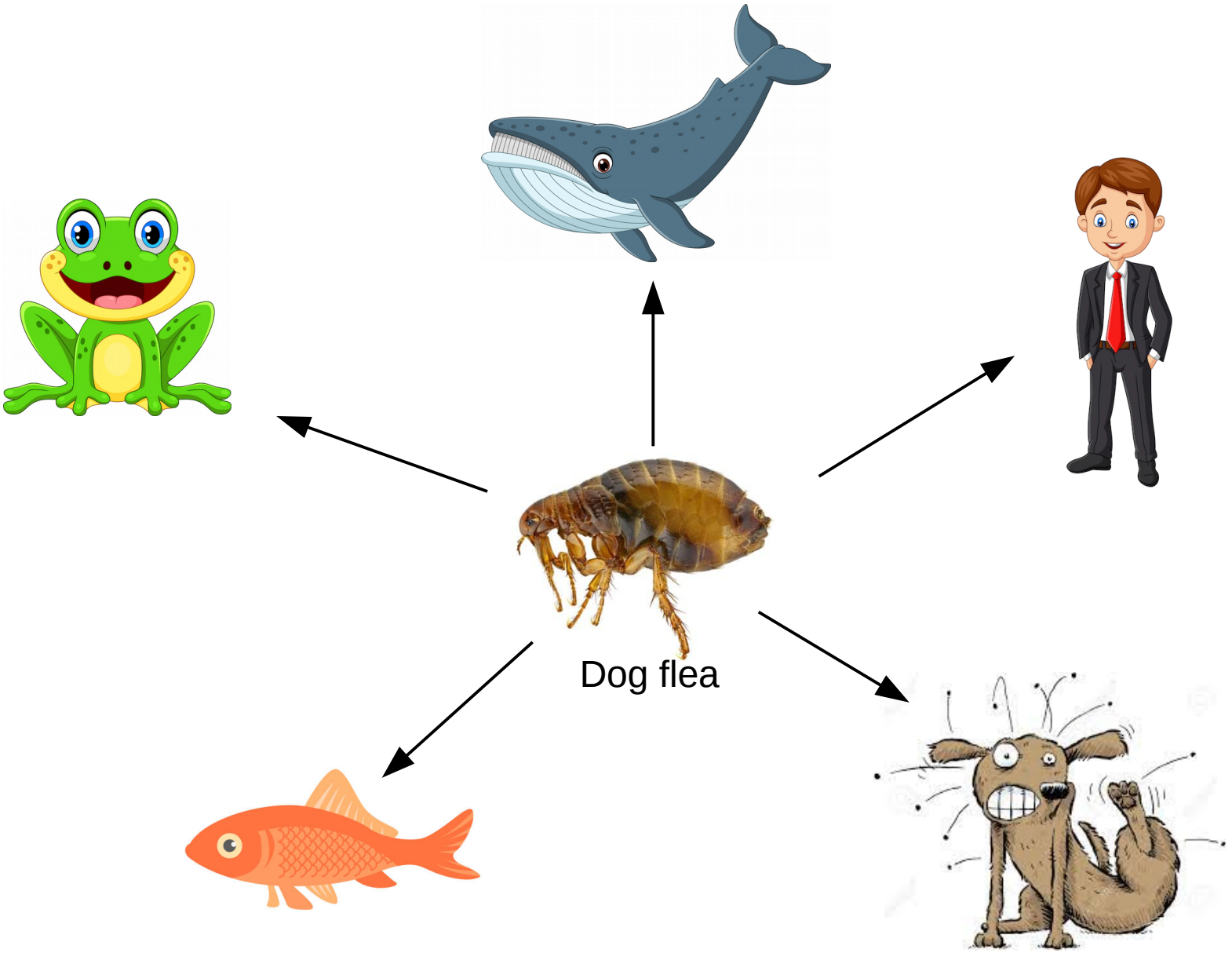
- Transfer constraints
- **Highways of transfers**
- Incomplete lineage sorting

Highway of transfers

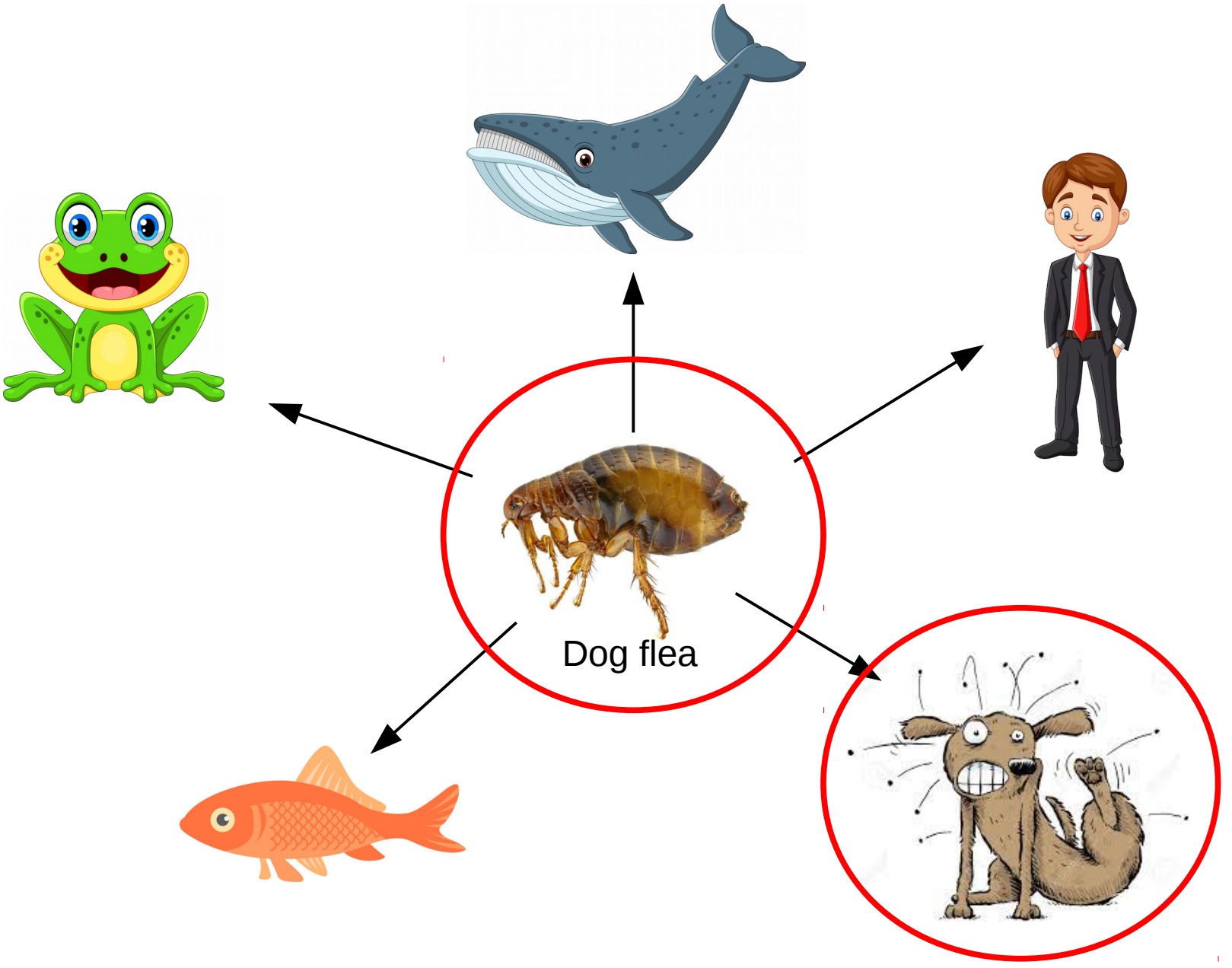
- Pair of species that exchanged many genes via HGT



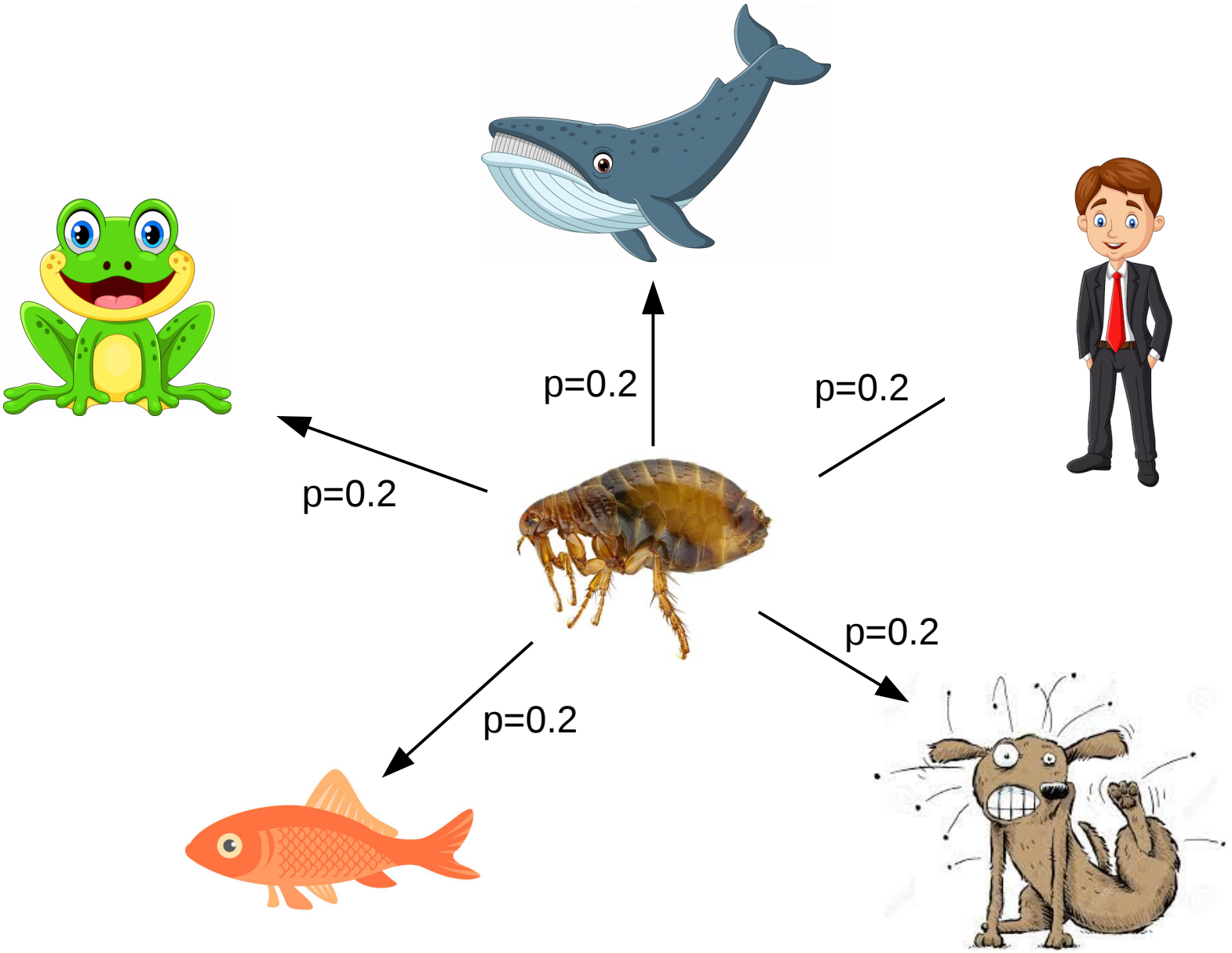
Fictional example



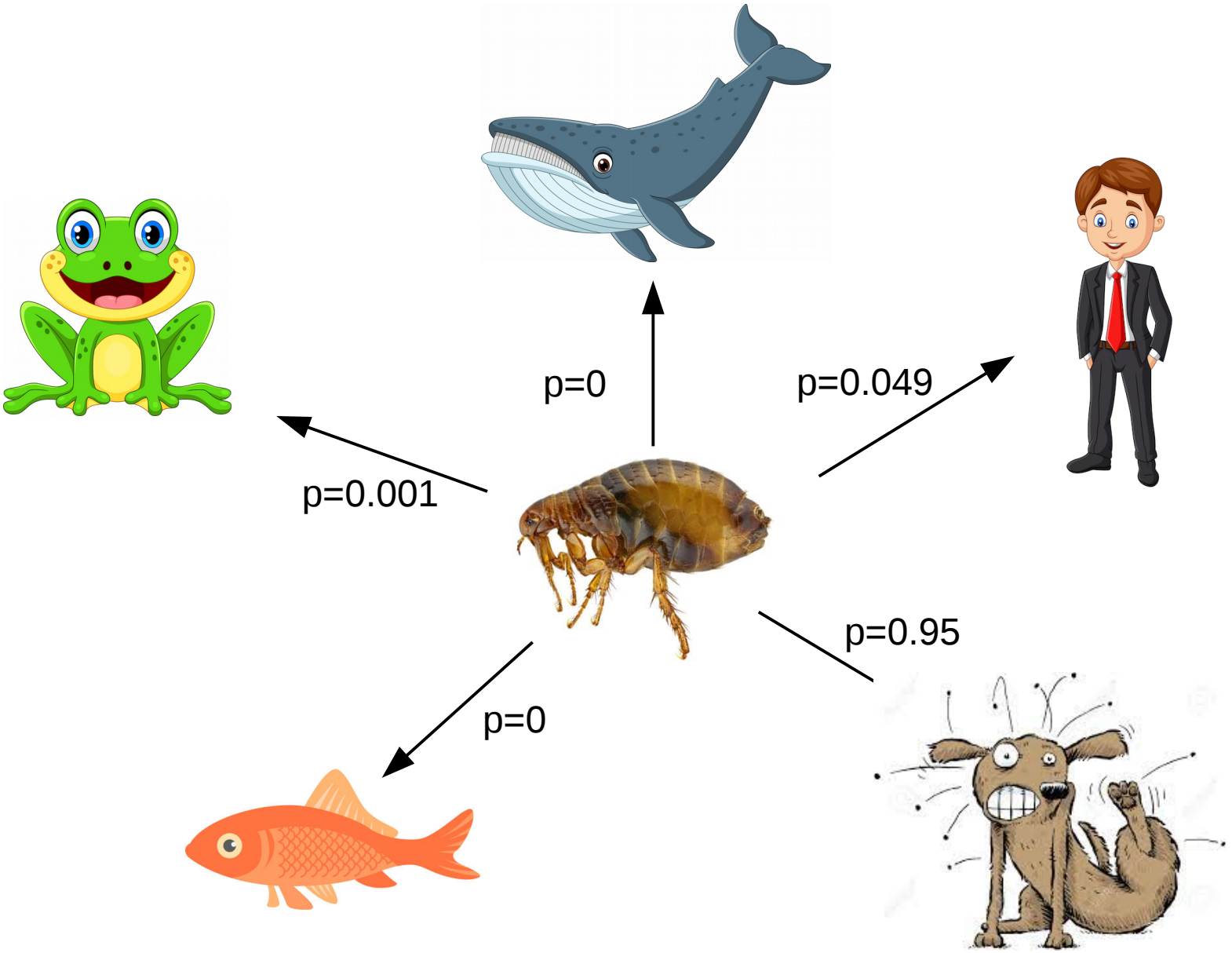
Fictional example



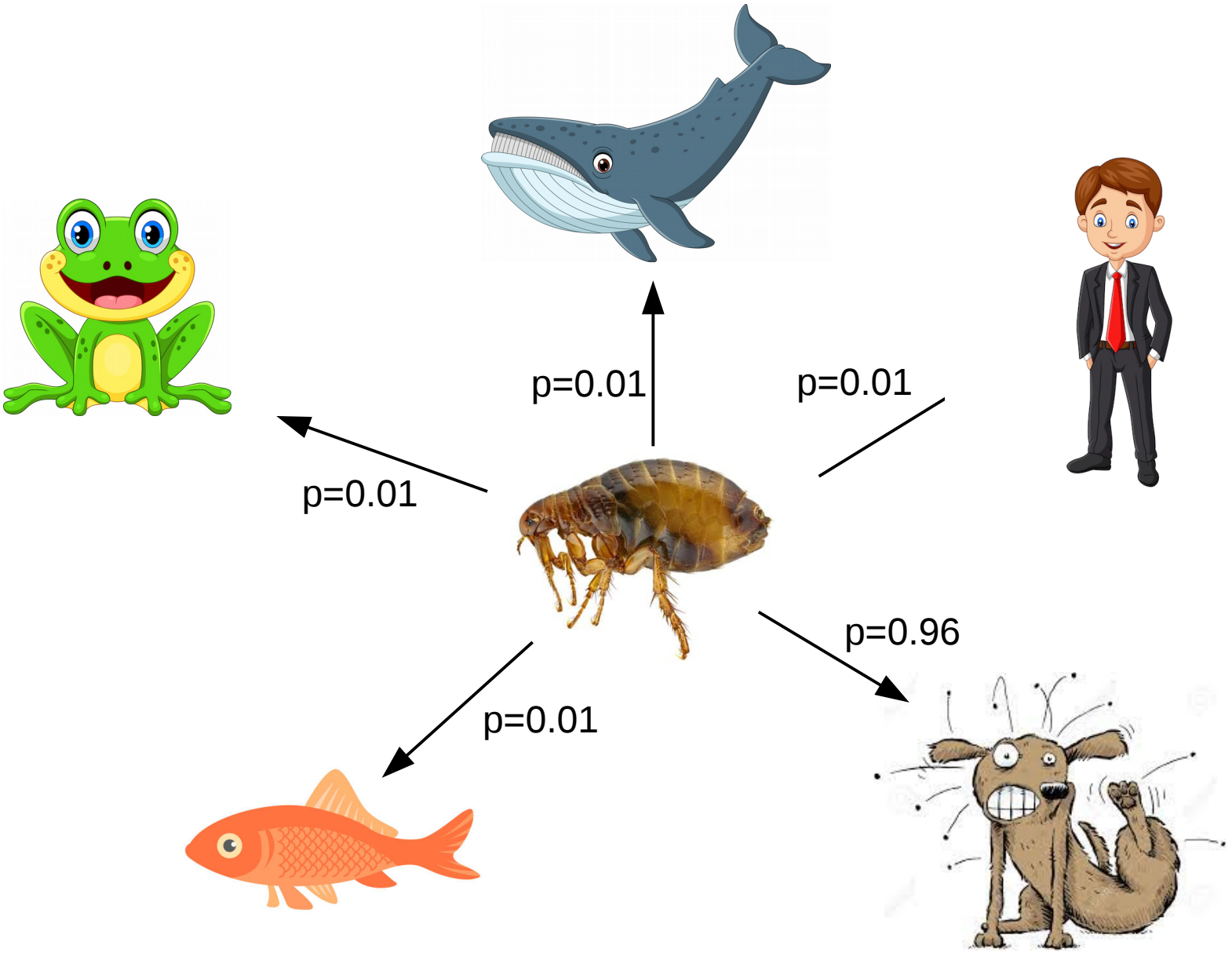
Uniform transfer probabilities



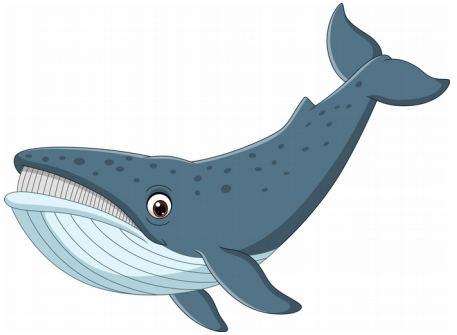
Per-species transfer probabilities



Uniform + highway



Uniform + highway



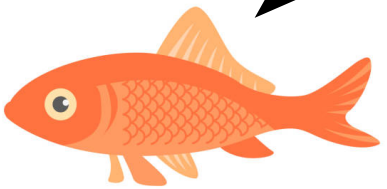
p= 0.2 - x

p= 0.2 - x

p= 0.2 - x

p = 0.2 + 4 * x

p= 0.2 - x

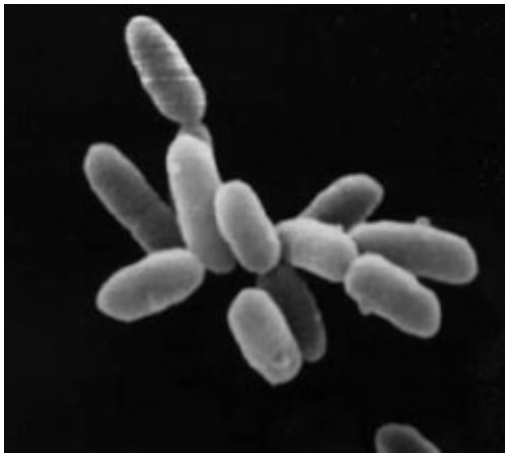


Detecting highways of transfers

AleRax provides a feature to either:

- Automatically identify potential highways
- Test highways from a list of candidate pairs of species

Example: origin of eukaryotes



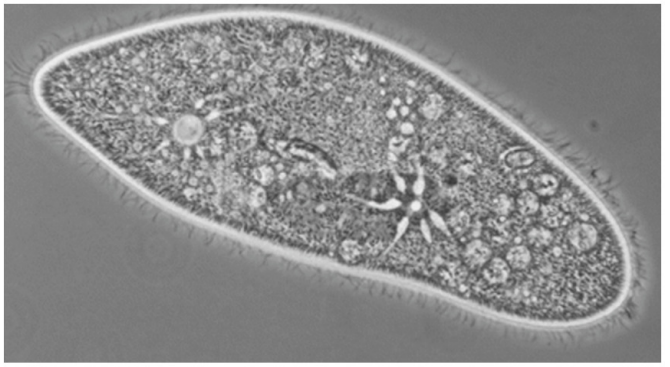
Archaea

Proteobacteria

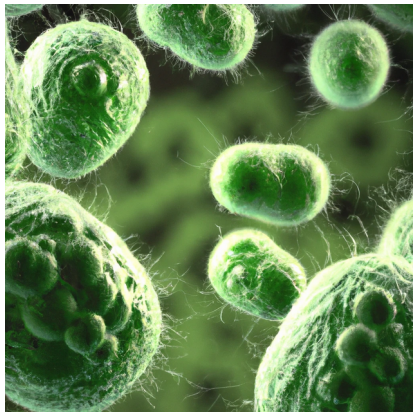
Endosymbiosis

Eukaryotes

Example: origin of green plants



Some old eukaryote



Cyanobacteria

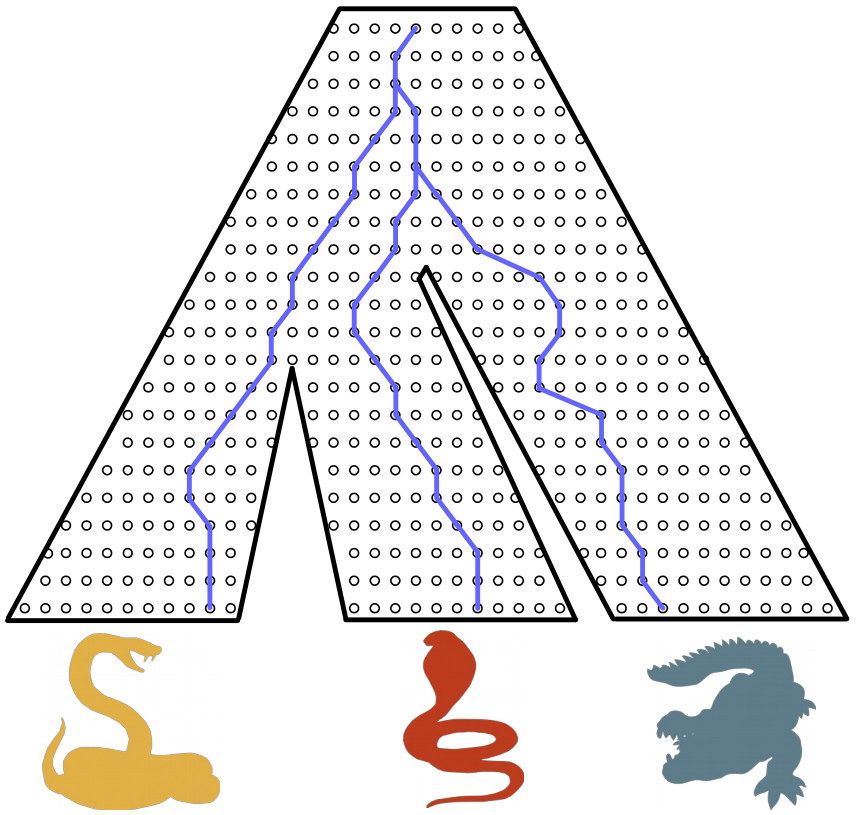
Endosymbiosis

Green plants

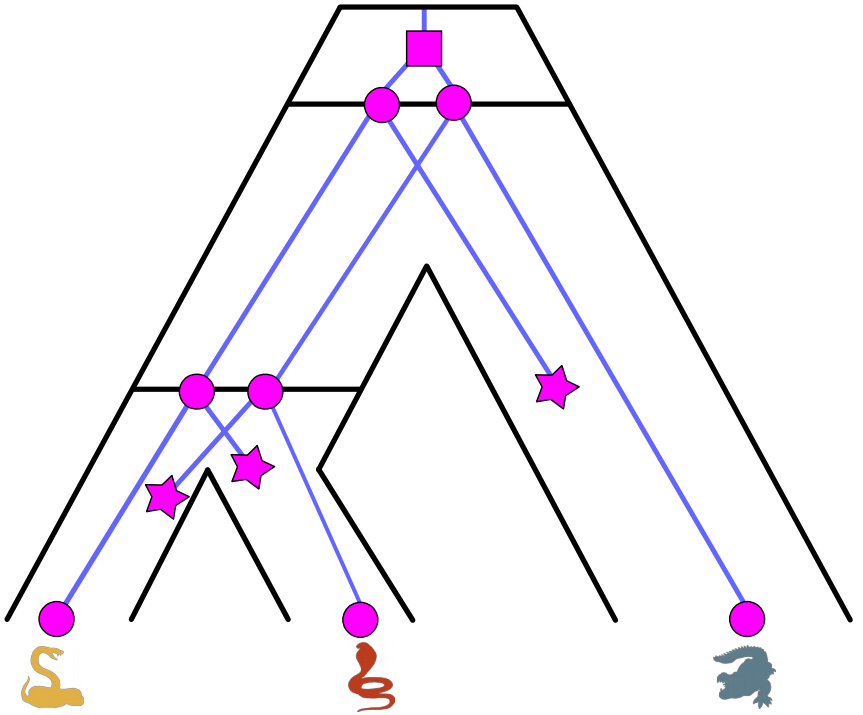
Challenges

- Transfer constraints
- Highways of transfers
- **Incomplete lineage sorting**

GeneRax and ALE do not model ILS



ILS scenario



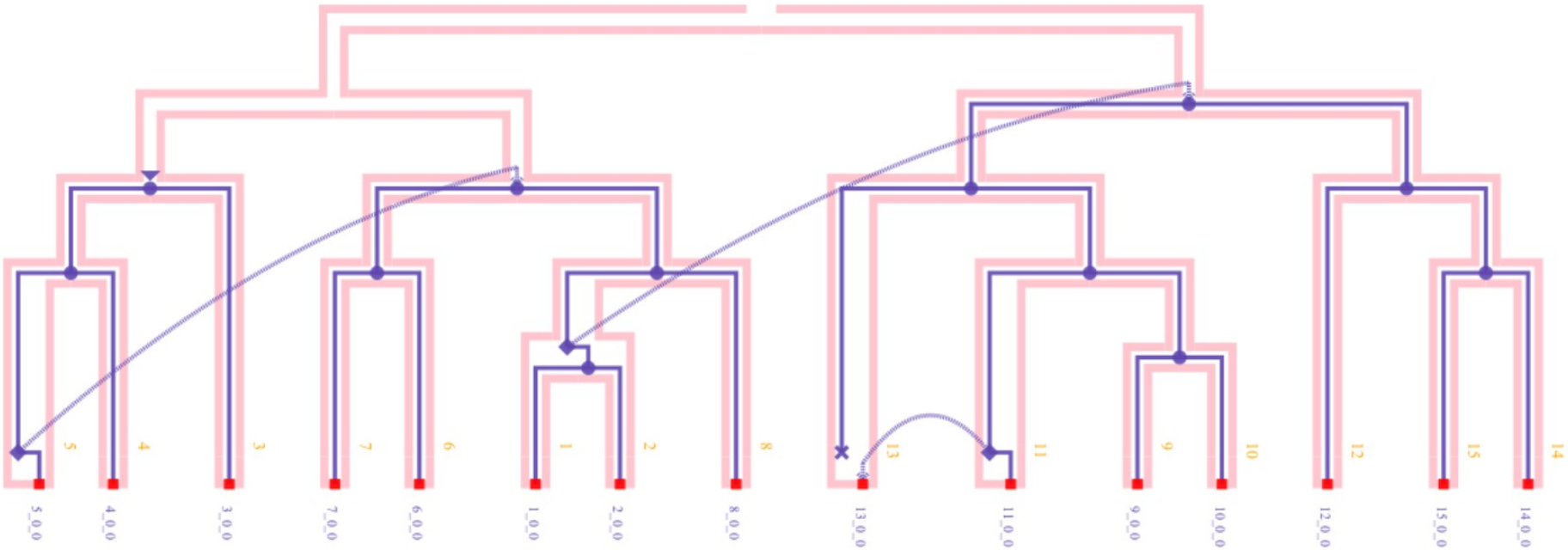
DTL scenario

ILS

- ILS inflates the number of DTL events

ILS

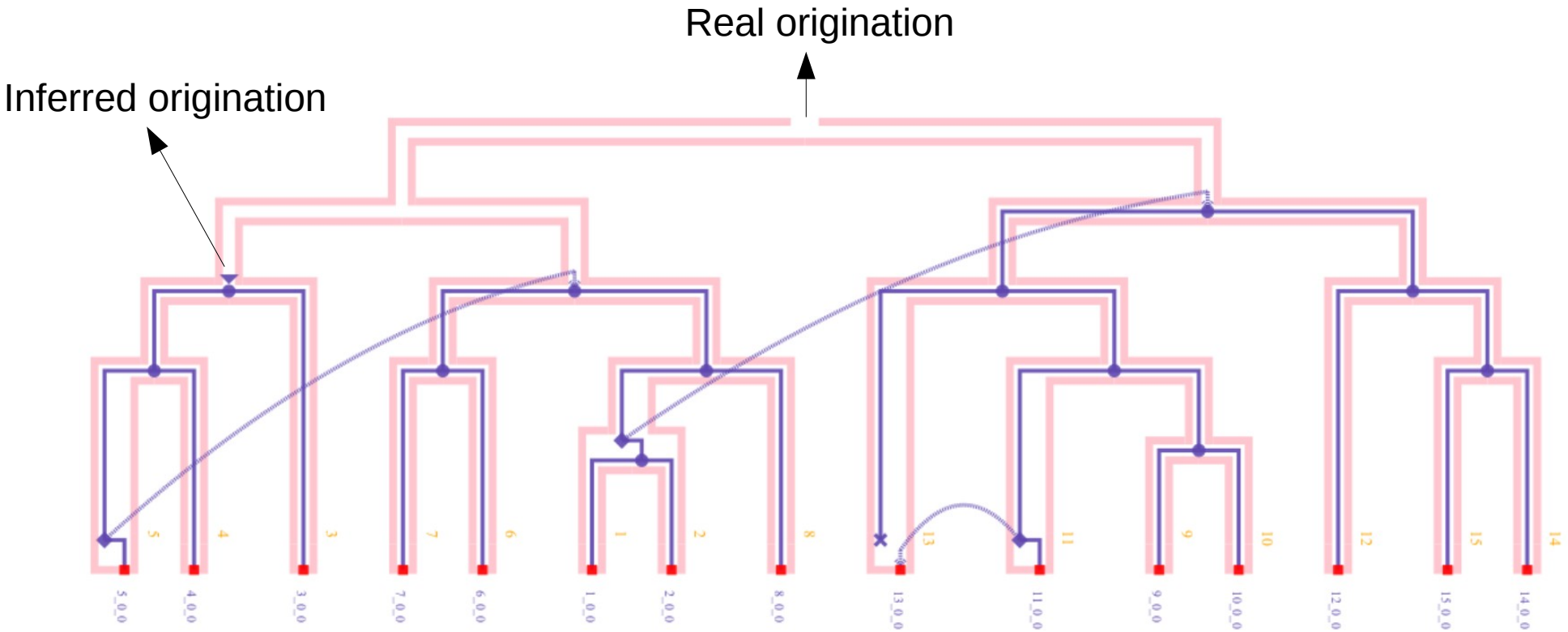
- ILS inflates the number of DTL events



GeneRax output for a dataset with ILS and without DTL events

ILS

- ILS inflates the number of DTL events



GeneRax output for a dataset with ILS and without DTL events

ILS

- ILS inflates the number of DTL events
- Solution: implement a model that accounts for both ILS and DTL events
- Future work...

Other challenges

- Missing genes
- Per-species origination probabilities
- Account for gene and species branch lengths

Gene tree species tree reconciliation

- Species-tree aware methods improve gene tree accuracy
- Methods:
  - Parsimony (Treerecs, Notung)
  - Maximum likelihood (GeneRax, Phyldog)
  - Gene tree distributions (Ale, AleRax)

Gene tree - species tree reconciliation

- We can already run very advanced analyses
- But there are still many open questions
- Active field of research
- I am always happy to discuss those questions with other scientists!

Time for a break!