Karlsruhe Institute of Technology

Phylogenetic species tree inference from gene trees despite paralogy

Master Thesis of

Paul Schade

At the Department of Informatics
Institute of Theoretical Computer Science

Reviewer:	Prof. Dr. Alexandros Stamatakis
	Prof. Dr. Peter Sanders
Advisor:	Benoit Morel

Time Period: 1. June 2020 – 30. November 2020

Statement of Authorship

I hereby declare that this document has been composed by myself and describes my own work, unless otherwise acknowledged in the text.

Karlsruhe, 30. November 2020

Abstract

In this thesis, newly developed distance-based methods to infer a species tree from gene family trees are studied.

The methods are based on species distances as calculated from the respective gene family trees. The distances from the gene family trees are merged via the novel *NJst+*, *mini*, and *tagging* methods into a single distance matrix each which summarises all pairwise distances among the species. Based on this merged distance matrix, each method computes a species tree.

We evaluate our methods via empirical and simulated data sets. The relative Robinson–Foulds distance between a given species tree and the newly calculated distance-based species tree is our main accuracy metric. We find that our best methods perform reasonably well with a mean relative Robinson-Foulds distance of 0.0470 (4.7 %) for miniNJ$_{w_s}$ and 0.0400 (4 %) for tagNJ$_{MAD}$ on the STANDARD data set (Tab. 5.1). For comparison, the most accurate tool (A-pro), we compared with, has a relative Robinson-Foulds distance of 0.0461 (4.6 %). The STANDARD data set has all used parameters on default setting, while other data sets have certain parameters varying. We used a total of 44 distinct sets of parameters with 3 different sequence lengths and 50 data sets each. Additionally, we evaluated our methods on 5 empirical data sets. The mean run times for the STANDARD data set are 0.10 s for miniNJ$_{w_s}$ and 4.23 s for tagNJ$_{MAD}$, while A-pro needed a mean run time of 7.17 s.

Deutsche Zusammenfassung

Diese Arbeit behandelt neu entwicklete distanzbasierte Methoden zur Berechnung von Speziesstammbäumen aus Genfamilienstammbäumen.

Die Methoden basieren auf Speziesabständen, die aus entsprechenden Genfamilienstammbäumen berechnet werden. Die Distanzen der Genfamilienstammbäume werden mit den neuartigen *NJst+*, *mini* und *tagging* Methoden zu jeweils einer einzigen Distanzmatrix zusammengführt, die alle paarweisen Distanzen zwischen den Spezies zusammenfasst. Basierend auf dieser Distanzmatrix, wird jeweils ein Speziesstammbaum berechnet.

Wir bewerten unsere Methoden anhand empirischer und simulierter Datensätze. Die relative Robinson-Foulds Distanz zwischen einem gebenen Speziesstammbaum und dem neu berechneten Speziesstammbaum stellt unsere wichtigste Genauigkeitsmetrik dar. Wir stellen fest, dass unsere besten Methoden, mit einer durchschnittlichen relativen Robinson-Foulds Distanz von 0,0470 (4,7 %) für miniNJ$_{w_s}$ und 0,0400 (4 %) für tagNJ$_{MAD}$ auf dem STANDARD Datensatz (Tab. 5.1), verhältnismäßig gut funktionieren. Zum Vergleich, A-pro, das bisher genaueste Programm, mit dem wir verglichen haben, erreicht ein arithmetisches Mittel bezüglich der relativen Robinson-Foulds Distanz von 0,0461 (4,6 %). Der STANDARD Datensatz ist mit den Standard-Parametern erstellt, während bestimmte Parameter bei anderen Datensätzen variieren. Wir haben insgesamt 44 unterschiedliche Parametersets zu je 3 Sequenzlängen und 50 Datensätzen. Zusätzlich haben wir unsere Methoden auf 5 empirische Datensätzen getestet. Die erforderliche Laufzeiten für den STANDARD Datensatz sind 0,10 s für miniNJ$_{w_s}$, 4,23 s für tagNJ$_{MAD}$ und 7,17 s für A-pro.

Acknowledgements

I want to thank Prof. Dr. Alexandros Stamatakis and Benoit Morel for excellent support and assistance throughout the whole time and across borders. They were always available for me and helped me with any question.

I want to thank Prof. Dr. Peter Sanders for being the second reviewer.

I also thank everyone else from the Computational Molecular Evolution group for helping me with templates, computer issues, and, situational, only a few social moments.

I want to thank my family for mental support and financial sponsorship during the many years of studying.

Thanks to all my friends for keeping my motivation up and giving me a good time.

I want to thank my Girlfriend for always making sure I will keep working, but making sure I have balance in my life. And of course for giving me accommodation for a long time during this work.

Contents

1. Introduction

1.1. Motivation

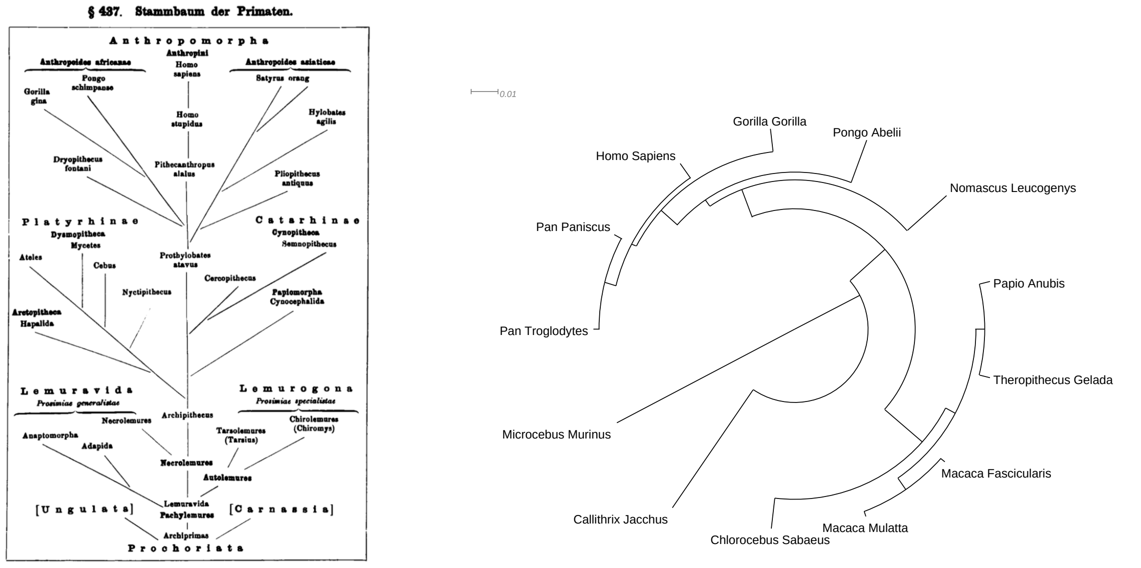
Species play an important role for numerous aspects of biology [1]. Representing their evolutionary relationships in species trees as in Fig. 1.1 is crucial for numerous analytical purposes, for instance to rise awareness to human induced mass extinction [2]. For over 100 years, scientist have estimated species trees based on so-called apomorphies. They considered external characteristics, that groups of species share or that are unique to a single species. In 1895, Ernst Haeckel estimated species trees for primates among other species. He distinguished between, for example, species with claws or fingernails and how the nostrils are arranged [3]. Fig. 1.1a shows one of his species trees.

Today, species trees are estimated on the basis of DNA-sequences. Fig. 1.1b shows a species tree from one of the DNA-sequence data sets we used. A widely used method for sequencing was the Sanger method [4][5] which was developed in the 1970s. In the 2000s, the development of next Generation Sequencing enabled scientists to obtain DNA-sequences substantially faster and cheaper than with the previous Sanger method [6]. With the faster methods it is possible to obtain longer sequences from more species.

To infer a species tree from DNA-sequences, often a two-step procedure is used. The first step estimates gene family trees from the sequence data. The second step combines these trees into a single species tree, that best explains the gene family tree. Tools such as ParGenes [7] [8] can infer gene family trees from large data sets with 'good' parallel efficiency. For the second step there exist methods and tools such as NJst [9], A-pro [10], DupTree [11], and FastMulRFS [12] among many others. The time complexities and associated execution times heavily depend on the number of species and gene family trees. For data sets with more than 100 species it takes some tools several days to estimate a species tree. Incomplete lineage sorting, duplication and loss events, and horizontal gene transfer make it difficult to correctly estimate species trees [13]. Thus, there is a need for new methods that handle these challenges better and run in a reasonable time.

1.2. Objectives of this thesis

The objectives of this thesis are to develop and evaluate new methods for the step of combining gene family trees into a single species tree. The methods focus on recognising paralogies in the gene family trees. To achieve this, we use distance based methods. We



(a) A primates species tree from 1895 [3].

(b) A primates species tree based on the primates data set (Tab. 5.1). Visualized using Dendroscope 3 [14].

Figure 1.1.: Species trees in comparison.

calculate distances in the gene family tree and use several techniques to compute a single pair-wise species distance matrix. We then use clustering methods to estimate species trees from the distance matrix. We evaluate our methods on simulated and empirical data sets with respect to their species tree reconstruction accuracy. Furthermore, we show the advantages of the theoretical time complexity of our methods in comparison to other methods.

1.3. Own contributions

My contribution was to combine existing techniques with our techniques into newly developed methods for species tree inference from gene family trees. Our techniques aim to filter paralogy. One approach is to decide, for every pair-wise distance in the gene family trees, if we consider it to be based on paralogy or not. Another technique is a heuristic, which only uses minimum distances. I implemented the methods in C, which can be executed in extremely short times compared to the time required for the inference of the gene family trees. Furthermore, we simulate data sets, to evaluate our methods on. We simulate large data sets using Simphy [15], INDELible [16], and ParGenes [7]. Among the new methods $miniNJ_{w_s}$ and $tagNJ_{MAD}$ showed the best accuracy.

2. Background

In this Chapter we introduce the necessary background knowledge, methods and terminology that will help the reader to understand the following Chapters more easily. Sec. 2.1 introduces biological background knowledge about genetics. Sec. 2.2 explains how phylogentic trees represent genetic relationships. Sec. 2.3 introduces methods for species tree inference. Sec. 2.4 describes existing methods for computing a distance matrix on a set of gene family trees. Sec. 2.5 explains the tagging procedure and the associated necessary rooting methods for phylogenetic trees.

2.1. Genetics

The thesis is motivated by biology, therefore, we introduce some relevant models and terminology in genetics first.

2.1.1. Species

In biology, a species is a unit of classification for organisms. Biologists still disagree on the species definition [17] and providing such a definition is not part of this thesis. In the scope of this work, we consider a species to be a group of organisms that share a pool of common genes and that are able to interbreed [18]. An example for some closely related species would be the following group of primates: Bonobo, Chimpanzee, Gorilla, Human, and Orangutan.

2.1.2. Genes, DNA, sequence and genome

There is still some disagreement on what a gene is [19]. For this thesis, we consider a gene to constitute information which children inherit from their parents. Biological properties of the offspring are derived from this information. The *Deoxyribonucleic acid* (*DNA*) *sequence* encodes the respective information. It is a sequence of the four nucleotides *cytosine (C)*, *guanine (G)*, *adenine (A)*, and *thymine (T)* [20]. A *sequence* is usually represented by the abbreviated nucleotides. The *genome* contains the entire *DNA* of a species including all genes and also the non-coding *DNA* [21].

2.1.3. Evolutionary events

The *genome* evolves with the species over time. It can evolve differently in distinct populations of the same *species*. Evolutionary events that occur are

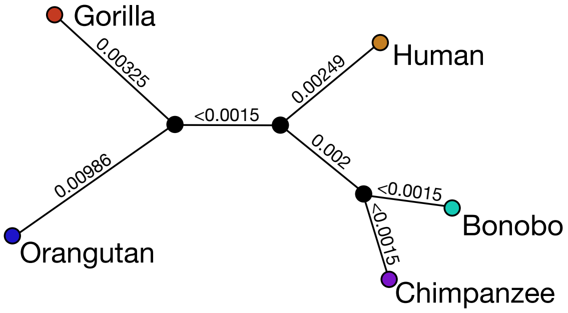


Figure 2.1.: Species tree of Bonobo, Chimpanzee, Gorilla, Human, and Orangutan [25]. A mapping assigns every leaf to exactly one species. The colours of the leaves represent the mapping. Branch lengths are not drawn to scale.

- *Speciation* (spec): the *genome* of two populations evolves differently and the populations form two new *species*. The genome of an individual belongs to one and only one of the two new *species*.
- *Duplication* (dup): a gene duplicates and is present more than once in the genome afterwards. The gene's copies can evolve differently, which makes them to not be exactly identical. Multiple duplications can also take place, yielding more than two copies of the gene in the genome.
- *Loss* (loss): a gene is erased from the genome. It is not present in the genome anymore. The species no longer inherits the gene.
- *Horizontal gene transfer* (hgt): a gene is not inherited by evolutionary descend but transferred from another species. This can be frequently observed between bacteria [22].

Gene pairs in distinct species that are a result of a speciation event are called *orthologous* genes [23]. Duplication events form pairs of *paralogous* genes [24]. A *paralogy* describes the relationship between paralogous genes.

2.2. Phylogenetic trees

A phylogenetic tree T represents the evolutionary relationships between species, genes, populations, or individuals [26]. It is an undirected graph, where any pair of vertices is connected by exactly one path. Fig. 2.1 shows an example of a species tree. Vertices are generally denoted as *nodes* p_i. Nodes with degree 1 are called *leaves* or *terminal nodes* and represent contemporary species or genes. Nodes with degree strictly greater than one are termed *internal nodes* and represent hypothetical ancestors. Fig. 2.1 shows internal nodes in black and leaves in colour. Edges are called *branches*. Branch lengths can represent true evolutionary time on the species trees or quantify the amount of sequence divergence in gene trees [26].
Trees are either rooted or unrooted. In rooted trees, the root node represents the hypothetical common ancestor of all species that are present in the tree. Unrooted trees do not contain information about the chronological order of speciation events.
The *topology* of a tree determines the arrangement of the branching pattern between the nodes of the tree. There are

$$\prod_{i=1}^{n-2} 2i+1 = \frac{(2n-3)!}{2n-2(n-2)!} \tag{2.1}$$

different possible rooted tree topologies and $\prod_{i=1}^{n-3} 2i+1$ possible unrooted tree topologies with n species [27]. A set of trees is a *forest.* A *bifurcating node* is an internal node of degree 3 or a root node of degree 2. Such a node assumes that speciations yield pairs of species. A *multifurcating node* or *polytomy* is an internal node that is not bifurcating. A polytomy assumes that simultaneous speciation into more than two species occurred. It can also be the result of insufficient phylogenetic signal for resolving the topology. A *bifurcating* or *fully resolved* tree is a tree where every internal node is bifurcating. A *multifurcating* tree is a tree that is not bifurcating. The *Lowest Common Ancestor* (*LCA*) of two leaves a_i and a_j, $LCA(a_i, a_j)$, in a rooted tree is the node q_i, which is furthest away from the root and that has both a_i and a_j as descendants.

2.2.1. Species tree

Species evolve through speciation and extinction events. A speciation event gives rise to two new descendant species that subsequently evolve independently. An extinction event does not give rise to any descendant. A species tree is the representation of such a history, where internal nodes correspond to speciation events and leaves correspond to contemporary species. In the scope of this work, extinction events do not appear on a species tree because they are not observable.

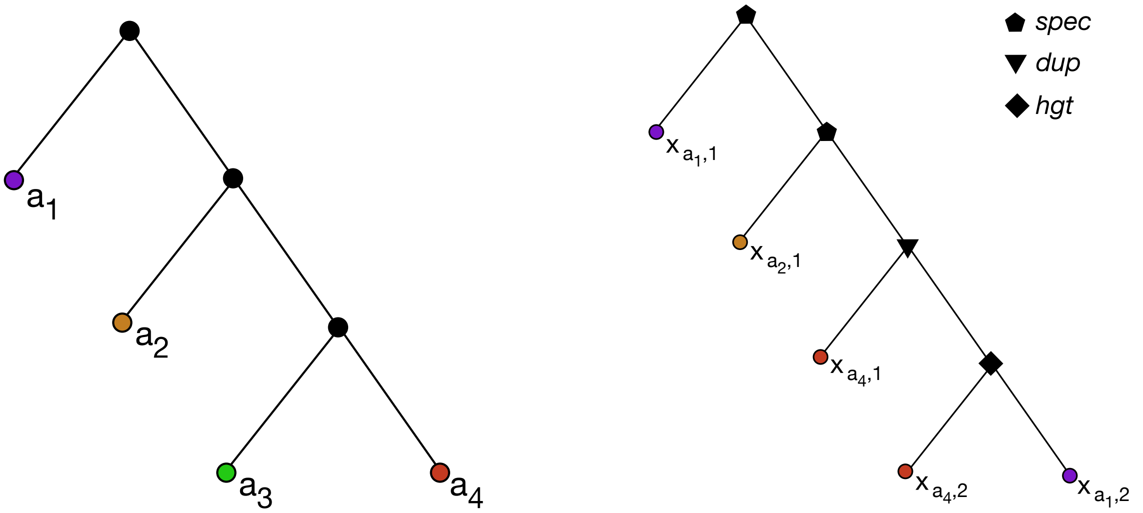
2.2.2. Gene family tree

Genes are present in the species' genomes and evolve along the species tree. A leaf $x_{a_i,s}$ in the gene family tree represents the sth copy of a gene in the genome of species a_i. A given mapping assigns each leaf to exactly one species. Fig. 2.2 shows a species tree and a corresponding gene family tree. In this work, we make the assumption that genes evolve through the evolutionary events described in Sec. 2.1.3, that we represent via a gene family tree. Internal nodes can have different origins. In Fig. 2.2b internal nodes represent different events which are distinguished by the nodes' shapes. The chain of events that generated the gene tree can be the following

- A *spec* event generated species a_1 and the ancestor of a_2. The gene evolved differently in the species. The root node represents this event.
- Another *spec* event resulted in a_2 and the ancestor of a_3. The second node marked as *spec* shows this event.
- A third *spec* event gave rise to species a_3 and a_4.
- But a *loss* event in species a_3 erased the gene from a_3's genome. That is the reason, why no leaf is mapped to a_3, and why there is no third *spec* node in the gene family tree.
- A *dup* event in species a_4 caused the gene to be present twice in a_4's genome. A *dup* node represents this phenomenon in the gene family tree. Furthermore, two leaves are mapped to species a_4.
- In a *hgt* event species a_4 transferred a gene copy to species a_1. The genes evolved differently in distinct species. Therefore, the two leaves mapped to species a_1 are located far apart in the gene family tree. The event is denoted by the *hgt* node in the gene family tree.

2.2.3. Incomplete lineage sorting

Apart from the aforementioned events that explain the differences between species tree and the gene family trees, studies [13] [28] [29] also conclude that large population sizes and short times between speciation events yield strong divergence between the trees. This effect is known as *Incomplete Lineage Sorting* (*ILS*).



(a) Rooted species tree for species a_1, a_2, a_3, and a_4. Colours represent the mapping to species.

(b) Rooted gene family tree for species tree 2.2a. Colours show the mapping of leaves to species. Node shapes indicate, which event gave rise to the internal node.

Figure 2.2.: Rooted species tree and corresponding gene family tree showing *duplication* (*dup*), *speciation* (*spec*), and *horizontal gene transfer* (*hgt*) events. *Loss* events are not visible in a gene family tree.

2.2.4. Gene family tree construction

To construct gene family trees the genomes of a set of species A are sequenced. Within each genome sequence, genes are detected that belong to a specific gene family. For each gene copy of a gene family in a species' genome, a leaf that is mapped to one of the species is added to the corresponding gene family tree. Tools like *RAxML-NG* [8] use maximum likelihood methods to estimate the tree topology of the gene family tree. The tool can work with different evolutionary models [8] and yields a topology which is likely to have generated the data under the selected model.

2.2.5. Tree comparison

A standard metric to objectively measure the similarity of two phylogenetic trees is the *Robinson-Foulds-Distance* (*RFD*) [30]. The two operations used to calculate the distance are α and α^{-1}. The operator α contracts two nodes p_i and p_j by removing the edge e_{p_i,p_j} between them from the tree. The new tree contains a new node p_{ij} instead of nodes p_i and p_j. The node contains the union of labels of the nodes p_i and p_j as labels and the union of edges of the nodes p_i and p_j as edges. The deconstruction operator α^{-1} inverses the contraction operator α. An arbitrary split of the union of edges and of the labels allocates edges and labels to the new nodes. Leaves must be allocated exactly one label. The minimum number of α and α^{-1} operations needed to transform tree T_1 into tree T_2 is the $RFD(T_1, T_2)$ [30].
The maximum number of operations needed for transforming an unrooted tree with n species into another is $2(n-3)$ [8]. The *relative RFD* (*rRFD*) is therefore

$$rRFD(T_1, T_2) = \frac{RFD(T_1, T_2)}{2(n-3)} \tag{2.2}$$

	a_1	a_2	a_3	a_4	a_5
a_1	×	2	2.5	3.8	5
a_2		×	2.6	4	5.3
a_3			×	2.7	4
a_4				×	3.4
a_5					×

Table 2.1.: Example distance matrix for species a_1, a_2, a_3, a_4, and a_5

2.3. Distance-based methods for species tree inference

Different methods to infer species trees exist. This Section introduces four of them. Distance based methods use a distance matrix D, which represents the distances between every pair of sequences or species (a_i,a_j), to calculate a phylogenetic tree. Sec. 2.4 outlines the calculation of the distance matrix D.

2.3.1. Neighbor Joining

The *neighbor-joining* (*NJ*) algorithm constructs a phylogenetic tree for n species from a given distance matrix D. *Neighbors* are nodes that are only separated by exactly one node. Starting from a star shaped tree in step 0 (Fig. 2.3a) the algorithm joins neighbors until a single unrooted bifurcating tree is obtained. *NJ* joins nodes p_i, p_j by adding an internal node q_s in between them. It chooses the pair of nodes such that the sum of branch lengths is minimized [31]. To do so, the *NJ* algorithm calculates a matrix Q at every step as follows:

$$Q_{p_i,p_j} = (n-2)D_{p_i,p_j} - \sum_{k=1}^{n} D_{p_i,p_k} - \sum_{k=1}^{n} D_{p_j,p_k} \tag{2.3}$$

The algorithm selects the pair of nodes (p_i,p_j), which corresponds to the minimum of matrix Q and joins them as follows. A new internal node q_s connects the node pair by being added in between them. The algorithm removes the pair of nodes from the pool of nodes and matrices and adds the new node. Distances from the joined nodes p_i, p_j to the new internal node q_s are calculated as follows:

$$D_{p_i,q_s} = \frac{1}{2}D_{p_i,p_j} + \frac{1}{2(n-2)}\left|\sum_{k=1}^{n} D_{p_i,p_k} - \sum_{k=1}^{n} D_{p_j,p_k}\right| \tag{2.4}$$

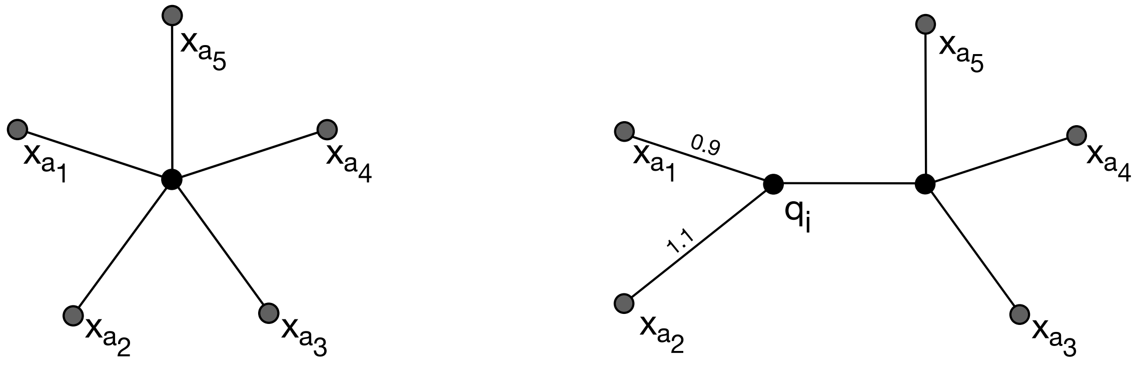
$$D_{p_j,q_s} = D_{p_i,p_j} - D_{p_i,q_i}. \tag{2.5}$$

Distances from other nodes p_k to the new node are calculated as follows:

$$D_{q_s,p_k} = \frac{1}{2}[D_{p_i,p_k} + D_{p_j,p_k} - D_{p_i,p_j}]. \tag{2.6}$$

NJ needs a total of $(n-3)$ joining steps. For each step i a quadratic matrix Q of size $(n-i)$ must be calculated. The time complexity is, therefore, $\mathcal{O}(n^3)$. *Fast Neighbor Joining* is an algorithm with a time complexity of $\mathcal{O}(n^2)$ that only induces a minimal loss of accuracy [32].

Tab. 2.1 shows an example distance matrix, Fig. 2.3 visualizes the first step of the *NJ* algorithm and Fig. 2.4 shows the resulting tree.



(a) *NJ* step 0, starting with a star-shaped tree

(b) *NJ* step 1, after joining the first pair

Figure 2.3.: First steps of *NJ* method for the distance matrix given in Tab. 2.1

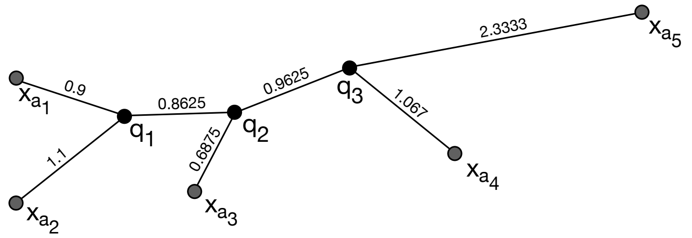


Figure 2.4.: Phylogenetic tree obtained from Tab. 2.1 with the *NJ* algorithm

2.3.2. UPGMA, WPGMA

The *unweighted pair group method with arithmetic mean* (*UPGMA*) and the *weighted pair group method with arithmetic mean* (*WPGMA*) are methods to build a phylogentic tree with n species from a given distance matrix D. The method joins two clusters of elements A_i and A_j, which have the lowest distance of all current clusters, at every step. A new internal node q_s in the phylogentic tree represents the new cluster. The method calculates distances to the new node q_s, such that all elements of the cluster have the same distance to q_s. Distances of the new cluster $A_i \cup A_j$ to an outside cluster A_k after joining are calculated as follows:

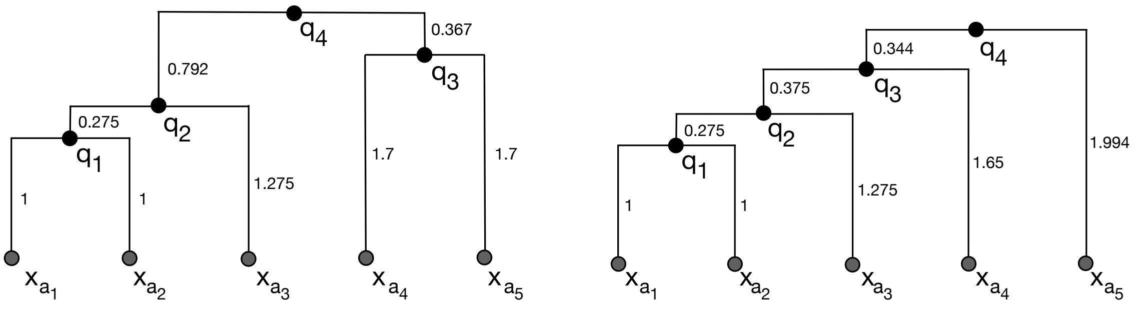
- $D_{A_i \cup A_j, A_k} = \dfrac{|A_i| \cdot D_{A_i,A_k} + |A_j| \cdot d_{A_j,A_k}}{|A_i| + |A_j|}$ for *UPGMA* and
- $D_{A_i \cup A_j, A_k} = \dfrac{D_{A_i,A_k} + d_{A_j,A_k}}{2}$ for *WPGMA* [33].

In *WPGMA* elements that were added early to the cluster receive a lower weight than elements that are added later on to the cluster. This step is repeated until all elements form one cluster.
Both *UPGMA* and *WPGMA* need $(n-1)$ joining steps. At every step i, the method updates the distance matrix of size $(n-i)$ and finds the minimum within this matrix. In a naïve implementation the time complexity is, therefore, $\mathcal{O}(n^3)$.

Fig. 2.5 shows trees inferred with *UPGMA* and *WPGMA* from the example distance matrix in Tab. 2.1.

2.3.3. Least squares method

The least squares method calculates branch lengths for a phylogenetic tree with n species for given distance matrix D. The branch lengths are calculated, such that the squares of



(a) *UPGMA* resulting tree

(b) *WPGMA* resulting tree

Figure 2.5.: Phylogenetic trees from distance matrix Tab. 2.1

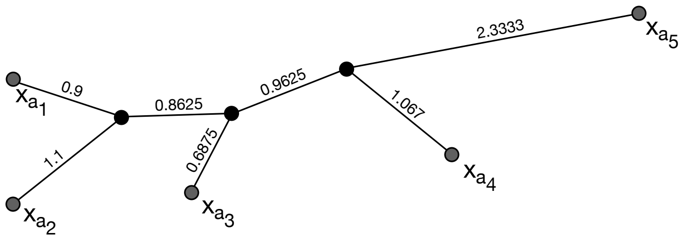


Figure 2.6.: Phylogenetic tree from Tab. 2.1 with *least squares method*

the differences between the given distances D_{a_i,a_j} and the estimated distances $\hat{D}_{a_i,a_j}$ on the tree are minimized. A key assumption is, that branch lengths are additive regarding the distance of species within a tree [27]. In Fig. 2.6 the distance between species a_1 and a_2 is therefore $\hat{D}_{a_1,a_2} = 1.1 + 0.9$. The branch lengths are calculated by minimizing the *least squares* score which is defined as follows

$$S = \sum_{i<j} (D_{a_i,a_j} - \hat{D}_{a_i,a_j})^2 \tag{2.7}$$

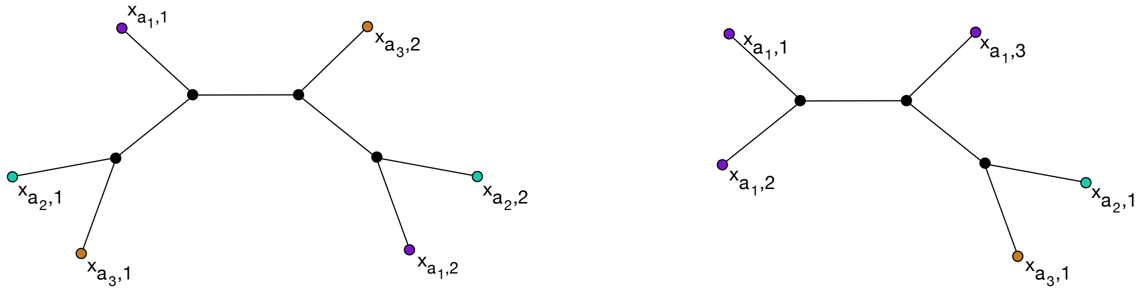
Minimizing the score requires solving a set of linear equations. Among different tree topologies, the one with the lowest score S is chosen. Since the number of possible topologies grows super-exponentially (Eq. 2.1) with the number of species, Cavalli-Sforza and Edwards proposed heuristics for choosing topologies [27]. Without using heuristics for choosing topologies, there is a factorial number of possible tree topologies. Therefore, the time complexity is $\mathcal{O}(n!)$. Fig. 2.6 shows the *least squares method* tree for the distance matrix of Tab. 2.1. It yields the same tree as the *NJ* method.

2.4. Computing distance matrices on sets of gene family trees

The main objective of this thesis is to develop and evaluate methods to infer species trees from sets of gene family trees. An essential step in these methods is to compute distance matrices on sets of gene family trees. This Section introduces existing methods. Sec. 2.4.1 shows the *NJst* method, Sec. 2.4.2 describes the *ustar* method.

2.4.1. NJst

The *NJst*-method [9] calculates a distance matrix D_{NJst} based on every pair of leaves of a set of gene family trees $\mathbb{T}$. $D(x_{a_i,s}, x_{a_j,t})$ denotes the distance between the s$^{\text{th}}$ leaf, which has a mapping to species a_i, and the t$^{\text{th}}$ leaf, which has a mapping to species a_j. The distance is defined as the number of internal nodes on the path connecting them [9]. This is equal to the path length minus 1, when every edge has a uniform length of 1. For example,



(a) Gene family tree k_1

(b) Gene family tree k_2

Figure 2.7.: A set K of gene family trees

in Fig. 2.7a the distance between leaves $x_{a_1,1}$ and $x_{a_2,1}$ is $D(x_{a_1,1}, x_{a_2,1}) = 2$. The variable m_{a_i,T_k} denotes the number of occurrences of species a_i in gene family tree T_k, whereas $\mathbb{T}$ is the set of all gene family trees. In the example in Fig. 2.7a, the number of occurrences of leaves mapped to certain species is $m_{a_1,T_k} = m_{a_2,T_k} = m_{a_3,T_k} = 2$. The overall distance $D_{NJst}(a_1, a_2)$ between two species a_1, a_2 is then:

$$D_{NJst}(a_i, a_j) = \frac{\sum_{T_k \in \mathbb{T}} \sum_{s=1}^{m_{a_i,T_k}} \sum_{t=1}^{m_{a_j,T_k}} D(x_{a_i,s,T_k}, x_{a_j,t,T_k})}{\sum_{T_k \in \mathbb{T}} m_{a_i,T_k} m_{a_j,T_k}} \tag{2.8}$$

The subscript T_k in the distance function indicates in which tree T_k the distance occurs. In the example in Fig. 2.7 the resulting distances are

$$\begin{aligned} D_{NJst}(a_1, a_2) &= \frac{18}{7} \approx 2.57 \ , \\ D_{NJst}(a_1, a_3) &= \frac{18}{7} \approx 2.57 \text{ and} \\ D_{NJst}(a_2, a_3) &= \frac{11}{5} = 2.2 \ . \end{aligned}$$

NJst uses the *NJ* algorithm to calculate a species tree from the distance matrix. The time complexity for calculating the distance matrix for a gene family tree with $\mathcal{O}(m)$ leaves is $\mathcal{O}(m^2)$. For a forest of K trees the time complexity for calculating the distance matrix is, therefore, $\mathcal{O}(m^2 K)$. Given n species, the complexity for *NJ* is $\mathcal{O}(n^3)$.

2.4.2. Ustar

USTAR WITH MULTIPLE SAMPLES PER TAXON [34] calculates a distance matrix D_{ustar} from a set of gene family trees $\mathbb{T}$. Like the *NJst* method (Sec. 2.4.1) it uses ever pair of leaves in the gene family trees and calculates an average. The difference is, that *ustar* calculates the average first per tree and then over all trees. The distances $D_{ustar}(a_i, a_j)$ is defined as follows

$$D_{ustar}(a_i, a_j) = \left(\sum_{T_k \in \mathbb{T}} \frac{1}{\Theta(m_{a_i,T_k} \cdot m_{a_j,T_k})} \right) \sum_{T_k \in \mathbb{T}} \frac{\sum_{s=1}^{m_{a_i,T_k}} \sum_{t=1}^{m_{a_j,T_k}} D(x_{a_i,s,T_k}, x_{a_j,t,T_k})}{m_{a_i,T_k} m_{a_j,T_k}} \tag{2.9}$$

where Θ is a step function:

$$\Theta(m_{a_i,T_k} \cdot m_{a_j,T_k})) = \begin{cases} 1 & \text{, if } m_{a_i,T_k} \neq 0 \wedge m_{a_j,T_k}) \neq 0 \\ 0 & \text{, else} \end{cases} \tag{2.10}$$

In the example in Fig. 2.7 the resulting distances are

$$\begin{aligned}
D_{ustar}(a_1, a_2) &= \frac{1}{2} \cdot \left(\frac{10}{4} + \frac{8}{3}\right) = \frac{86}{24} = 2.58\overline{3}\ , \\
D_{ustar}(a_1, a_3) &= \frac{1}{2} \cdot \left(\frac{10}{4} + \frac{8}{3}\right) = \frac{86}{24} = 2.58\overline{3} \text{ and} \\
D_{ustar}(a_2, a_3) &= \frac{1}{2} \cdot \left(\frac{10}{4} + \frac{1}{1}\right) = \frac{22}{8} = 1.75\ .
\end{aligned}$$

The averaging of the distances per species pair for every tree does not add time complexity to the averaging steps compared to *NJst*. The time complexity for calculating the distance matrix for K gene family trees with $\mathcal{O}(m)$ leaves per tree is $\mathcal{O}(m^2K)$.

2.5. Gene tree tagging

To develop methods that do not use every distance of pairs of leaves in gene family trees, we need a classification of gene leaf pairs in *orthologous* and *paralogous* pairs. A method for classifying such pairs is to label every internal node and classify each pair by a *LCA* label. In this context the labels are called *tags*. Assigning tags to nodes is called *tagging*.

2.5.1. Astral-Pro tagging and rooting

The *ASTRAL* method for paralogs and orthologs (*A-Pro*) [10] is a refinement of the *Accurate Species TRee ALgorithm* (*ASTRAL*) tool. Among other approaches, *A-Pro* uses *tagging and rooting* of untagged and unrooted gene family trees. The root is chosen by calculating a score S for every possible root r. The root with the lowest score is then selected. The score S_{q_i} for a node q_i is recursively defined as

$$S_{q_i} = \begin{cases} 0, & \text{if } q_i \text{ is leaf node, else} \\ S_{q_i,l} + S_{q_i,r} + & \begin{cases} 1, & \text{if } N(q_i,l) = N(q_i,r) \\ 2, & \text{else if } N(q_i,l) \subset N(q_i,r) \vee N(q_i,r) \subset N(q_i,l) \\ 3, & \text{else if } N(q_i,l) \cap N(q_i,r)) \neq \emptyset \\ 0, & \text{else} \end{cases} \end{cases} \tag{2.11}$$

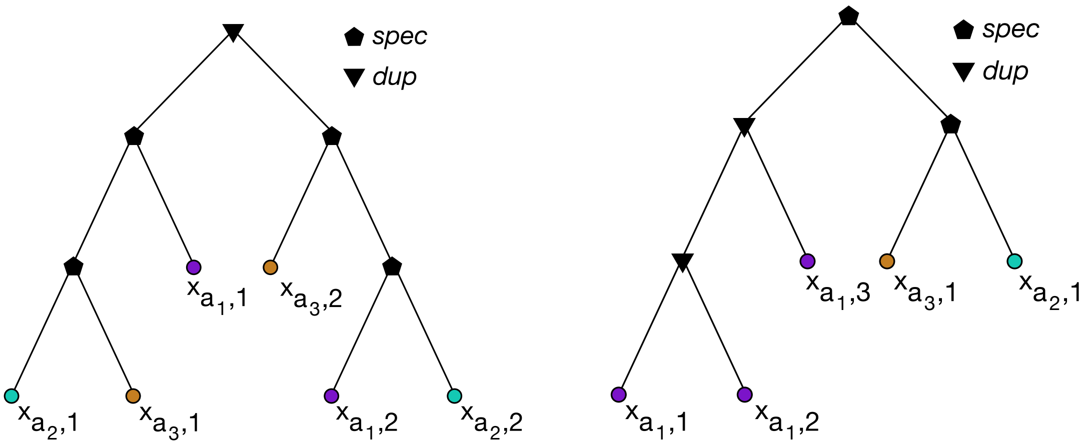
where $S_{q_i,l}$ and $S_{q_i,r}$ are the scores of the left and right children of node q_i. The function $N(q_i)$ describes the set of species present in the subtree of q_i.

The method tags every node as being either a *speciation* (*spec*) or a *duplication* (*dup*). The node q_i is tagged as *spec*, if its score is equal to the sum of scores of its child nodes $q_{i,l}$ and $q_{i,r}$: $S_{q_i} = S_{q_i,l} + S_{q_i,r}$ and as *dup* otherwise. This means that every node that has at least one species represented in two different child nodes will be tagged as *dup*. There are $\mathcal{O}(m)$ possible roots for a tree with $\mathcal{O}(m)$ leaves. The score calculation needs to compare $\mathcal{O}(m^2)$ gene leaf mappings. The time complexity for rooting and tagging a set of K gene family trees in a naïve implementation is, therefore, $\mathcal{O}(m^3K)$. Using memoization, the time complexity can be reduced to $\mathcal{O}(m^2K)$ [10]. Fig. 2.8 shows the tags of internal nodes for two rooted gene family trees.

2.5.2. MAD rooting

Another rooting method is *the Minimal Ancestor Deviation* (*MAD*) approach. It uses all pairwise *LCA*s and branch lengths in unrooted trees to root them [35]. It calculates a score S for every branch in the tree. The score S is the mean-square of the relative deviation

$$S = \left(\overline{r^2_{ab,\alpha}}\right)^{\frac{1}{2}} \tag{2.12}$$



(a) $\mathrm{A}_{\mathrm{Pro}}$ rooted gene family tree k_1

(b) $\mathrm{A}_{\mathrm{Pro}}$ rooted gene family tree k_2

Figure 2.8.: A set K of $\mathrm{A}_{\mathrm{Pro}}$ rooted gene family trees. Same colours represent that leaves are mapped to the same species.

for a possible root $r_{<p_i \circ p_j>}$ at the branch between nodes p_i and p_j. The method then chooses the root with the lowest score. The relative deviation $r_{p_s,p_t,r_{<p_i \circ p_j>}}$ between two nodes p_s and p_t is defined relative to their LCA given a root $r_{<p_i \circ p_j>}$.

$$r_{p_s,p_t,r_{<p_i \circ p_j>}} = \left| \frac{2d^b_{p_s,p_t}}{d^b_{p_s,LCA(p_s,p_t)}} - 1 \right| = \left| \frac{2d^b_{p_s,LCA(p_s,p_t)}}{d^b_{p_t,LCA(p_s,p_t)}} - 1 \right| \tag{2.13}$$

Where $d^b_{p_s,p_t}$ is the sum of branch lengths along the path between nodes p_s and p_t. The superscript b shows that real branch length values are used and not uniform branch length values. The root $r_{<p_i \circ p_j>}$ has distances $d^b_{r_{<p_i \circ p_j>},p_i} = \rho d^b_{p_i,p_j}$ and $d^b_{r_{<p_i \circ p_j>},p_j} = (1-\rho)d^b_{p_i,p_j}$ to node p_i and p_j, where ρ is chosen such as to minimize the squared relative deviations

$$r(\rho) = \sum_{p_s \in I} \sum_{p_t \in J} \left(\frac{2(d_{p_s,r_{<p_i \circ p_j>}})^2}{d^b_{p_s,p_t}} - 1 \right)^2 \tag{2.14}$$

where $I = \{\text{leaves } x_k : d_{x_k,p_i} < d_{x_k,p_j}\}$, $J = \{\text{leaves } x_k : x_k \notin I\}$. In a tree with $\mathcal{O}(m)$ leaves there are $\mathcal{O}(m)$ possible roots. For each root the *MAD* algorithm calculates $\mathcal{O}(m^2)$ relative derivations. For a set of K gene family trees, the time complexity is therefore $\mathcal{O}(m^3K)$.

2.6. Non distance-based methods

There also exist methods that do not use a distance matrix D. They will be briefly introduced here. *DupTree* [11] is a tool to estimate species trees from a set of gene family trees using a parsimony procedure. It evaluates a species tree based on the number of gene duplications, that are needed to explain the gene family trees. The species tree with the lowest such number is the estimated species tree.
A-pro [10] uses a tagging strategy on the gene family trees as described in Sec. 2.5.1 to generate gene family trees without duplications. It then uses the *Accurate Species TRee ALgorithm* (*ASTRAL*) [36] to calculate a species tree. *ASTRAL* divides the gene family trees into sets of quartets which consist of 4 leaves and finds the species tree that explains most of the quartets. *FastMulRFS* [12] is an implementation that solves the *Robinson-Foulds supertree problem* to estimate a species tree. It tries to find the species tree that has the lowest RFD between itself and the gene family trees.

3. Properties of empirical data

For a better understanding of the empirical data sets, we did a prestudy on them. For every species in each of the data sets we counted the number of gene family trees it appears in (Sec. 3.1). We calculated the average number of leaves per tree that are mapped to a species for each species (Sec. 3.2). Finally, we show the distribution of numbers of leaves per gene family tree (Sec. 3.3).

3.1. Coverage of species

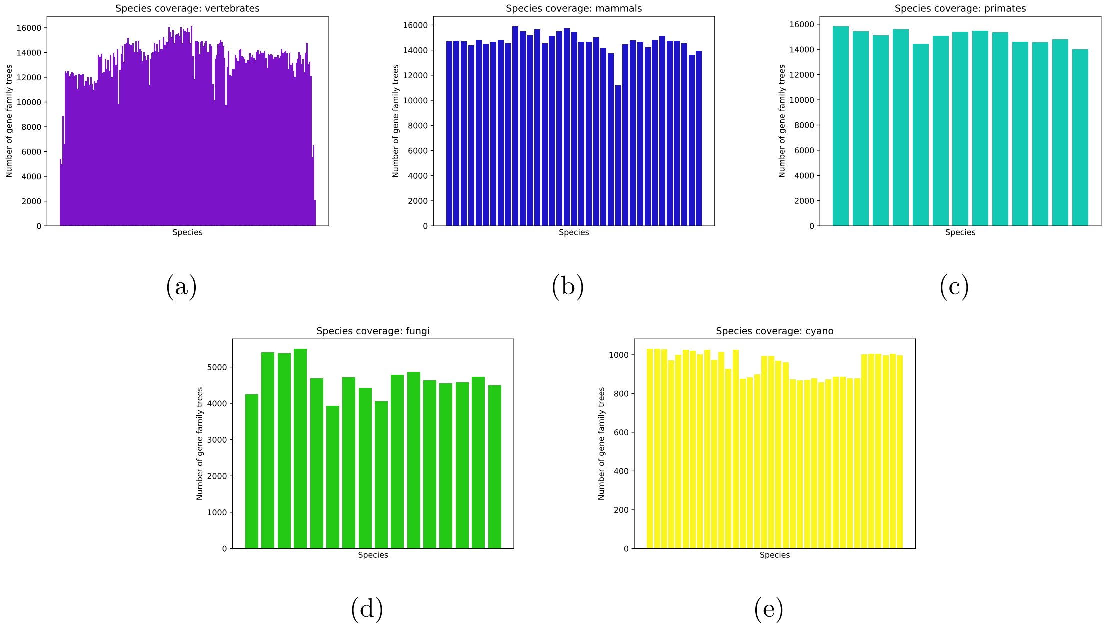
First, we counted the number of gene family trees each species appears in. Fig. 3.1 shows how the species coverage among the gene family trees are. We observe that not all species are present in the same number of gene family trees. For instance, in the vertebrates data set (Fig. 3.1a) one species is represented in 2 000 gene family trees and some other species are present in up to 16 000 gene family trees. Since this difference is substantial, our inference methods should take the difference in coverage over all gene family trees into account.

3.2. Average coverage per gene family tree

We now study the coverage of species within the gene family trees. For each species we compute the average number of leaves that are mapped to the species per gene family tree. We only consider a gene family tree for the species if at least one leaf is mapped to the species. Fig. 3.2 shows the distributions on the different data sets. There are differences among the species. Fig. 3.2a shows the biggest differences with several species just above 1 and one species going over 3. Our methods consequently have to take that into account. All averages are strictly above 1.0. Therefore, for each species, there is at least one gene family tree with more than one copy.

3.3. Gene family tree sizes

We define the size of a gene family tree by the number of its gene leaves. Because of different coverages (Sec. 3.1) and average numbers of leaves per gene family tree (Sec. 3.2) we expect different gene family tree sizes. Fig. 3.3 shows the distribution of gene family tree sizes on the empirical data sets. The distributions have a clear peak around the number of species in the data set. But there is also a large variance in the sizes. Fig. 3.3a shows that



(a)

(b)

(c)

(d)

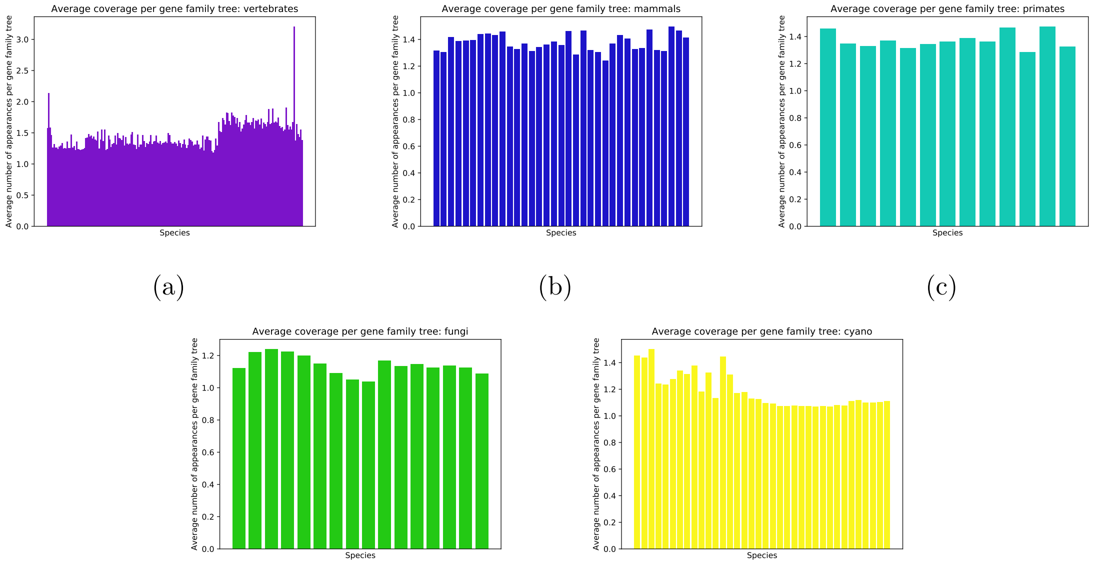
(e)

Figure 3.1.: Absolute coverage

there are small gene family trees with less than 10 leaves and large trees with more than 300 leaves. Thus, our species tree inference methods should take into account the sizes of the gene family trees.

3.4. Summary

We observed that the empirical data sets have very heterogeneous gene family trees with respect to species coverage, average coverage per tree, and tree sizes. In the next Chapter, we will present several solutions to account for this heterogeneity when inferring a species tree from gene family trees.



(a)

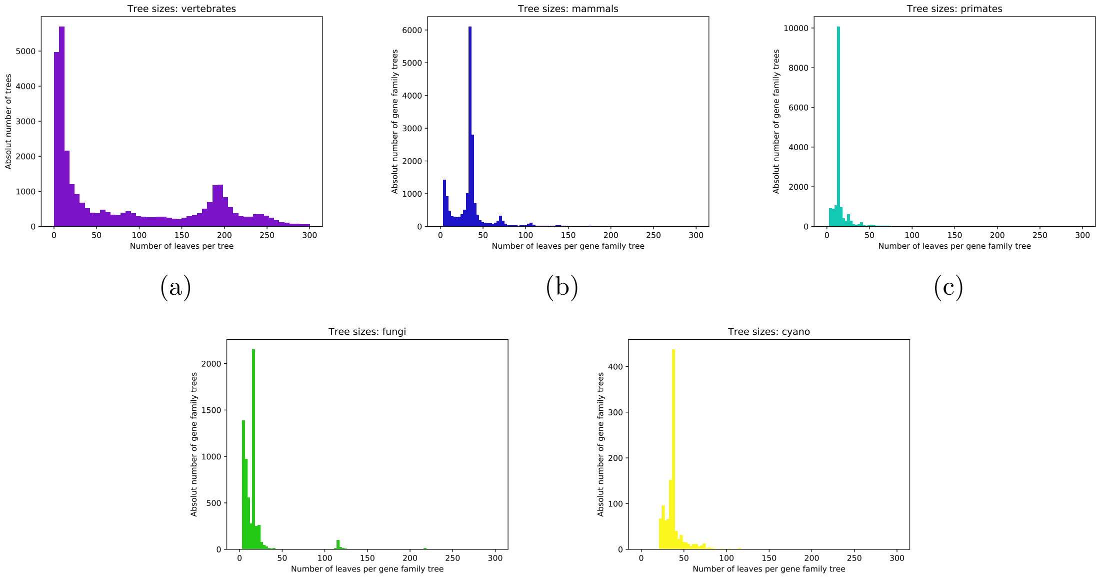
(b)

(c)

(d)

(e)

Figure 3.2.: Average coverage



(a)

(b)

(c)

(d)

(e)

Figure 3.3.: Gene family tree sizes

4. Methods

As part of this thesis we developed novel methods for calculating distance matrices. We categorise our techniques into 3 groups:

- **Picking** distances from gene family trees. Here, we do not take all distances between the leaves in the gene family trees into consideration for calculating the distance matrix.
- **Norming** and **weighting** of distances. We norm distances between leaves, such that they are better comparable among different gene family tree sizes. We also weight distances, because some distances are more probable to be close to the distance in the true species tree.
- Furthermore, we use different statistical methods for **averaging** over a set of distances within a gene family tree and over different gene family trees.

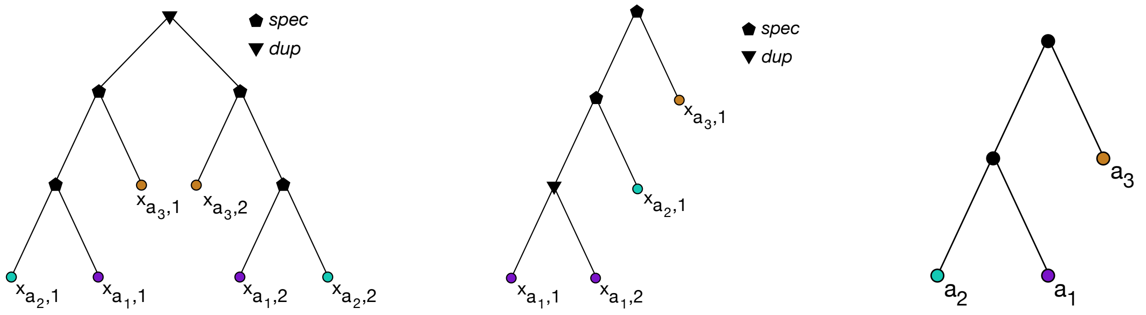
We combined the above methods of calculating the distance matrices with different methods for species tree inference, which we describe in Sec. 2.3.

4.1. Picking distances

For our distance matrix, we only want to consider orthologous gene pairs. Therefore, we attempt to filter out paralogous gene pairs. Paralogous gene pairs are a result of duplication events. The more speciations happened, after a duplication, the more additional leaf pairs will appear in the gene family tree compared to the species tree. Fig. 4.1 shows an example for that. Fig. 4.1a shows an early duplication resulting in 12 distance pairs between leaves that are mapped to different species. In contrast to this, Fig. 4.1b shows a late duplication, which results in only 5 distance pairs between leaves of different species. Under both scenarios there are pairs of paralogous genes among pairs of orthologous genes in the gene family tree as well as orthologous genes. To filter paralogous gene pairs, we use the following tagging technique.

4.1.1. Tagging

To identify paralogous gene pairs, we decide for every internal node if it is the result of a duplication event or not. Every node that is not categorised as a duplication is assumed to be a speciation. To achieve this, we use the *A-Pro* tagging technique described in Sec. 2.5.1. We classify every pair of leaves in the gene family tree that has a *LCA* tagged as *dup* as a



(a) Gene family tree T_0 with early duplication event

(b) Gene family tree T_1 with late duplication event

(c) Corresponding rooted species tree

Figure 4.1.: Species tree and corresponding gene family tree showing the impact of an early and a late *duplication*

pair of paralogous genes. Therefore, we do not use its distance for our distance matrix. To be able to tag a gene family tree it has to be rooted. We use *A-Pro* rooting (Sec. 2.5.1) and *MAD* rooting (Sec. 2.5.2) to root the unrooted gene family trees. The tagged distance function D_{tag} is defined as follows:

$$D_{tag}(a_i, a_j) = \{D(x_{a_i,s}, x_{a_j,t}) \quad | \quad (x_{a_i,s}, x_{a_j,t}) \text{ has } LCA \text{ tagged as } spec\} \tag{4.1}$$

Where $x_{a_i,s}$ is the leaf corresponding to the s$^{\text{th}}$ gene copy in species a_i's genome, and $D(x_{a_i,s}, x_{a_j,t})$ is the distance between the leaves $x_{a_i,s}$ and $x_{a_j,t}$ measured in the number of internal nodes along the path between them. Fig. 4.1b shows how duplication events create additional internal nodes between pairs. To take these nodes into account, we developed an additional approach $D_{tag,spec}$. For this approach, we consider as well only gene pairs which we consider to be orthologous. Furthermore, we count the distance without internal nodes that are tagged as *dup*. We define the number of nodes tagged as *dup* along the path from $x_{a_i,s}$ to $x_{a_j,t}$ as $m^{dup}(x_{a_i,s}, x_{a_j,t})$. The function $D_{tag,spec}$ is defined as follows:

$$D_{tag,spec}(a_i, a_j) = \{D(x_{a_i,s}, x_{a_j,t}) - m^{dup}(x_{a_i,s}, x_{a_j,t}) \,|\, (x_{a_i,s}, x_{a_j,t}) \text{ has } LCA \text{ tagged as } spec\} \tag{4.2}$$

In the example in Fig. 4.1 the resulting distances for tree T_0 are:

$$\begin{aligned} D_{tag}(a_1, a_2) &= \{1, 1\} & D_{tag,spec}(a_1, a_2) &= \{1, 1\} \\ D_{tag}(a_1, a_3) &= \{2, 2\} & D_{tag,spec}(a_1, a_3) &= \{2, 2\} \\ D_{tag}(a_2, a_3) &= \{2, 2\} & D_{tag,spec}(a_2, a_3) &= \{2, 2\} \end{aligned}$$

The distances for tree T_1 are:

$$\begin{aligned} D_{tag}(a_1, a_2) &= \{2, 2\} & D_{tag,spec}(a_1, a_2) &= \{1, 1\} \\ D_{tag}(a_1, a_3) &= \{3, 3\} & D_{tag,spec}(a_1, a_3) &= \{2, 2\} \\ D_{tag}(a_2, a_3) &= \{2\} & D_{tag,spec}(a_2, a_3) &= \{2\} \end{aligned}$$

The time complexity for using this method on K gene family trees with $\mathcal{O}(m)$ leaves per gene family tree is dominated by the rooting time complexity $\mathcal{O}(m^3K)$ as described in Sec. 2.5.1.

4.2. Norming and weighting

Gene family trees come with a different number of total leaves as Sec. 3.3 shows. They also have different number of species that are represented and different accuracy towards the real gene family tree. This is because of the following reasons:

- Certain genes are only present in a subgroup of species, because there were loss events.
- Genes are more or less often duplicated and lost.
- There is missing or imprecise data, which is caused by sequencing errors or gene family tree inference inaccuracies.

With norming and weighting techniques we try to alleviate these differences. All norming and weighting techniques take constant time per distance and, therefore, do not add time complexity to the method.

4.2.1. Using branch lengths

Instead of counting the number of internal nodes along a path, we can also use the path length by taking into account the corresponding branch length. The distance D_b is then defined as

$$D^b(x_{a_i,s}, x_{a_j,t}) = \sum_{e_{p_u,p_v} \in E_{x_{a_i,s},x_{a_j,t}}} |e_{p_u,p_v}| \tag{4.3}$$

Where $|e_{p_u,p_v}|$ is the branch length of the branch e_{p_u,p_v} and $E_{x_{a_i,s},x_{a_j,t}}$ is the set of branches along the path from $x_{a_i,s}$ to $x_{a_j,t}$. Branch lengths represent the change in the gene sequence along the path. Duplication and gene loss events along the path add and delete internal nodes, but should not affect the branch length.

4.2.2. Normalizing by gene family tree size

We normalize distances by multiplying with $\dfrac{1}{s(T_k)}$, where $s(T_k)$ is the number of leaves in a gene family tree T_k. Thereby the distances have a maximum value of 1. With this approach we attempt to represent distances between leaves on an equal scale independent of the tree size.

4.2.3. Normalizing by logarithm of gene family tree size

In a similar manner we 'normalize' distances by $\dfrac{1}{log(s(T_k))}$. So they are normalized by the minimum possible maximum depth of a tree with $s(T_k)$ species.

4.2.4. Weighting by the gene family tree size

For this technique, we assume that larger trees contain less missing data. Therefore, we multiply distances by the number of leaves $s(T_k)$ of the gene family tree T_k. When averaging the weighting factor $s(T_k)$ is also considered. For example, the *mini* technique with additional *weighting by the gene family tree size* function $D^{w_s}_{mini}$, where the superscript w_s indicates the corresponding weighting, is:

$$D^{w_s}_{mini}(a_i, a_j) = \frac{\sum_{T_k \in \mathbb{T}} \min_{s=1}^{m_{a_i,T_k}} \min_{t=1}^{m_{a_j,T_k}} D(x_{a_i,s,T_l}, x_{a_j,t,T_k}) s(T_k)}{\sum_{T_k \in \mathbb{T}} \Theta(m_{a_i,T_k} m_{a_j,T_k}) s(T_k)} \tag{4.4}$$

4.2.5. Weighting by the number of covered species in the gene family tree

Weighting by the number of leaves gives more weight to gene family trees with more duplication events. But we attempt to not give too much weight to paralogous gene pair distances. Because of that, in this technique we weight by the number of species $|N(T_k)|$ present in a gene family tree T_k. Thereby we intend to weight trees that cover more species stronger than trees that only cover a small subgroup. We also consider the weighting in the averaging. For example, the *mini* technique with additional *weighting by number of covered species in gene family tree* function $D^{w_{|N|}}_{mini}$, where the superscript $w_{|N|}$ indicates the corresponding weighting, is:

$$D^{w_{|N|}}_{mini}(a_i, a_j) = \frac{\sum_{T_k \in \mathbb{T}} \min_{s=1}^{m_{a_i,s,T_k}} \min_{t=1}^{m_{a_j,T_k}} D(x_{a_i,s,T_k}, x_{a_j,t,T_k})|N(T_k)|}{\sum_{T_k \in \mathbb{T}} \Theta(m_{a_i,T_k} m_{a_j,T_k})|N(T_k)|} \tag{4.5}$$

4.3. Statistical average

As results in Sec. 3.1 and Sec. 3.2 show, each species is present in a different number of gene family trees and in a different number per gene family tree. To take that into account we use statistical averaging per gene family tree and over all gene family trees. The *NJst* method uses the arithmetic mean over all distances per species pair as statistic for averaging. But there are more statistics that can potentially be better representations of the average distance per species pair. We distinguish between finding a statistical average per gene family tree (Sec. 4.3.1) and over the set of all gene family trees (Sec. 4.3.2).

4.3.1. Statistical average per gene family tree

For the statistical average per gene family tree techniques we average the set of distances and per tree per species pair, before adding them to the set of distances per species tree over all trees. We use two different techniques for that.

4.3.1.1. *ustar* averaging

We calculate the arithmetic mean of the set of distances per species pair as described in Sec. 2.4.2. The distance between two species a_i, a_j per tree is

$$D^{T_k}_{ustar}(a_i, a_j) = \frac{\sum_{s=1}^{m_{a_i,T_k}} \sum_{t=1}^{m_{a_j,T_k}} D(x_{a_i,s,T_k}, x_{a_j,t,T_k})}{m_{a_i,T_k} m_{a_j,T_k}} \tag{4.6}$$

4.3.1.2. Minimum distance

Furthermore, we use the smallest distance $D(x_{a_i,s}, x_{a_j,t})$ within a gene family tree between leaves mapped to species a_i and a_j. With this technique we filter out all paralogous gene pairs as they yield larger distances than pairs of orthologous genes. The resulting distance $D_{min}(a_1, a_2)$ for species a_1, a_2 in a tree T_k is:

$$D^{T_k}_{min}(a_i, a_j) = \min_{s=1}^{m_{a_i,T_k}} \min_{t=1}^{m_{a_j,T_k}} D(x_{a_i,s,T_k}, x_{a_j,t,T_k}) \tag{4.7}$$

The step function Θ and number of occurrences m_{a_i,T_k} are defined as in Sec. 2.4. Finding the minimum takes linear time in the number of pairs and does, therefore, not increase the time complexity of the method. In the example in Fig. 4.1, the resulting distances for tree T_0 are

$$D^{T_0}_{min}(a_1, a_2) = 1$$
$$D^{T_0}_{min}(a_1, a_3) = 2$$
$$D^{T_0}_{min}(a_2, a_3) = 2$$

The distances for tree T_1 are

$$D^{T_1}_{min}(a_1, a_2) = 2$$
$$D^{T_1}_{min}(a_1, a_3) = 3$$
$$D^{T_1}_{min}(a_2, a_3) = 2$$

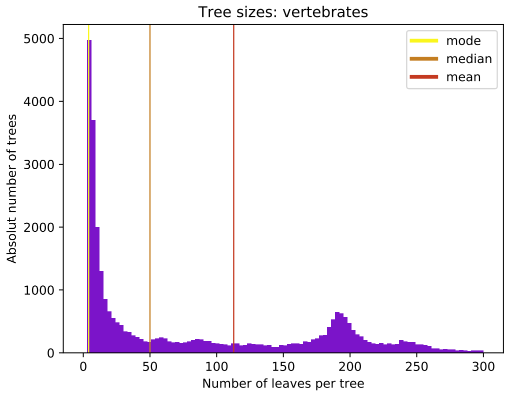


Figure 4.2.: Graphical representation of arithmetic mean, median and mode for the distribution of tree sizes for the vertebrates data set 5.1. The mode is at 4, the median is at 50, and the mean is at 112.8.

4.3.2. Statistical averaging over all gene family trees

The *NJst* and *ustar* methods use the arithmetic mean as the statistical average over the set of distances per species pair from all gene family trees. In the following we use two different statistical values, the median (Sec. 4.3.3) and the mode (Sec. 4.3.4).

4.3.3. Median

The *median* is the value which lies in the middle of a sorted list of the data. Fig. 4.2 shows that in a graphical representation of a distribution, the *median* cuts the area in two equal sized halves. It is more robust against outliers than the arithmetic mean. The distance between species a_i and a_j is chosen as the *median* from the set $\{D^{T_k}_{a_i,a_j}\}_{\mathbb{T}} = \{D^{T_k}(a_i, a_j)|T_k \in \mathbb{T}\}$ of distances from all gene family trees $\mathbb{T}$.

$$D_{median} = median(\{D^{T_k}_{a_i,a_j}\}_{\mathbb{T}}) \tag{4.8}$$

4.3.4. Mode

Additionally to the arithmetic mean and the median, we also used the *mode* for averaging. It is the value that appears most often in a data set. The distance between species a_i and a_j is chosen as the *mode* from the set of distances of all gene family trees. Fig. 4.2 shows that the mode is represented by the highest peak of a histogram of a distribution.

$$D_{mode} = mode(\{D^{T_k}_{a_i,a_j}\}_{\mathbb{T}}) \tag{4.9}$$

5. Experiments & Results

In this Chapter we describe the simulated data sets (Sec. 5.1.2) and empirical data sets (Sec. 5.1.1), which we used to test our methods. Sec. 5.2.1 presents the accuracy and execution times for our methods and for existing tools on the simulated data sets. Sec. 5.2.2 contains the results for the empirical data sets.

5.1. Experimental setup

To test our new techniques, we implemented them in *C*, which is accessible online[1]. Our tool can combine several of the techniques, which we describe in Ch. 4. We tested the following methods which are based on the *NJst* (Sec. 2.4.1) method. We call them *NJst+* methods:

- $NJst_b$: using branch lengths (Sec. 4.2.1) and *Neighbor Joining* (*NJ*) (Sec. 2.3.1)
- $NJst_{n_s}$: normed by gene family size (Sec. 4.2.2) and *NJ* (Sec. 2.3.1)
- $NJst_{n_{log(s)}}$: normed by the logarithm of gene family size (Sec. 4.2.3) and *NJ* (Sec. 2.3.1)
- *ustarNJ*: ustar (Sec. 2.4.2) averaging and *NJ* (Sec. 2.3.1)
- *miniNJ*: mini (Sec. 4.3.1.2) technique and *NJ* (Sec. 2.3.1)
- $tagNJ_{A-pro}$: tagging (Sec. 4.1.1) with A-pro rooting (Sec. 2.5.1) and *NJ* (Sec. 2.3.1)

We also evaluate the following methods, which all contain the *mini* technique. We call them *mini* methods:

- *miniNJ*: mini (Sec. 4.3.1.2) technique and *NJ* (Sec. 2.3.1)
- $miniNJ_{w_s}$: mini (Sec. 4.3.1.2) technique, weighting by gene family size (Sec. 4.2.4), and *NJ* (Sec. 2.3.1)
- $miniNJ_{w_{|N|}}$: mini (Sec. 4.3.1.2) technique, weighting by number of covered species in gene family (Sec. 4.2.5), and *NJ* (Sec. 2.3.1)
- *miniUPGMA*: mini (Sec. 4.3.1.2) technique and *UPGMA* (Sec. 2.3.2)
- *miniWPGMA*: mini (Sec. 4.3.1.2) technique and *WPGMA* (Sec. 2.3.2)
- *miniMedianNJ*: mini (Sec. 4.3.1.2) technique, median (Sec. 4.3.3) and *NJ* (Sec. 2.3.1)

[1] `https://github.com/SchadePaul/MasterThesis`

- *miniModeNJ*: mini (Sec. 4.3.1.2) technique, mode (Sec. 4.3.4), and *NJ* (Sec. 2.3.1)

We also assess the following methods, which all contain the *tagging* technique. We call them *tagging* methods:

- $tagNJ_{A-pro}$: tagging (Sec. 4.1.1) with A-pro rooting (Sec. 2.5.1) and *NJ* (Sec. 2.3.1)
- $tagSpecNJ_{A-pro}$: tagging (Sec. 4.1.1) only counting *spec* with A-pro rooting (Sec. 2.5.1), and *NJ* (Sec. 2.3.1)
- $tagNJ_{MAD}$: tagging (Sec. 4.1.1) with MAD rooting (Sec. 2.5.2) and *NJ* (Sec. 2.3.1)
- $tagNJ_{MAD,n_s}$: tagging (Sec. 4.1.1) with MAD rooting (Sec. 2.5.2), normed by gene family tree size (Sec. 4.2.2), and *NJ* (Sec. 2.3.1)
- $tagNJ_{MAD,n_{log(s)}}$: tagging (Sec. 4.1.1) with MAD rooting (Sec. 2.5.2), normed by gene family tree size (Sec. 4.2.3), and *NJ* (Sec. 2.3.1)
- $tagNJ_{MAD,w_s}$: tagging (Sec. 4.1.1) with MAD rooting (Sec. 2.5.2), weighted by gene family tree size (Sec. 4.2.4), and *NJ* (Sec. 2.3.1)
- $tagSpecNJ_{MAD}$: tagging (Sec. 4.1.1) only counting *spec* with MAD rooting (Sec. 2.5.2) and *NJ* (Sec. 2.3.1)

We used our tool variations to estimate a species tree from a set of gene trees. We calculated the *rRFD* of the estimated tree to the true species tree to assess accuracy. We also measured running times of the tools. We did the same with the existing tools *A-Pro* [10], *DupTree* [11], and *FastMulRFS* [12][37][38][39] for the sake of comparison.

5.1.1. Empirical data sets

Tab. 5.1 describes the empirical data sets, including the number of species in the species tree and the number of gene family trees. The data sets are taken from the HOGENOM [40] and ENSEMBL [41] databases, and the *A-Pro* [10] paper.

Name	Number of species	Number of gene family trees	Database
Fungi	16	7 180	A-pro [10]
Cyano	36	1 099	HOGENOM [40]
Primates	13	16 670	ENSEMBL [41]
Mammals	35	18 525	ENSEMBL [41]
Vertebrates	193	33 396	ENSEMBL [41]

Table 5.1.: Empirical data sets

5.1.2. Simulated data sets

To generate additional data sets to test the methods on, we also simulated data sets in 3 steps.

- First, we simulated sets of species trees and their gene family trees using the *SimPhy* software package [15]. We show the parameters, which we used for SimPhy, in Tab. 5.2.
- On the gene family trees, we simulated DNA-sequences with *INDELible* [16]. The parameters for INDELible are provided in Tab. 5.3.
- In the last step we inferred gene family trees from the DNA-sequences using *ParGenes* [7][8] with parameters as specified in Tab. 5.4.

Parameter name	Parameter value
Standard parameters (STANDARD)	
Speciation rate	5e-9
Number of gene family trees per species tree	1 000
Number of species	25 + an outgroup
Duplication rate (events/generation)	4.9e-10
Loss rate relative to duplication rate	1
Effective population size	4.7e+8
Ingroup divergence to the ingroup ratio	1.0
Generations	LogN(21.25,0.2)
Global substitution rate	LogN(-21.9,0.1)
Lineage specific rate gamma shape	LogN(1.5,1)
Gene family specific rate gamma shape	LogN(1.551533,0.6931472)
Gene tree branch specific rate gamma shape	LogN(1.5,1)
Seed	9644
Controlling Duplication and Loss Rates (5 × 4 conditions) (DUPLOS)	
Duplication rate (events/generation)	4.9e-10, 2.7e-10, 1.9e-10, 5.2e-11, 0
Loss rate relative to duplication rate	1, 0.5, 0.1, 0
Controlling Duplication and ILS rate (3 × 4 conditions) (ILS)	
Duplication rate (events/generation)	4.9e-10, 1.9e-10, 0
Effective population size	4.7e+8, 1.9e+8, 4.8e+7, 1e+4
Controlling number of species (SPECIES)	
Number of taxa	10, 25, 35, 50, 100 + an outgroup
Controlling number of gene family trees per species tree (GENES)	
Number of locus trees per species tree	100, 250, 1 000, 2 500, 10 000

Table 5.2.: Parameters used for *SimPhy*

We used different parameters for different data sets. We have a STANDARD data set with the default parameters. Based on that we simulated a data set DUPLOS with varying duplication and loss rates. Additionally we simulated the data set ILS with varying duplication rates and population sizes. The data set SPECIES has different number of species per species tree. Last, there is the data set GENES with different number of gene family trees per species tree. For every set of parameters, in Simphy, we simulated 50 sets of species trees with their corresponding gene family trees. For every set of gene family trees we simulated DNA-sequences with a length of 100 and 500 base pairs (bp). The simulating of DNA sequences and inferring of gene family trees induces inaccuracies to the gene family trees compared to the true gene family trees. With a shorter sequence length of 100 bp, the inaccuracy increases. In addition, we also conducted additional experiments

Parameter name	Parameter value
Sequence length	100, 500
Sequence base frequencies	Dirichlet(A=36,C=26,G=28,T=32)
Sequence transition rates	Dirichlet(TC=16,TA=3,TG=5,CA=5,CG=6,AG=15)

Table 5.3.: Parameters used for *INDELible*

Parameter name	Parameter value
Sequence type	nucleotide
Model	RAxML global parameter option GTR

Table 5.4.: Parameters used for *ParGenes*

directly on the gene family trees (true gt) from Simphy without simulating DNA sequences and inferring gene trees.

The CUT data set

To model missing data in the simulated data sets, we removed data from the data sets. To do so, we selected a random node from the rooted true gene trees and removed the DNA data contained in the subtree below that node. We repeated this to get different level of missing data. Including the trees without missing data the CUT data set contains 10 levels of missing data. We limited all data sets to the same total number of gene family tree leaves to not let the total numbers of gene family tree leaves influence the results. To do so, we removed all gene family trees that added more gene family tree leaves to the data set than needed.

5.2. Results

In this Section we provided results for the simulated (Sec. 5.2.1) and empirical data sets (Sec. 5.2.2).

5.2.1. Results on simulated data sets

In this Subsection we present the results for the different simulated data sets STANDARD (Sec. 5.2.1.1), SPECIES (Sec. 5.2.1.2), GENES (Sec. 5.2.1.3), ILS (Sec. 5.2.1.4), DUPLOS (Sec. 5.2.1.5), and CUT (Sec. 5.2.1.6).

5.2.1.1. The STANDARD data set

The STANDARD data set provides an overview over the results of all methods we intend to test. Fig. 5.1 shows the rRFD for the *NJst+* methods.

Norming

The different norming techniques behave very differently. *Using branch lengths* (Sec. 4.2.1), increases the rRFD to ≈ 0.081 (8.1 %) compared to ≈ 0.052 (5.2 %) for the *NJst* method. Using branch lengths does not improve the method. *Normalizing by gene family tree size* (Sec. 4.2.2) yields an even lower accuracy with a rRFD of ≈ 0.164 (16.4 %). Thus, norming seems to not be able to handle distances well. *Normalizing by the logarithm of the gene family tree size* (Sec. 4.2.3) yields a marginally improved rRFD of ≈ 0.050 (5 %) than *NJst*. But its influence is not substantial, since the logarithm of gene family tree sizes does not vary substantially among the trees.

Averaging per gene family tree

The *ustar* averaging yields higher rRFDs than *NJst*. Having gene family trees with many distance pairs contribute the same weight as gene family trees with few distance pairs to the overall distance appears to be less accurate. *MiniNJ* improves the accuracy despite the large discrepancy in the averaging per gene family trees. Therefore, we will evaluate

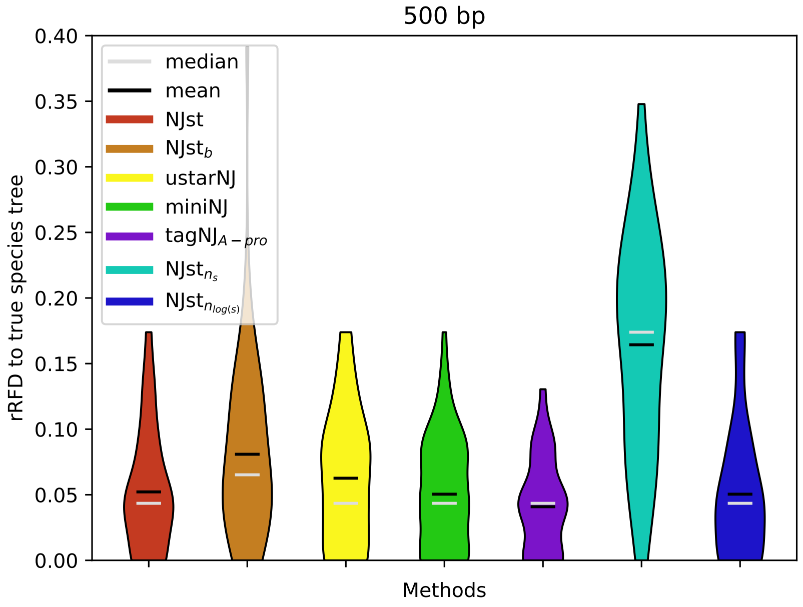


Figure 5.1.: Comparison of the rRFD for *NJst+* methods on the STANDARD (500 bp) data set. Additional plots are available in App. A.

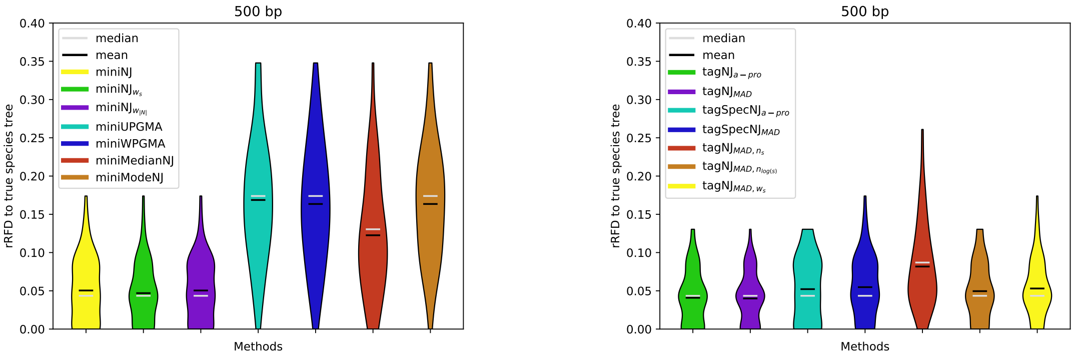
some *mini* methods on this data set. Fig. 5.2a shows the rRFD of the *mini* methods on the STANDARD data set. Using other clustering methods than *NJ* and different averaging approaches yield much worse rRFDs. All of their rRFD values are at least twice the value of *miniNJ*. The weighting techniques have low impact on the accuracy. *Weighting by gene family tree size* improves the rRFD and is the most accurate method among the *mini* methods.

Tagging

The $tagNJ_{A-pro}$ technique can lower the rRFD even though many distances between gene tree leaves are not considered for the overall distance calculation at all. Therefore, we will also evaluate additional *tagging* methods on this data set. Fig. 5.2b shows the rRFD of the *tagging* methods. Using *tagSpec* instead of *tag* yields lower accuracy. It appears that counting nodes that are labeled *dup* improves accuracy. Different norming and weighting techniques do not improve the results further. Using *MAD* rooting instead of *A-pro* rooting yields a slightly lower rRFD and is, therefore, the most accurate method among the *tagging* methods.

Number of pairs used

With the mini and tag techniques we attempted to not count the distances among paralogous gene pairs in the distance matrix. Fig. 5.3 shows how many distances in all gene family trees the mini and tag techniques used to calculate the distance matrix. The number of distances is relative to the number of all leave pairs, which is equal to the number of distances *NJst* uses. While mini only uses less than 10% of the pairs, the tag techniques with different rooting strategies use around 20% of the pairs. We conclude that the mini technique filters out most paralogous gene pairs, but also filters out many orthologous gene pairs. The tag technique filters less orthologous gene pairs. This is why the tagging methods show better results than the mini methods. We conclude that because of the low number of used pairs, weighting helps to improve the accuracy of the mini technique. The



(a) Comparison of the rRFD for the *mini* methods on the STANDARD data set

(b) Comparison of the rRFD for the *tagging* methods on the STANDARD data set

Figure 5.2.: Comparison of the rRFD for the *mini* and *tagging* methods on the STANDARD (500 bp) data set. Additional plots are provided in App. A.

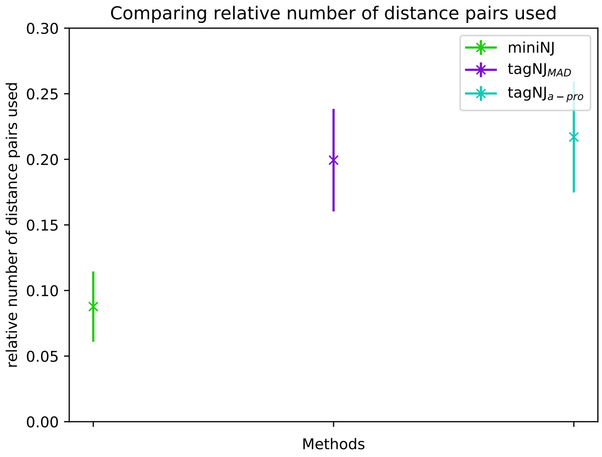


Figure 5.3.: Comparison of the mean number of distance pairs with standard deviation used relative to NJst method on the STANDARD data set

tag technique on the other hand, already uses enough distances such that is does not profit from the weighting.

Sequence length

For all following results we will only consider at *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$. While *NJst* is used as a reference, *miniNJ*$_{w_s}$ and *tagNJ*$_{MAD}$ are the best among our methods. Fig. 5.4 shows the results for different sequence lengths used in the simulations. *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$ yield worse accuracy for smaller sequence length. This is expected since there is less information.

Comparison with other tools

Fig. 5.5a shows the accuracy of our methods compared to other tools. *tagNJ*$_{MAD}$ has better accuracy than any other tool we tested. *A-pro* has the second best accuracy with a mean of ≈ 0.0461. The *tagNJ*$_{MAD}$ and *miniNJ*$_{w_s}$ methods outperform *DupTree* and *FastMulRFS*, which have a mean rRFD of ≈ 0.0583 and ≈ 0.1603 respectively.

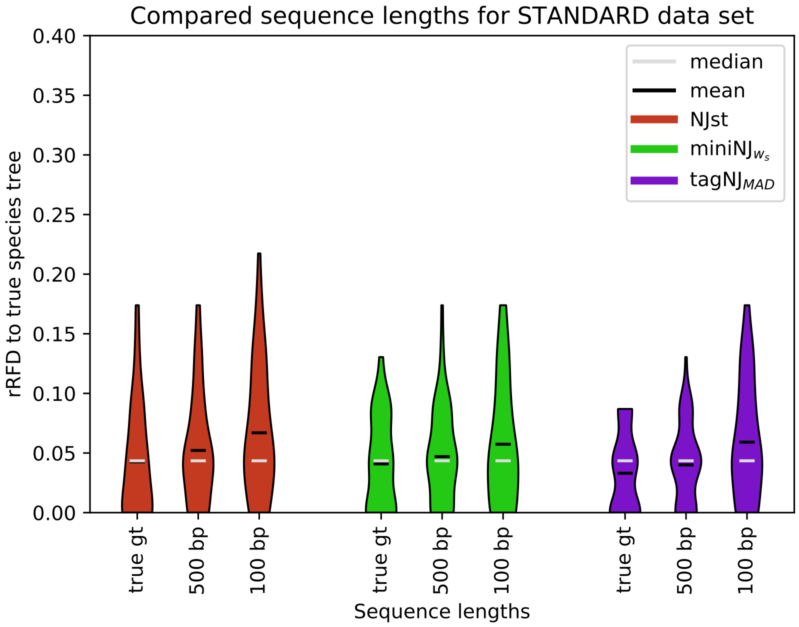
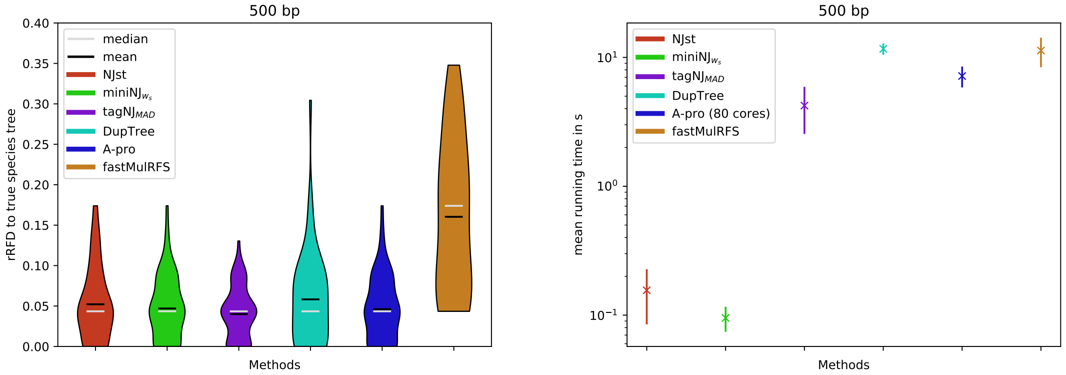


Figure 5.4.: Comparison of the rRFD for true gene trees and sequence lengths of 100 bp and 500 bp

Fig. 5.5b shows mean running times for our best methods compared to those of other tools. The *miniNJ*$_{w_s}$ method outperforms all other tools with a lower run time for calculating the distance matrix. Our implementation of *tagNJ*$_{MAD}$ is also faster than the other tools we tested. The competing tools take around 1 *s* of runtime with *A-pro* being faster than the other two. While *DupTree* and *FastMulRFS* are implemented as sequential algorithms, *A-pro* is parallelized and uses 80 cores at the same time. Despite the parallelization of *A-pro*, sequential *miniNJ*$_{w_s}$ attains a speedup of 75 compared to *A-pro*.

5.2.1.2. The SPECIES data set

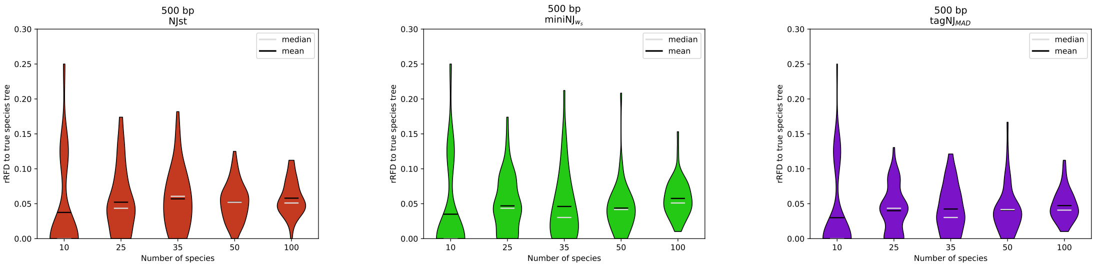
Fig. 5.6 shows the accuracy for *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$ with different number of species in the species tree, but all other parameters fixed. For increasing number of species our methods yield lower accuracy. This can explained by a the decreasing ratio of gene family trees per species. To estimate a species tree with a higher number of species with the same accuracy, a larger number of gene family trees would be needed as results in Sec. 5.2.1.3 suggest. Fig. 5.7a shows mean running times for *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$ with different numbers of species per species tree n. We calculated best fitting curves of shape $f(n) = a \cdot n^b$ to assess the runtime increase. We base the empirical run time estimate on the fitted exponent b. *MiniNJ*$_{w_s}$ has an empirical run time estimate of $\mathcal{O}(n^{2.0})$, and *NJst* has an empirical run time estimate of $\mathcal{O}(n^{2.1})$. *TagNJ*$_{MAD}$ behaves worse with an empirical run time estimate of $\mathcal{O}(n^{2.3})$. The NJ algorithm has a theoretical time complexity of $\mathcal{O}(n^3)$ (Sec. 2.3.1). Therefore, the distance matrix calculation dominates the time. Theoretical complexities depend on the number of gene family trees K and the number of leaves per gene family tree m. The theoretical complexities for distance matrix calculation for miniNJ$_{w_s}$ and NJst is $\mathcal{O}(m^2K)$ and $\mathcal{O}(m^3K)$ for tagNJ$_{MAD}$. Since we do not have information about the relationship between the number of species n and the number of leaves per gene family tree m, we can not compare the empirical run time estimates with the theoretical time complexities.



(a) Comparison of the rRFD for different tools on the STANDARD (500 bp) data set

(b) Comparison of the running time for different tools on the STANDARD (500 bp) data set

Figure 5.5.: Results for different tools on the STANDARD (500 bp) data set. Additional plots are provided in App. A.

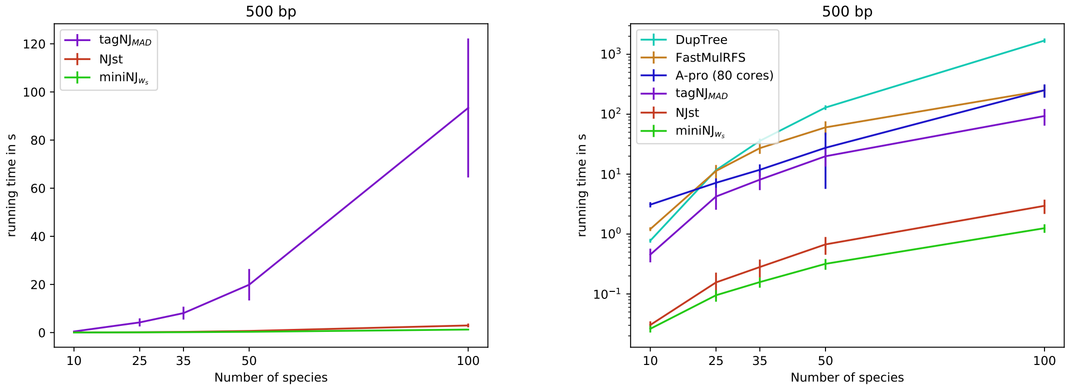


(a) Comparison of the rRFD for different numbers of species using the NJst method on the SPECIES (500 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the SPECIES (500 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the SPECIES (500 bp) data set

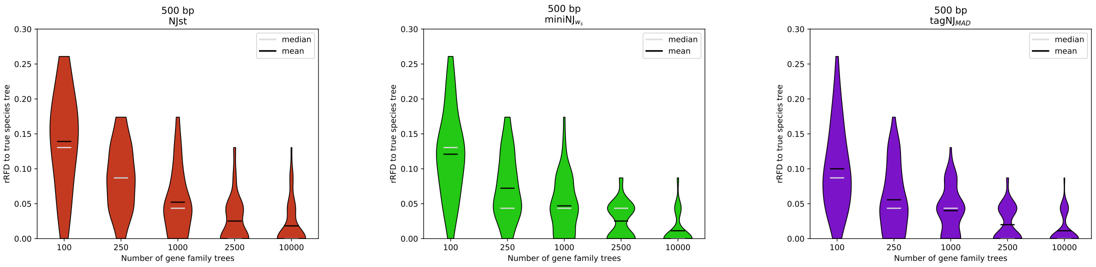
Figure 5.6.: Comparison of the rRFD for different methods on the SPECIES (500 bp) data set. Additional plots are provided in App. B.



(a) Comparing mean running times with standard deviation for different methods on the SPECIES (500 bp) data set

(b) Comparing mean running times with standard deviation for different tools on the SPECIES (500 bp) data set on a logarithmic y-axis scale

Figure 5.7.: Running times for different methods on the SPECIES (500 bp) data set



(a) Comparison of the rRFD for different numbers of species using the NJst method on the GENES (500 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the GENES (500 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the GENES (500 bp) data set

Figure 5.8.: Comparison of the rRFD for different methods on the GENES (500 bp) data set. Additional plots are provided in in App. C.

Fig. 5.7b shows also the mean running time on a logarithmic scale for different number of species for our methods and the other tools we tested. The empirical run time estimates are *FastMulRFS*: $\mathcal{O}(n^{2.1})$, *A-pro*: $\mathcal{O}(n^{3.1})$, and *DupTree*: $\mathcal{O}(n^{3.7})$. We see that *miniNJ*$_{w_s}$ is by far the fastest followed by *NJst*. The tagging technique outperforms the *DupTree*, *A-pro*, and *FastMulRFS* for the SPECIES data sets.

5.2.1.3. The GENES data set

Fig. 5.8 shows the accuracy for *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$ for different numbers of gene family trees and all other parameters being fixed. For an increasing number of gene family trees our methods yields higher accuracy. A higher number of gene family trees yields more information. Therefore, the higher accuracy is not surprising. Fig. 5.9 shows the mean running time of *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$ with different numbers of gene family trees. All methods show linear running time as predicted by the theoretical time complexity.

5.2.1.4. The ILS data set

Fig. 5.10 shows the accuracy of *NJst*, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$ for different population sizes. With smaller population size the rRFD decreases for our methods. For a very small

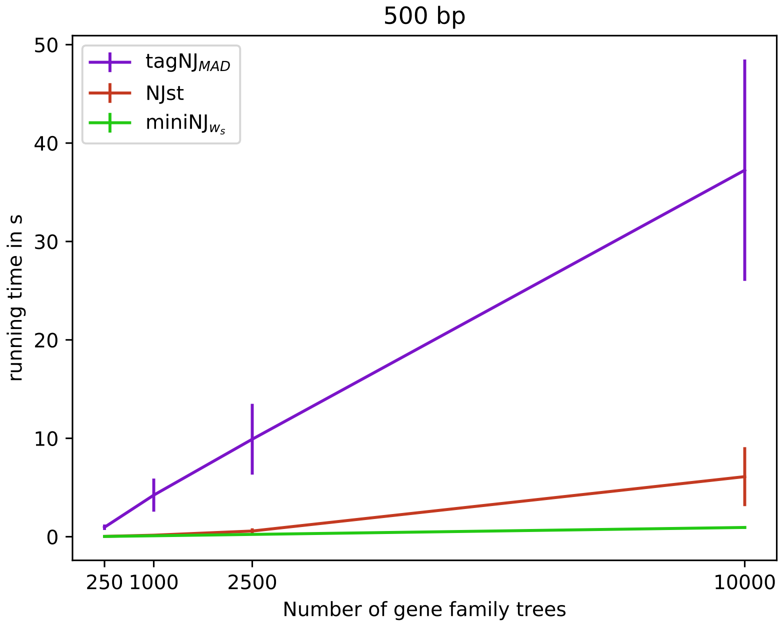


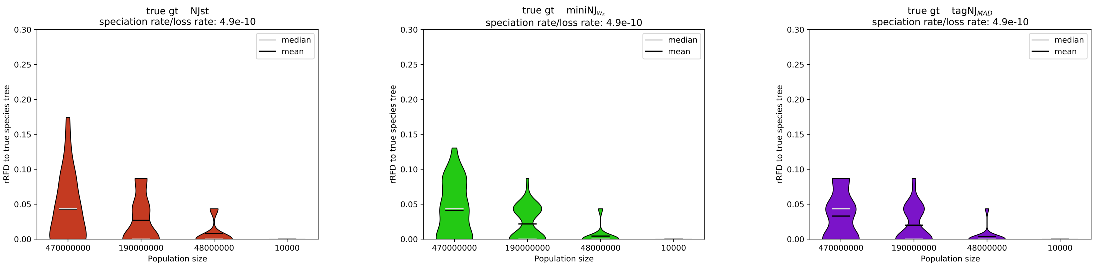
Figure 5.9.: Comparison of the mean running times with standard deviation for different methods on the GENES (500 bp) data set

population of 10 000 all estimated species trees are identical to the real species trees. The increase in accuracy is expected, since the *ILS* rate decreases with smaller population sizes.

We chose the true gene family trees (true gt) for generating the results with this data set, because the inferred trees did not show increasing accuracy with decreasing ILS so clearly. Fig. 5.11 shows the results for the (500 bp) data set. The differences between the inferred gene family trees and the true gene family trees appear to increase with decreasing population size.

5.2.1.5. The DUPLOS data set

Fig. 5.12a shows accuracy of our methods for different gene loss rate to gene duplication rate ratios with a constant gene duplication rate of $4.9 \cdot 10^{-10}$. Accuracy tends to increase

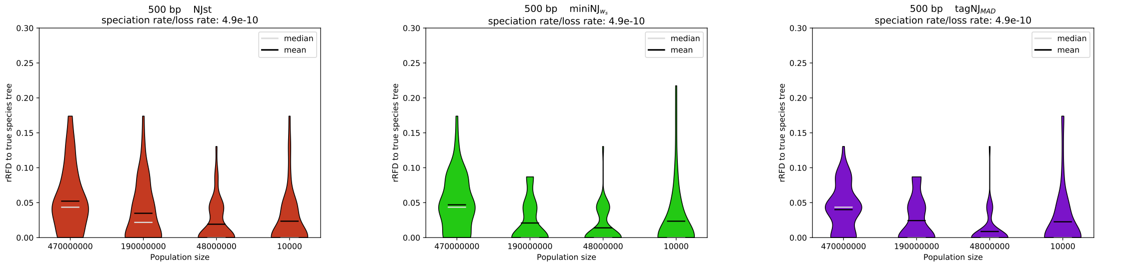


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (true gt) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (true gt) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (true gt) data set

Figure 5.10.: Comparison of the rRFD for different methods on the ILS (true gt) data set. Additional plots are provided in App. D.

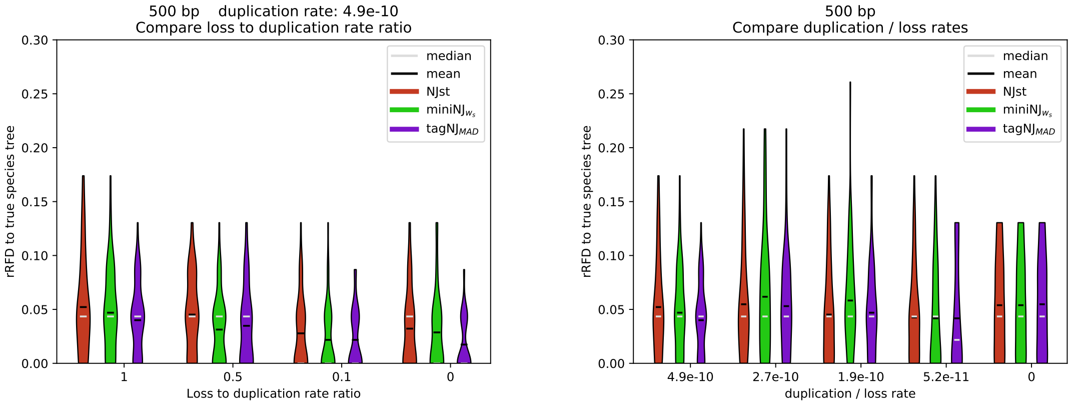


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (500 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (500 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (500 bp) data set

Figure 5.11.: Comparison of the rRFD for different methods on the ILS (500 bp) data set. Additional plots are provided in App. D.



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different duplication and loss rates

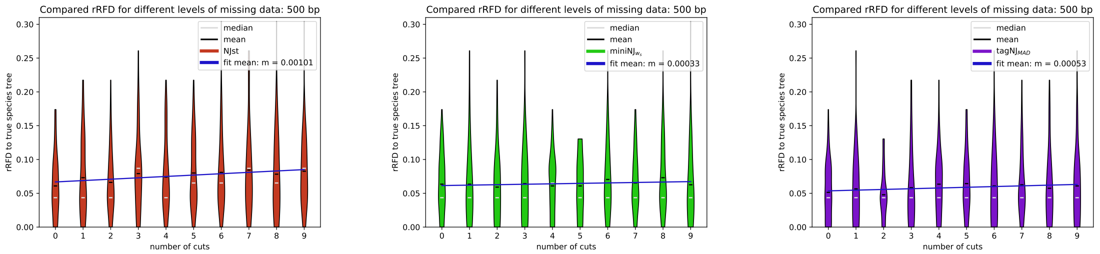
Figure 5.12.: Results for the DUPLOS data set. Additional plots are provided in App. E.

with lower gene loss rate. This could mean that *loss* events are problematic for the methods. The analyses of the CUT (Sec. 5.2.1.6) data set reveals more details, since loss and missing data can not be distinguished from each other by our methods.

Fig. 5.12b shows the accuracy for *NJst*, $miniNJ_{w_s}$, and $tagNJ_{MAD}$ for different gene duplication and gene loss rates. The gene duplication rate and gene loss rate are equal for this data set. Accuracy varies for different rates. At rate 0 all methods yield the same results. With no gene pairs in the data set no method will identify a paralogous gene pair and all methods use all gene pairs.

5.2.1.6. The CUT data set

Fig. 5.13 shows the accuracy of the *NJst* (Fig. 5.13a), $miniNJ_{w_s}$ (Fig. 5.13b), and $tagNJ_{MAD}$ methods for the CUT data set. All methods tend to show decreasing accuracy with increasing number of missing data. The linear fit line shows the highest gradient for the *NJst* method. This could mean that $miniNJ_{w_s}$ and $tagNJ_{MAD}$ better handle missing data. A reason could be that *NJst* uses every distance pair, so every missing sequence will affect the resulting distance matrix. In contrast, $miniNJ_{w_s}$ and $tagNJ_{MAD}$ do not consider every pair for the distance matrix, so there is a probability that the missing data was not used for the distance matrix.

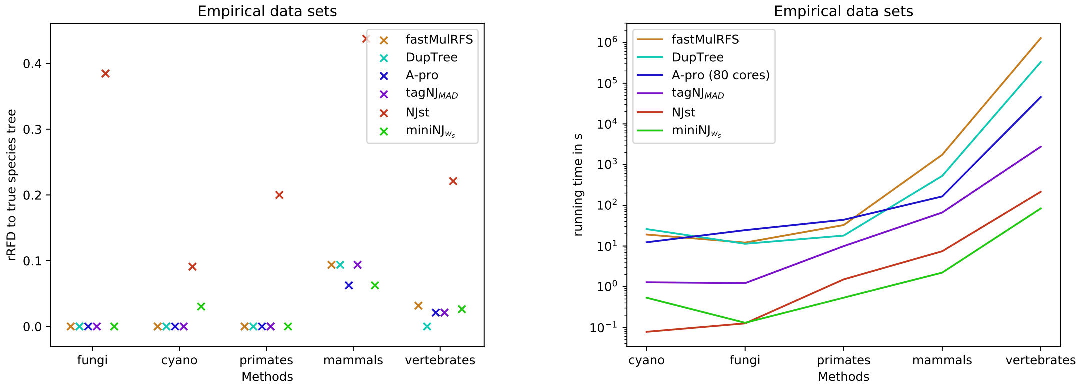


(a) Comparison of the rRFD for different numbers of cuts with $NJst$ method

(b) Comparison of the rRFD for different numbers of cuts with miniNJ$_{w_s}$ method

(c) Comparison of the rRFD for different numbers of cuts with tagNJ$_{MAD}$ method

Figure 5.13.: Comparison of the rRFD for different numbers of cuts with different methods. Additional plots are provided in App. F.



(a) Comparison of the rRFD for different methods and tools on empirical data sets

(b) Comparisons of running times on a logarithmic scale for different methods and tools on empirical data sets

Figure 5.14.: Results of the empirical data set

5.2.2. Results on empirical data sets

Fig. 5.14a shows the accuracy of our *NJst*, *miniNJ*$_{w_s}$, *tagNJ*$_{MAD}$ methods, and the *A-pro*, *DupTree*, and *FastMulRFS* tools on empirical data sets. *NJst* shows the worst accuracy over all data sets. *miniNJ*$_{w_s}$ and *tagNJ*$_{MAD}$ can compete with the other tools regarding accuracy. They almost always attain the same accuracy.

Fig. 5.14a shows the running times of *NJst*, *miniNJ*$_{w_s}$, *tagNJ*$_{MAD}$, and *A-pro*, *DupTree*, and *FastMulRFS* on empirical data sets on a logarithmic scale. While *NJst* and *miniNJ*$_{w_s}$ can estimate species trees within seconds for all data sets, *tagNJ*$_{MAD}$ takes about 45 min to calculate a tree for the vertebrates data set. The tested tools have higher running times on all empirical data sets than all of our methods. Requiring more than 12 h (*A-pro*) and several days (*DupTree*, *FastMulRFS*) for the vertebrates data sets. While all of our methods, as well as *DupTree* and *FastMulRFS* are implemented sequentially, *A-pro* runs parallelized on 80 cores. *MiniNJ*$_{w_s}$'s speedup of 547 towards *A-pro* is a crucial difference, especially when taking *A-pro*'s parallelization into account.

6. Conclusion & future work

6.1. Conclusion

We developed new distance-based methods for species tree inference from gene family trees. The best performing methods are miniNJ$_{w_s}$ and tagNJ$_{MAD}$. Both methods are able to filter out paralogous gene pairs (Fig.5.3), which is the main goal of this work.

Their accuracy is analogous to that of the existing tools A-pro [10] and DupTree [11]. The rRFD on the STANDARD data set (Sec. 5.2.1.1) is 0.0470 (4.7 %) for miniNJ$_{w_s}$, 0.0400 (4 %) for tagNJ$_{MAD}$, 0.0461 (4.6 %) for A-pro, and 0.0583 (5.3 %) for DupTree. In particular on the simulated data sets our methods were much more accurate than FastMulRFS [12] (rRFD 0.1603 (16 %)). The execution time of miniNJ$_{w_s}$ is substantially lower than the execution times of any other tool. The mean run times for the SPECIES data set (Sec. 5.2.1.2) with 100 species are 1.3 s for miniNJ$_{w_s}$, 93.3 s for tagNJ$_{MAD}$, 251 s for A-pro, 1 696 s for DupTree, and 250 s for FastMulRFS. MiniNJ$_{w_s}$ attains substantial speedups of 75 on the STANDARD data set (Sec.5.2.1.1) and of up to 547 on the vertebrates data set (Sec.5.2.2) compared to the parallelized tool *A-pro*.

The experiments on simulated data sets showed that our methods behave well. An increase of information in the gene family trees per species (Fig. 5.6, 5.8), a low ILS rate (Fig. 5.10), and a low gene loss rate (Fig. 5.12) increase accuracy. The empirical run time estimate for miniNJ$_{w_s}$ is $\mathcal{O}(n^{2.0})$ and for tagNJ$_{MAD}$ $\mathcal{O}(n^{2.3})$ where n is the number of species. This shows that the time complexity is dominated by calculating the distance matrix and not by the *Neighbor Joining* algorithm.

However both, *miniNJ*$_{w_s}$, and *tagNJ*$_{MAD}$, are sensitive to missing data (Fig.5.13). Which constitutes the main limitation.

The *miniNJ*$_{w_s}$ method as well as other variants of the mini technique can be used to quickly generate a starting tree for maximum likelihood tree search methods. *MiniNJ*$_{w_s}$ is useful for generating starting trees as it is very fast and shows 'good' accuracy.

6.2. Future work

Our methods do not always estimate the true species tree yet. Further studies of the impact of ILS and gene loss could possibly help to further improve the methods and increase the accuracy of the estimated species trees.

Refining our filtering techniques for identifying paralogous gene pairs for polytomies in gene family trees could also improve accuracy.

Our tool is able to combine several of the techniques, which are grouped as picking distances (Sec. 4.1), norming and weighting (Sec. 4.2), and statistical averaging (Sec. 4.3). A dedicated, highly optimized implementation just for miniNJ$_{w_s}$ and tagNJ$_{MAD}$ could outperform our implementation regarding running time and memory usage. Furthermore, the distance matrix calculation on the set of gene family trees could easily be parallelized. Since this is the part that dominates run-times, a parallelization could substantially improve efficiency.

An evaluation on more simulated and empirical data will help to improve our understanding of how accurate the methods are.

Bibliography

[1] Q. D. Wheeler, “The phylogenetic species concept (sensu wheeler and platnick),” *Species concepts and phylogenetic theory: a debate. Columbia University Press, New York*, pp. 55–69, 2000.

[2] G. Ceballos, P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle, and T. M. Palmer, “Accelerated modern human–induced species losses: Entering the sixth mass extinction,” *Science advances*, vol. 1, no. 5, p. e1400253, 2015.

[3] E. Haeckel, *Systematische Phylogenie: Wirbelthiere*, vol. 3. G. Reimer, 1895.

[4] F. Sanger and A. R. Coulson, “A rapid method for determining sequences in dna by primed synthesis with dna polymerase,” *Journal of molecular biology*, vol. 94, no. 3, pp. 441–448, 1975.

[5] F. Sanger, S. Nicklen, and A. R. Coulson, “Dna sequencing with chain-terminating inhibitors,” *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[6] S. C. Schuster, “Next-generation sequencing transforms today’s biology,” *Nature methods*, vol. 5, no. 1, pp. 16–18, 2008.

[7] B. Morel, A. M. Kozlov, and A. Stamatakis, “ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes,” *Bioinformatics*, vol. 35, pp. 1771–1773, 10 2018.

[8] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, “Raxml-ng: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference,” *bioRxiv*, 2019.

[9] L. Liu and L. Yu, “Estimating Species Trees from Unrooted Gene Trees,” *Systematic Biology*, vol. 60, pp. 661–667, 03 2011.

[10] C. Zhang, C. Scornavacca, E. K. Molloy, and S. Mirarab, “Astral-pro: quartet-based species tree inference despite paralogy,” *bioRxiv*, 2019.

[11] A. Wehe, M. S. Bansal, J. G. Burleigh, and O. Eulenstein, “DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony,” *Bioinformatics*, vol. 24, pp. 1540–1541, 05 2008.

[12] E. K. Molloy and T. Warnow, “Fastmulrfs: Fast and accurate species tree estimation under generic gene duplication and loss models,” *Bioinformatics*, vol. 36, no. 1, pp. i57–i65, 2020.

[13] W. P. Maddison, “Gene Trees in Species Trees,” *Systematic Biology*, vol. 46, pp. 523–536, 09 1997.

[14] D. H. Huson and C. Scornavacca, “Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks,” *Systematic Biology*, vol. 61, pp. 1061–1067, 09 2012.

[15] D. Mallo, L. De Oliveira Martins, and D. Posada, "SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees ," *Systematic Biology*, vol. 65, pp. 334–344, 11 2015.

[16] W. Fletcher and Z. Yang, "INDELible: A Flexible Simulator of Biological Sequence Evolution," *Molecular Biology and Evolution*, vol. 26, pp. 1879–1888, 05 2009.

[17] J. Mallet, "A species definition for the modern synthesis," *Trends in Ecology & Evolution*, vol. 10, no. 7, pp. 294–299, 1995.

[18] J. L. Gittleman, "Species," 2019. [Online; accessed 28-September-2020].

[19] H. Pearson, "What is a gene?," *Nature*, vol. 441, no. 7092, pp. 398–401, 2006.

[20] T. E. of Encyclopaedia Britannica, "Dna," 2020. [Online; accessed 28-September-2020].

[21] H. Winkler, *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche.* Gustav Fischer, 1920.

[22] C. M. Thomas and K. M. Nielsen, "Mechanisms of, and barriers to, horizontal gene transfer between bacteria," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 711–721, 2005.

[23] D. Moreira, "Orthologous gene," in *Encyclopedia of Astrobiology* (R. Amils, M. Gargaud, J. Cernicharo Quintanilla, H. J. Cleaves, W. M. Irvine, D. Pinti, and M. Viso, eds.), pp. 1–1, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[24] D. Moreira and P. López-García, "Paralogous gene," in *Encyclopedia of Astrobiology* (R. Amils, M. Gargaud, J. Cernicharo Quintanilla, H. J. Cleaves, W. M. Irvine, D. Pinti, and M. Viso, eds.), pp. 1–1, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[25] G. Nickel, D. Tefft, K. Goglin, and M. Adams, "An empirical test for branch-specific positive selection," *Genetics*, vol. 179, pp. 2183–93, 09 2008.

[26] Z. Yang *et al.*, *Computational molecular evolution*, vol. 284. Oxford University Press Oxford, 2006.

[27] L. L. Cavalli-Sforza and A. W. Edwards, "Phylogenetic analysis. models and estimation procedures," *American journal of human genetics*, vol. 19, no. 3 Pt 1, p. 233, 1967.

[28] P. Pamilo and M. Nei, "Relationships between gene trees and species trees.," *Molecular biology and evolution*, vol. 5, no. 5, pp. 568–583, 1988.

[29] R. Nichols, "Gene trees and species trees are not the same," *Trends in Ecology & Evolution*, vol. 16, no. 7, pp. 358–364, 2001.

[30] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[31] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[32] I. Elias and J. Lagergren, "Fast neighbor joining," in *Automata, Languages and Programming* (L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, eds.), (Berlin, Heidelberg), pp. 1263–1274, Springer Berlin Heidelberg, 2005.

[33] R. Sokal and C. Michener, "U. of kansas, a statistical method for evaluating systematic relationships," *University of Kansas science bulletin (University of Kansas, 1958)*, 1958.

[34] E. S. Allman, J. H. Degnan, and J. A. Rhodes, "Species tree inference from gene splits by unrooted star methods," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 1, pp. 337–342, 2018.

[35] F. Tria, G. Landan, and T. Dagan, "Phylogenetic rooting using minimal ancestor deviation," *Nature Ecology & Evolution*, vol. 1, p. 0193, 06 2017.

[36] S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow, "ASTRAL: genome-scale coalescent-based species tree estimation," *Bioinformatics*, vol. 30, pp. i541–i548, 08 2014.

[37] N. Moshiri, "Treeswift: a massively scalable python package for trees," *SoftwareX*, vol. 11, p. 100436, 2020.

[38] P. Vachaspati and T. Warnow, "Fastrfs: fast and accurate robinson-foulds supertrees using constrained exact optimization," *Bioinformatics*, vol. 33, no. 5, pp. 631–639, 2017.

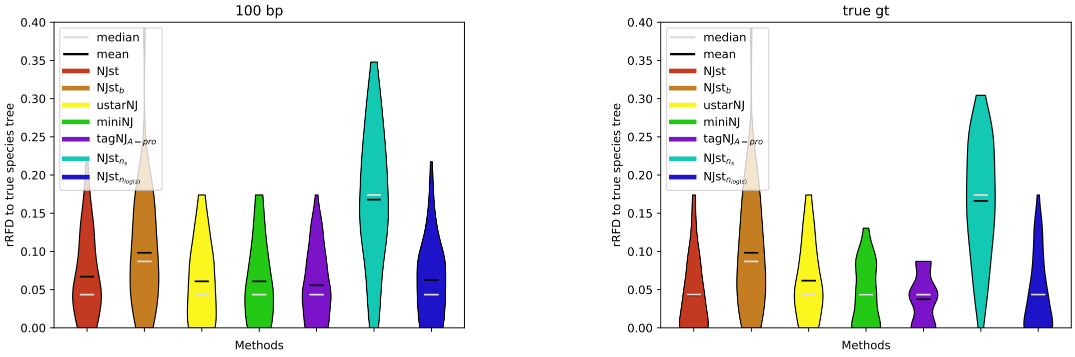
[39] C. Zhang, M. Rabiee, E. Sayyari, and S. Mirarab, "Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees," *BMC Bioinformatics*, vol. 19, no. Suppl 6, p. 153, 2018.

[40] S. Penel, A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perrière, "Databases of homologous gene families for comparative genomics," in *BMC bioinformatics*, vol. 10, p. S3, Springer, 2009.

[41] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, *et al.*, "Ensembl 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D754–D761, 2018.

Appendix

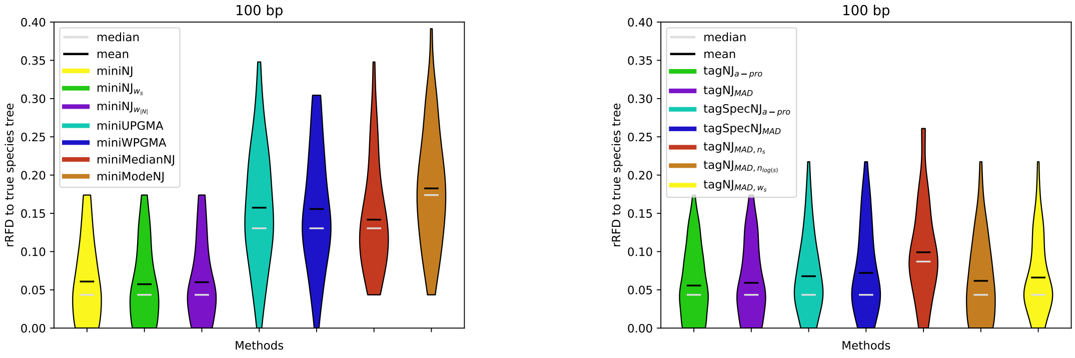
A. STANDARD data set



(a) Comparison of the rRFD for the *NJst+* methods on the STANDARD (100 bp) data set

(b) Comparison of the rRFD for the *NJst+* methods on the STANDARD (true gt) data set

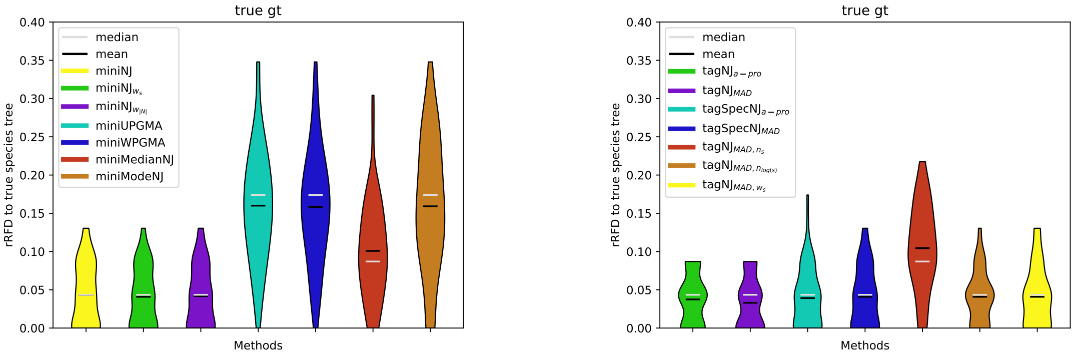
Figure A.1.: Comparison of the rRFD for the *NJst+* methods on the STANDARD (100 bp, true gt) data set



(a) Comparison of the rRFD for different methods using the mini technique on the STANDARD data set

(b) Comparison of the rRFD for different methods using the tag technique on the STANDARD data set

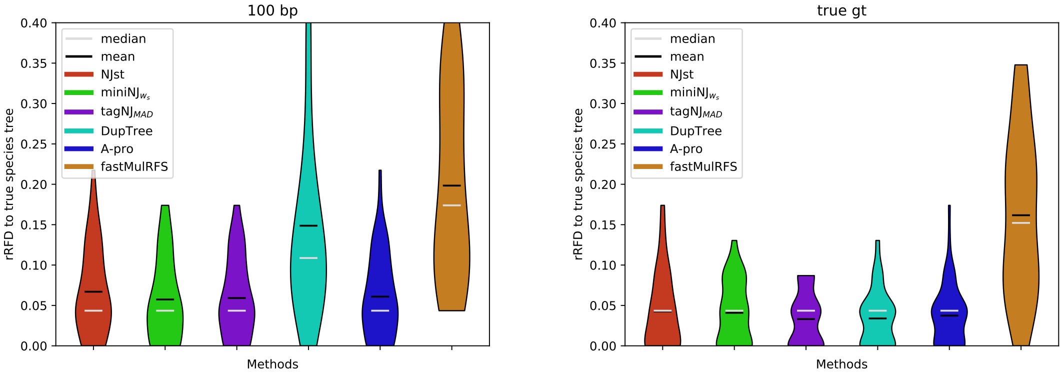
Figure A.2.: Comparison of the rRFD for the *mini* and *tagging* methods on the STANDARD (100 bp) data set



(a) Comparison of the rRFD for different methods using the mini technique on the STANDARD data set

(b) Comparison of the rRFD for different methods using the tag technique on the STANDARD data set

Figure A.3.: Comparison of the rRFD for the *mini* and *tagging* methods on the STANDARD (true gt) data set

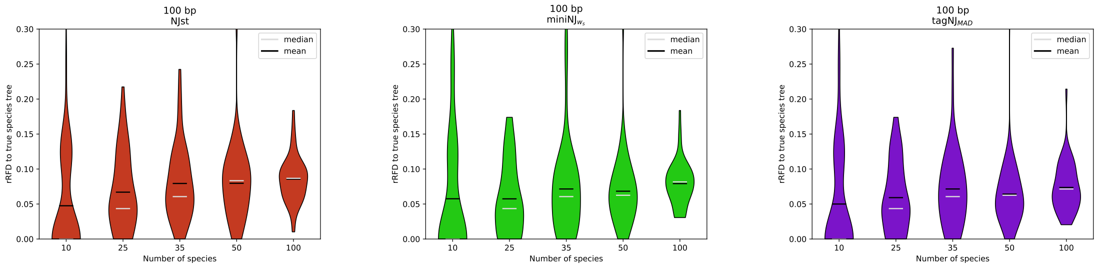


(a) Comparison of the rRFD for different tools on the STANDARD (100 bp) data set

(b) Comparison of the rRFD for different tools on the STANDARD (true gt) data set

Figure A.4.: Comparison of the rRFD for different tools on the STANDARD (100 bp, true gt) data set

B. SPECIES data set

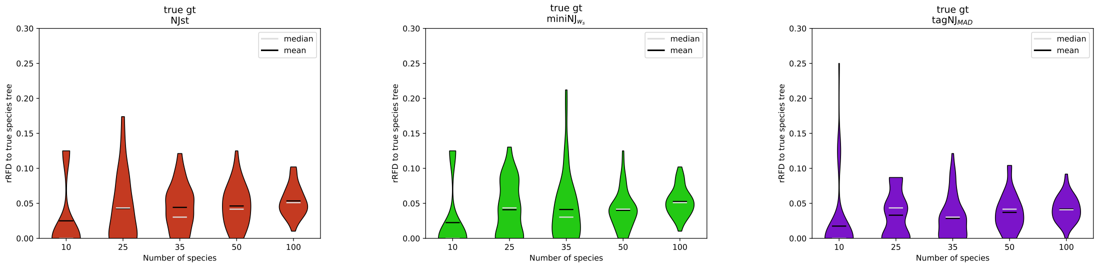


(a) Comparison of the rRFD for different numbers of species using the NJst method on the SPECIES (100 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the SPECIES (100 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the SPECIES (100 bp) data set

Figure B.5.: Comparison of the rRFD for different methods on the SPECIES (100 bp) data set



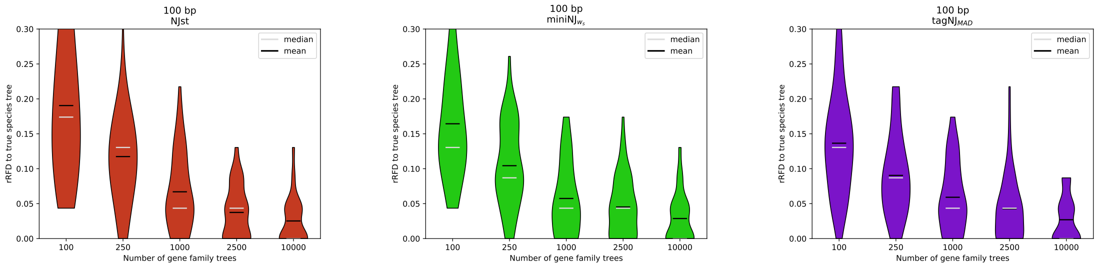
(a) Comparison of the rRFD for different numbers of species using the NJst method on the SPECIES (true gt) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the SPECIES (true gt) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the SPECIES (true gt) data set

Figure B.6.: Comparison of the rRFD for different methods on the SPECIES (true gt) data set

C. GENES data set

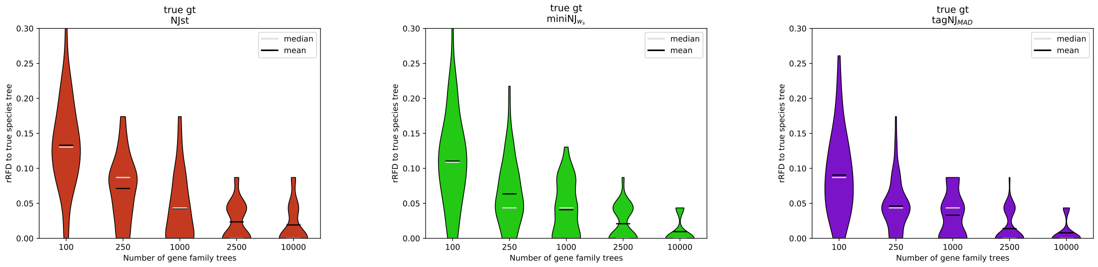


(a) Comparison of the rRFD for different numbers of species using the NJst method on the GENES (100 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the GENES (100 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the GENES (100 bp) data set

Figure C.7.: Comparison of the rRFD for different methods on the GENES (100 bp) data set



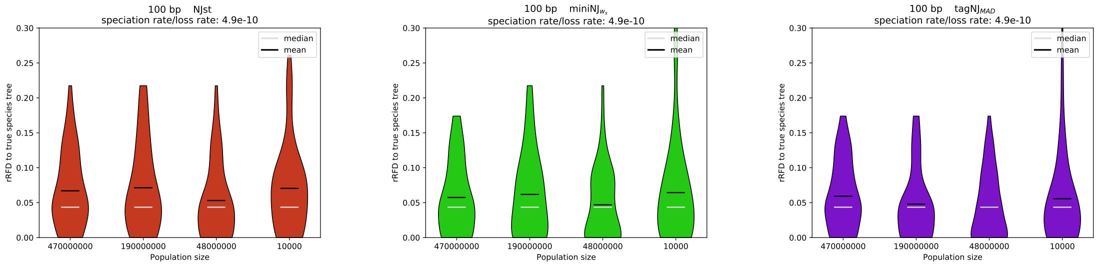
(a) Comparison of the rRFD for different numbers of species using the NJst method on the GENES (true gt) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the GENES (true gt) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the GENES (true gt) data set

Figure C.8.: Comparison of the rRFD for different methods on the GENES (true gt) data set

D. ILS data set

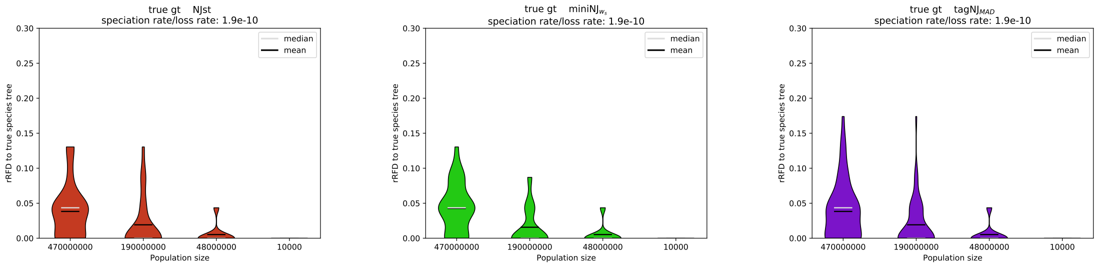


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (100 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (100 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (100 bp) data set

Figure D.9.: Comparison of the rRFD for different methods on the ILS (100 bp) data set

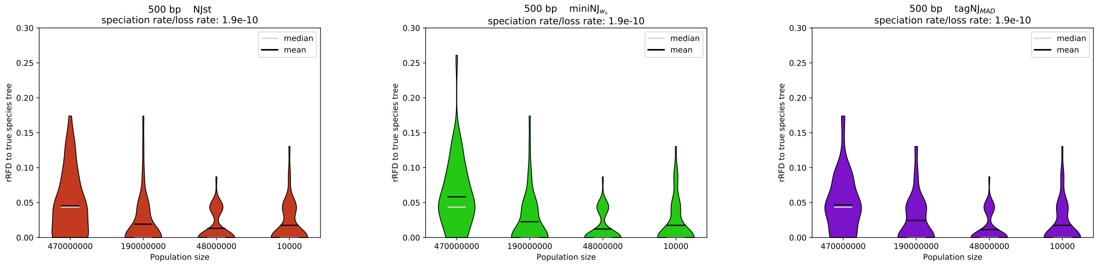


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (true gt) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (true gt) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (true gt) data set

Figure D.10.: Comparison of the rRFD for different methods on the ILS (true gt) data set

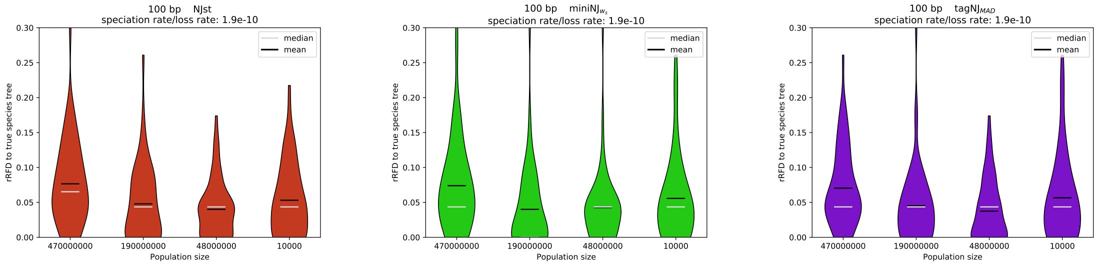


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (500 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (500 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (500 bp) data set

Figure D.11.: Comparison of the rRFD for different methods on the ILS (500 bp) data set

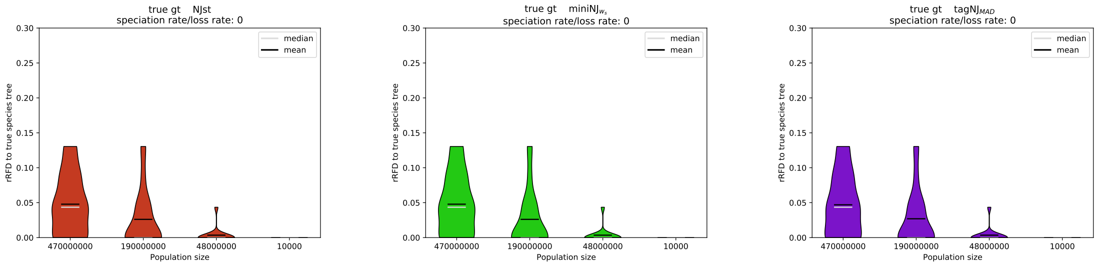


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (100 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (100 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (100 bp) data set

Figure D.12.: Comparison of the rRFD for different methods on the ILS (100 bp) data set

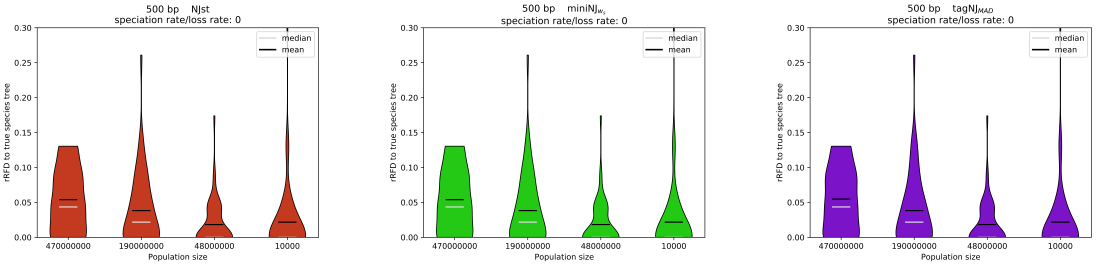


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (true gt) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (true gt) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (true gt) data set

Figure D.13.: Comparison of the rRFD for different methods on the ILS (true gt) data set

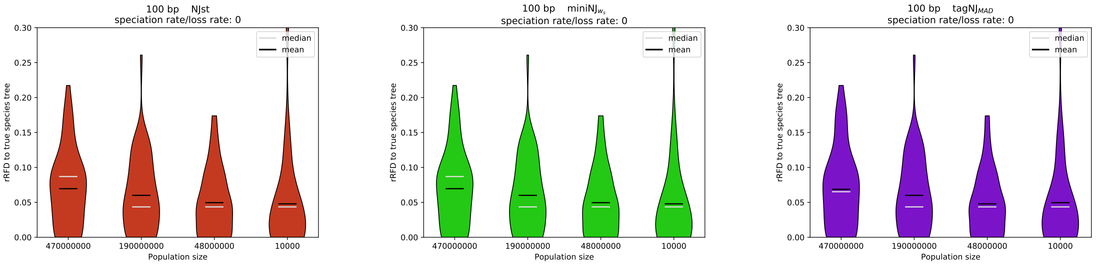


(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (500 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (500 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (500 bp) data set

Figure D.14.: Comparison of the rRFD for different methods on the ILS (500 bp) data set



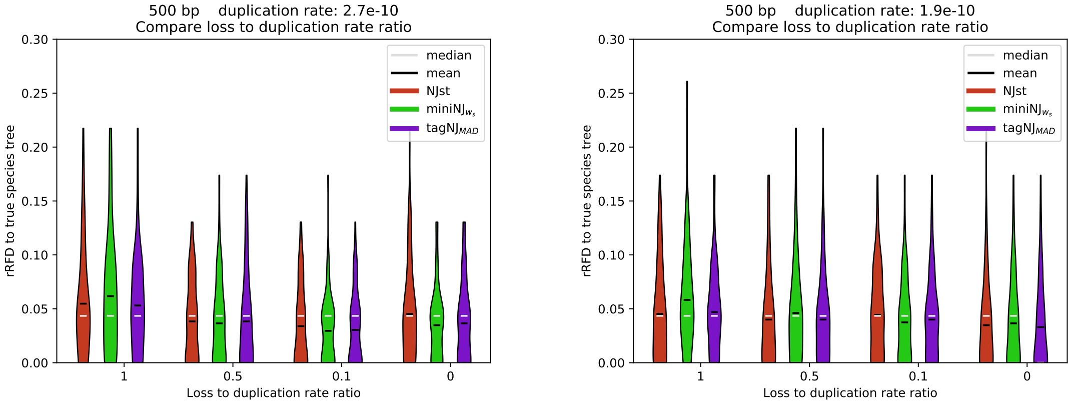
(a) Comparison of the rRFD for different numbers of species using the NJst method on the ILS (100 bp) data set

(b) Comparison of the rRFD for different numbers of species using the miniNJ$_{w_s}$ method on the ILS (100 bp) data set

(c) Comparison of the rRFD for different numbers of species using the tagNJ$_{MAD}$ method on the ILS (100 bp) data set

Figure D.15.: Comparison of the rRFD for different methods on the ILS (100 bp) data set

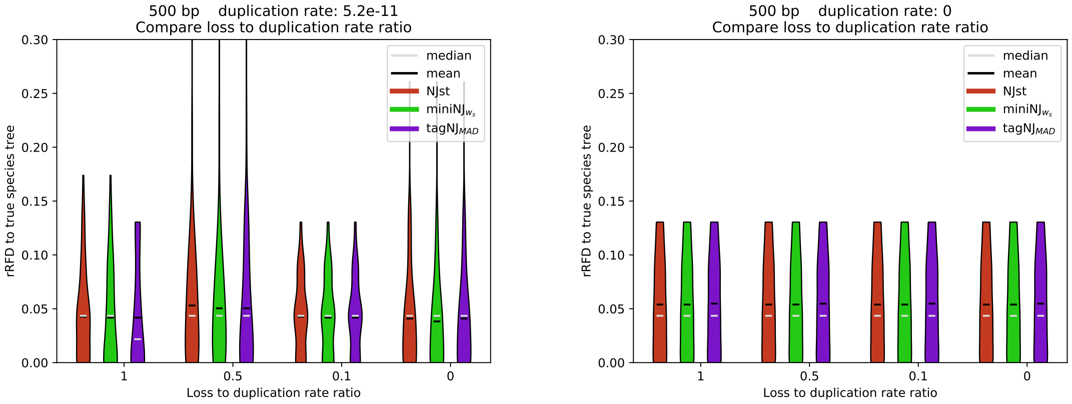
E. DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

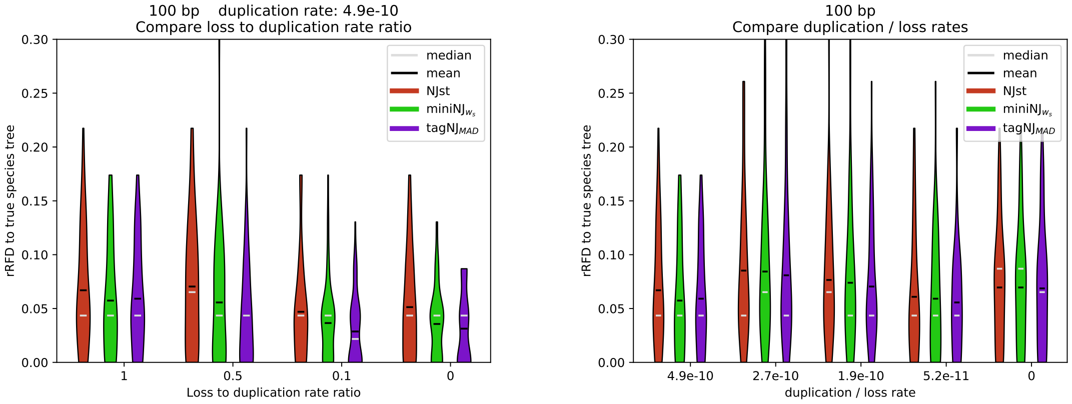
Figure E.16.: Results for the DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

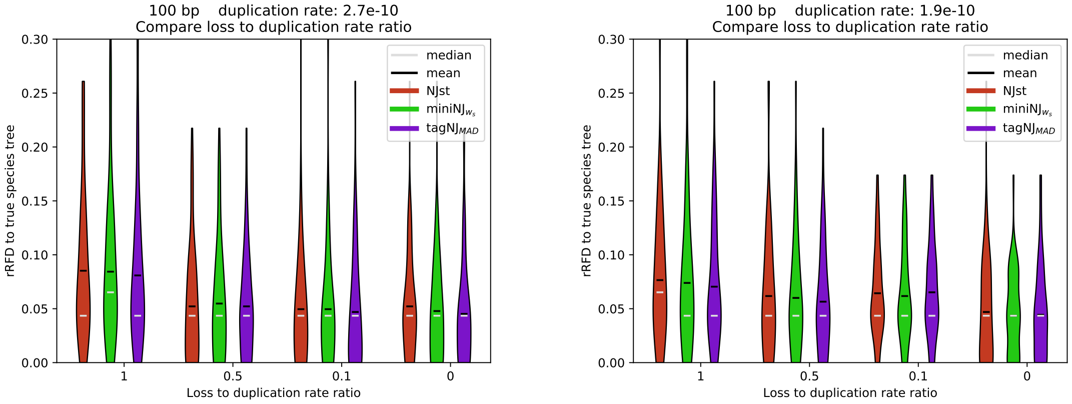
Figure E.17.: Results for the DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different duplication and loss rates

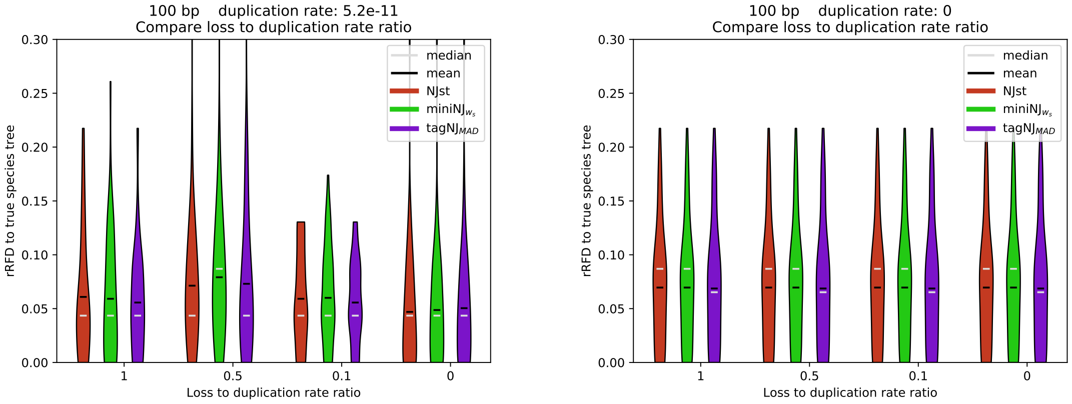
Figure E.18.: Results for the DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

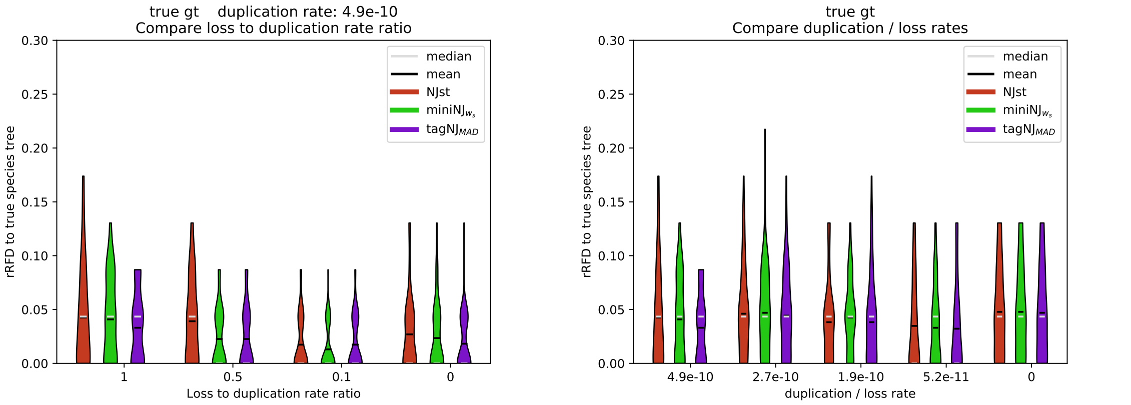
Figure E.19.: Results for the DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

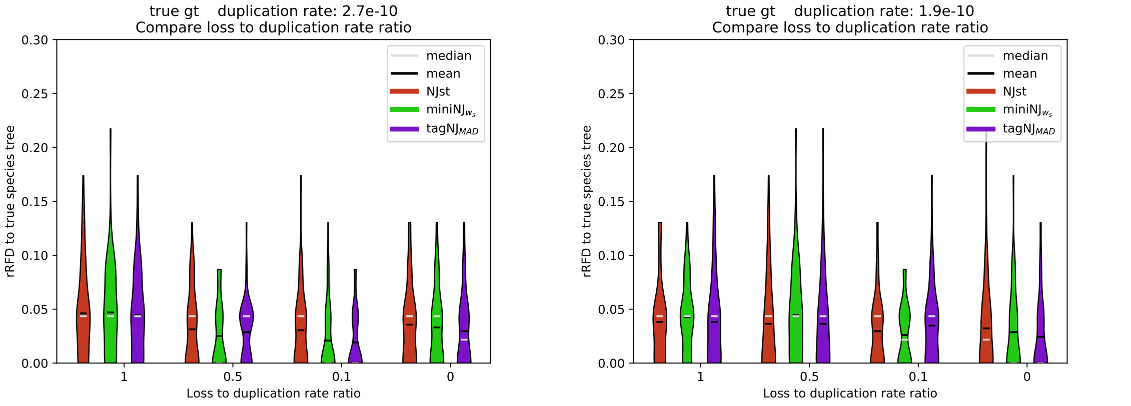
Figure E.20.: Results for the DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different duplication and loss rates

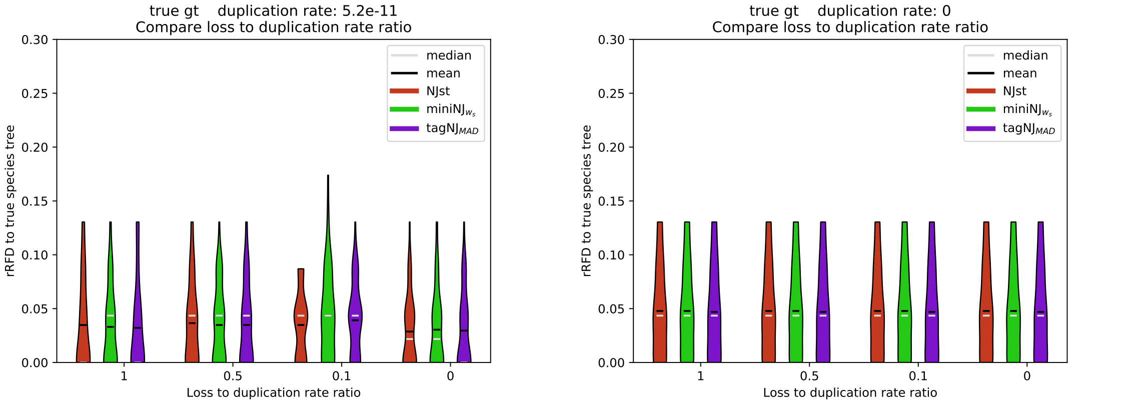
Figure E.21.: Results for the DUPLOS data set



(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

Figure E.22.: Results for the DUPLOS data set

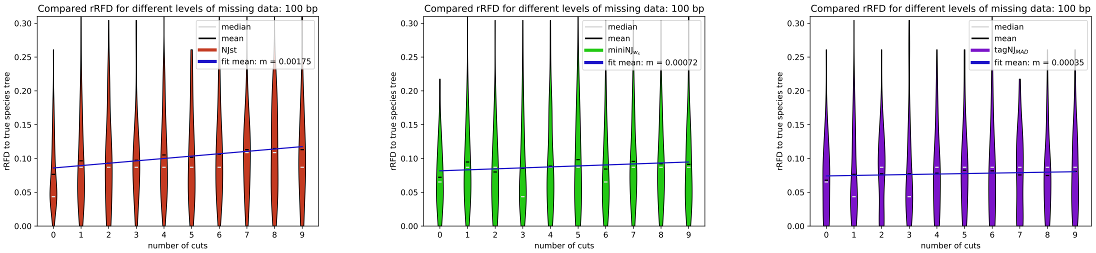


(a) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

(b) Comparison of the rRFD for different methods on the DUPLOS data set with different loss to duplication rate ratios

Figure E.23.: Results for the DUPLOS data set

F. CUTS data set

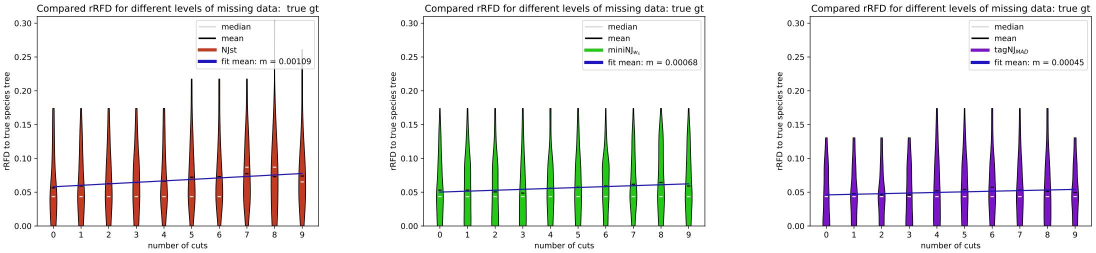


(a) Comparison of the rRFD different numbers of cuts with the NJst method

(b) Comparison of the rRFD different numbers of cuts with the miniNJ$_{w_s}$ method

(c) Comparison of the rRFD different numbers of cuts with the tagNJ$_{MAD}$ method

Figure F.24.: Comparison of the rRFD different numbers of cuts with different methods



(a) Comparison of the rRFD different numbers of cuts with the NJst method

(b) Comparison of the rRFD different numbers of cuts with the miniNJ$_{w_s}$ method

(c) Comparison of the rRFD different numbers of cuts with the tagNJ$_{MAD}$ method

Figure F.25.: Comparison of the rRFD for different numbers of cuts with different methods