A probabilistic algorithm for gene-species reconciliation with segmental duplications

Yao-ban Chan[1]

Celine Scornavacca[2]
Michael Charleston[3]

[1] School of Mathematics and Statistics / Melbourne Integrative Genomics

THE UNIVERSITY OF MELBOURNE

[2] Institut des Sciences de l'Évolution
Université Montpellier
[3] School of Natural Sciences
University of Tasmania

Guttmann 2025

INTRODUCTION

In recent years there has been an explosion of genome sequence data, across many species.

INTRODUCTION

In recent years there has been an explosion of genome sequence data, across many species.

The first genome sequencing (completed 2003) took $3 billion and 13 years. Nowadays, a human genome can be sequenced for $500 in a few days.

INTRODUCTION

In recent years there has been an explosion of genome sequence data, across many species.

The first genome sequencing (completed 2003) took $3 billion and 13 years. Nowadays, a human genome can be sequenced for $500 in a few days.

Modern human genomic databases (such as the UK Biobank) contain full genomes of 500,000 individuals.

INTRODUCTION

In recent years there has been an explosion of genome sequence data, across many species.

The first genome sequencing (completed 2003) took $3 billion and 13 years. Nowadays, a human genome can be sequenced for $500 in a few days.

Modern human genomic databases (such as the UK Biobank) contain full genomes of 500,000 individuals.

This hugely increasing amount of information requires sophisticated and efficient mathematical methods for its analysis.

INTRODUCTION

In phylogenetics, we study the evolutionary history of species and how they relate to each other.

INTRODUCTION

In phylogenetics, we study the evolutionary history of species and how they relate to each other.

It is thought that all species are descended from a single prehistoric species, but we are not yet able to analyse all 8.7 million species at once.

Introduction

In phylogenetics, we study the evolutionary history of species and how they relate to each other.

It is thought that all species are descended from a single prehistoric species, but we are not yet able to analyse all 8.7 million species at once.

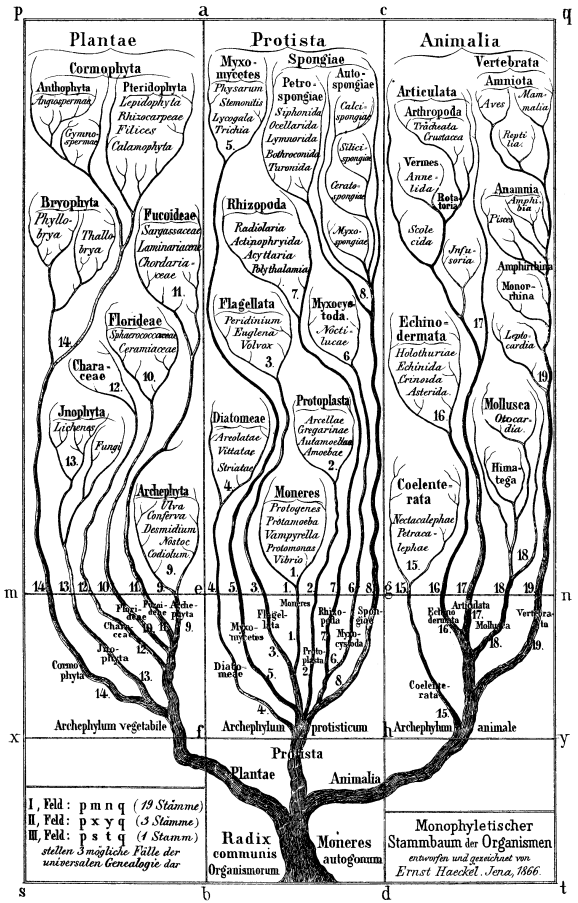
Instead, we concentrate on a species family — a group of species that are (supposedly) descendants of a common ancestor.

PHYLOGENIES

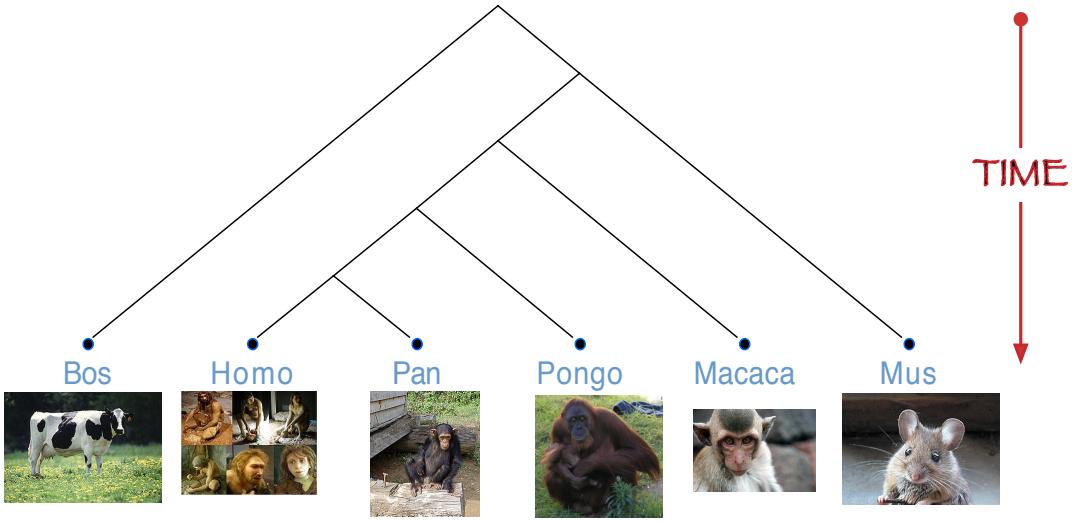
The evolution of a species family is depicted by a phylogeny or phylogenetic tree.

PHYLOGENIES

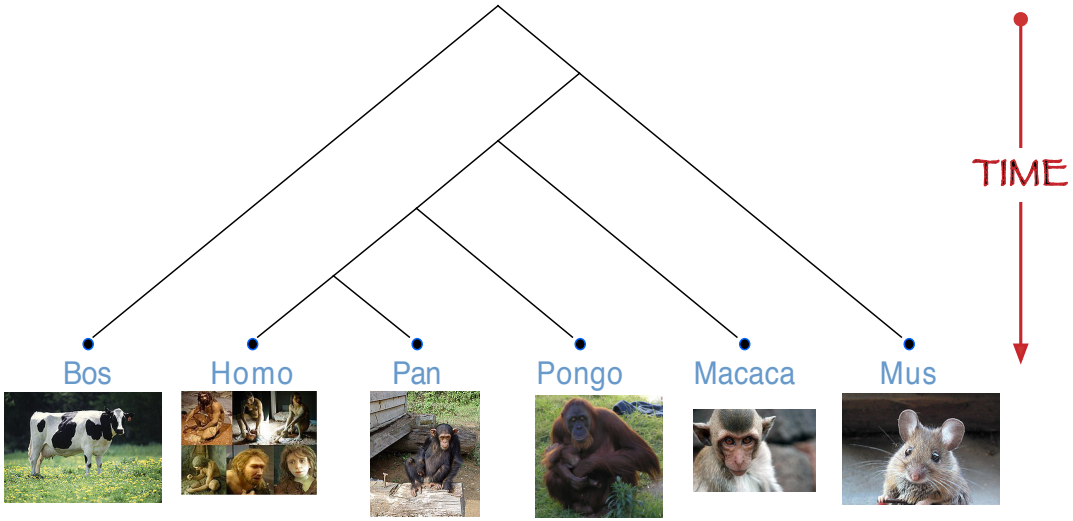
The evolution of a species family is depicted by a phylogeny or phylogenetic tree.



PHYLOGENIES



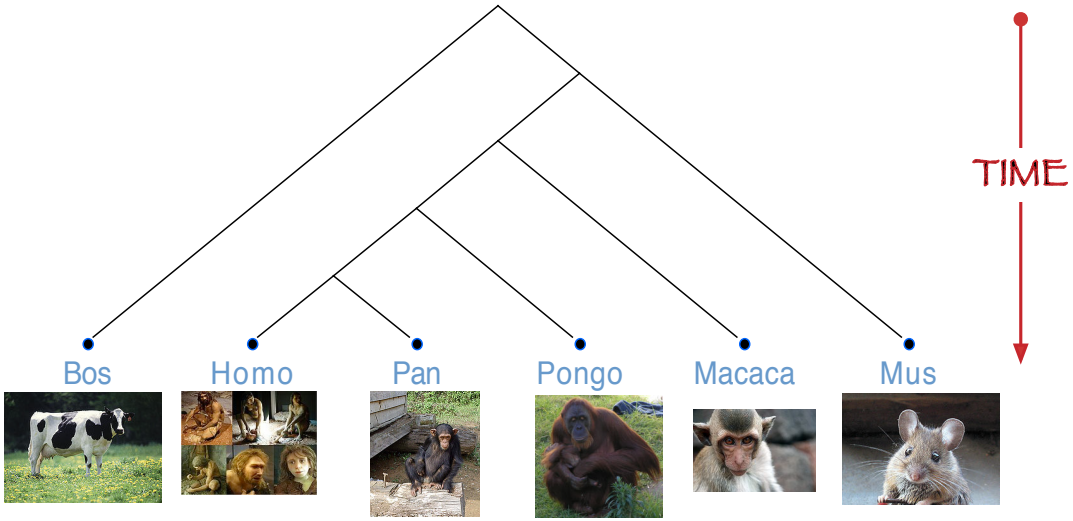
PHYLOGENIES



In a phylogenetic tree:

- Leaves represent existing species;

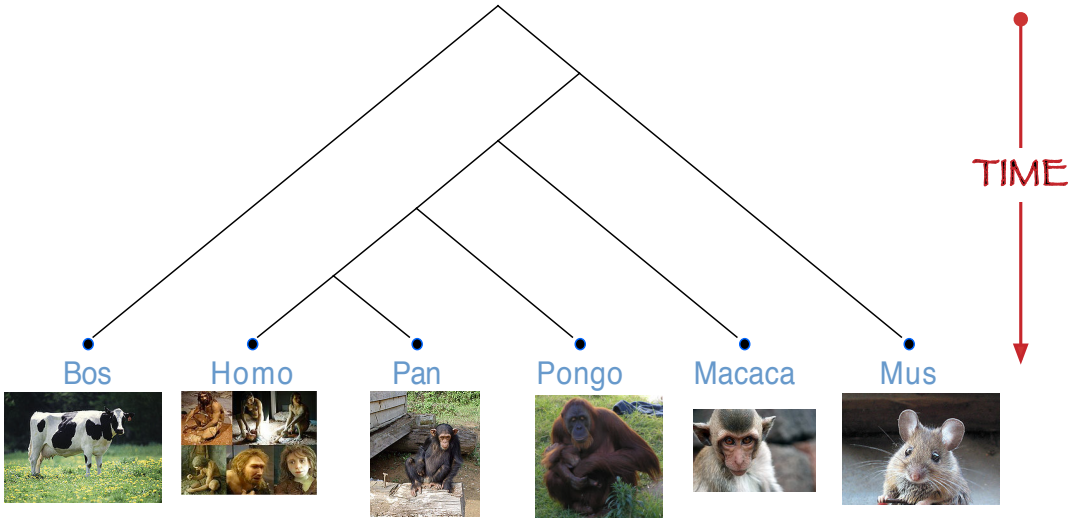
PHYLOGENIES



In a phylogenetic tree:

- Leaves represent existing species;
- Internal nodes represent (hypothetical) ancestors;

PHYLOGENIES



In a phylogenetic tree:

- Leaves represent existing species;
- Internal nodes represent (hypothetical) ancestors;
- Branch lengths represent the times that these ancestors existed.

PHYLOGENY INFERENCE

The earliest phylogenies were constructed by considering physical characteristics of the species.

PHYLOGENY INFERENCE

The earliest phylogenies were constructed by considering physical characteristics of the species.

These days, we look at differences between the genomes of the species.

PHYLOGENY INFERENCE

The earliest phylogenies were constructed by considering physical characteristics of the species.

These days, we look at differences between the genomes of the species.

```
Mouse     ...CTGCGATTACACCGGAGTCGACCTAG...
Dog       ...CTGCTATTACCCGGGAGTAGACCTAG...
Bat       ...ATTCAATGACATTGGGATTACCCTAG...
Rat       ...ATTCCATGACATTCGCTAGCGTCTAG...
```

PHYLOGENY INFERENCE

The earliest phylogenies were constructed by considering physical characteristics of the species.

These days, we look at differences between the genomes of the species.

```
Mouse     ...CTGCGATTACACCGGAGTCGACCTAG...
Dog       ...CTGCTATTACCCGGGAGTAGACCTAG...
Bat       ...ATTCAATGACATTGGGATTACCCTAG...
Rat       ...ATTCCATGACATTCGCTAGCGTCTAG...
```

Sequences which are more similar to each other are considered to have diverged more recently.

PHYLOGENY INFERENCE

The earliest phylogenies were constructed by considering physical characteristics of the species.

These days, we look at differences between the genomes of the species.

```
Mouse     ...CTGCGATTACACCGGAGTCGACCTAG...
Dog       ...CTGCTATTACCCGGGAGTAGACCTAG...
Bat       ...ATTCAATGACATTGGGATTACCCTAG...
Rat       ...ATTCCATGACATTCGCTAGCGTCTAG...
```

Sequences which are more similar to each other are considered to have diverged more recently.

Modern-day methods maximise the phylogeny likelihood according to stochastic models of sequence evolution.

Genes

Genomes can be subdivided into thousands of smaller sequences with specific functions — genes.

Genes

Genomes can be subdivided into thousands of smaller sequences with specific functions — genes.

Like species, genes evolve through time and we can depict evolutionary relationships between them.

Genes

Genomes can be subdivided into thousands of smaller sequences with specific functions — genes.

Like species, genes evolve through time and we can depict evolutionary relationships between them.

We consider gene families, and construct their gene phylogenies, using methods similar to those for species trees.

Gene and species phylogenies

Gene and species trees can be very similar — when a species diverges into two (speciation), this will also create two new genes.

Gene and species phylogenies

Gene and species trees can be very similar — when a species diverges into two (speciation), this will also create two new genes.

As such, gene phylogenies are often used in methods to construct or improve species phylogenies.

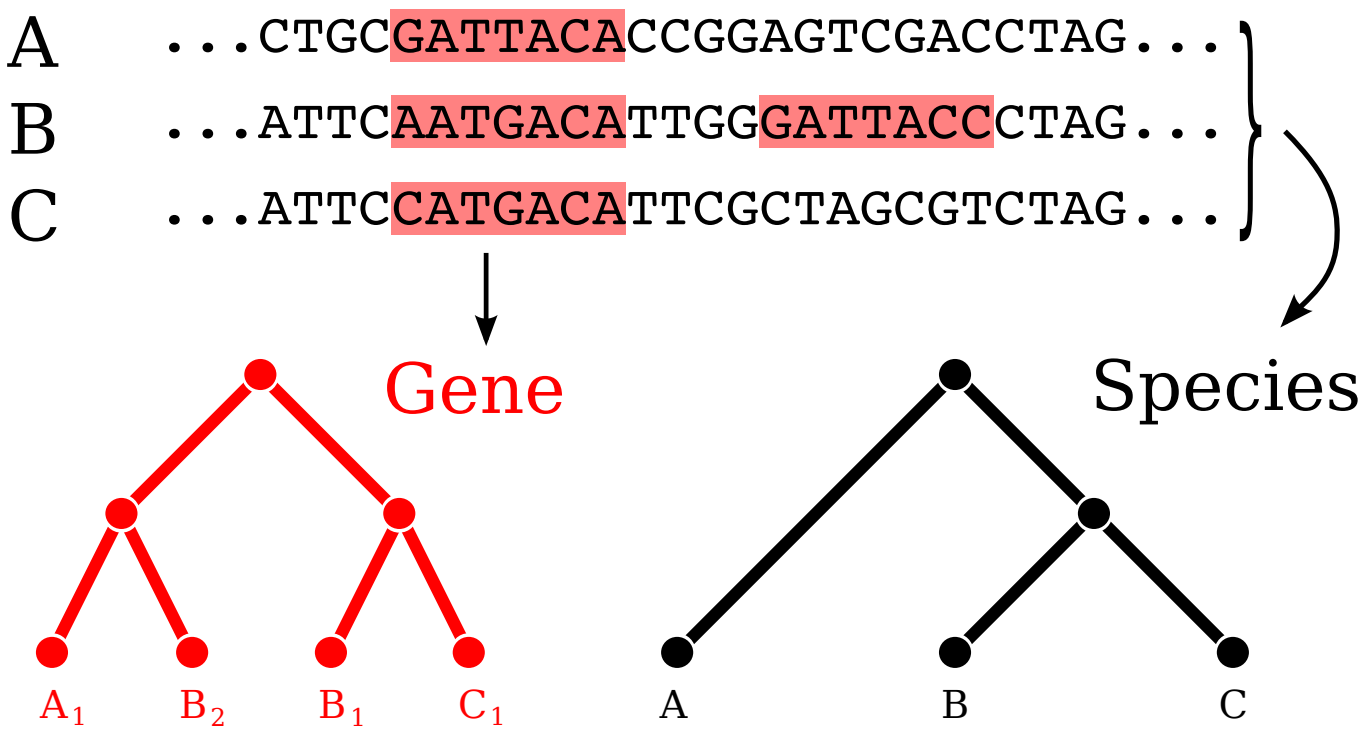
Gene and species phylogenies

Gene and species trees can be very similar — when a species diverges into two (speciation), this will also create two new genes.

As such, gene phylogenies are often used in methods to construct or improve species phylogenies.

But gene trees can differ significantly from their species tree!

Gene and species phylogenies



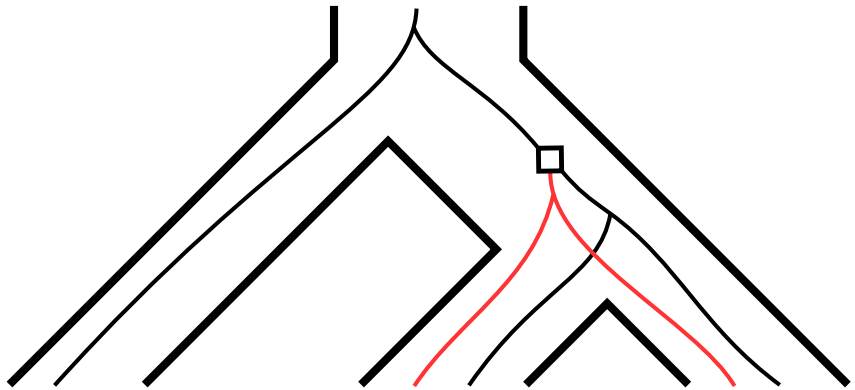
Genetic evolutionary events

This difference occurs because of evolutionary events which affect genes but do not create new species. These include:

Genetic evolutionary events

This difference occurs because of evolutionary events which affect genes but do not create new species. These include:

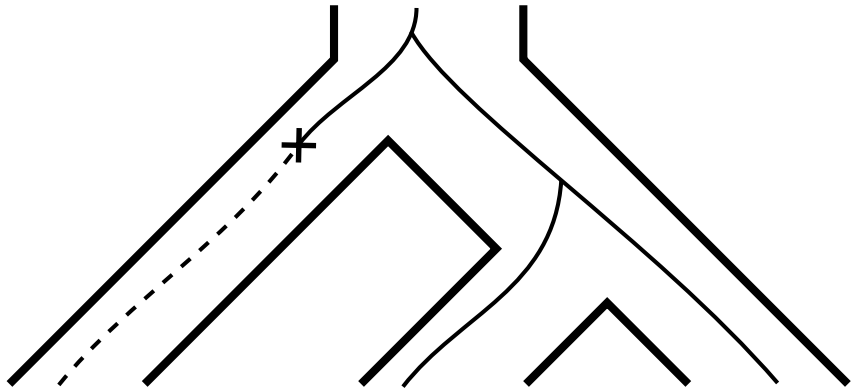
- Duplications;
-
-
-



Genetic evolutionary events

This difference occurs because of evolutionary events which affect genes but do not create new species. These include:

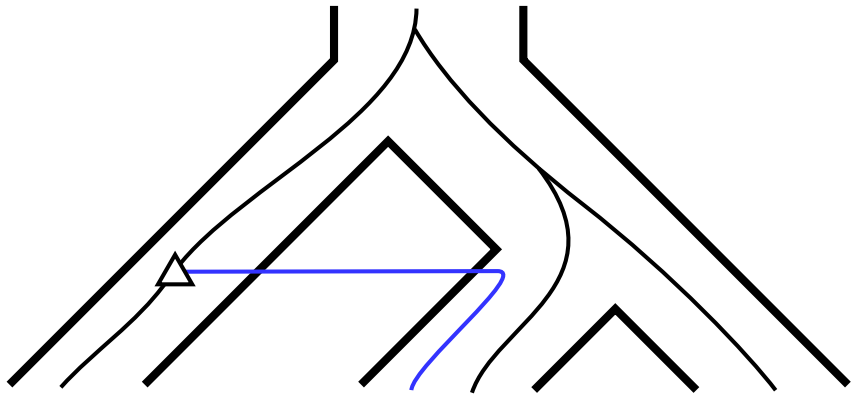
- Duplications;
- Losses;
-
-



Genetic evolutionary events

This difference occurs because of evolutionary events which affect genes but do not create new species. These include:

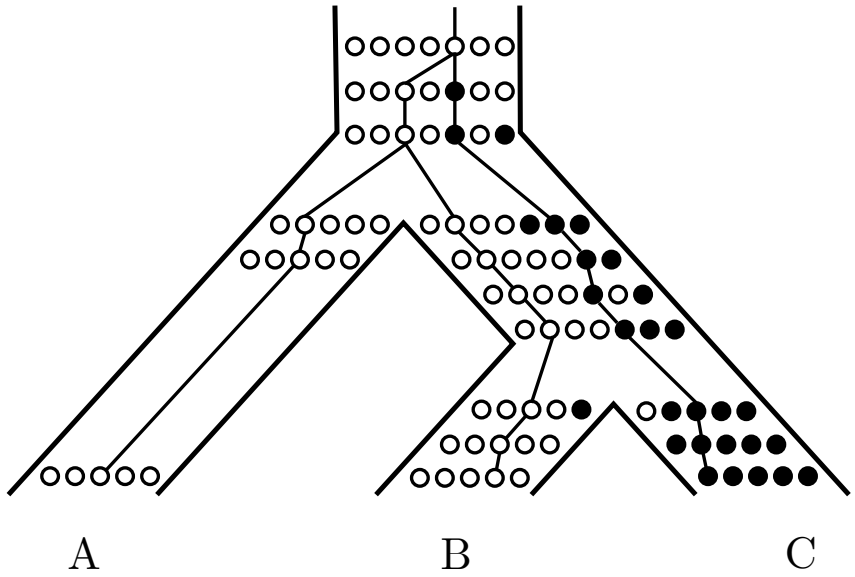
- Duplications;
- Losses;
- Lateral genetic transfer;
-



Genetic evolutionary events

This difference occurs because of evolutionary events which affect genes but do not create new species. These include:

- Duplications;
- Losses;
- Lateral genetic transfer;
- Incomplete lineage sorting (ILS).



RECONCILIATIONS

Reconciliations (Goodman 1979) map a gene tree into a species tree, explaining the discrepancy by these events.

RECONCILIATIONS

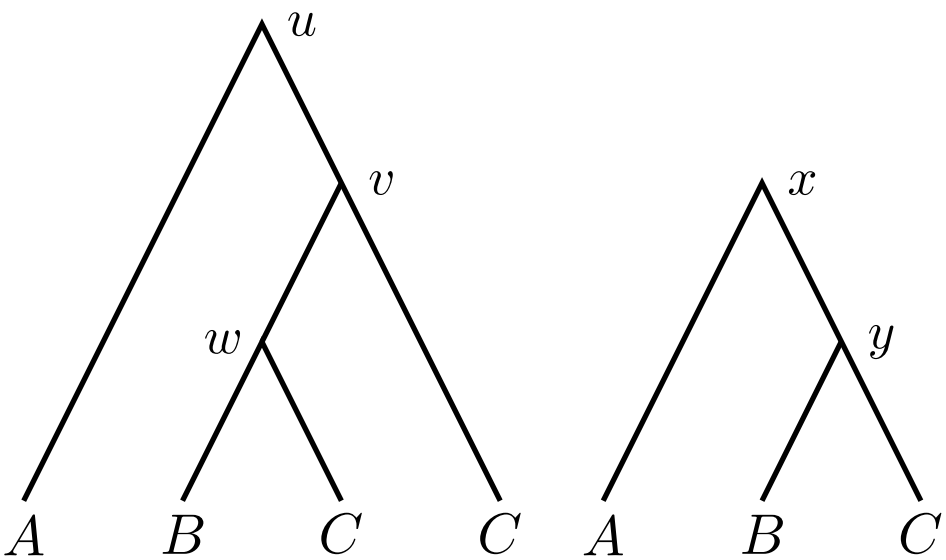
Reconciliations (Goodman 1979) map a gene tree into a species tree, explaining the discrepancy by these events.

The simplest case is the DL model: only duplications and losses are allowed.

RECONCILIATIONS

Reconciliations (Goodman 1979) map a gene tree into a species tree, explaining the discrepancy by these events.

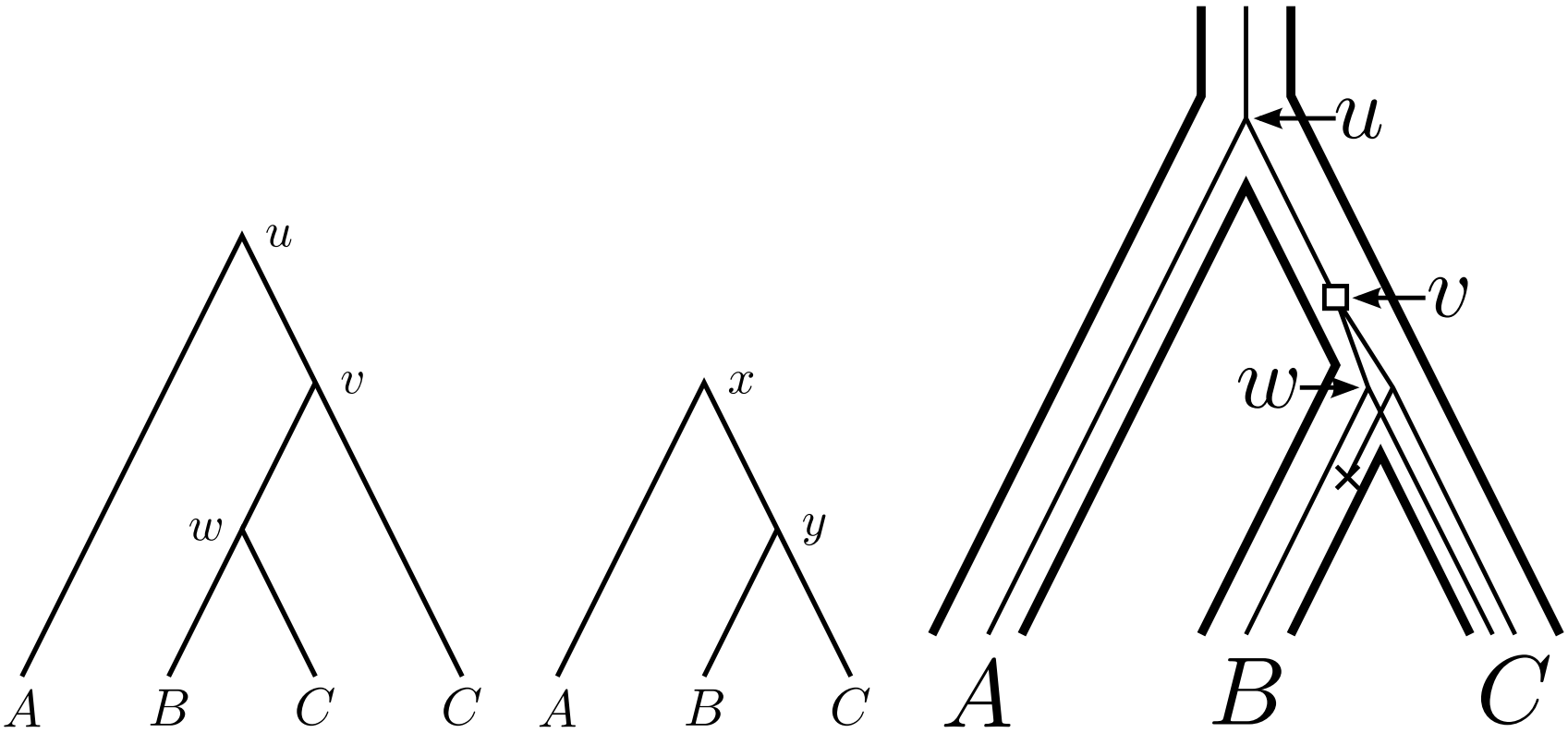
The simplest case is the DL model: only duplications and losses are allowed.



RECONCILIATIONS

Reconciliations (Goodman 1979) map a gene tree into a species tree, explaining the discrepancy by these events.

The simplest case is the DL model: only duplications and losses are allowed.



PARSIMONY VS PROBABILISTIC METHODS

There are two main paradigms for finding (the best) reconciliation:

PARSIMONY VS PROBABILISTIC METHODS

There are two main paradigms for finding (the best) reconciliation:

- Parsimony assigns a cost to each event and minimises the total cost.

PARSIMONY VS PROBABILISTIC METHODS

There are two main paradigms for finding (the best) reconciliation:

- Parsimony assigns a cost to each event and minimises the total cost.

  This is fast, but less realistic.

PARSIMONY VS PROBABILISTIC METHODS

There are two main paradigms for finding (the best) reconciliation:

- Parsimony assigns a cost to each event and minimises the total cost.

  This is fast, but less realistic.

- Probabilistic methods search for the maximum-likelihood reconciliation under a probabilistic model of gene evolution within species.

PARSIMONY VS PROBABILISTIC METHODS

There are two main paradigms for finding (the best) reconciliation:

- Parsimony assigns a cost to each event and minimises the total cost.

  This is fast, but less realistic.

- Probabilistic methods search for the maximum-likelihood reconciliation under a probabilistic model of gene evolution within species.

  This is more realistic, but slower.

PARSIMONY

The most parsimonious reconciliation for the DL model can be found in linear time with the lowest common ancestor (LCA) reconciliation.

PARSIMONY

The most parsimonious reconciliation for the DL model can be found in linear time with the lowest common ancestor (LCA) reconciliation.

Including other events makes the problem harder, but sometimes still solvable.

PARSIMONY

The most parsimonious reconciliation for the DL model can be found in linear time with the lowest common ancestor (LCA) reconciliation.

Including other events makes the problem harder, but sometimes still solvable.

For example, the DTL model (DL with transfers) can be solved in polynomial time with dynamic programming.

PARSIMONY

The most parsimonious reconciliation for the DL model can be found in linear time with the lowest common ancestor (LCA) reconciliation.

Including other events makes the problem harder, but sometimes still solvable.

For example, the DTL model (DL with transfers) can be solved in polynomial time with dynamic programming.

However, more complicated models can be (NP-)hard to optimise.

SEGMENTAL DUPLICATIONS

Traditionally, gene trees are reconciled to species trees independently.

SEGMENTAL DUPLICATIONS

Traditionally, gene trees are reconciled to species trees independently.

However, they are not independent: when genes are adjacent to each other on the chromosome, segmental duplications can occur which affect many genes at once.

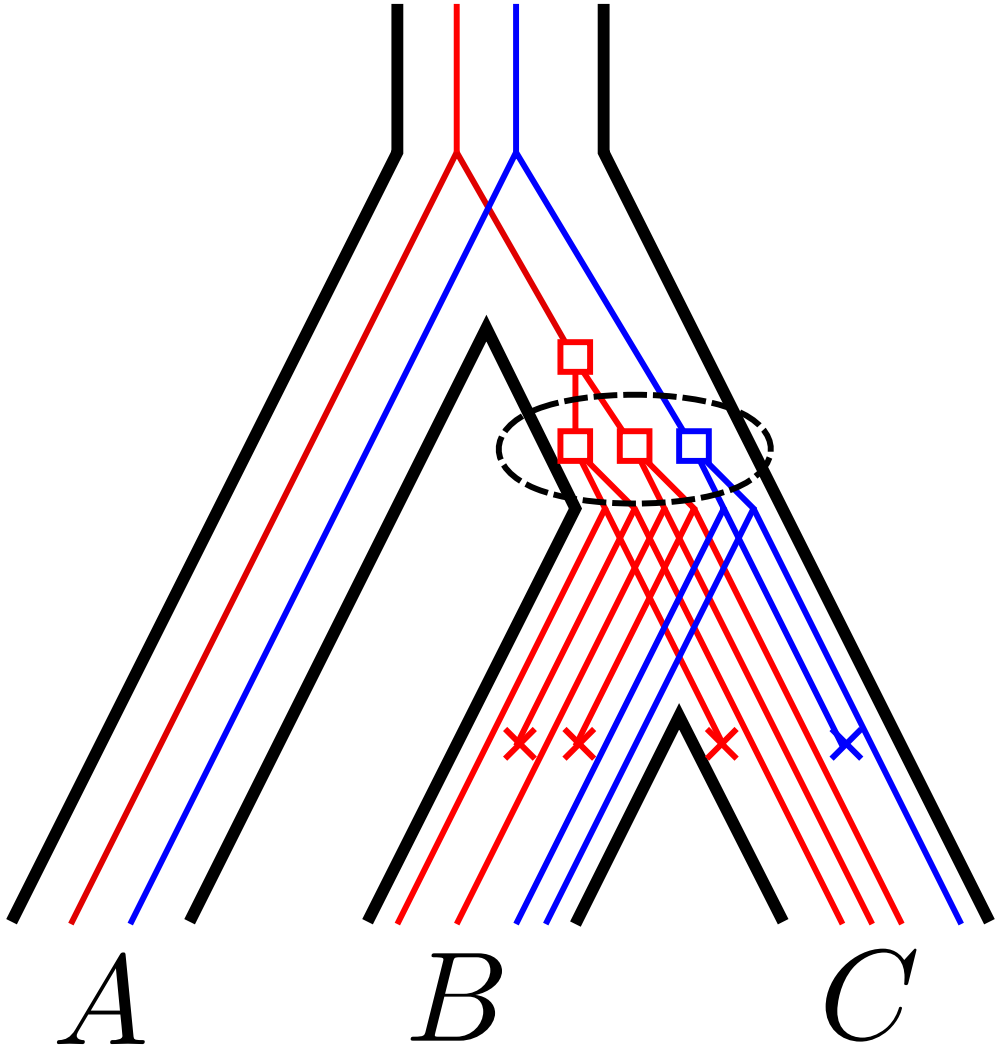
SEGMENTAL DUPLICATIONS

Traditionally, gene trees are reconciled to species trees independently.

However, they are not independent: when genes are adjacent to each other on the chromosome, segmental duplications can occur which affect many genes at once.

This can go up to whole genome duplications.

SEGMENTAL DUPLICATIONS



SEGMENTAL DUPLICATIONS

Suppose that a reconciliation α has $d_{SD}(\alpha)$ segmental duplications and $l(\alpha)$ losses.

SEGMENTAL DUPLICATIONS

Suppose that a reconciliation α has $d_{SD}(\alpha)$ segmental duplications and $l(\alpha)$ losses.

Problem (Most parsimonious reconciliation, MPR)

Given gene trees $\mathcal{G}$, a species tree S, duplication cost δ, and loss cost λ, find the reconciliation α that minimises

$$c_{SD}(\alpha) := \delta \cdot d_{SD}(\alpha) + \lambda \cdot l(\alpha).$$

SEGMENTAL DUPLICATIONS

Suppose that a reconciliation α has $d_{SD}(\alpha)$ segmental duplications and $l(\alpha)$ losses.

Problem (Most parsimonious reconciliation, MPR)

Given gene trees $\mathcal{G}$, a species tree S, duplication cost δ, and loss cost λ, find the reconciliation α that minimises

$$c_{SD}(\alpha) := \delta \cdot d_{SD}(\alpha) + \lambda \cdot l(\alpha).$$

Dondi *et al.* (2019) studied this problem and gave an algorithm that was exponential in the number of genes, with a base of $\lfloor \frac{\delta}{\lambda} \rfloor$.

SEGMENTAL DUPLICATIONS

Suppose that a reconciliation α has $d_{SD}(\alpha)$ segmental duplications and $l(\alpha)$ losses.

Problem (Most parsimonious reconciliation, MPR)

Given gene trees $\mathcal{G}$, a species tree S, duplication cost δ, and loss cost λ, find the reconciliation α that minimises

$$c_{SD}(\alpha) := \delta \cdot d_{SD}(\alpha) + \lambda \cdot l(\alpha).$$

Dondi *et al.* (2019) studied this problem and gave an algorithm that was exponential in the number of genes, with a base of $\lfloor \frac{\delta}{\lambda} \rfloor$.

They also showed that this problem is NP-hard.

SEGMENTAL DUPLICATIONS

Suppose that a reconciliation α has $d_{SD}(\alpha)$ segmental duplications and $l(\alpha)$ losses.

Problem (Most parsimonious reconciliation, MPR)

Given gene trees $\mathcal{G}$, a species tree S, duplication cost δ, and loss cost λ, find the reconciliation α that minimises

$$c_{SD}(\alpha) := \delta \cdot d_{SD}(\alpha) + \lambda \cdot l(\alpha).$$

Dondi *et al.* (2019) studied this problem and gave an algorithm that was exponential in the number of genes, with a base of $\lfloor \frac{\delta}{\lambda} \rfloor$.

They also showed that this problem is NP-hard. ☹

THE BOLTZMANN DISTRIBUTION

We approach the problem from a slightly different way by imposing a probability distribution on reconciliations.

THE BOLTZMANN DISTRIBUTION

We approach the problem from a slightly different way by imposing a probability distribution on reconciliations.

For a reconciliation α with cost $c_{SD}(\alpha)$, we set its probability to be:

$$P(\alpha) \propto e^{-c_{SD}(\alpha)/kT},$$

where k is Boltzmann's constant and T is a temperature parameter.

THE BOLTZMANN DISTRIBUTION

We approach the problem from a slightly different way by imposing a probability distribution on reconciliations.

For a reconciliation α with cost $c_{SD}(\alpha)$, we set its probability to be:

$$P(\alpha) \propto e^{-c_{SD}(\alpha)/kT},$$

where k is Boltzmann's constant and T is a temperature parameter.

This is the well-known Boltzmann distribution from statistical mechanics.

THE BOLTZMANN DISTRIBUTION

We approach the problem from a slightly different way by imposing a probability distribution on reconciliations.

For a reconciliation α with cost $c_{SD}(\alpha)$, we set its probability to be:

$$P(\alpha) \propto e^{-c_{SD}(\alpha)/kT},$$

where k is Boltzmann's constant and T is a temperature parameter.

This is the well-known Boltzmann distribution from statistical mechanics.

There are known to be a lot of reconciliations for given gene and species trees, so calculating the normalising constant is hard.

THE BOLTZMANN DISTRIBUTION

By changing the temperature, we can access different parts of the reconciliation space.

THE BOLTZMANN DISTRIBUTION

By changing the temperature, we can access different parts of the reconciliation space.

When the temperature is high, reconciliations occur with (near-)uniform probability.

THE BOLTZMANN DISTRIBUTION

By changing the temperature, we can access different parts of the reconciliation space.

When the temperature is high, reconciliations occur with (near-)uniform probability.

When the temperature is low, cheaper reconciliations are much more probable and we almost always get (near-)optimal reconciliations.

Sampling from the Boltzmann distribution

To sample from the Boltzmann distribution, we use a Gibbs sampling scheme; starting from a valid reconciliation, we change the position of each gene node one at a time.

Sampling from the Boltzmann distribution

To sample from the Boltzmann distribution, we use a Gibbs sampling scheme; starting from a valid reconciliation, we change the position of each gene node one at a time.

The new position of the node is resampled from its conditional distribution based on the rest of the reconciliation.

SAMPLING FROM THE BOLTZMANN DISTRIBUTION

To sample from the Boltzmann distribution, we use a Gibbs sampling scheme; starting from a valid reconciliation, we change the position of each gene node one at a time.

The new position of the node is resampled from its conditional distribution based on the rest of the reconciliation.

This only depends on the relative costs of the reconciliations (and not on the normalising factor!).

Sampling from the Boltzmann distribution

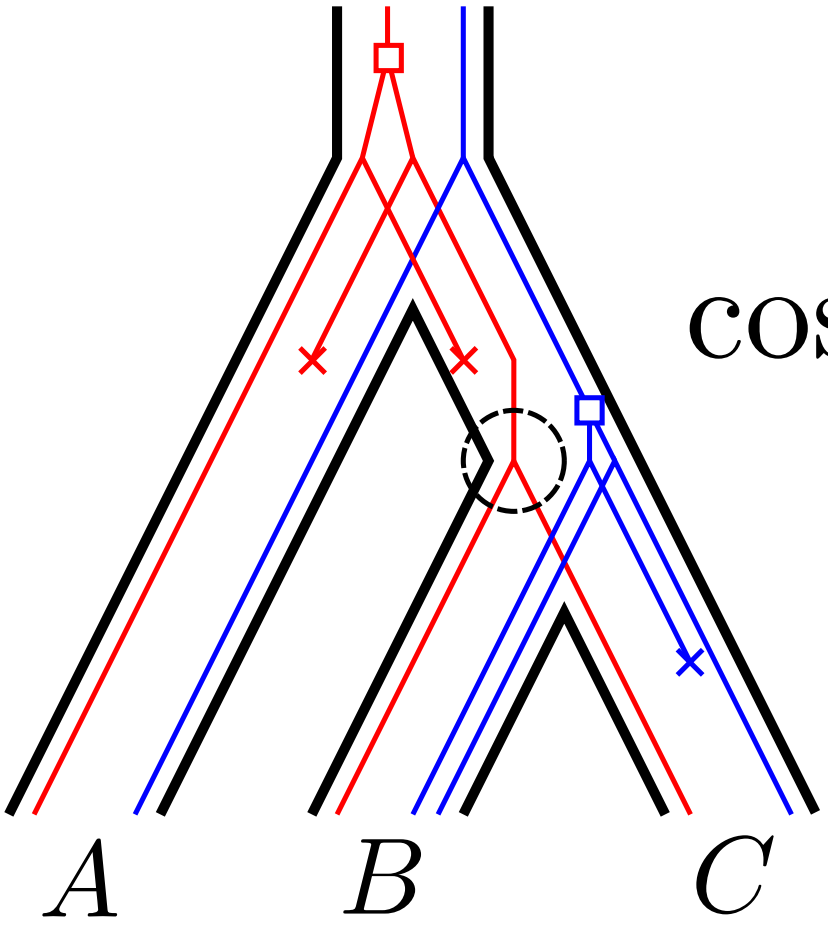
To sample from the Boltzmann distribution, we use a Gibbs sampling scheme; starting from a valid reconciliation, we change the position of each gene node one at a time.

The new position of the node is resampled from its conditional distribution based on the rest of the reconciliation.

This only depends on the relative costs of the reconciliations (and not on the normalising factor!).

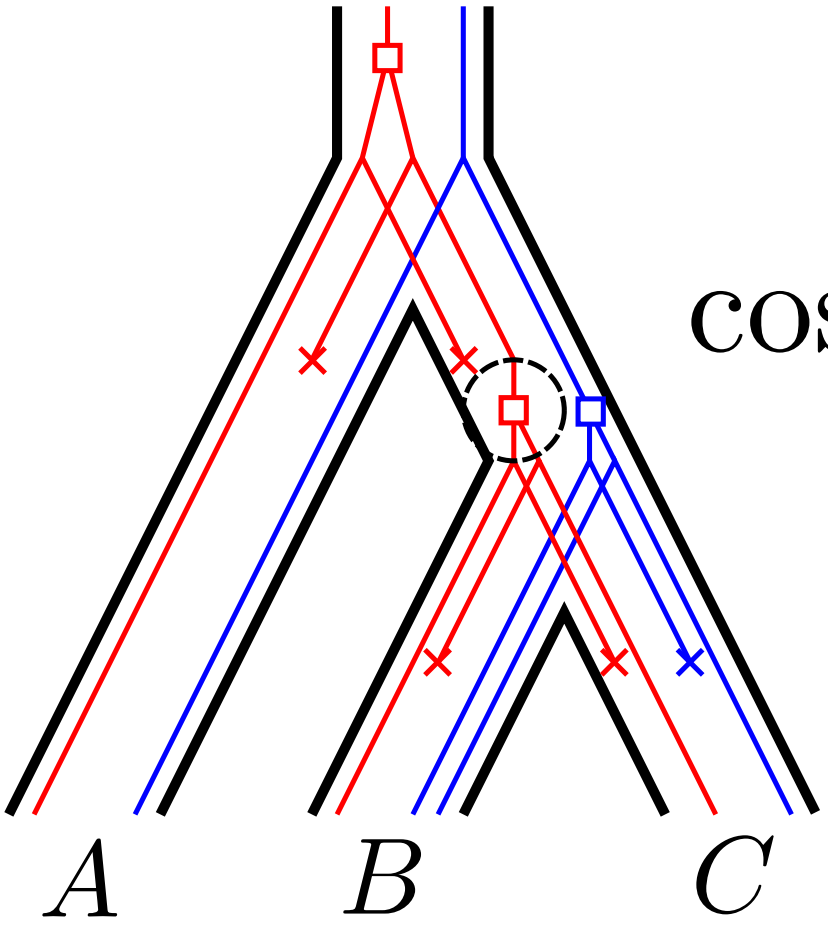
Thus each move can be performed very quickly — $O(1)$ time!

GIBBS SAMPLING



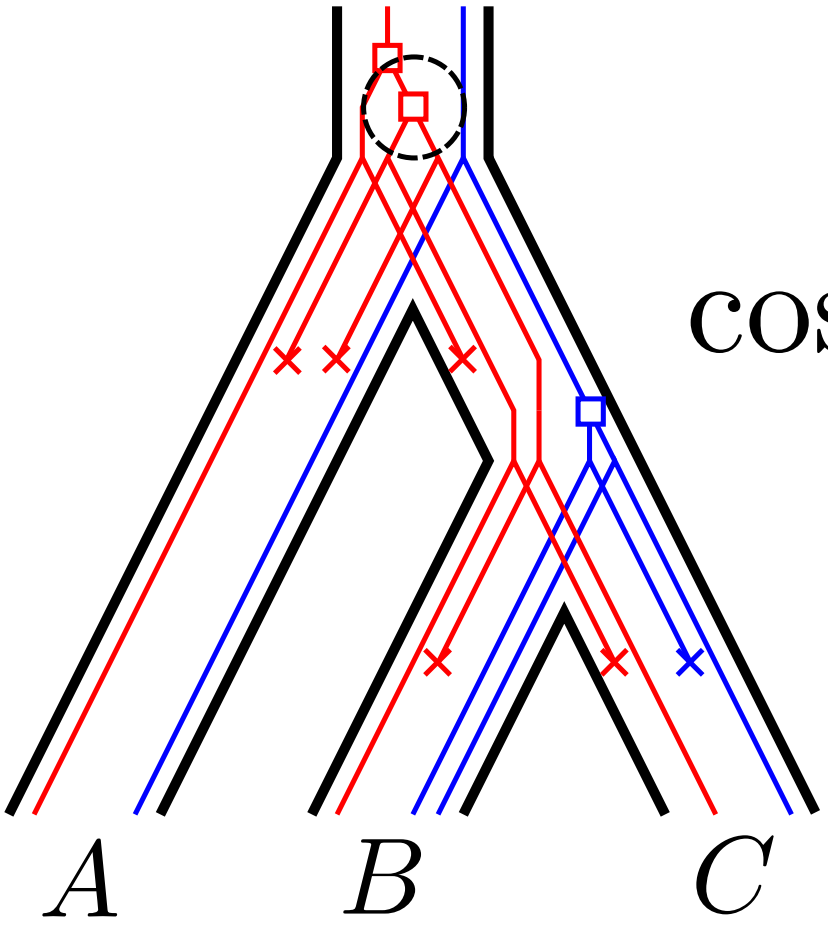
$$\text{cost} = ?$$

GIBBS SAMPLING



$$\text{cost} = ? + 2\lambda$$

GIBBS SAMPLING



$$\text{cost} = ? + \delta + 3\lambda$$

SIMULATED ANNEALING

To produce an optimal reconciliation, we start at a high temperature to access all parts of the reconciliation space.

SIMULATED ANNEALING

To produce an optimal reconciliation, we start at a high temperature to access all parts of the reconciliation space.

Then we slowly lower the temperature to (near-)zero.

SIMULATED ANNEALING

To produce an optimal reconciliation, we start at a high temperature to access all parts of the reconciliation space.

Then we slowly lower the temperature to (near-)zero.

This will produce an optimal reconciliation with probability approaching 1.

THE ALGORITHM

1: Set α to be the LCA mapping from $\mathcal{G}$ to S

2: **for** $t = 1, \ldots, t_{\max}$ **do**
3: $T \leftarrow T_0 \left(1 - \frac{t-1}{t_{\max}}\right)$
4: **for** each internal vertex u of $\mathcal{G}$ in a fixed order **do**
5: set M_u to the set of all possible images of u

6: **for** each image m in M_u **do**
7: set α_m to be α with the image of u set to m
8: calculate $c_{SD}(\alpha_m)$
9: **end for**

10: resample $\alpha(u)$ from $m \in M_u$ with probability $\propto e^{-c_{SD}(\alpha_m)/kT}$
11: **end for**
12: **end for**

13: **return** α

SIMULATIONS

We simulate 50 gene trees over 20 species.

SIMULATIONS

We simulate 50 gene trees over 20 species.

Species trees are generated by a pure birth process with birth rate 1.

SIMULATIONS

We simulate 50 gene trees over 20 species.

Species trees are generated by a pure birth process with birth rate 1.

Gene trees start with a single copy, then:

- Each lineage duplicates with rate r_B;

SIMULATIONS

We simulate 50 gene trees over 20 species.

Species trees are generated by a pure birth process with birth rate 1.

Gene trees start with a single copy, then:

- Each lineage duplicates with rate r_B;
- A duplication is segmental with probability 0.5 for each contemporary lineage;

SIMULATIONS

We simulate 50 gene trees over 20 species.

Species trees are generated by a pure birth process with birth rate 1.

Gene trees start with a single copy, then:

- Each lineage duplicates with rate r_B;
- A duplication is segmental with probability 0.5 for each contemporary lineage;
- At each speciation, one lineage is lost with probability 0.5.

SIMULATIONS

We simulate 50 gene trees over 20 species.

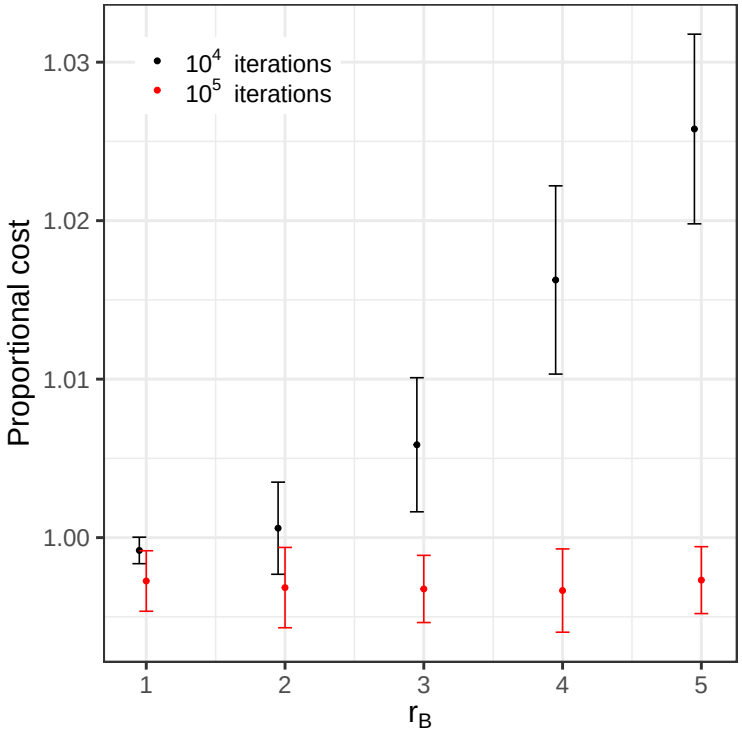
Species trees are generated by a pure birth process with birth rate 1.

Gene trees start with a single copy, then:

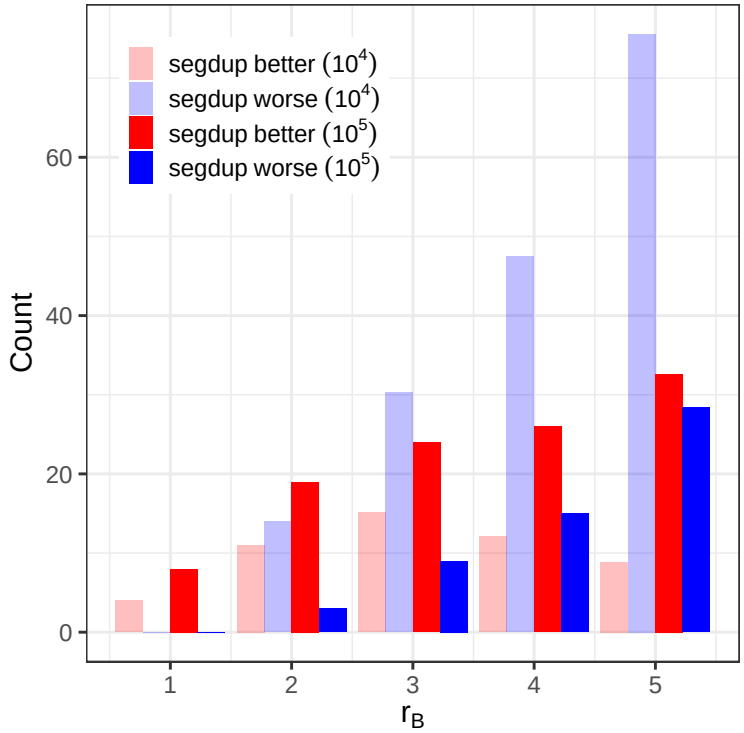
- Each lineage duplicates with rate r_B;
- A duplication is segmental with probability 0.5 for each contemporary lineage;
- At each speciation, one lineage is lost with probability 0.5.

We infer reconciliations with $\delta = 10, \lambda = 1$ (a challenging scenario) and compare our algorithm to MultRec (Dondi *et al.* 2019).

SIMULATION RESULTS

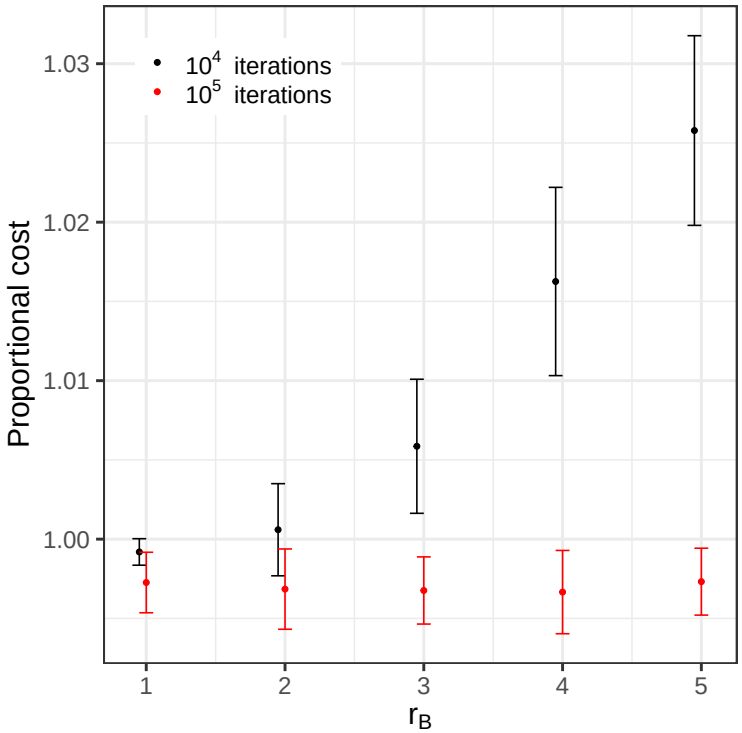


(a) Relative cost of `segdup` vs MultRec

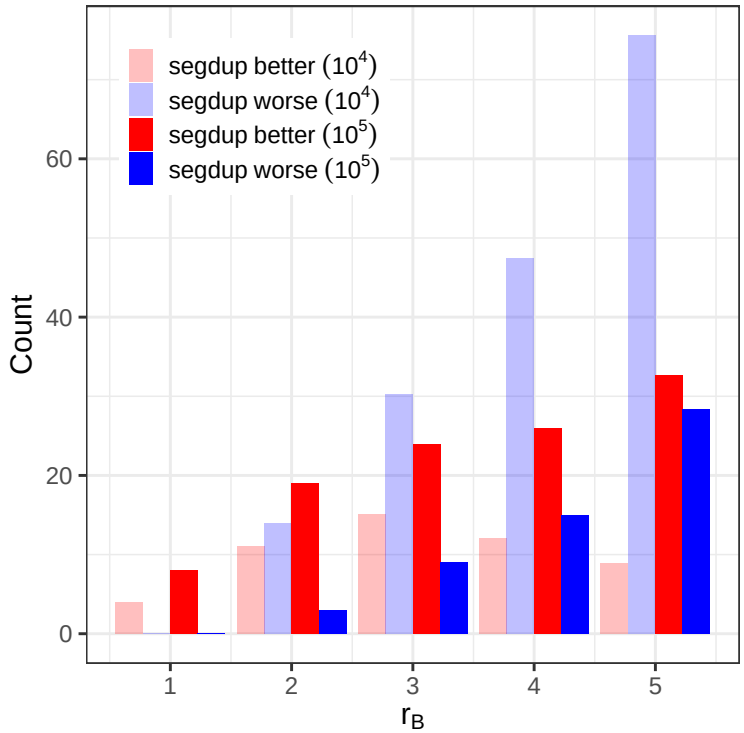


(b) Frequencies of `segdup` vs MultRec

SIMULATION RESULTS



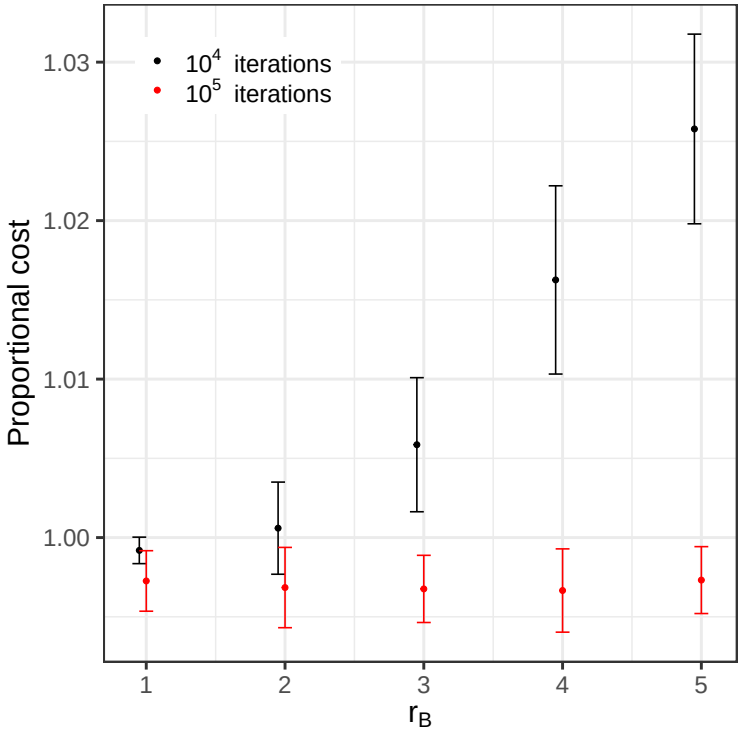
(a) Relative cost of `segdup` vs MultRec



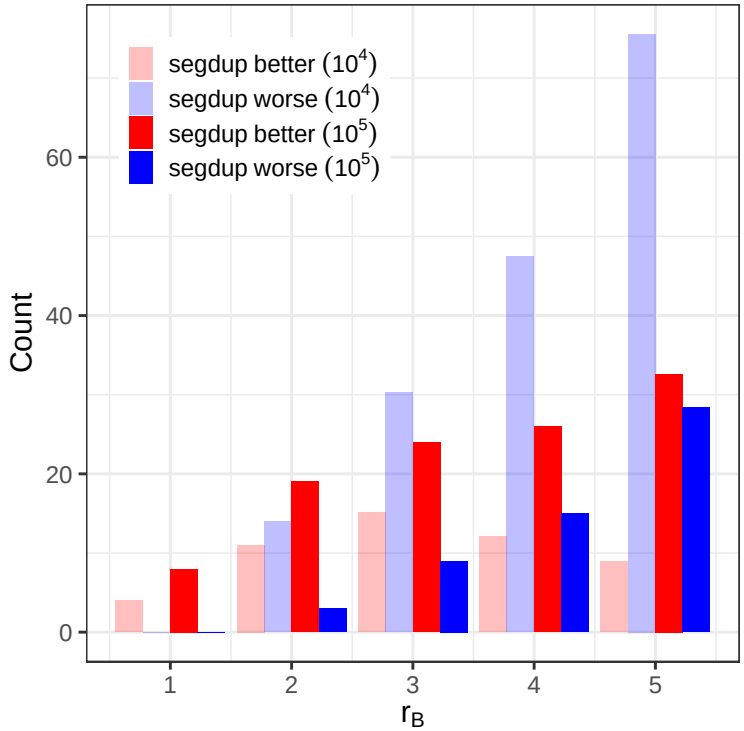
(b) Frequencies of `segdup` vs MultRec

`segdup` consistently finds a better reconciliation on average than MultRec, for 10^5 iterations. For 10^4 iterations, it is slightly worse for high duplication rate.

SIMULATION RESULTS



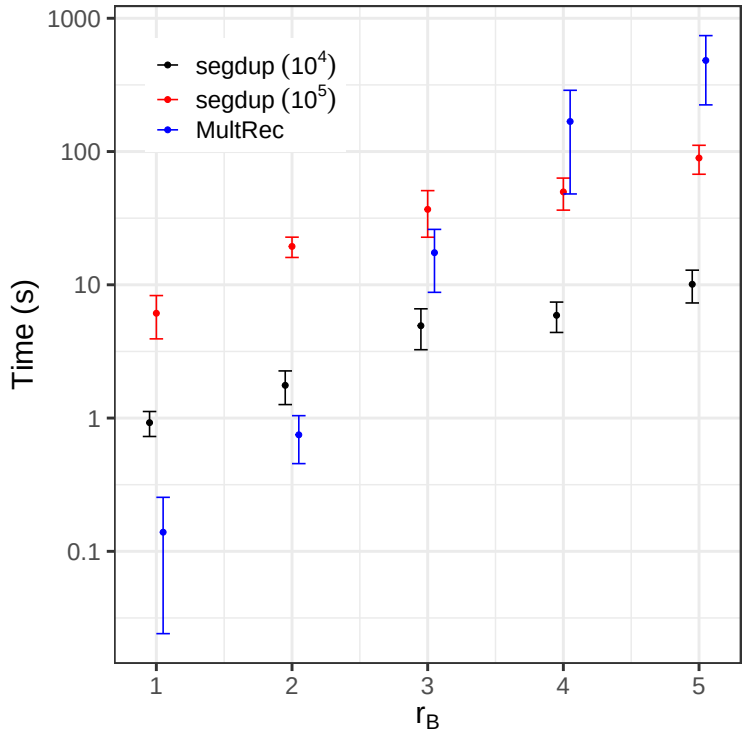
(a) Relative cost of `segdup` vs MultRec



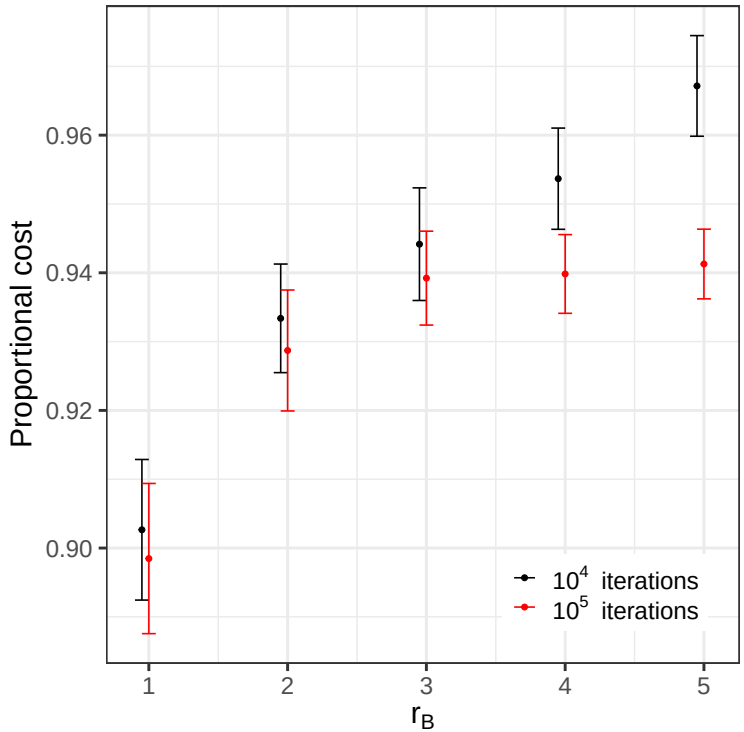
(b) Frequencies of `segdup` vs MultRec

`segdup` and MultRec often find the same (optimal) reconciliation. When they disagree, `segdup` usually does better for 10^5 iterations.

SIMULATION RESULTS

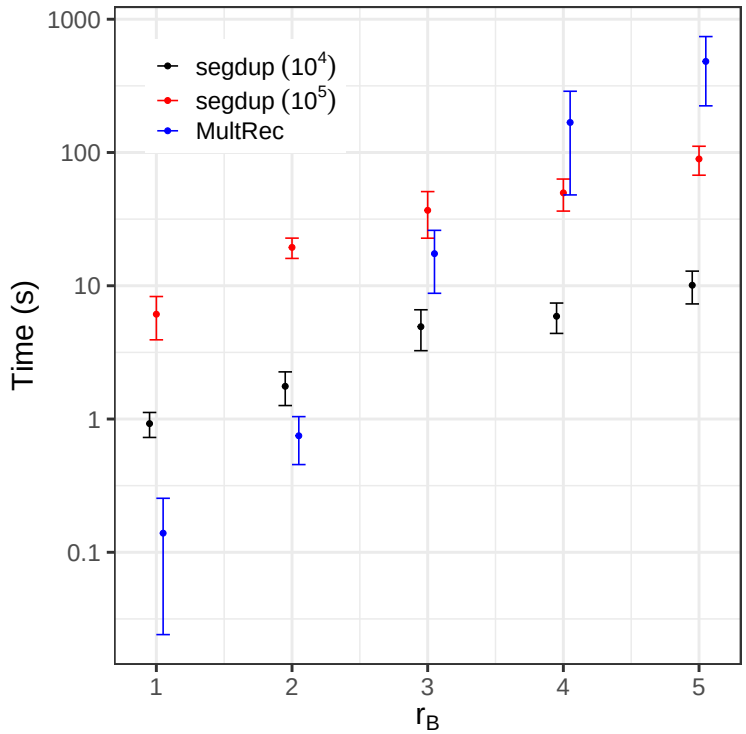


(a) Running time

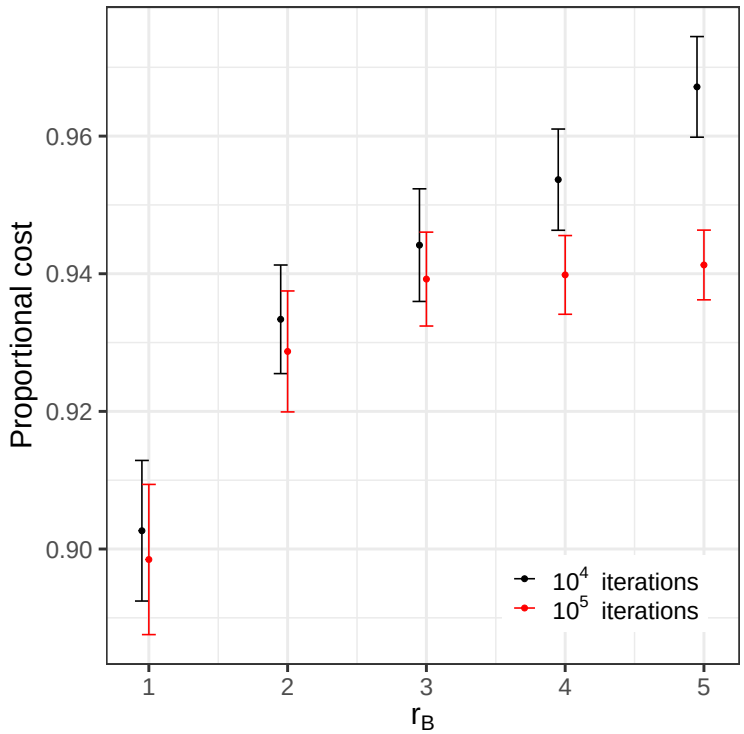


(b) Relative cost of `segdup` vs true reconciliation

SIMULATION RESULTS



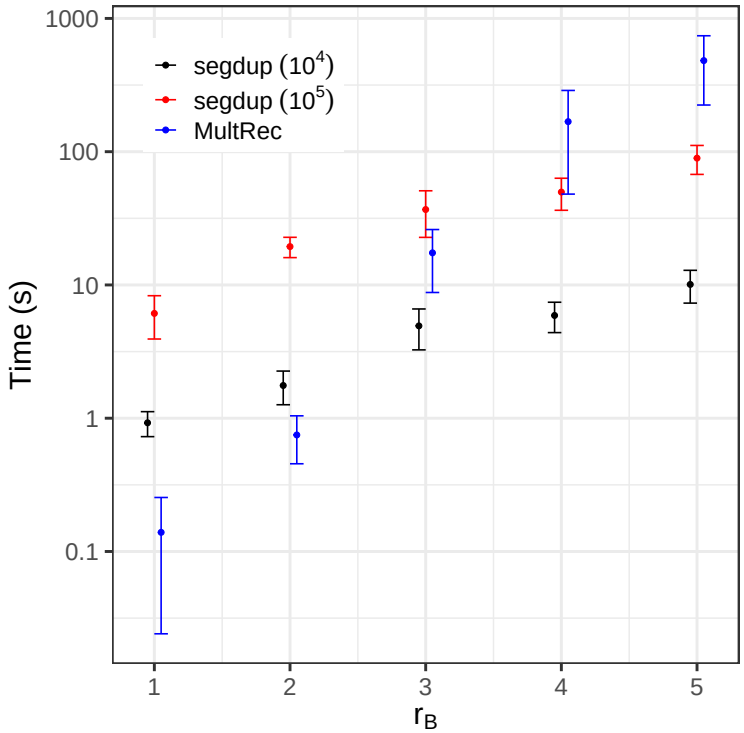
(a) Running time



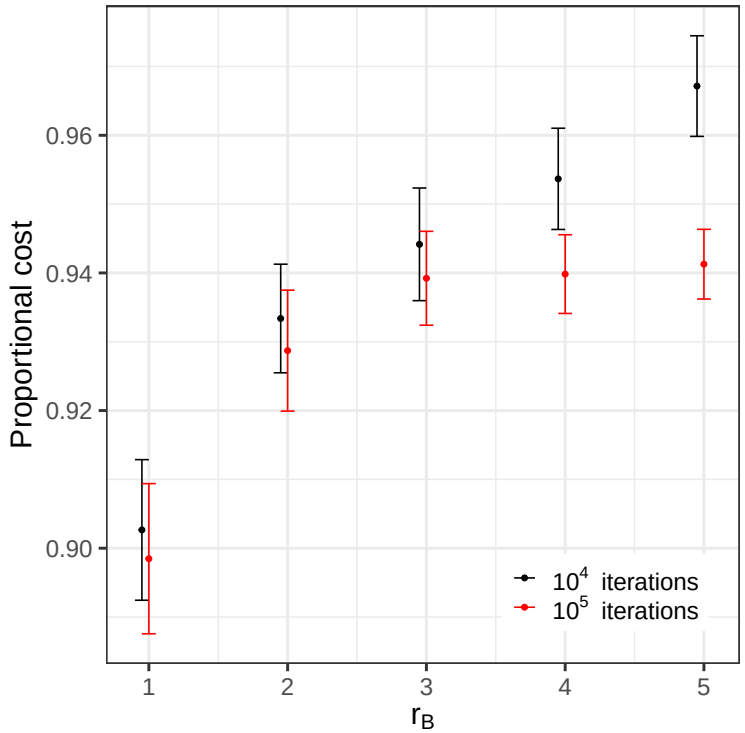
(b) Relative cost of `segdup` vs true reconciliation

`segdup` is slower than MultRec for low duplication rate, but this changes as the duplication rate increases.

SIMULATION RESULTS



(a) Running time



(b) Relative cost of `segdup` vs true reconciliation

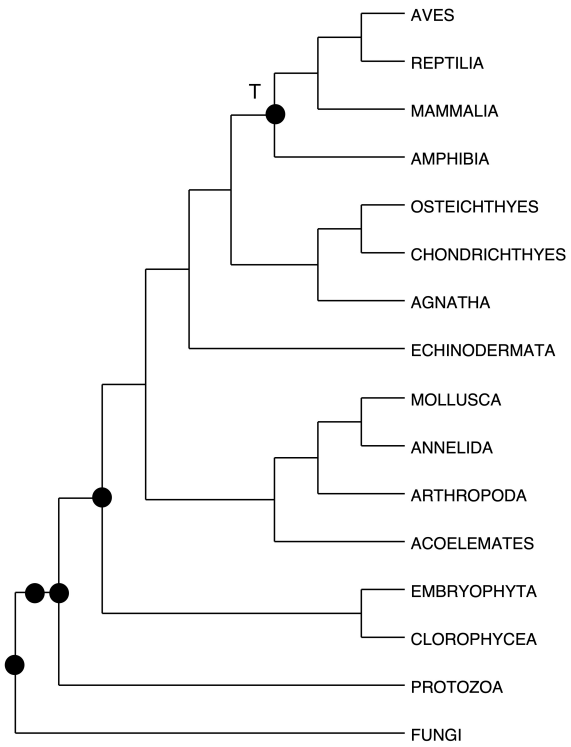
`segdup` always finds a slightly better reconciliation than the true scenario.

GUIGÓ *et al.* TREES

We re-analysed a real dataset of 53 gene trees over a species tree of 16 eukaryotes (Guigó *et al.* 1996).

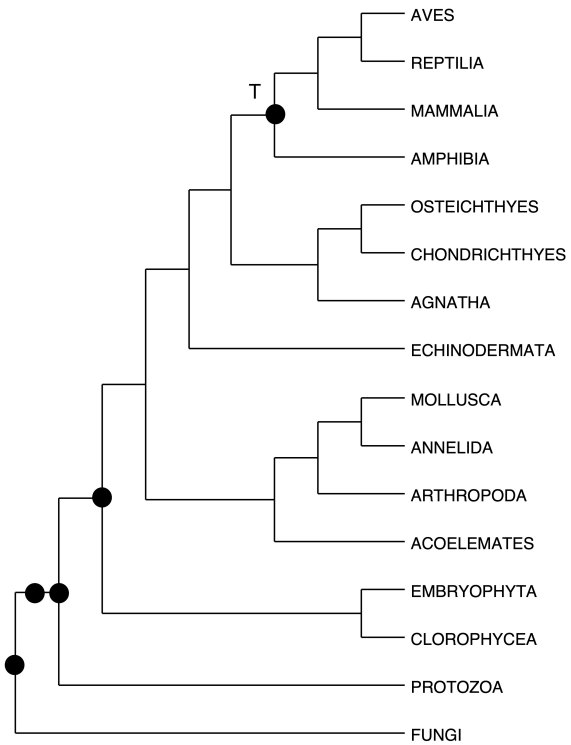
GUIGÓ *et al.* TREES

We re-analysed a real dataset of 53 gene trees over a species tree of 16 eukaryotes (Guigó *et al.* 1996).



GUIGÓ *et al.* TREES

We re-analysed a real dataset of 53 gene trees over a species tree of 16 eukaryotes (Guigó *et al.* 1996).



Previous analysis for $\delta = 50, \lambda = 1$ found 5 segmental duplications (black circles), including one above the Tetrapoda clade (marked T).

GUIGÓ *et al.* TREES

`segdup` finds either the MPR (468 losses) or a near-optimal reconciliation with 469 losses.

GUIGÓ *et al.* TREES

`segdup` finds either the MPR (468 losses) or a near-optimal reconciliation with 469 losses.

In the second reconciliation, the duplication at the Tetrapoda clade is one branch higher.

GUIGÓ *et al.* TREES

`segdup` finds either the MPR (468 losses) or a near-optimal reconciliation with 469 losses.

In the second reconciliation, the duplication at the Tetrapoda clade is one branch higher.

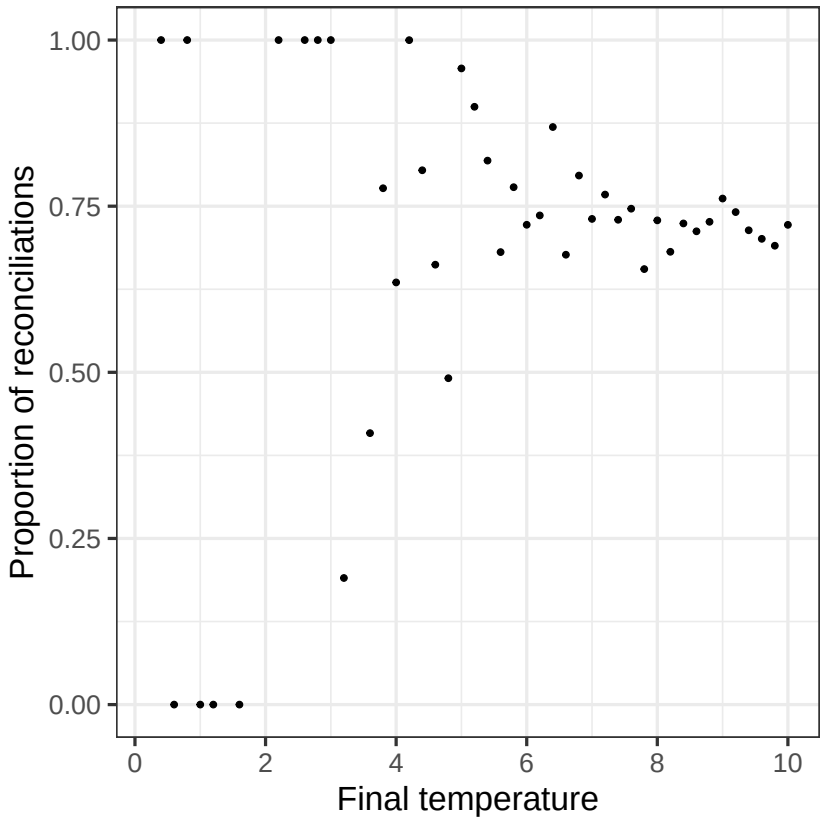
These two solutions are close in terms of cost, but are not easy to obtain from each other.

GUIGÓ *et al.* TREES

To explore this further, we study the relative occurences of these duplications for varying temperatures.

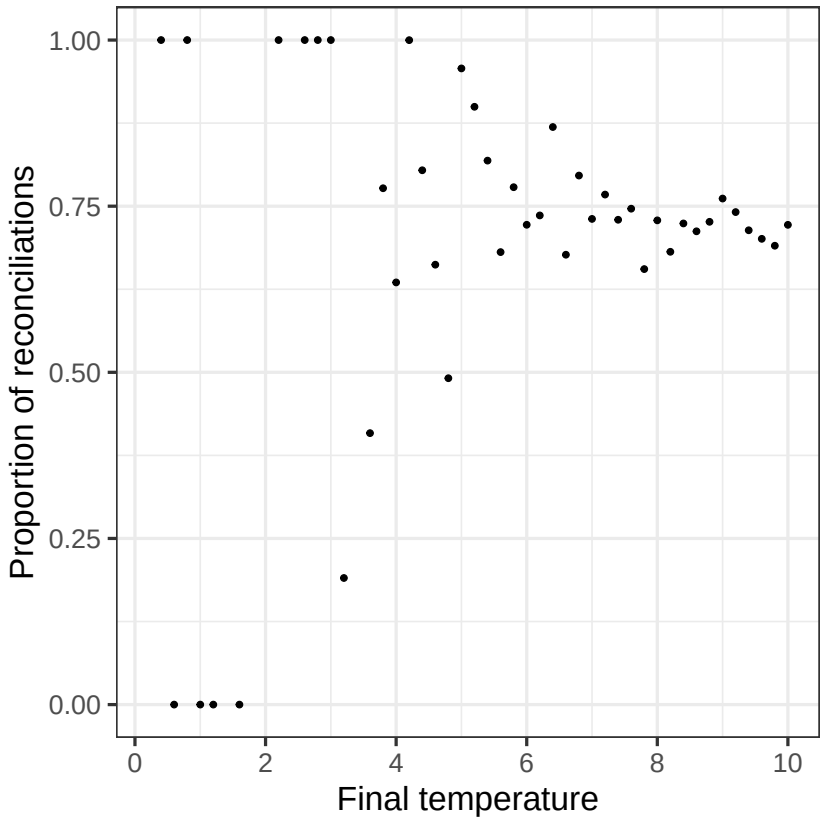
GUIGÓ *et al.* TREES

To explore this further, we study the relative occurences of these duplications for varying temperatures.



GUIGÓ *et al.* TREES

To explore this further, we study the relative occurences of these duplications for varying temperatures.



Both these duplications are feasible, but the Boltzmann distribution favours the higher duplication.

SUMMARY AND FUTURE WORK

- We have developed an algorithm to solve the MPR problem for segmental duplications.

SUMMARY AND FUTURE WORK

- We have developed an algorithm to solve the MPR problem for segmental duplications.

- This algorithm outperforms the current methods in simulations.

SUMMARY AND FUTURE WORK

- We have developed an algorithm to solve the MPR problem for segmental duplications.
- This algorithm outperforms the current methods in simulations.
- By imposing a Boltzmann distribution, we can study the space of reconciliations instead of a single MPR.

SUMMARY AND FUTURE WORK

- We have developed an algorithm to solve the MPR problem for segmental duplications.
- This algorithm outperforms the current methods in simulations.
- By imposing a Boltzmann distribution, we can study the space of reconciliations instead of a single MPR.
- This algorithm combines the speed of a probabilistic algorithm with the accuracy of a parsimonious method.

SUMMARY AND FUTURE WORK

- We have developed an algorithm to solve the MPR problem for segmental duplications.
- This algorithm outperforms the current methods in simulations.
- By imposing a Boltzmann distribution, we can study the space of reconciliations instead of a single MPR.
- This algorithm ~~combines the speed of a probabilistic algorithm with the accuracy of a parsimonious method~~ provides a new way to attack NP-hard reconciliation problems.

SUMMARY AND FUTURE WORK

- We have developed an algorithm to solve the MPR problem for segmental duplications.
- This algorithm outperforms the current methods in simulations.
- By imposing a Boltzmann distribution, we can study the space of reconciliations instead of a single MPR.
- This algorithm ~~combines the speed of a probabilistic algorithm with the accuracy of a parsimonious method~~ provides a new way to attack NP-hard reconciliation problems.
- This could be applied to many problems where gene (subtree) dependence prevents standard dynamic programming approaches.

References

- R. Dondi, M. Lafond, and C. Scornavacca. Reconciling multiple genes trees via segmental duplications and losses. *Alg. Mol. Biol.* 14, 1–19, 2019.
- M. Goodman *et al.*. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* 28, 1979.
- R. Guigó, I. Muchnik, and T. F. Smith. Reconstruction of ancient molecular phylogeny. *Mol. Phyl. Evol.* 6(2), 189–213, 1996.
- C. Chauve, Y. Ponty, and J. P. P. Zanetti. Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. *BMC Bioinform.* 16(19), 1–9, 2015.