Reconstructing the duplication history of tandemly repeated genes

Olivier Elemento (1,2), Olivier Gascuel(1), Marie-Paule Lefranc(2)

(1) LIRM Montpellier, Méthodes et Algorithmes pour l'Analyse de Séquences

(2) LIGM, IMGT the International ImMunoGeneTics Database, http ://imgt.cines.fr

1. Introduction

2. Mathematical model

3. Reconstructing duplication trees

4. Experimental results

5. Perspectives

1. Introduction

Tandemly repeated sequences

– two or more adjacent copies of a stretch of DNA

Tandemly repeated sequences

- two or more adjacent copies of a stretch of DNA
- they exist in several forms :
  - microsatellites (neurodegenerative diseases), minisatellites
  - larger sequences (genes)

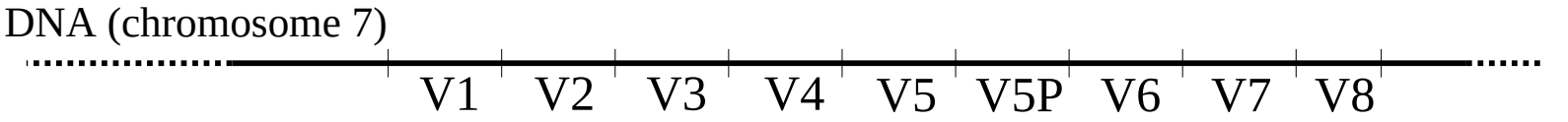
Tandemly repeated genes

Example : the human TRGV locus

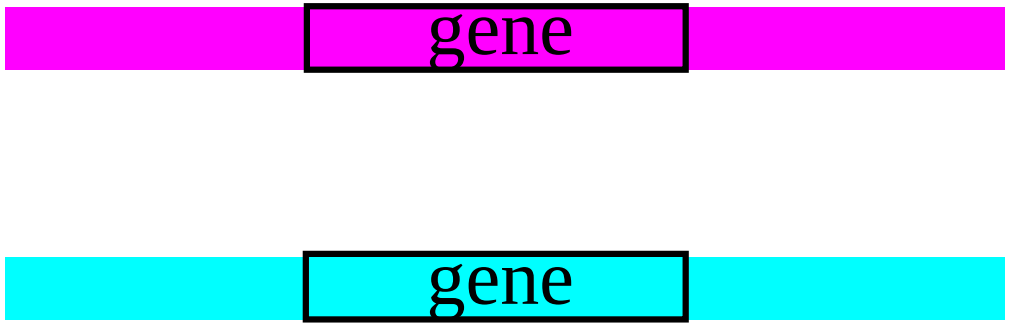
– it contains 9 adjacent copies of the same gene

– each of them is 4-5kb long

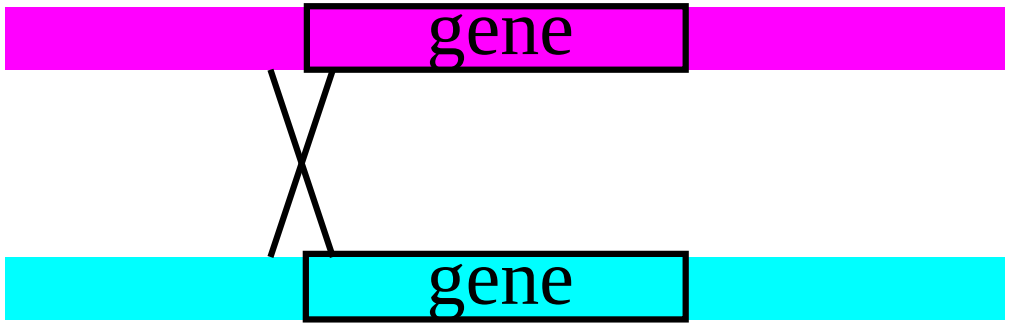
– they share between 85 and 97% of identity



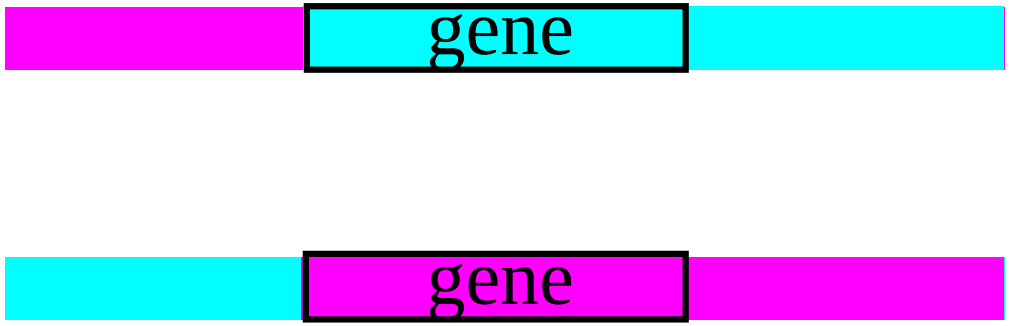
Recombination



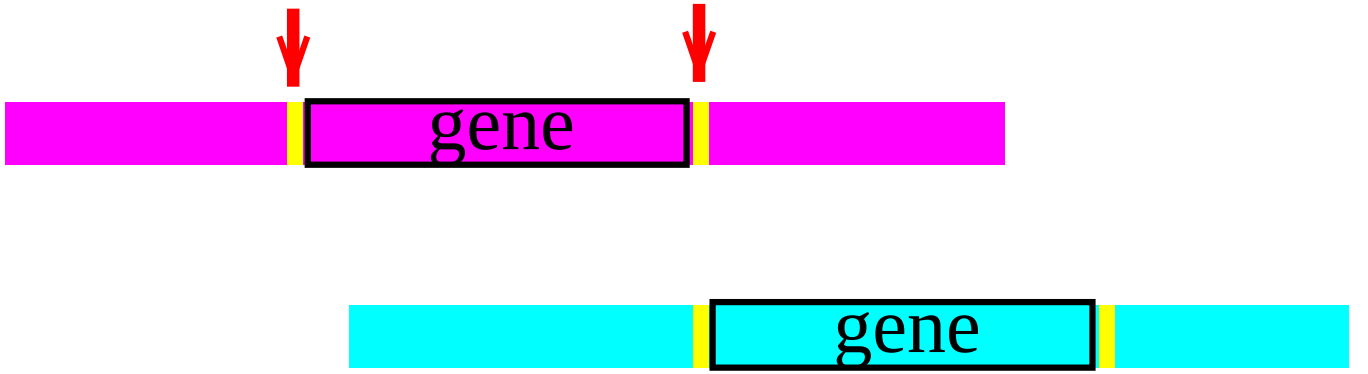
Recombination



Recombination

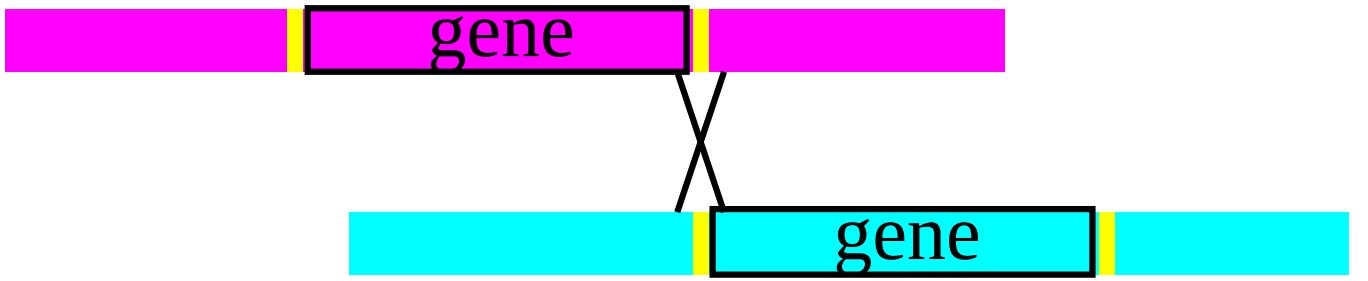


Unequal recombination (step 1)



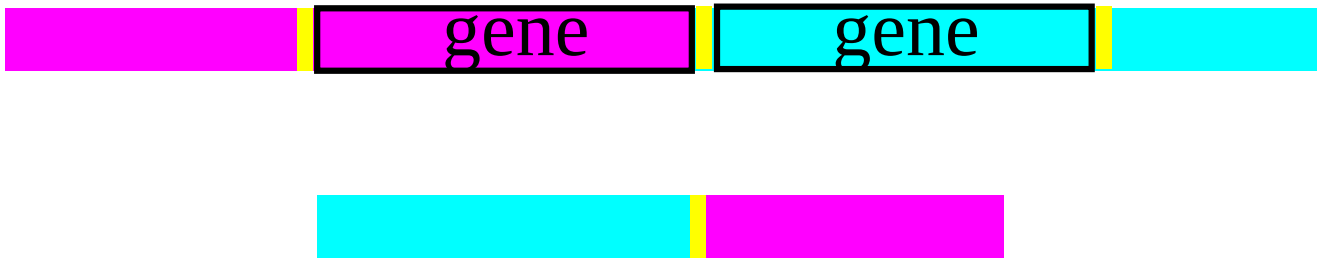
initial duplication caused by the presence of short repeated sequences

Unequal recombination (step 1)



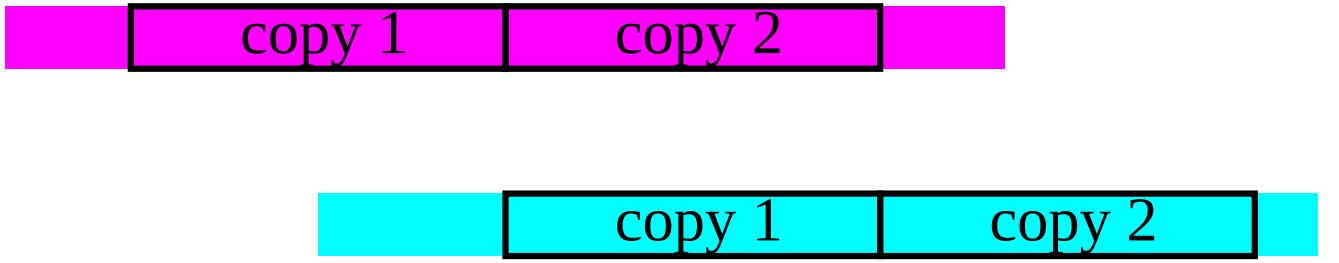
initial duplication caused by the presence of short repeated sequences

Unequal recombination (step 1)



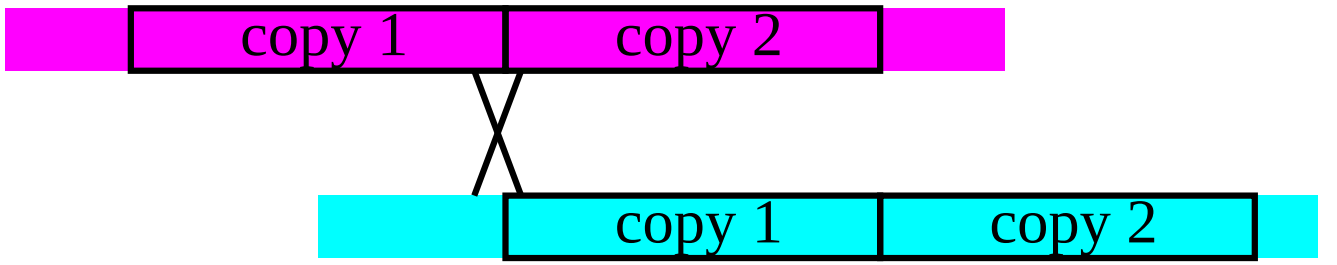
initial duplication caused by the presence of short repeated sequences

Unequal recombination (step 2)



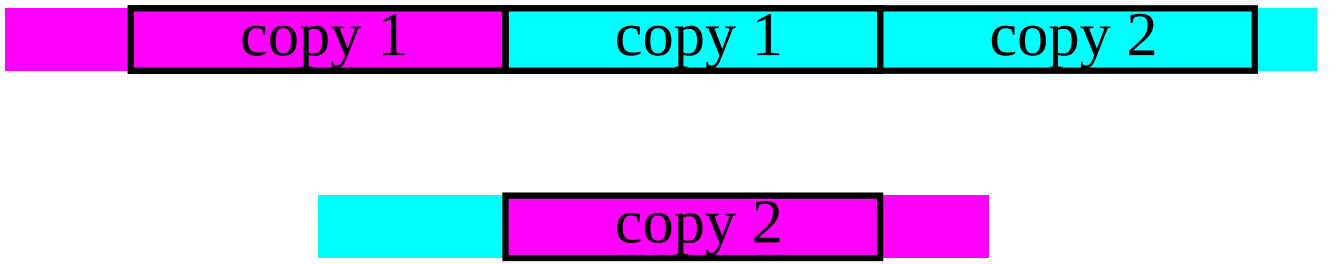
the presence of several times the same copy favors additional duplications

Unequal recombination (step 2)



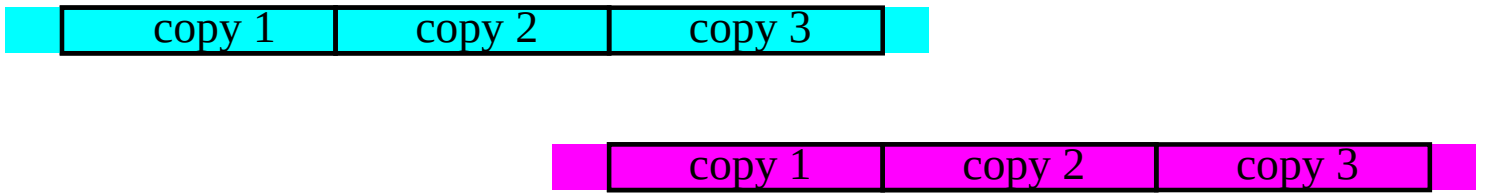
the presence of several times the same copy favors additional duplications

Unequal recombination (step 2)



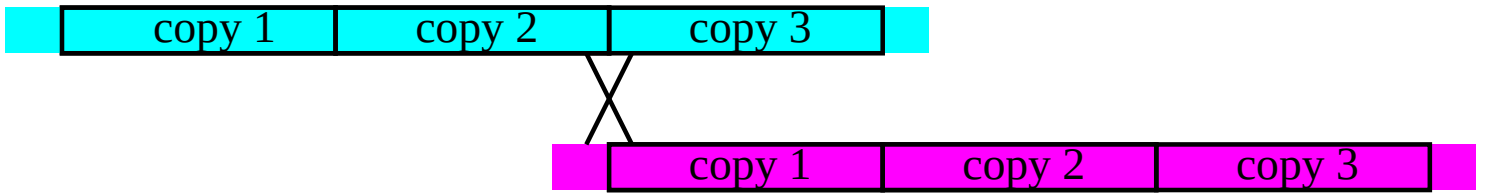
the presence of several times the same copy favors additional duplications

Unequal recombination (step 3)



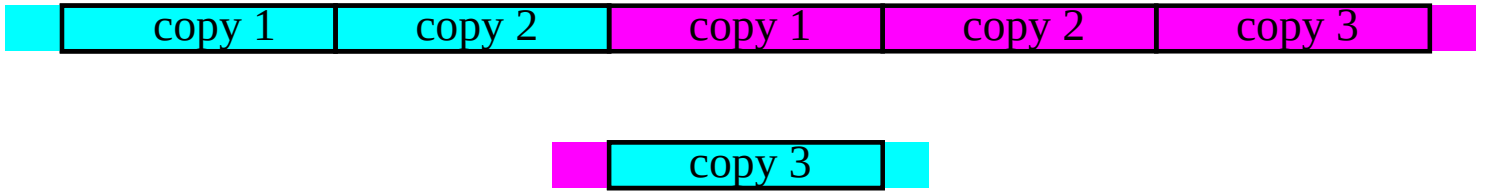
“block” duplication, i.e. simultaneous duplication of several copies

Unequal recombination (step 3)



“block” duplication, i.e. simultaneous duplication of several copies

Unequal recombination (step 3)



“block” duplication, i.e. simultaneous duplication of several copies

Preliminary hypothesis

– unequal recombination is the sole generating mechanism

Preliminary hypothesis

– unequal recombination is the sole generating mechanism

– there was no gene conversions

Preliminary hypothesis

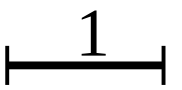
– unequal recombination is the sole generating mechanism

– there was no gene conversions

– there was “no gene deletions”

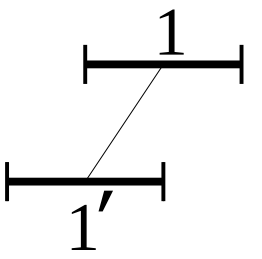
The duplication events

1-duplication



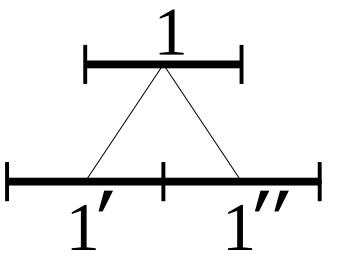
The duplication events

1-duplication



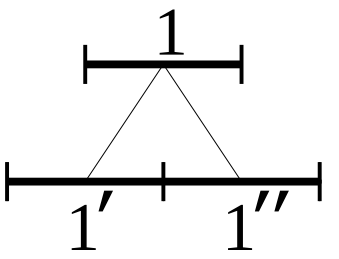
The duplication events

1-duplication

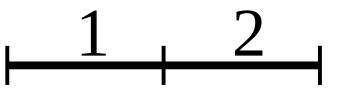


The duplication events

1-duplication

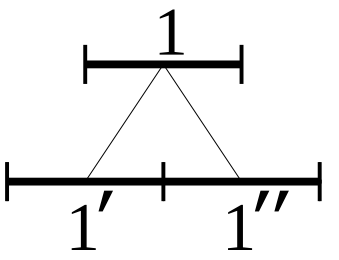


2-duplication

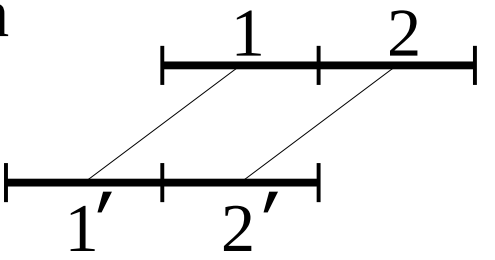


The duplication events

1-duplication

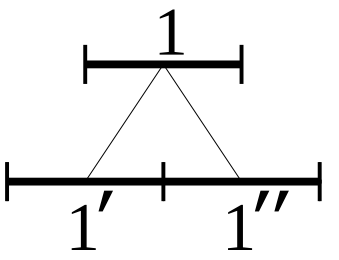


2-duplication

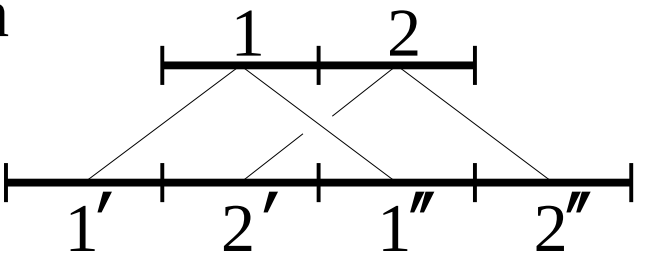


The duplication events

1-duplication

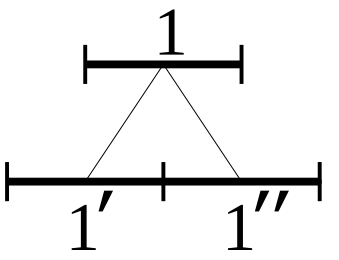


2-duplication

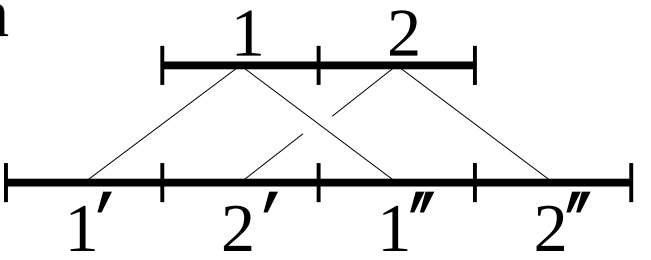


The duplication events

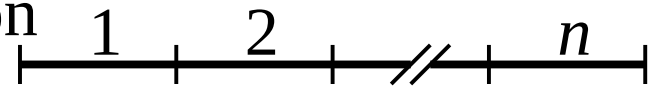
1-duplication



2-duplication

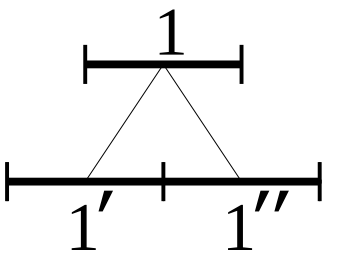


n-duplication

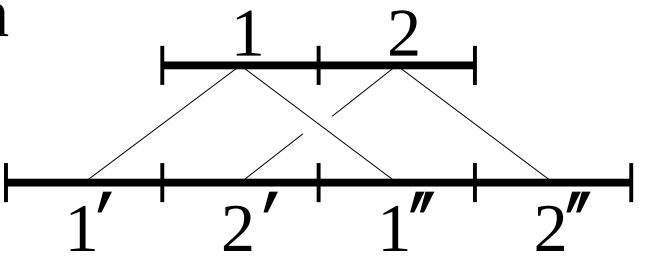


The duplication events

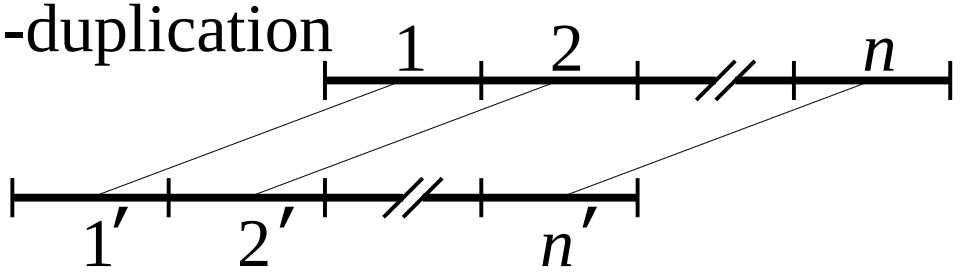
1-duplication



2-duplication

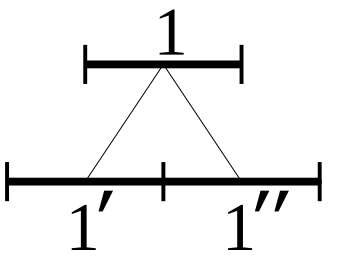


n-duplication

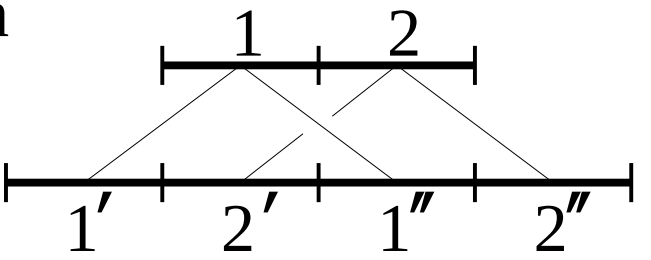


The duplication events

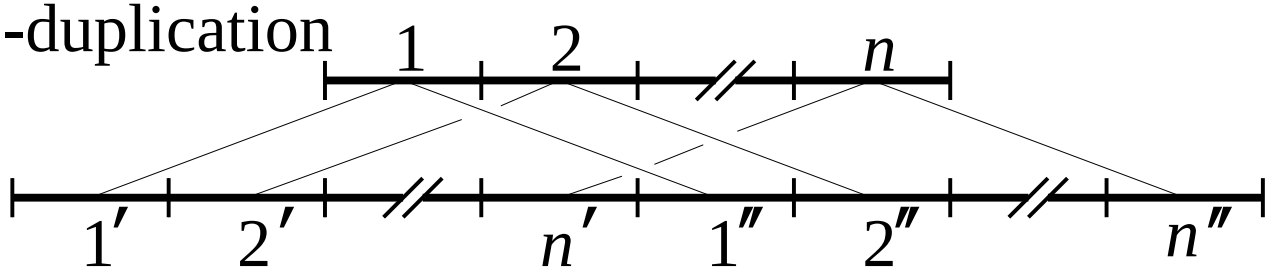
1-duplication



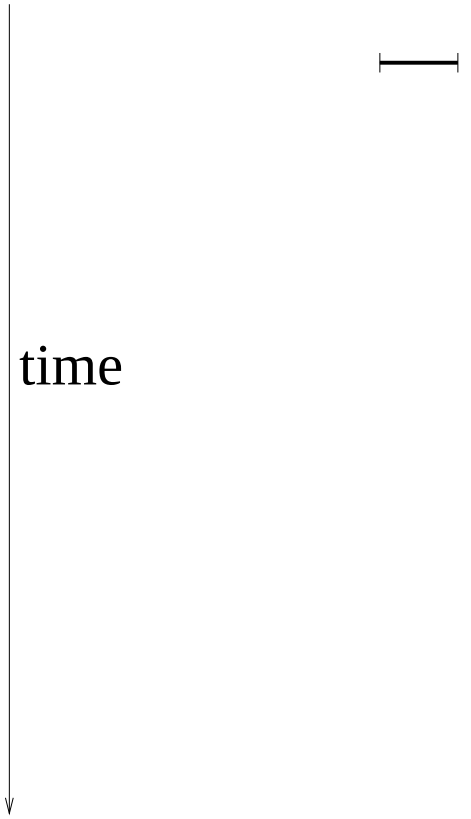
2-duplication



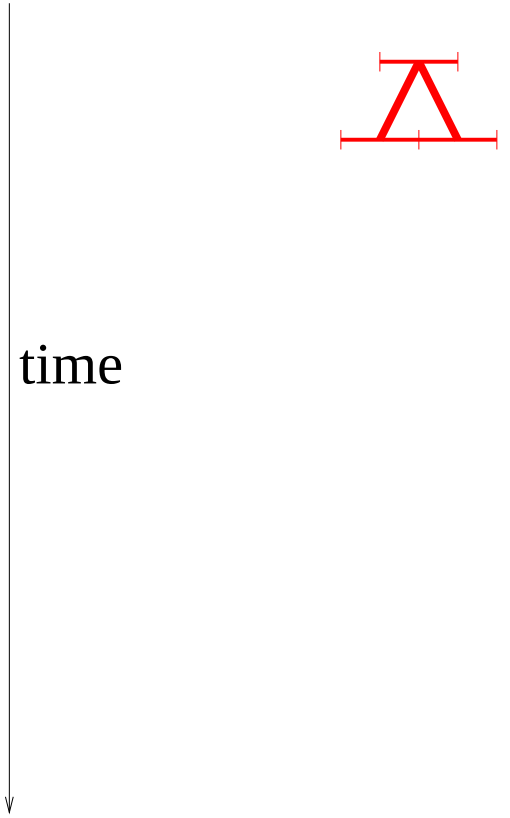
n-duplication



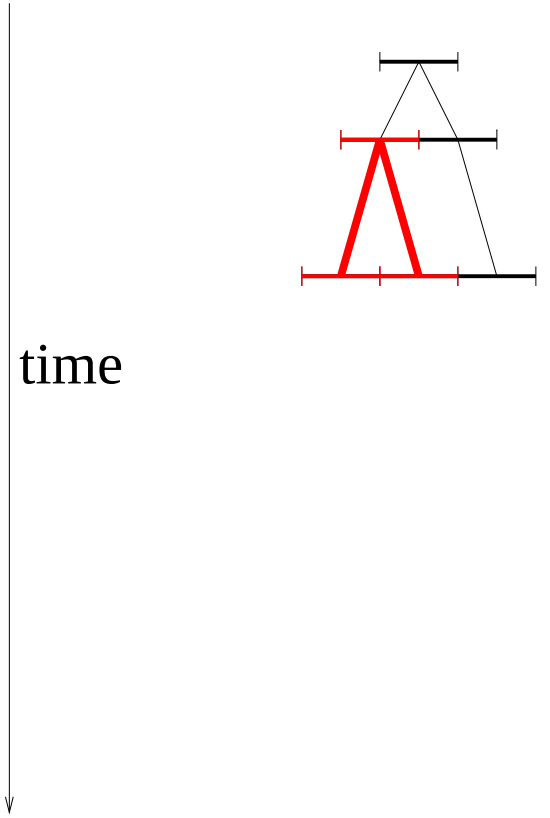
Time valued duplication history (reality)



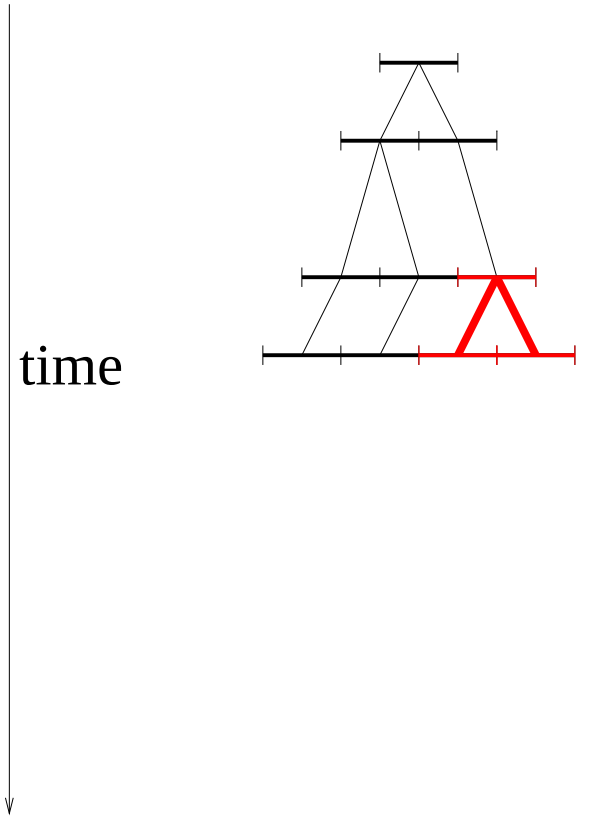
Time valued duplication history (reality)



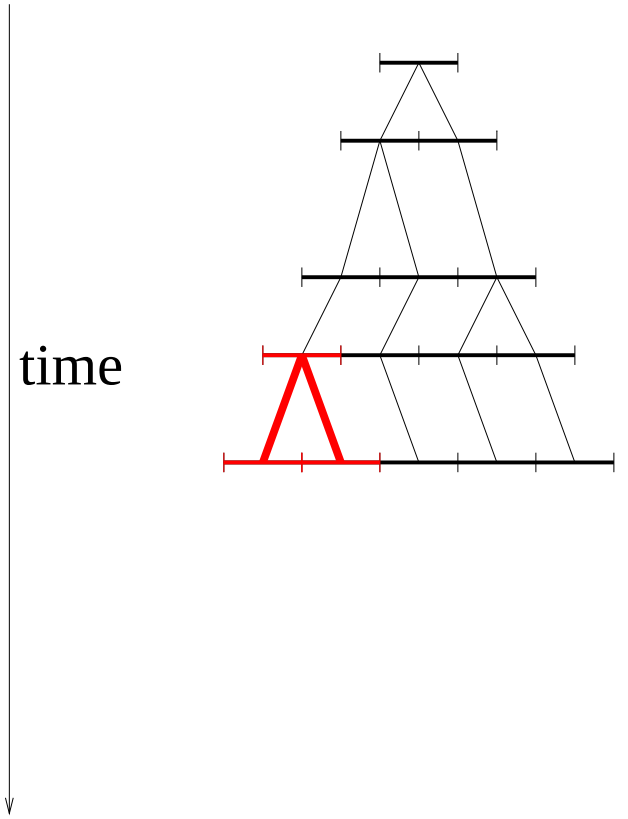
Time valued duplication history (reality)



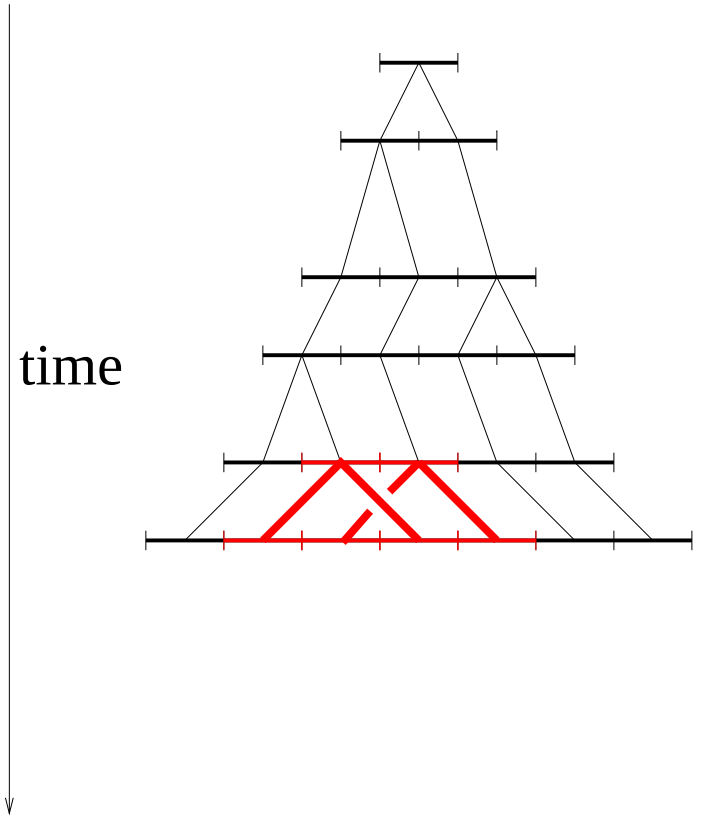
Time valued duplication history (reality)



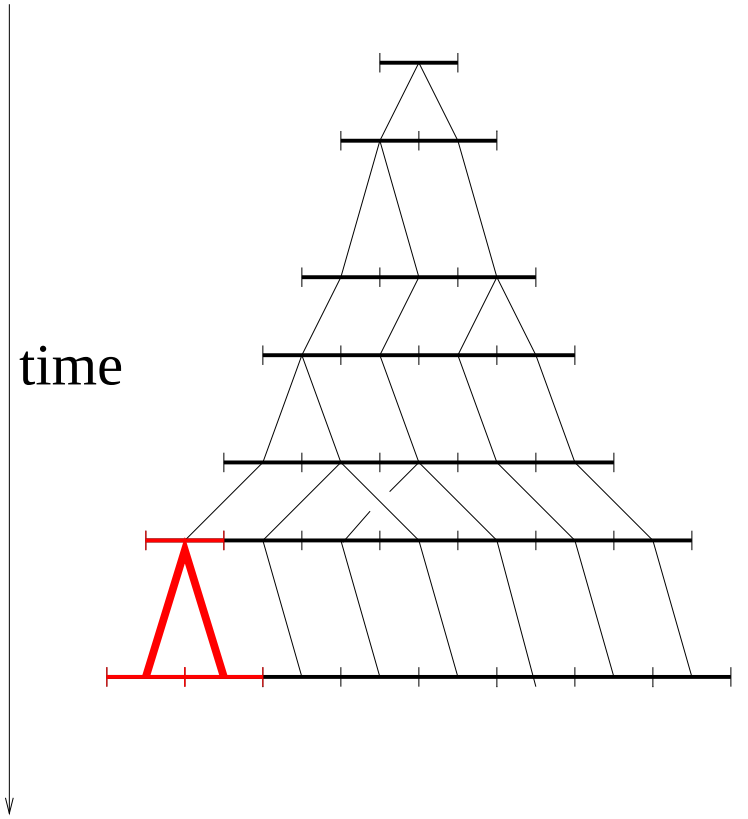
Time valued duplication history (reality)



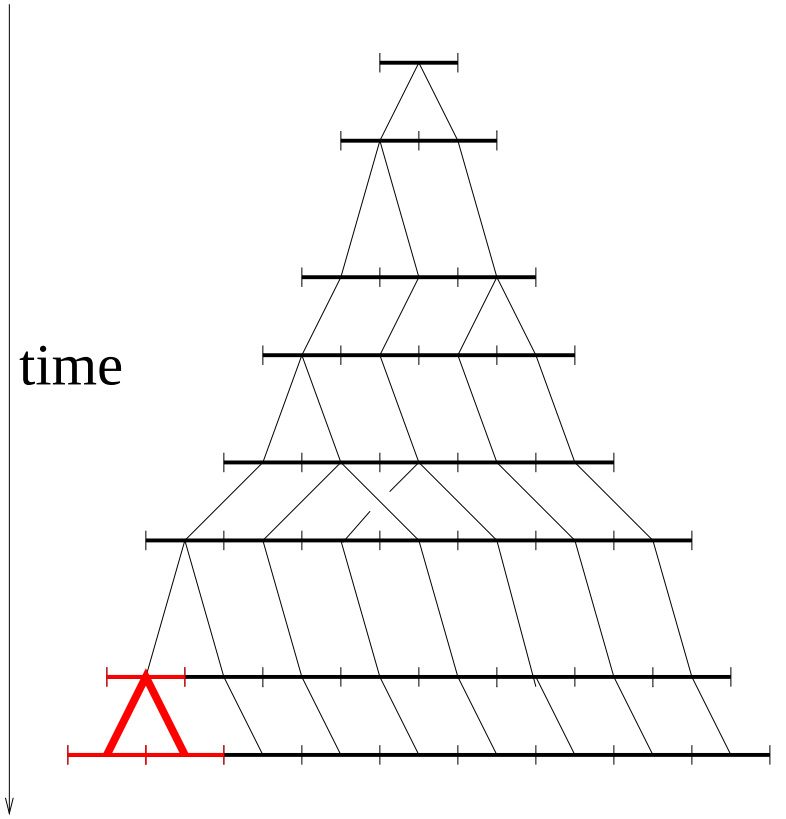
Time valued duplication history (reality)



Time valued duplication history (reality)

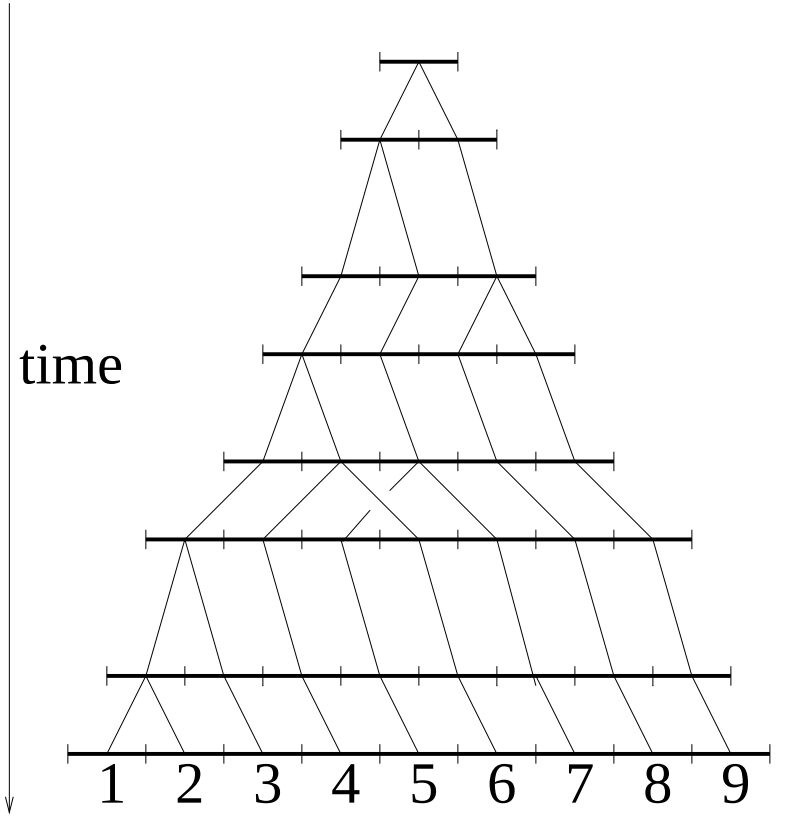


Time valued duplication history (reality)

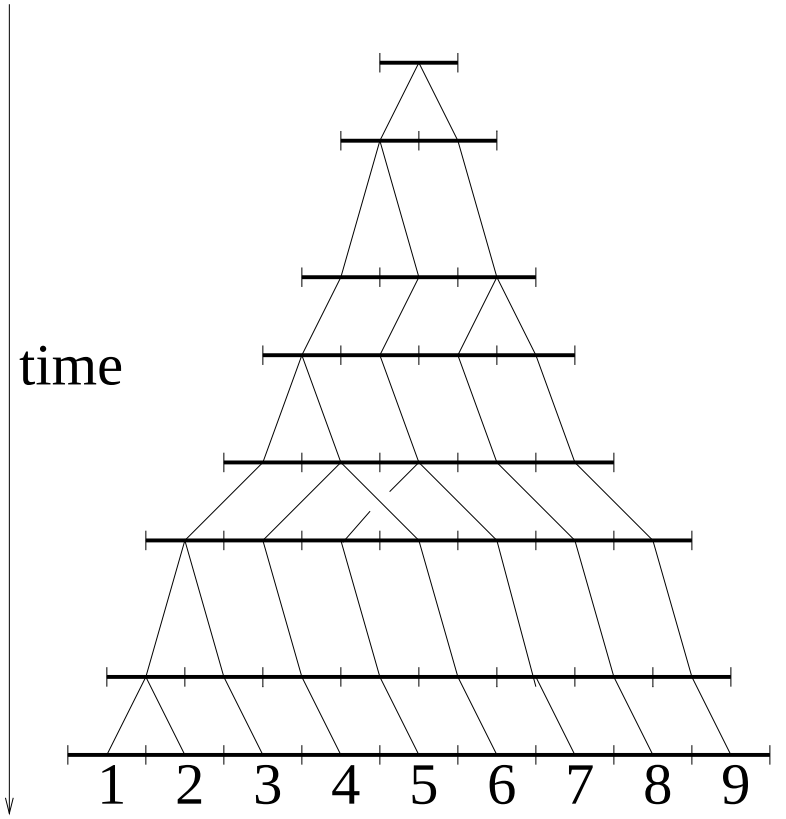


Time valued duplication history (reality)

– it implies a rooted phylogeny

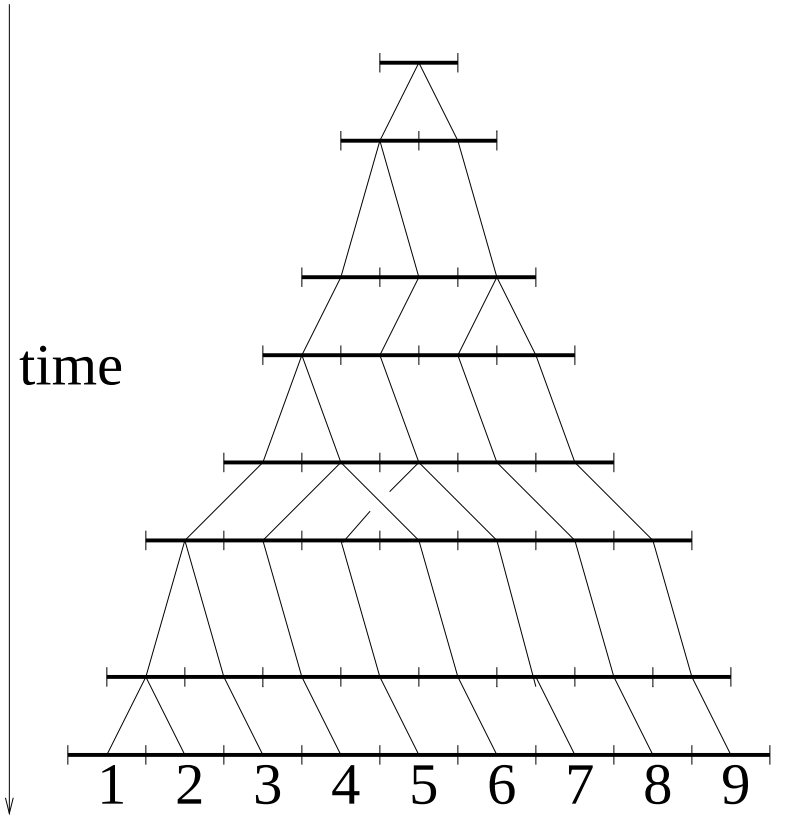


Time valued duplication history (reality)



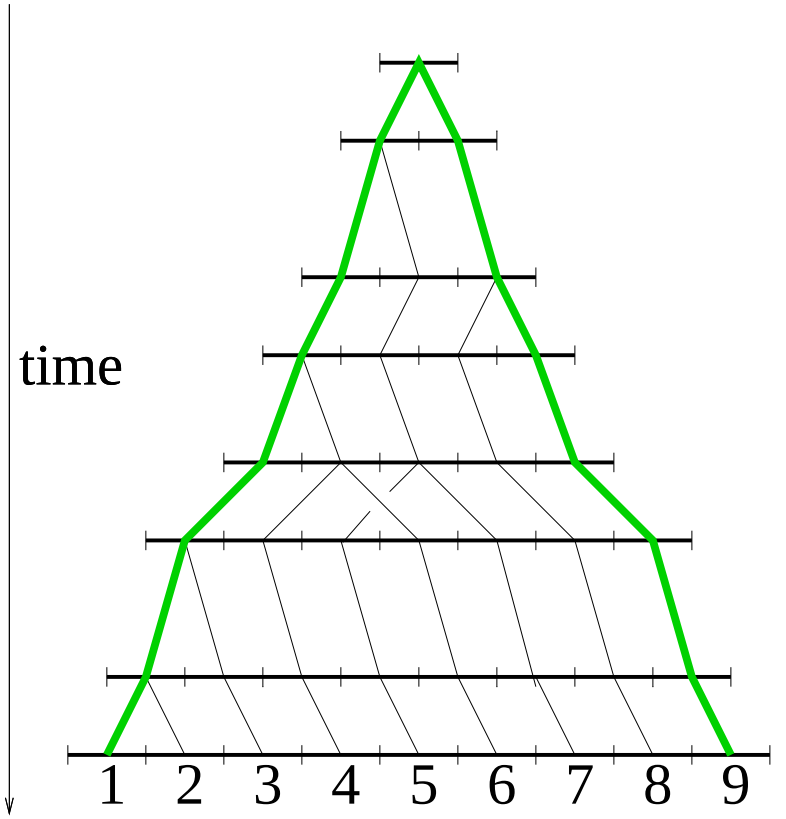
- it implies a rooted phylogeny
- its taxa are ordered

Time valued duplication history (reality)



- it implies a rooted phylogeny
- its taxa are ordered
- its branches are time valued

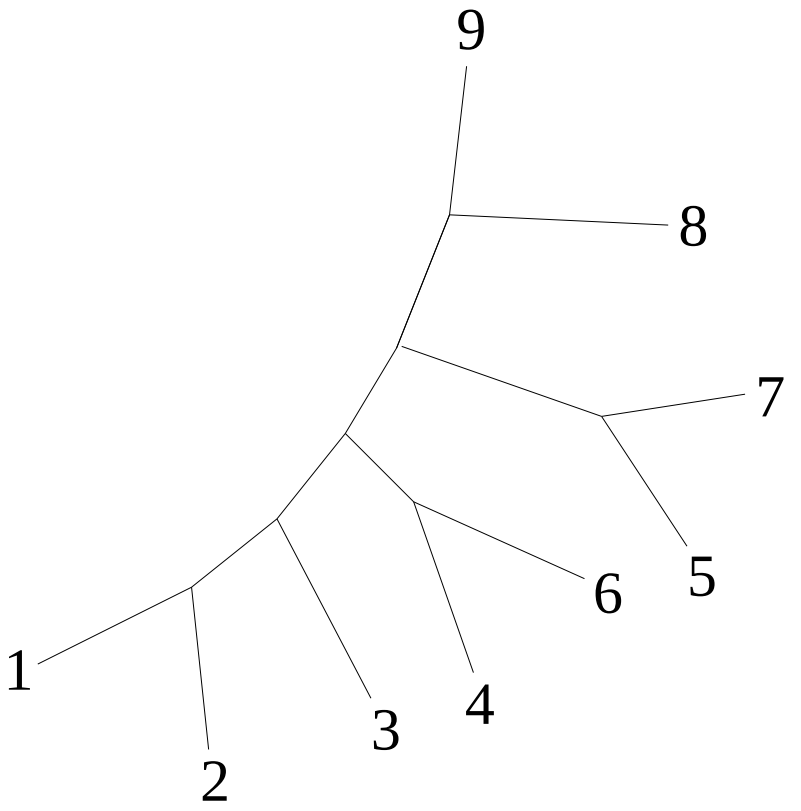
Time valued duplication history (reality)



- it implies a rooted phylogeny
- its taxa are ordered
- its branches are time valued
- the root is situated between the most distant taxa

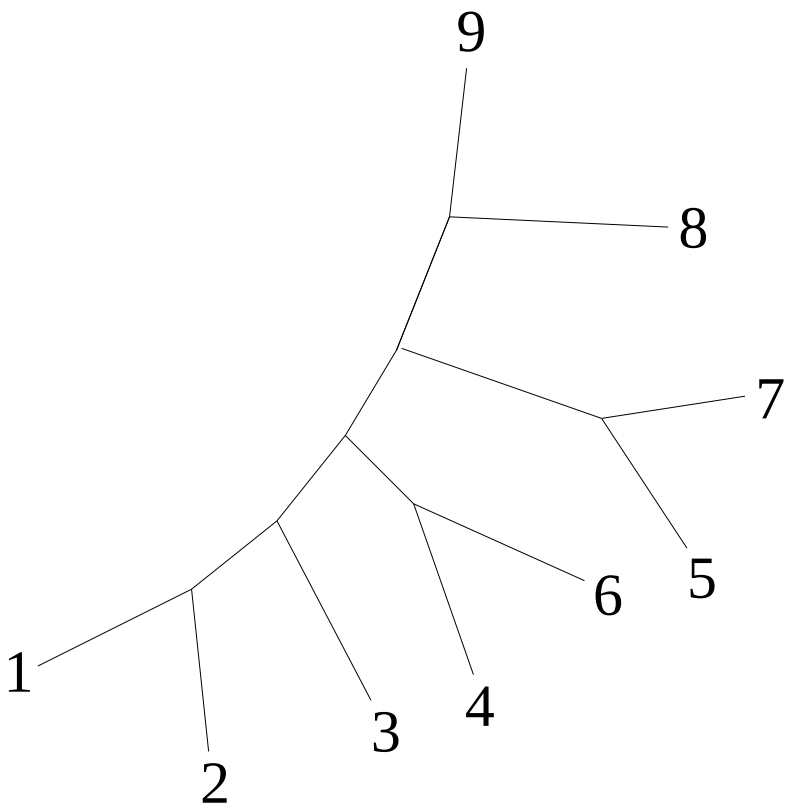
Duplication tree (what can be infered)

– it is an unrooted phylogeny

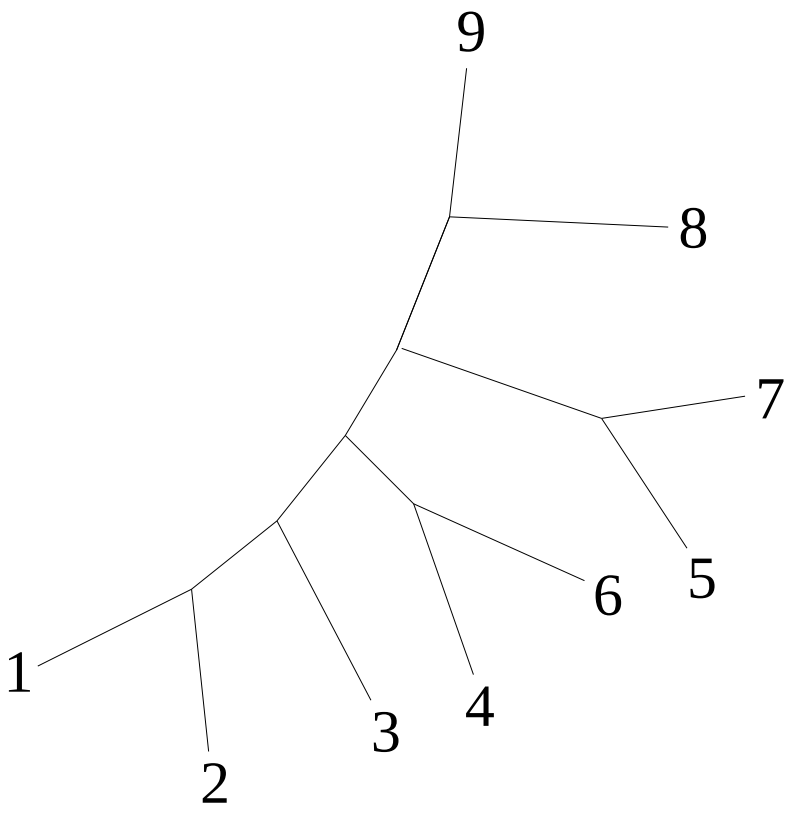


Duplication tree (what can be infered)

- it is an unrooted phylogeny
- its taxa are ordered

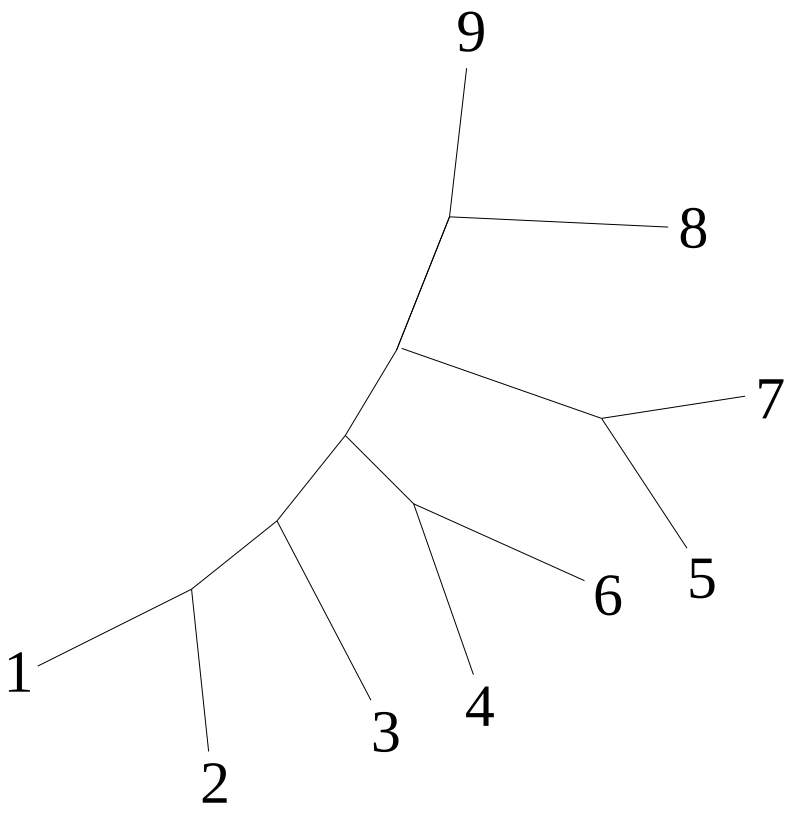


Duplication tree (what can be infered)



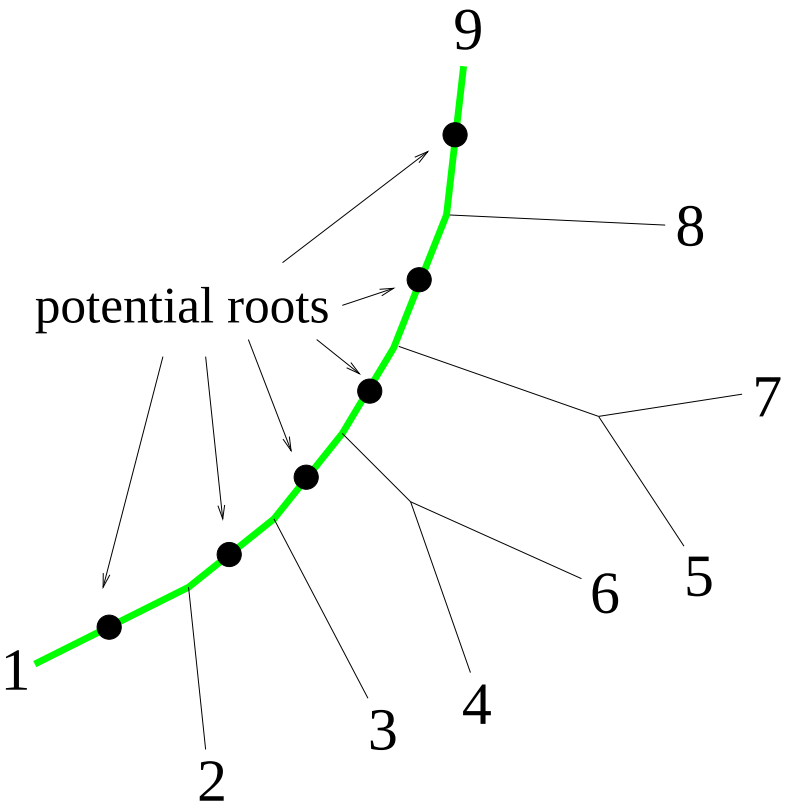
- it is an unrooted phylogeny
- its taxa are ordered
- its branches are mutation rate-valued

Duplication tree (what can be infered)



- it is an unrooted phylogeny
- its taxa are ordered
- its branches are mutation rate-valued
- its topology is compatible with at least one duplication history

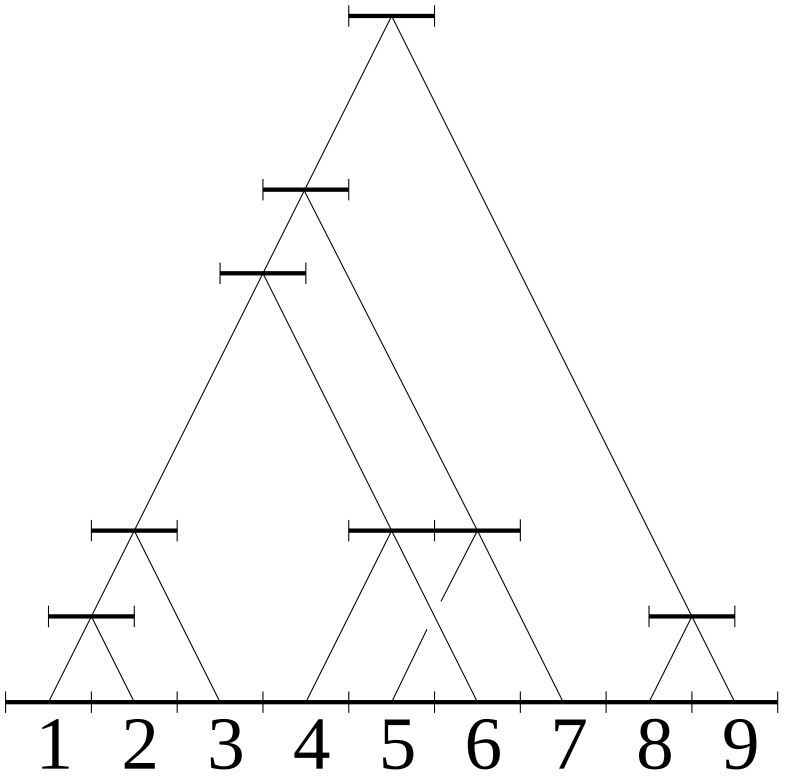
Duplication tree (what can be infered)



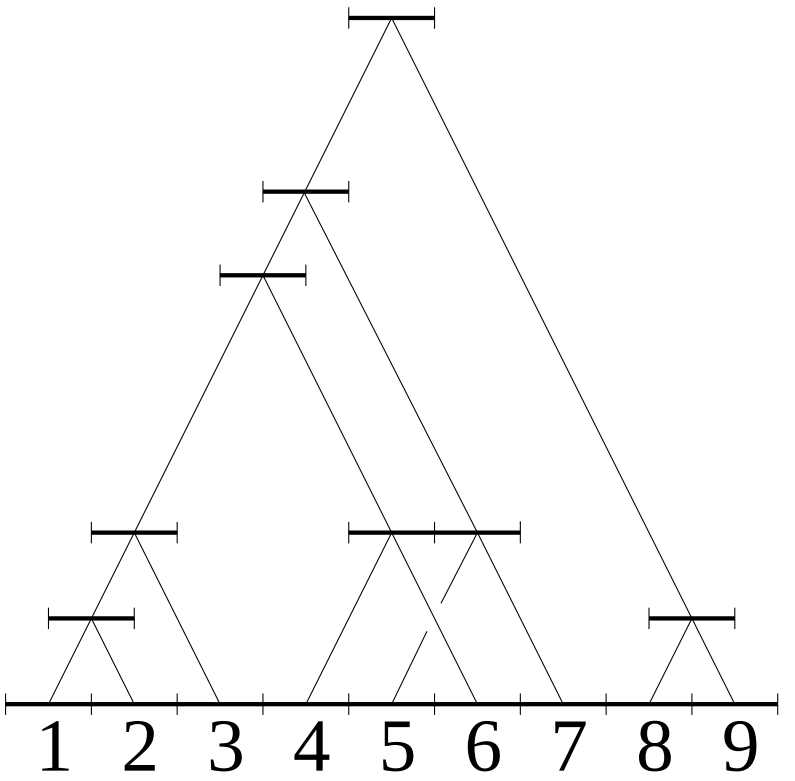
- it is an unrooted phylogeny
- its taxa are ordered
- its branches are mutation rate-valued
- its topology is compatible with at least one duplication history
- the root is situated somewhere in the tree between the most distant taxa

Ordinal duplication history

– obtained when rooting a duplication tree

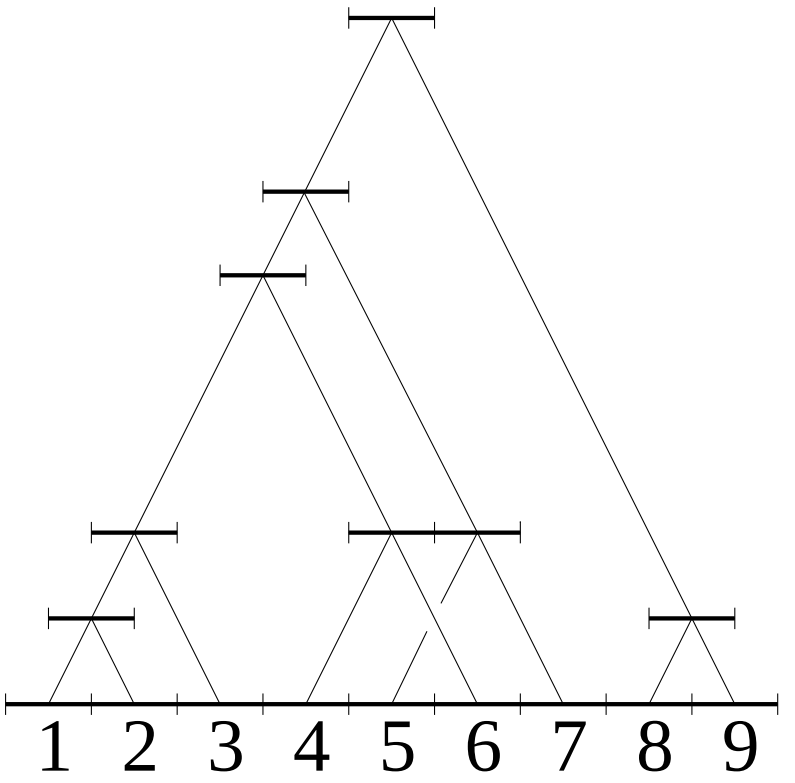


Ordinal duplication history



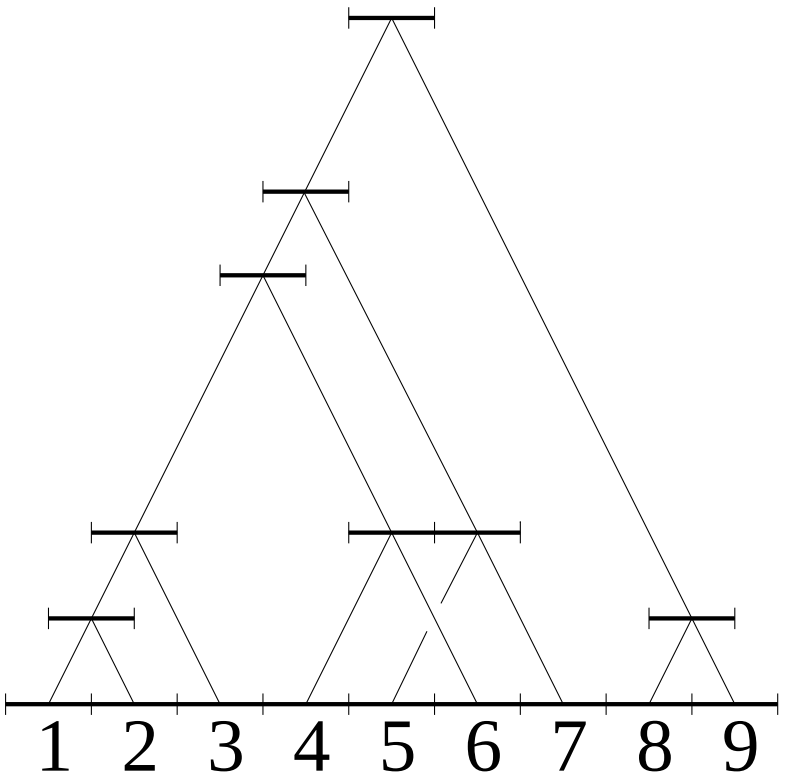
- obtained when rooting a duplication tree
- it is the topological version of the time valued duplication history

Ordinal duplication history



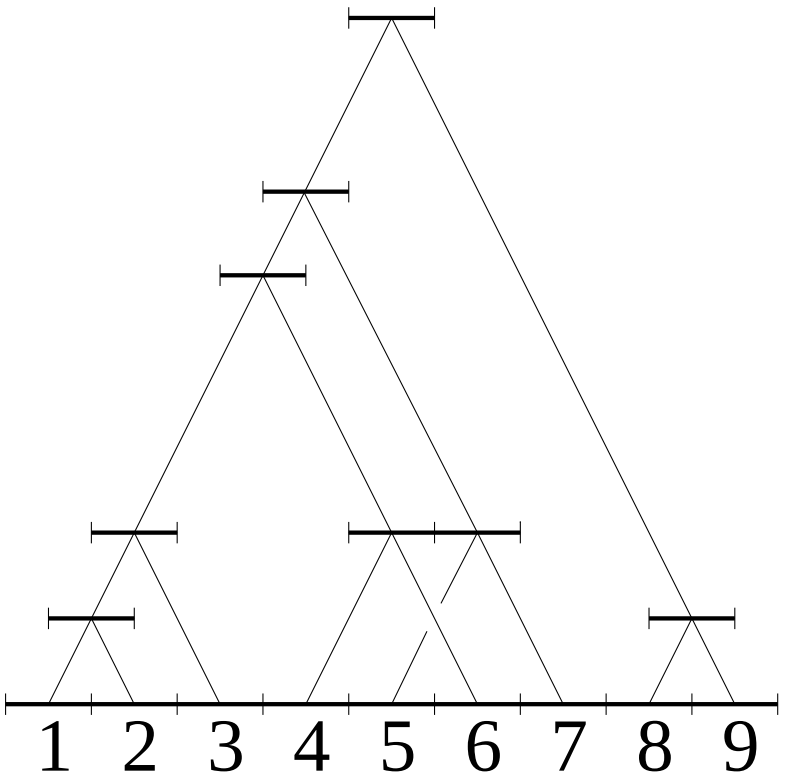
- obtained when rooting a duplication tree
- it is the topological version of the time valued duplication history
- it is a rooted phylogeny

Ordinal duplication history



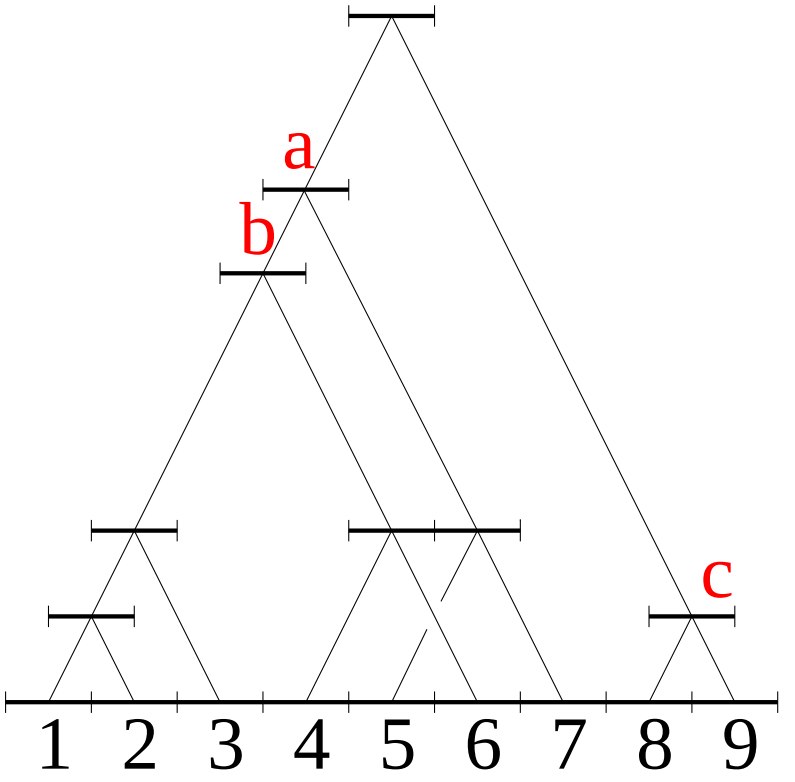
- obtained when rooting a duplication tree
- it is the topological version of the time valued duplication history
- it is a rooted phylogeny
- its taxa are ordered

Ordinal duplication history



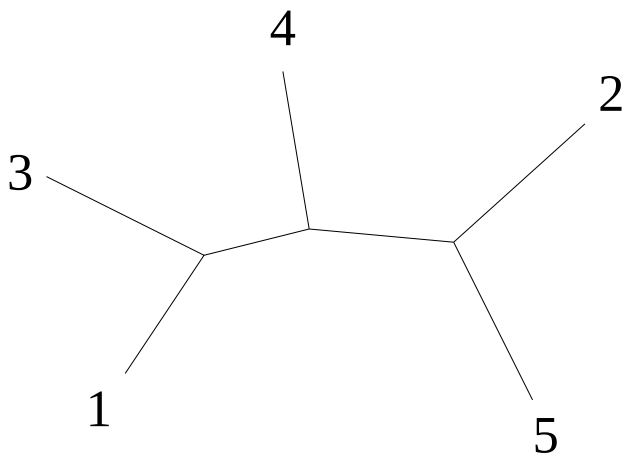
- obtained when rooting a duplication tree
- it is the topological version of the time valued duplication history
- it is a rooted phylogeny
- its taxa are ordered
- its branch lengths have no special meaning

Ordinal duplication history

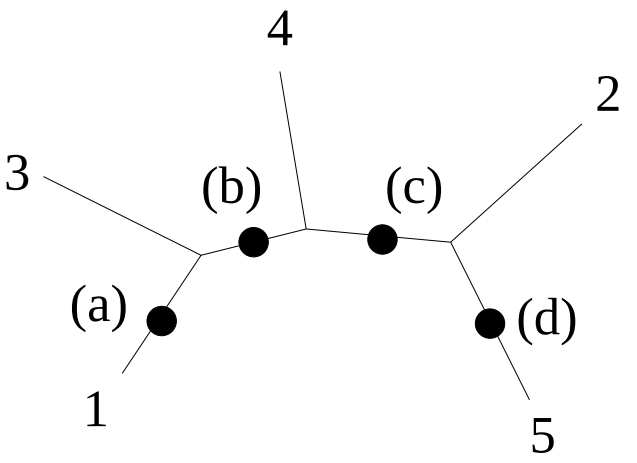


- obtained when rooting a duplication tree
- it is the topological version of the time valued duplication history
- it is a rooted phylogeny
- its taxa are ordered
- its branch lengths have no special meaning
- the duplication events are partially ordered

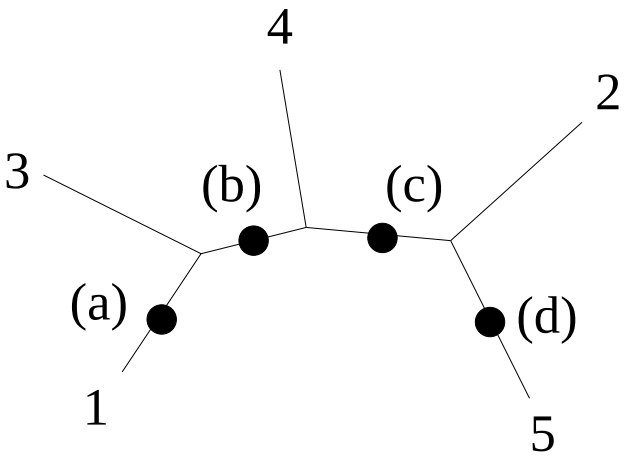
Not all phylogenies are duplication trees

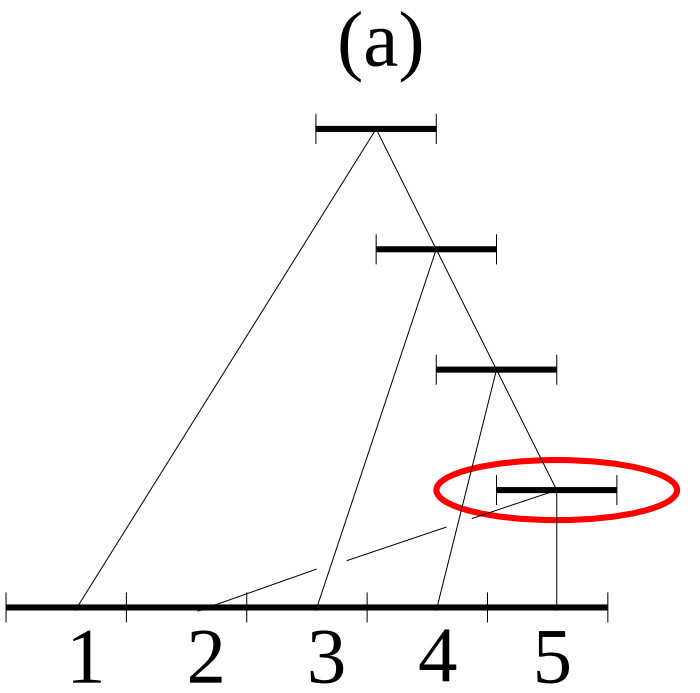


Not all phylogenies are duplication trees



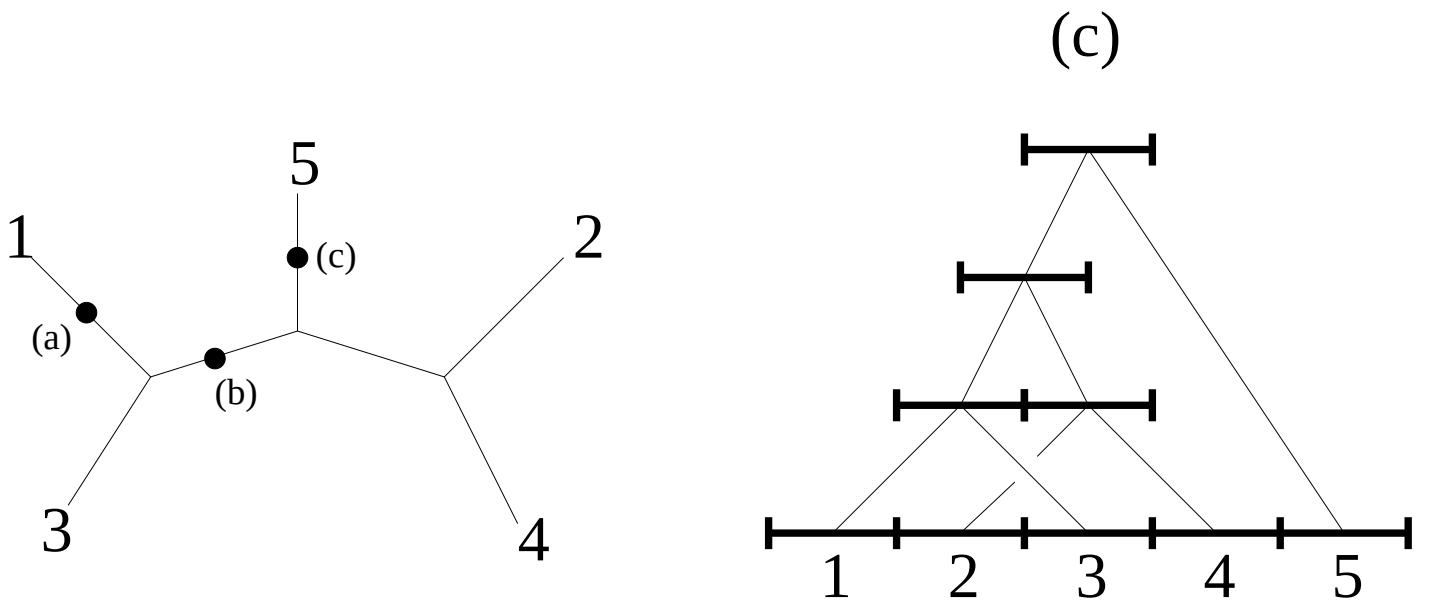
Not all phylogenies are duplication trees





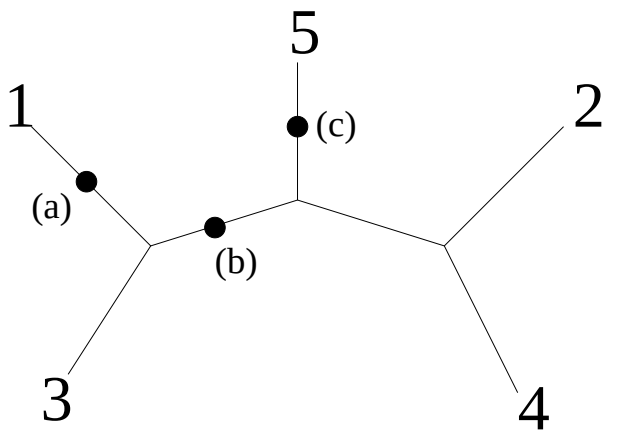
2 and 5 are not adjacent !

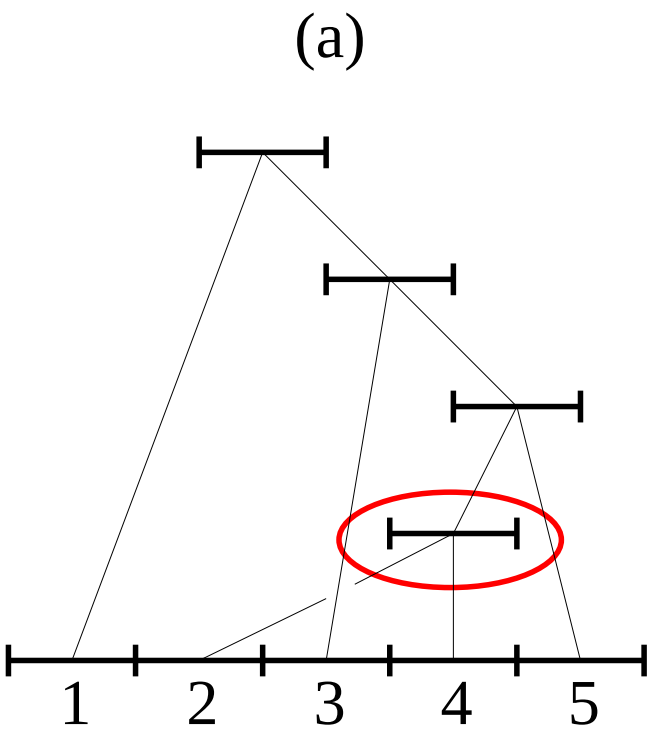
Not all potential roots lead to correct ordinal duplication histories



correct ordinal duplication history

Not all potential roots lead to correct ordinal duplication histories





incorrect ordinal duplication history

Definition

A phylogeny is a duplication tree if, among its potential roots, at least one of them leads to a correct ordinal duplication history

The PDT algorithm

– it takes as input a rooted phylogeny with ordered leaves

The PDT algorithm

– it takes as input a rooted phylogeny with ordered leaves

– it recursively agglomerates each terminal pair belonging to correct duplication events

The PDT algorithm

– it takes as input a rooted phylogeny with ordered leaves

– it recursively agglomerates each terminal pair belonging to correct duplication events

– it stops and returns :
(true) when the tree has been reduced to its root
(false) when it cannot go further

The PDT algorithm

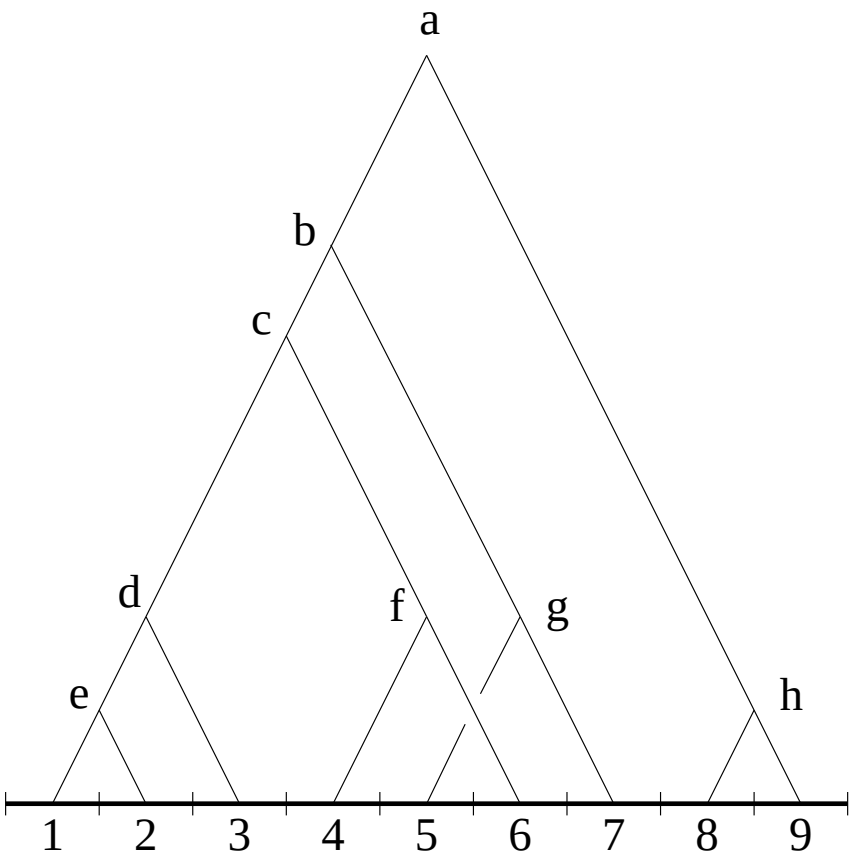
– we apply the PDT algorithm to each potential root of the considered phylogeny

The PDT algorithm

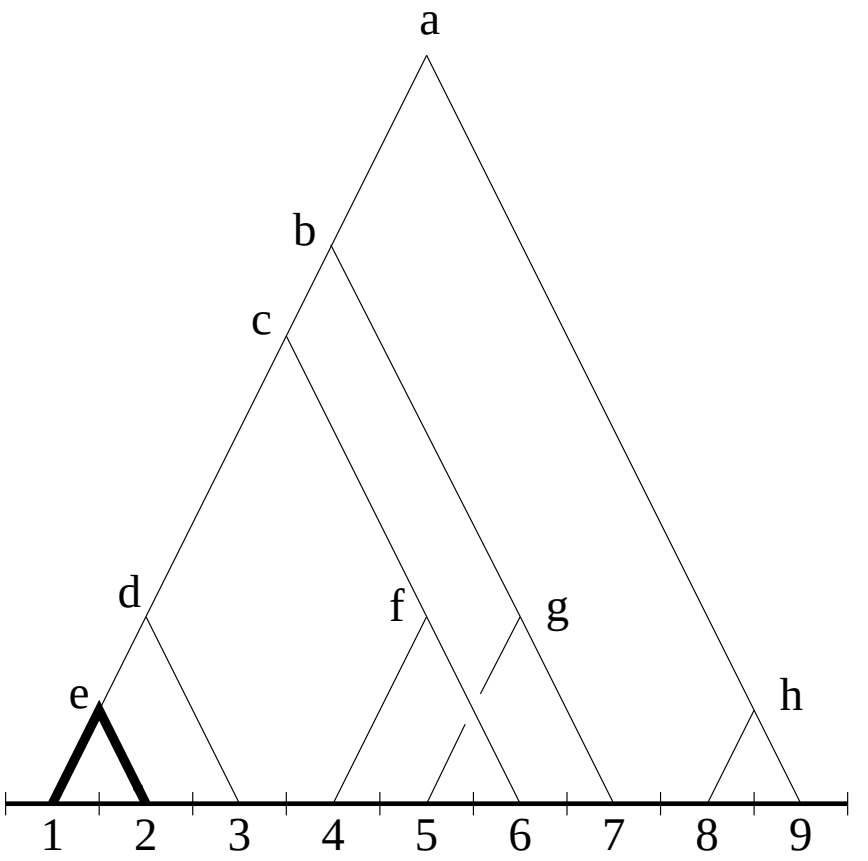
– we apply the PDT algorithm to each potential root of the considered phylogeny

– if PDT return “true” at least once, the phylogeny is a duplication tree

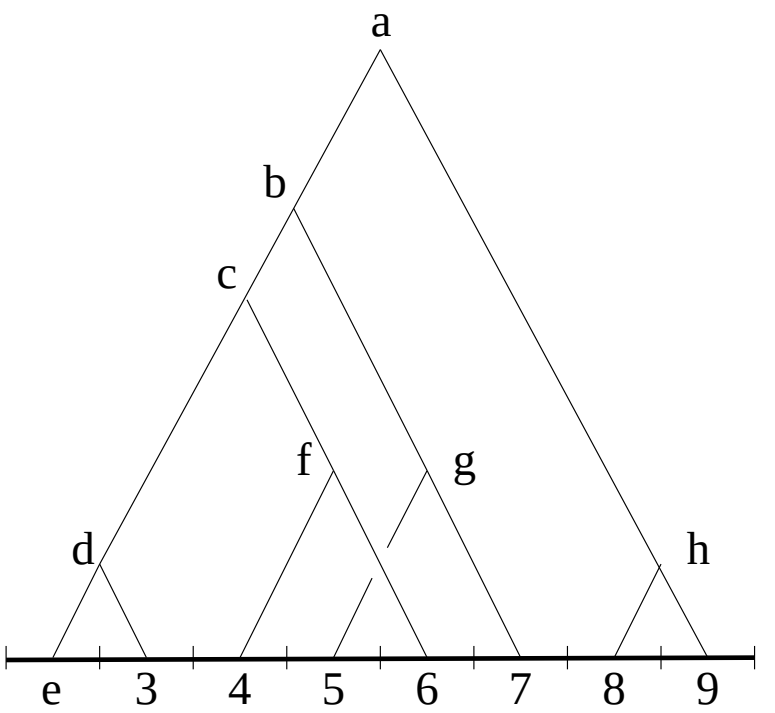
The PDT algorithm



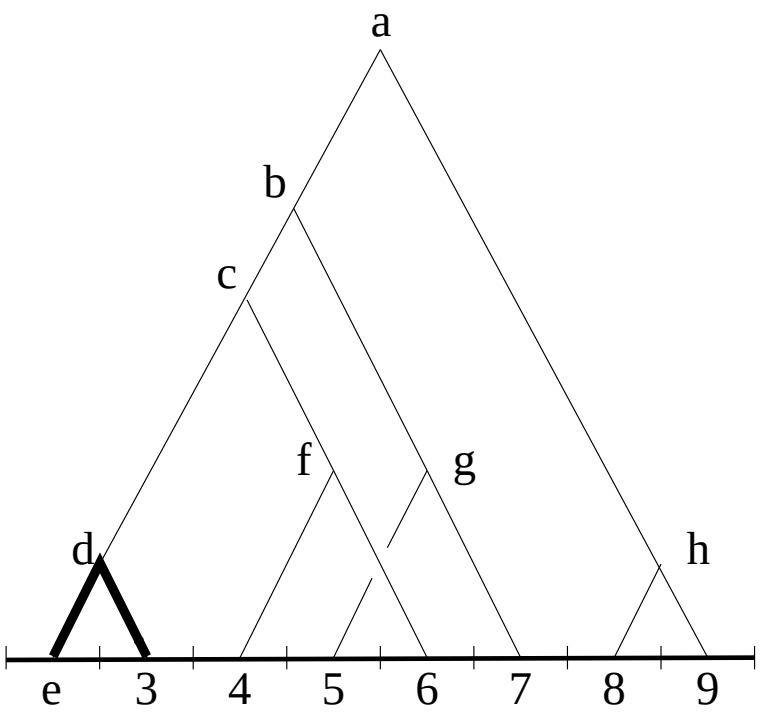
The PDT algorithm



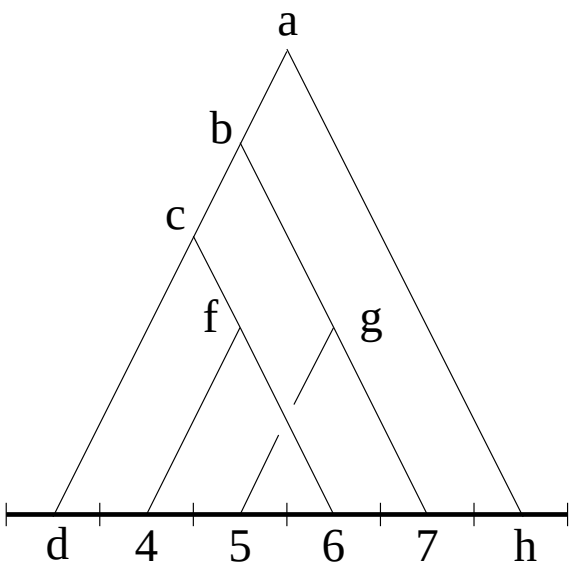
The PDT algorithm



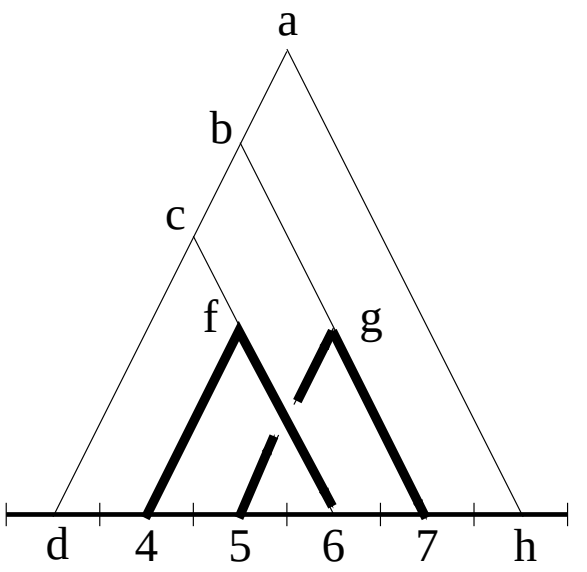
The PDT algorithm



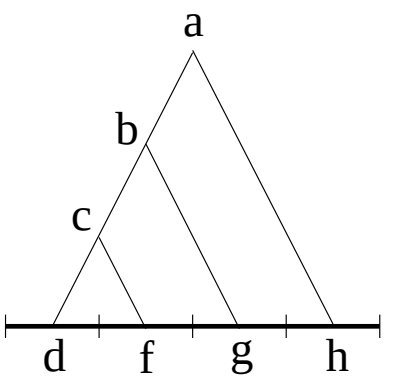
The PDT algorithm



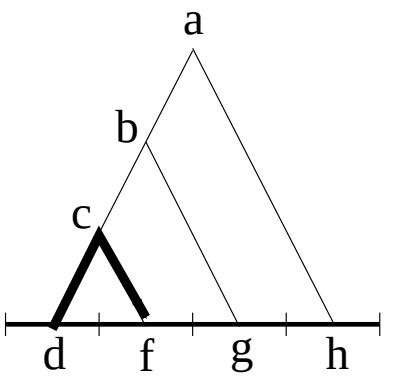
The PDT algorithm



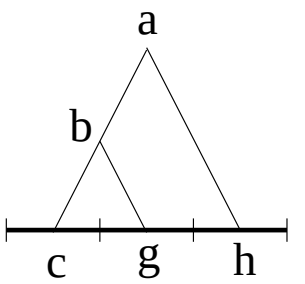
The PDT algorithm



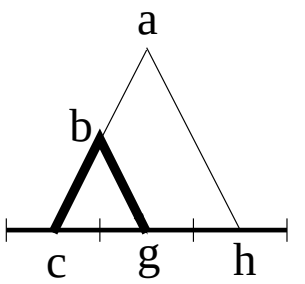
The PDT algorithm



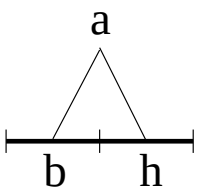
The PDT algorithm



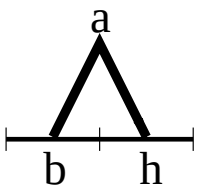
The PDT algorithm



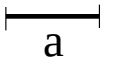
The PDT algorithm



The PDT algorithm

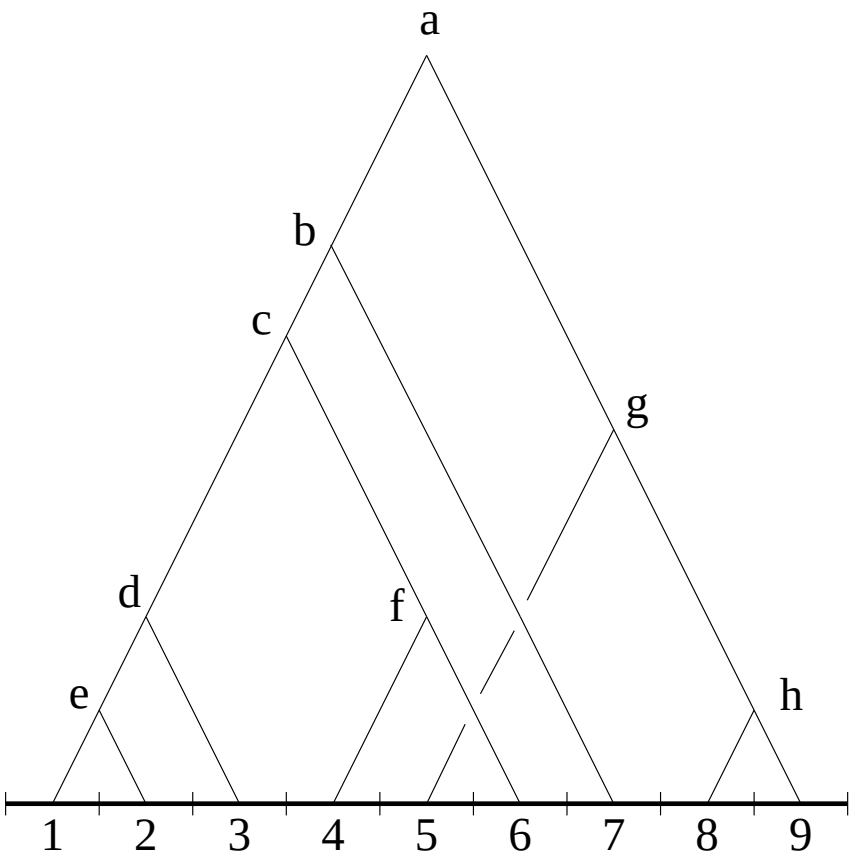


The PDT algorithm

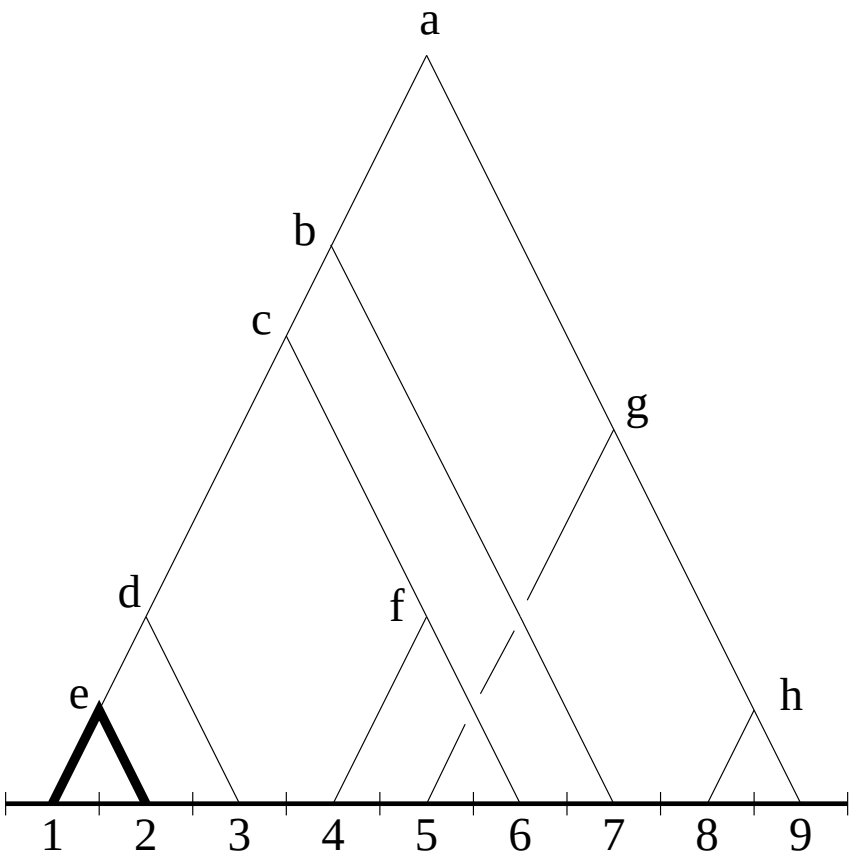


true!

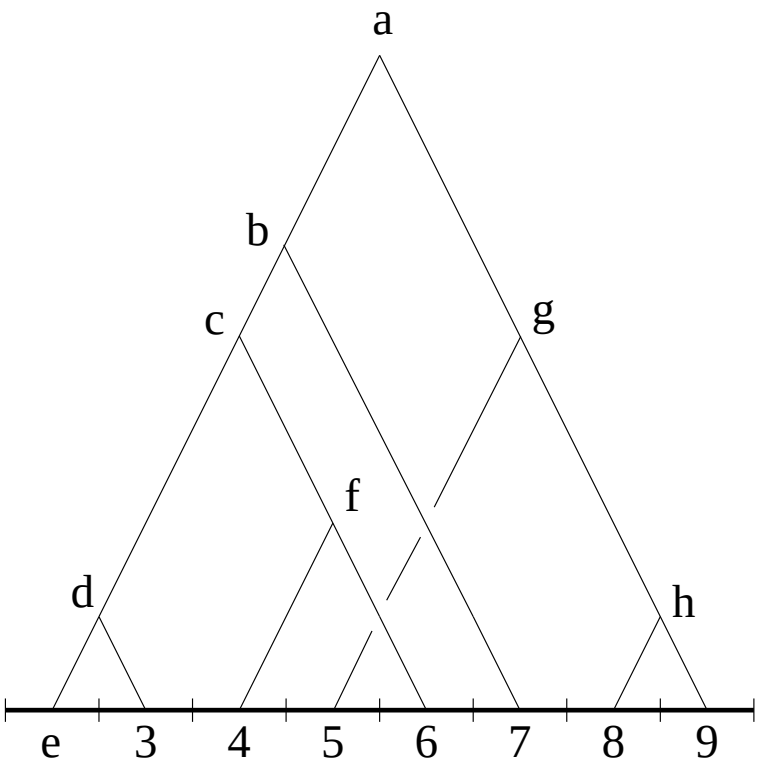
The PDT algorithm



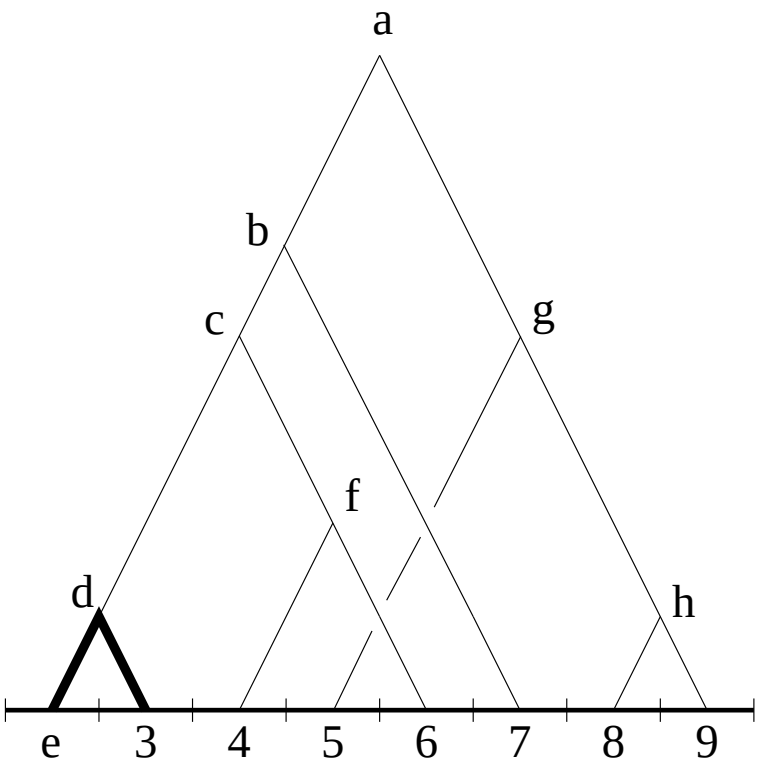
The PDT algorithm



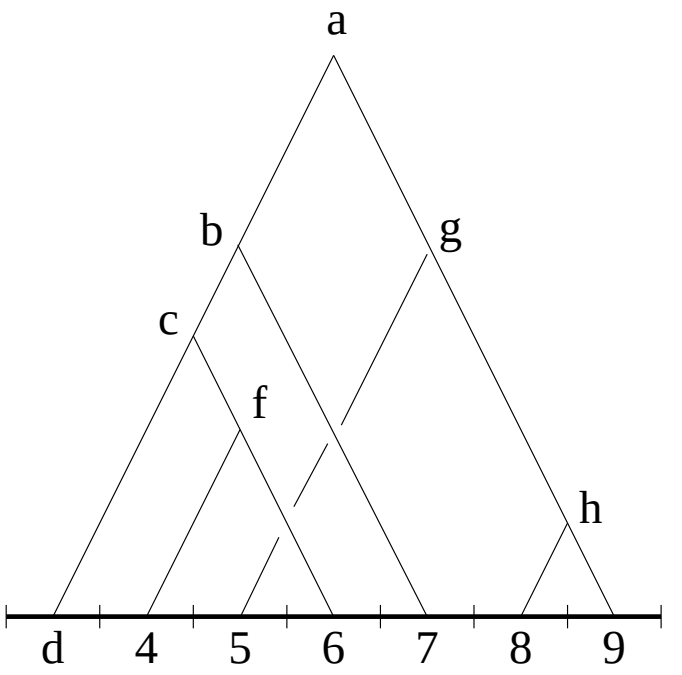
The PDT algorithm



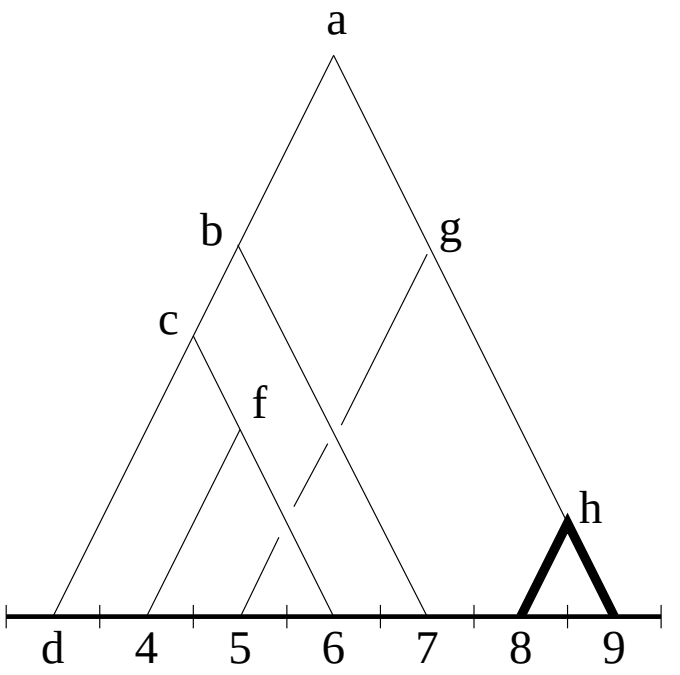
The PDT algorithm



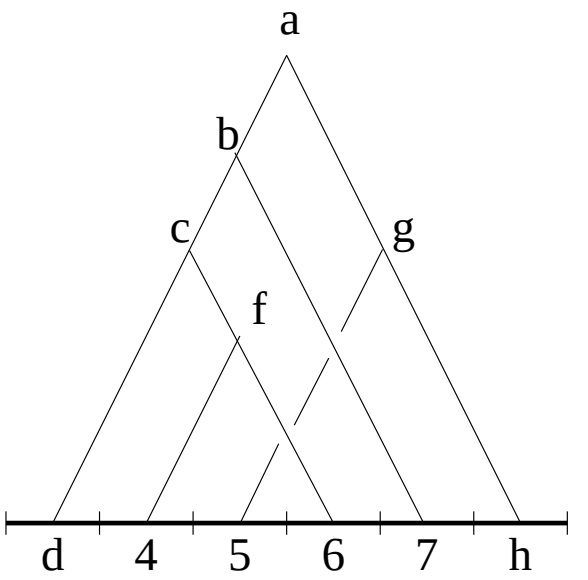
The PDT algorithm



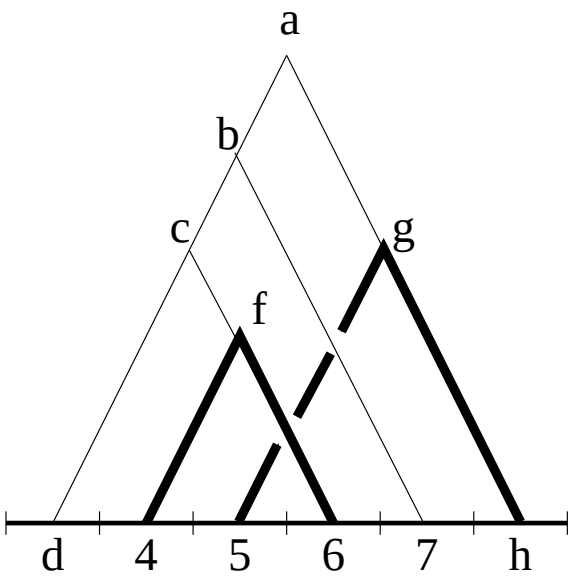
The PDT algorithm



The PDT algorithm



The PDT algorithm



false!

7 is between 6 and h

Counting duplication trees

– we used PDT to count (or estimate) the number of duplication trees

Counting duplication trees

– we used PDT to count (or estimate) the number of duplication trees

– the number of duplication trees is largely inferior to the number of distinct phylogenies

Counting duplication trees

– we used PDT to count (or estimate) the number of duplication trees

– the number of duplication trees is largely inferior to the number of distinct phylogenies

– the number of phylogenies expands approximately 3^n faster than the number of duplication trees

Reconstructing duplication trees

– the goal is to recontruct the optimal duplication tree(s) from a given set of aligned and ordered DNA sequences

Reconstructing duplication trees

– the goal is to recontruct the optimal duplication tree(s) from a given set of aligned and ordered DNA sequences

– we use an exhaustive search approach

Reconstructing duplication trees

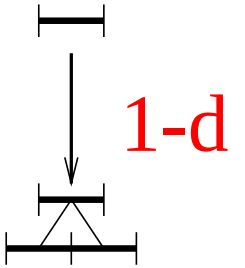
– the goal is to recontruct the optimal duplication tree(s) from a given set of aligned and ordered DNA sequences

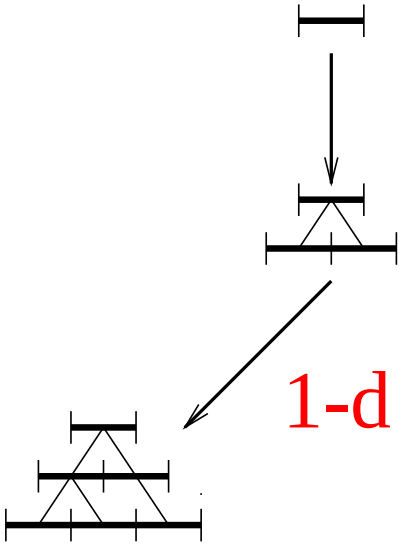
– we use an exhaustive search approach

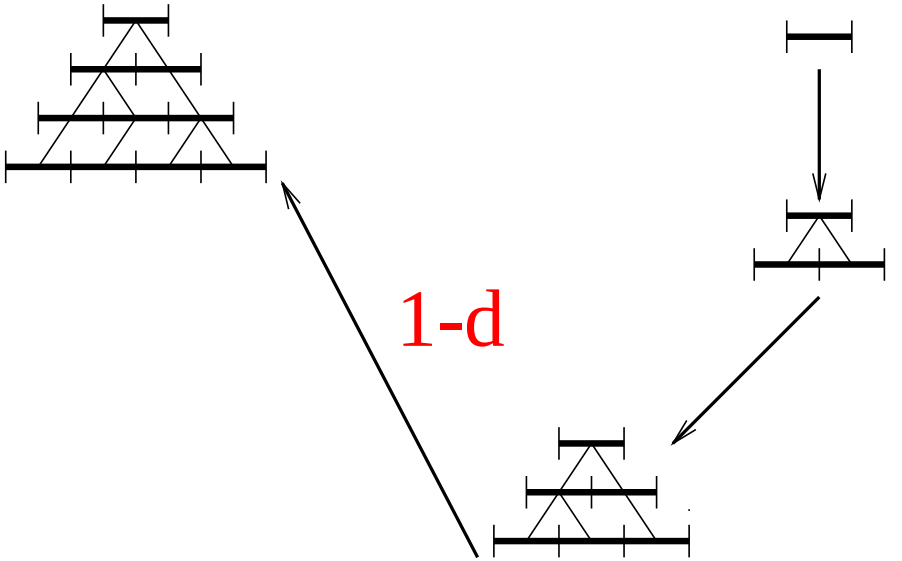
– we assess the optimality of the reconstruction using a parcimony criterion

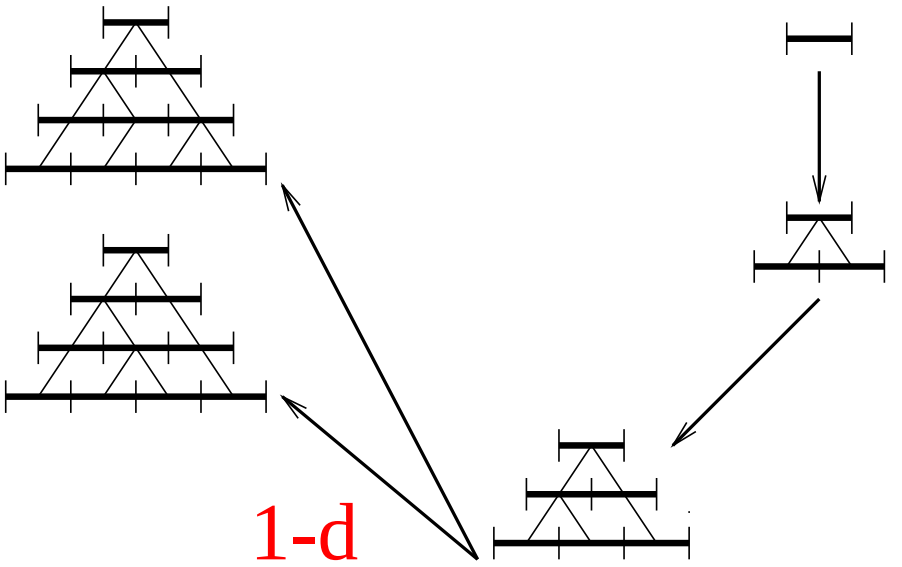
Exhaustive approach

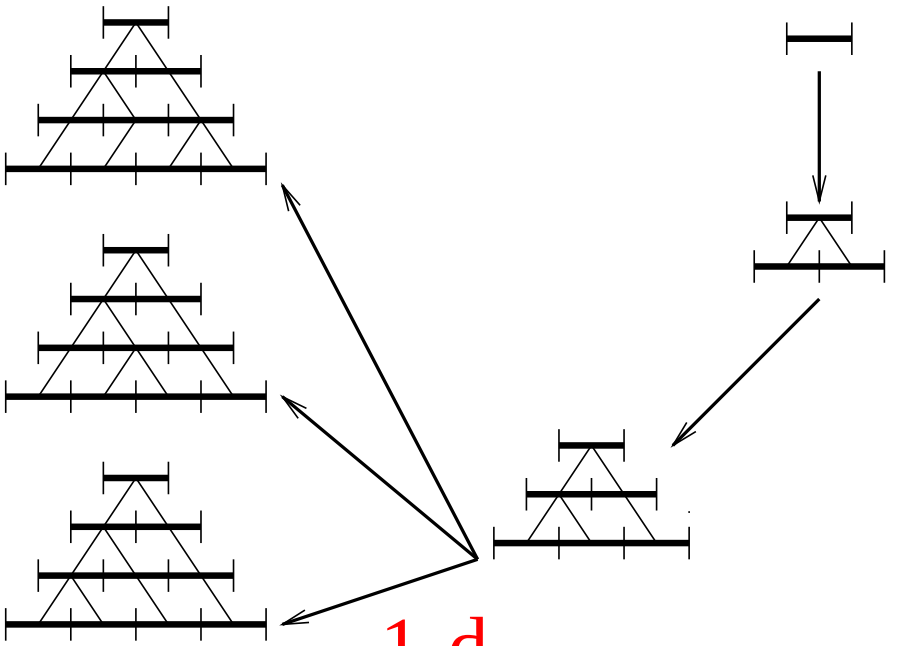
– we generate every possible duplication tree, using a simulation of the duplication process

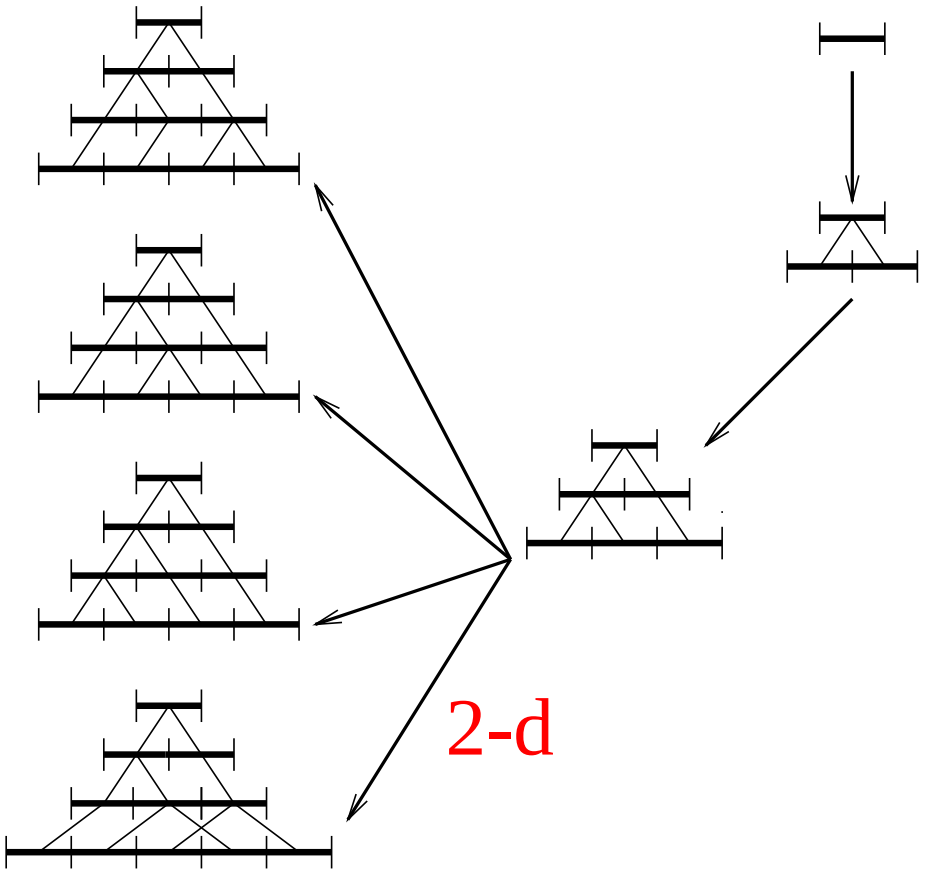

1-d


1-d

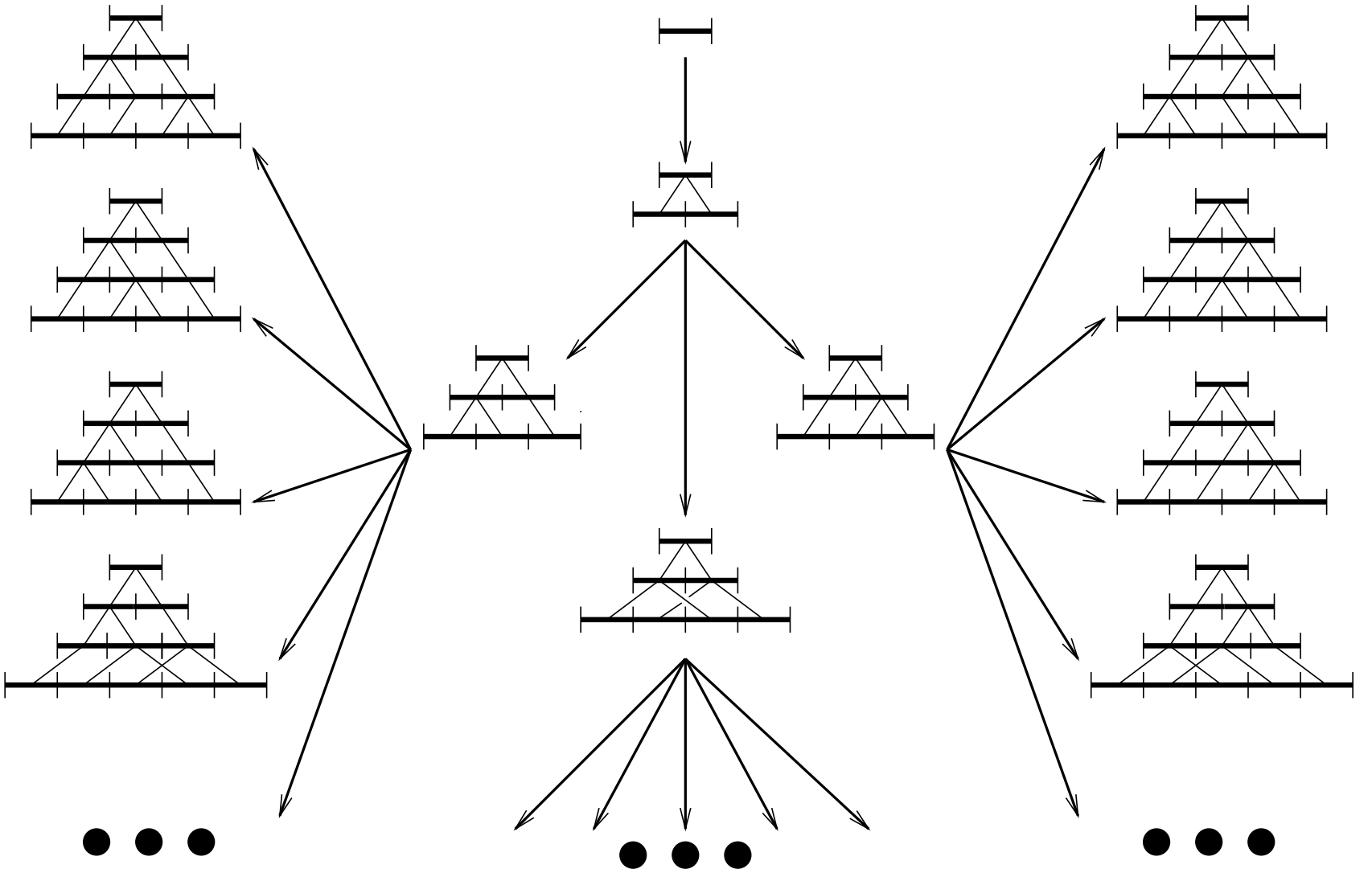

1-d


1-d


1-d


2-d

3. Reconstructing duplication trees



Exhaustive approach

– we generate every possible duplication tree, using a simulation of the duplication process

– we select the trees that minimize the parcimony criterion

Experimental results

– we applied this reconstruction procedure to the TRGV locus

Experimental results

– we applied this reconstruction procedure to the TRGV locus

– only 1 duplication tree is found by exhaustive search

First validation

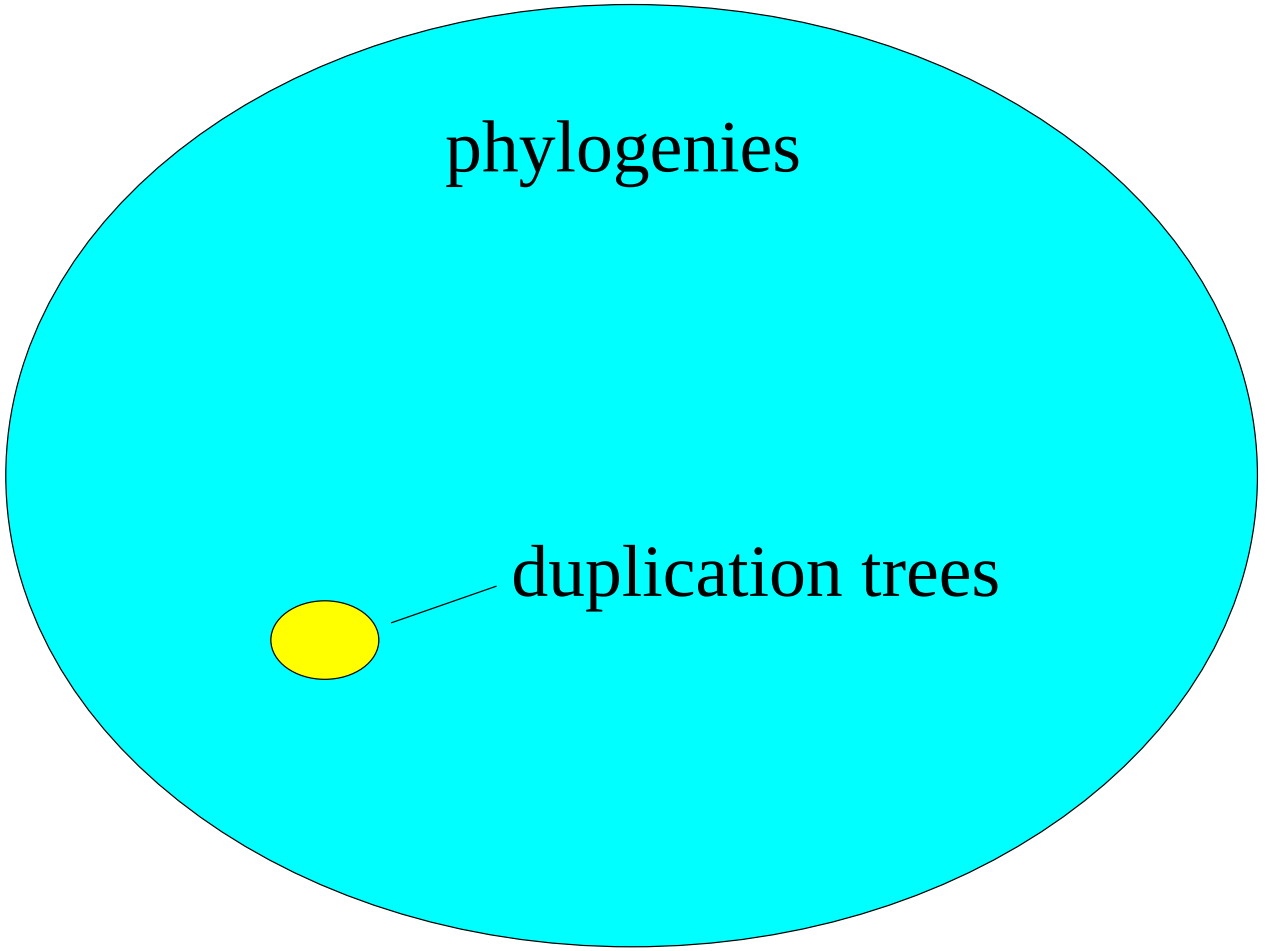
– this duplication tree is identical to the most parcimonious phylogeny, reconstructed from the same data, but without restriction to duplication trees

First validation

– this duplication tree is identical to the most parcimonious phylogeny, reconstructed from the same data, but without restriction to duplication trees

– the probability of a phylogeny to be a duplication tree is less than 0.04 for 9 taxa

First validation



Second validation

– we root the duplication tree using both the molecular clock hypothesis on functional genes and an outgroup

Second validation

– we root the duplication tree using both the molecular clock hypothesis on functional genes and an outgroup

– both methods root the tree at the same branch

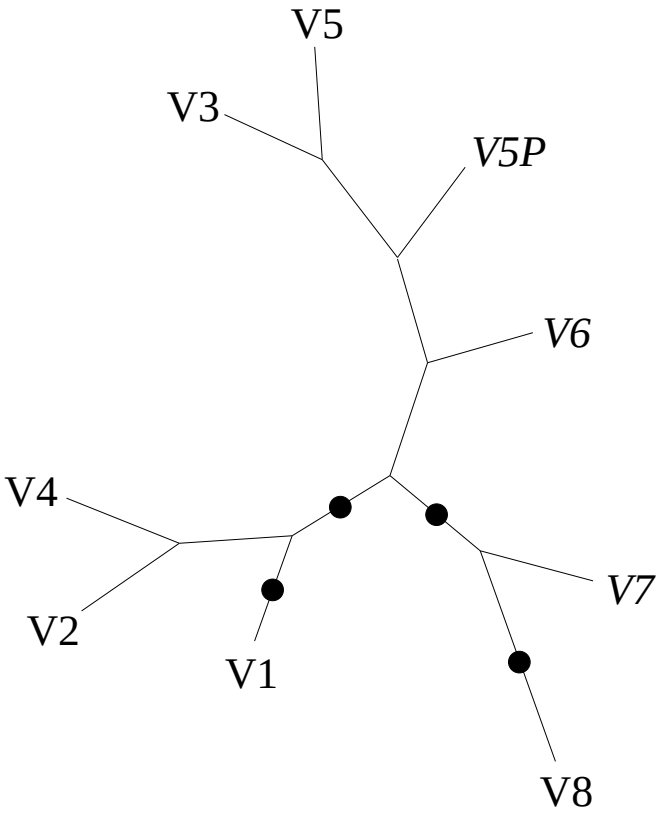
Second validation

– we root the duplication tree using both the molecular clock hypothesis on functional genes and an outgroup

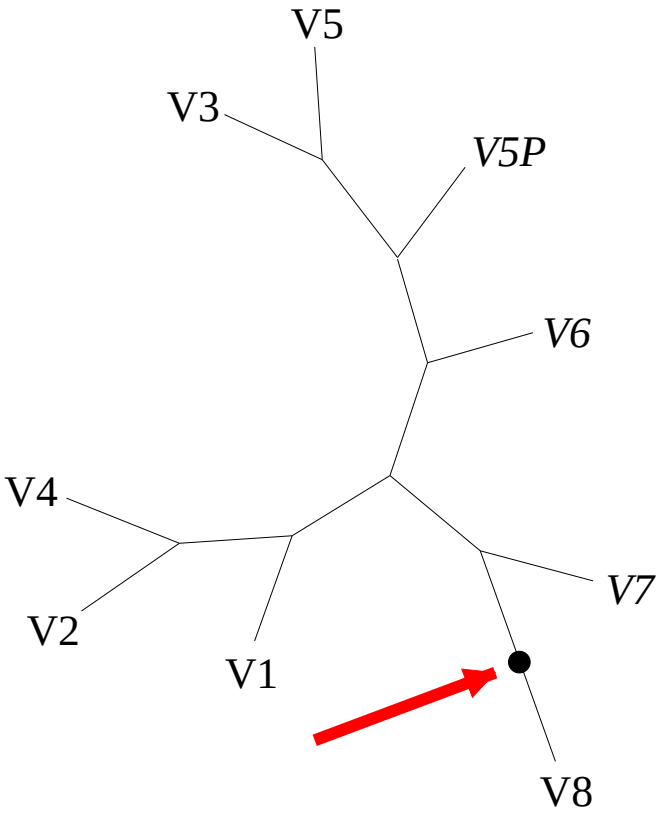
– both methods root the tree at the same branch

– the root belongs to the “potential roots”

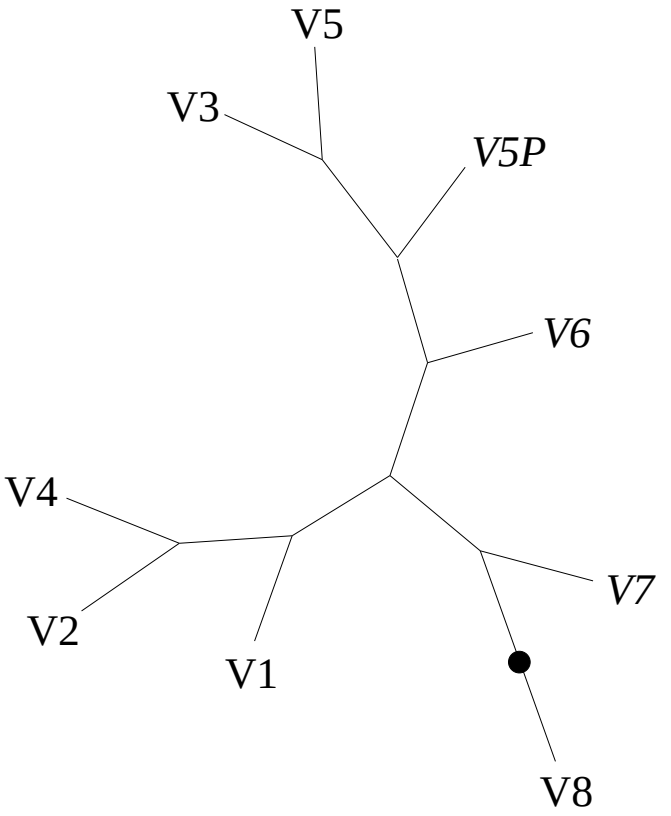
Second validation

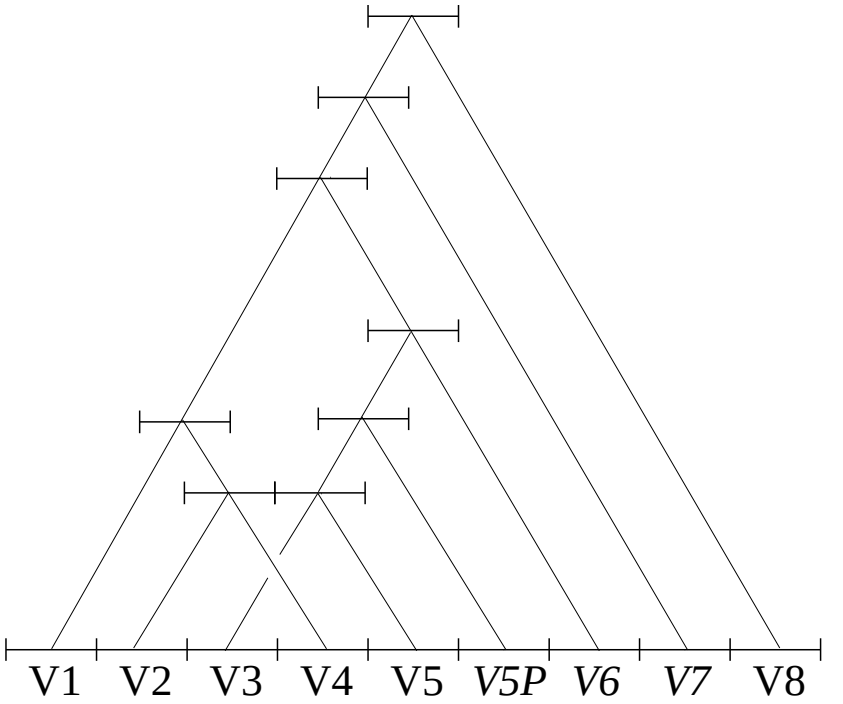


Second validation



Second validation

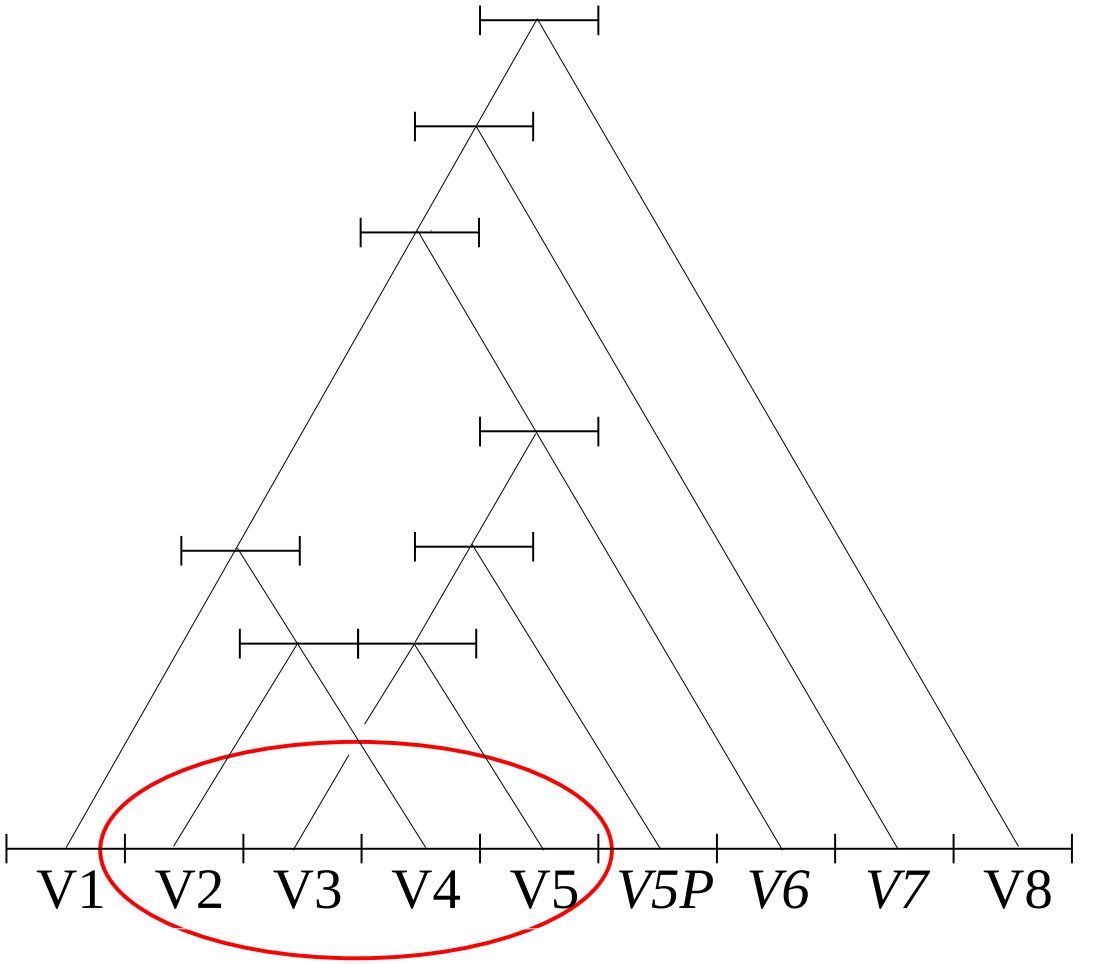




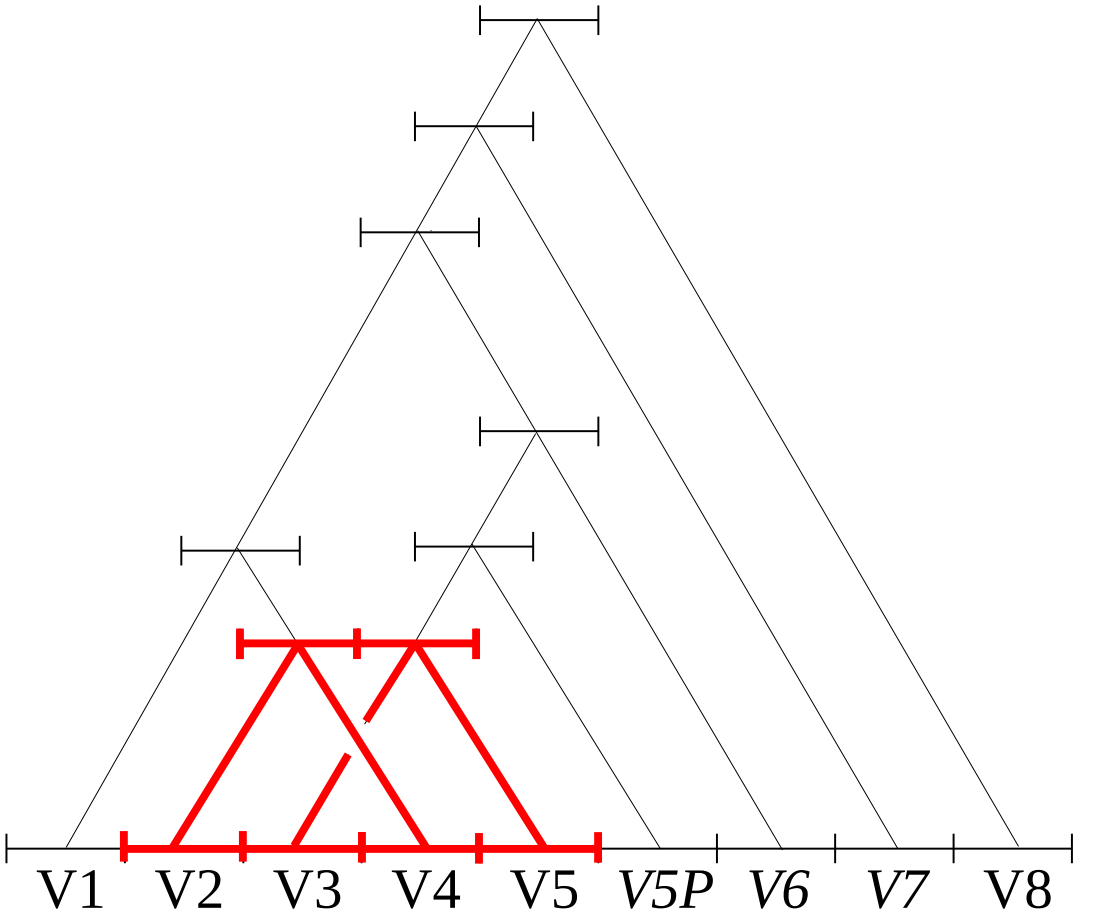
Third validation

– the ordinal duplication history is in agreement with a polymorphism that exists for this locus

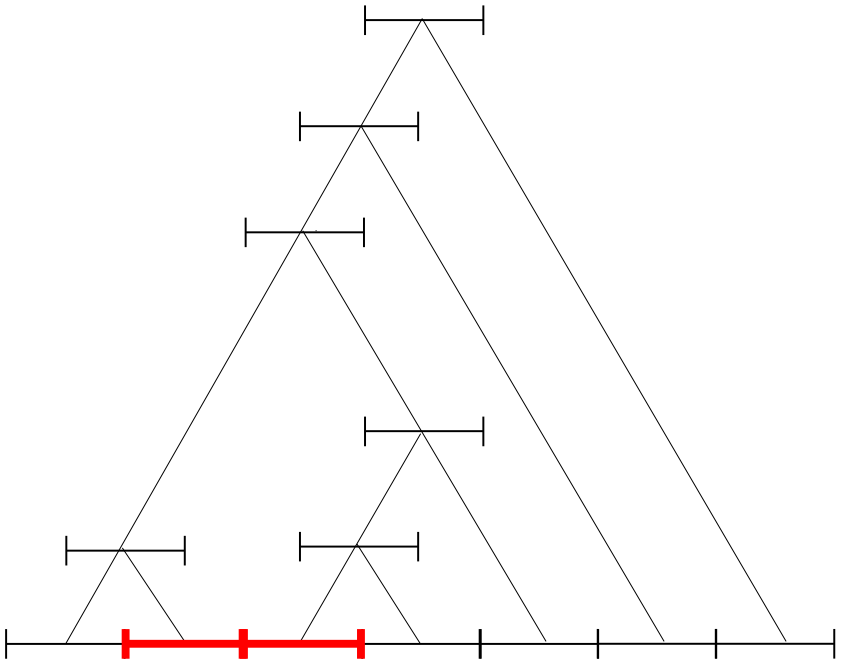
Polymorphism for the TRGV locus



Polymorphism for the TRGV locus



Polymorphism for the TRGV locus



Conclusion

– these results validate the duplication model

Conclusion

– these results validate the duplication model

– they show that our reconstruction procedure can provide a valid solution

Conclusion

– these results validate the duplication model

– they show that our reconstruction procedure can provide a valid solution

– they are robust to gene deletions (most duplications are 1-duplications)

Perspectives

– development of a fast heuristics to improve the reconstruction speed

Perspectives

– development of a fast heuristics to improve the reconstruction speed
– comparison with other criteria such as minimum evolution, ...

Perspectives

– development of a fast heuristics to improve the reconstruction speed
– comparison with other criteria such as minimum evolution, ...
– better mathematical characterisation of duplication trees (for example, can we enumerate them ?)

Perspectives

– development of a fast heuristics to improve the reconstruction speed
– comparison with other criteria such as minimum evolution, ...
– better mathematical characterisation of duplication trees (for example, can we enumerate them ?)
– applying our methods and algorithms to other datasets

Perspectives

- development of a fast heuristics to improve the reconstruction speed
- comparison with other criteria such as minimum evolution, ...
- better mathematical characterisation of duplication trees (for example, can we enumerate them ?)
- applying our methods and algorithms to other datasets
- more complex duplication models