

An analysis of maximum parsimony algorithms to predict parasitism in Eukaryota

using a large multiurcated phylogenetic synthesis tree

Abstract

This study focuses on the ancestral state reconstruction of parasitism in the tree of life of Eukaryota. We predict unknown states of species and estimate origins and losses of parasitism.

The challenge here is the size of the tree and the little information about it.

Such a large phylogenetic tree does not completely exist and therefore we work with a synthesis tree of OTL [1] which is highly multifurcated.

For the 2,535,437 leaf nodes we could not gather much data. From the GloBI database [2] which we used, we could only collect 25,992 parasitic and 34,879 free-living species. It follows that we have only $\approx 2.4\%$ state information.

So far, especially small scale studies have been carried out or highly manual. In this scale, it requires different data sources to be interconnected.

We performed an analysis of existing algorithms and selected a Sankoff maximum parsimony algorithm using the R package *Castor* [3].

Nevertheless, the results are convincing and even though purely computational approach which did not include human experts input, results coincide with prior knowledge. Also regarding the number of events, our estimates coincide with previous results by human experts, e.g. the study by Weinstein and Kuris [4].

Anmerkung: Klassisch packt man keine Referenzen in Abstracts (bzw. wenn das meist als Kurzreferenz also (Author et al., 2017). (Bernhard)

We have compared the results of some subtrees with known knowledge (Chordata, Nematoda, Platyhelminthes and Apicomplexa) and, except for the Nematoda, the results looked

very good. In the case of Nematoda, the data situation is strongly shifted to the few parasites.

We could partly compare our number of origins with the results of Sara B. Weinstein and Armand M. Kuris from their article *Independent origins of parasitism in Animalia* and have come to a similar magnitude. They identified 223 parasitic origins in Metazoa and we were able to estimate about 300 origins.

Contents

1	Introduction	6
1.1	Definitions	8
1.1.1	Parasite	9
1.1.2	Maximum parsimony	9
2	Methods	10
2.1	Description of data sets	11
2.1.1	OTL	11
2.1.2	GloBI	12
2.2	Metadata analysis	13
2.2.1	Transition probabilities	13
2.2.2	Missing information	14
2.3	Simulation	15
2.3.1	random binary tree	16
2.3.2	simulating states and transitions between them	16
2.3.3	simulating loss of information of the tree topology	17
2.4	Ancestral state reconstruction methods	17
2.4.1	Fitch maximum parsimony	19
2.5	Implementation	22
3	Results	23
3.1	Metadata analysis	23
3.1.1	Missing information	23
3.1.2	Results of simulation / Influence of different parameters	28
3.2	Results of the real data analysis created with Sankoff	30
3.2.1	Biological view	30

3.2.2	Origins and Losses	34
3.2.3	Cross evaluation - leave 100 out	36
3.3	Effects of Taxa in the different models	37
4	Discussion	41
4.1	Simulation	41
	Bibliography	43
5	Appendices	46
5.1	Methods overview	46
5.2	OTL analysis	46
5.2.1	List of all phyla	46

1 Introduction

This paper is about the analysis of ancestral state reconstruction algorithms for non-binary trees, applied to the currently largest phylogeny synthesis tree of Open Tree Of Life, with the application of prediction of parasitism.

Anmerkung: Mein Vorschlag einer Gliederung (jeweils ca. ein Absatz) (Bernhard)

i) Motivation:

- Was ist das große Ziel?

Das Ziel dieser Arbeit ist die Anwendung von maximum parsimony algorithmen auf nicht binäre Bäume und auf sehr große Datensätze. Insbesondere auf das Beispiel 'Entstehung des Parasitismus' im ganzen Eukaryotischen Tree of Life.

- Was soll erreicht werden?

Wir wollen vorhandene Algorithm (Sankoff/castor [3]) auf diese Aufgabenstellung hin testen und ihre Vorhersagekraft abschätzen. Außerdem wollen wir den Fitch algorithmus für binäre Bäume auf unser Problem erweitern und mit dem Sankoff Algorithmus vergleichen.

- Warum ist das relevant? Was könnte man dann tun?

predict states of species...

TODO: !!!

ii) Hintergrund:

- Was gab es in dieser Richtung bereits als ganze Ansätze oder wenn nicht, warum nicht? Woran ist es bisher gescheitert?

For about 50 years, people have been working on ancestral state reconstruction.

The first paper mentioned above is by Camin and Sokal, who in 1965 were working on algorithms for discrete-state data [5]. Different methods have been developed and the question is which method is the most suitable for our problem. For this purpose, various studied methods and their advantages and disadvantages were compared.

Bisher wurden vorallem Algorithmen für das binäre Problem entwickelt, da man wesentlich kleinere Teilbäume betrachtet hat, von welchen man auch alle Aufspaltungen kennt. Durch die Entwicklung von OTL, eines gesamten Baum des Lebens, ergibt sich das Problem, dass dieser bei weitem nicht binär ist.

Researchers of the phylogenies have been dealt with the ancestral state reconstruction in the 60s. The first methods were only brute force **TODO: Quelle, siehe Fitch: Camin and Sokal 1965** . Next came a set of parsimony algorithms such as: Fitch-parsimony [6], Wagner-parsimony [7] ... **TODO: weitere?** .

With more and more data, there is now the possibility to use more information to calculate the probabilities of the ancestral states. In addition to the states of the leafs, algorithms could also use branch lengths. The likelihood based algorithms came more in interest.

Our focus came with another 'data extension'. We wanted to work with the biggest phylogenetic tree that exists at this moment, which goes over all observed species. For most **TODO: most?** species there is no phylogeny, but only a taxonomic classification.

- Welche Grundlagen sind notwendig:
 - open tree of life: Was ist das, warum relevant und überlegen als reine Ansätze?
TODO: !!!

So the biggest 'phylogenetic tree' is a synthesis of phylogenetic trees filled with a taxonomic tree given by Open Tree of Life [1]. This tree is not binary

and therefore the developed algorithms are not directly applicable.

- Algorithmen: Was gibt es? Ruhig ausführlicher als hier bereits und vor allem auch nach einer Darstellung am Ende ableiten, was für uns relevant ist. Also beschreiben, wie Methode a, b, c funktionieren und dann abwägen, was daher für Dich am relevantesten ist.

TODO: !!!

Anmerkung: GloBI und OTL in der Einleitung vorstellen. (Emanuel)

iii) Outlook/Structure of this work

In this work, we have looked at the algorithms that are generally suited to our data, to develop them further for the not binary case, and finally to compare their usability with our sythesis tree.

We have decided to consider only parsimony algorithms since we have no information on branch lengths and no other additional information like different transition probabilities of our states.

In den Bacteria und Archaea wurde das mal allgemein für binary traits gemacht: "A total 90 % of all binary traits described molecular functions, specifically the presence or absence of a gene or a set of gene involved in biochemical pathways." [8]

Unterschied bei uns ist parasitismus komplexer.. da sind meist nicht einzelne Gene zuständig..

1.1 Definitions

- Multifurkation - binär (binary or bifurcating... [9])
- height (min, max, mean), depth of a tree/node (Distanz zur Wurzel vs distanz zum Blatt)
- OTL, OTT, GloBI

1.1.1 Parasite

Since we use GloBi to classify species as parasitic or free-living, we use their definition of parasitism. In GloBi, ontobee definitions are used. The interaction has parasite is defined as: "An interaction relationship between two organisms living together in more or less intimate association in a relationship in which association is disadvantageous or destructive to one of the organisms." [?]. This definition includes: ecto,- and endoparasites, parasitoids, kleptoparasites and pathogenes.

TODO: Mungall, C., (2017). Definition for the interaction-term: "parasitised by; has parasite" ob ontobee.org. Last checked: 24.07.2017 at ontobee.org/ontology/...

1.1.2 Maximum parsimony

Maximum parsimony methods Maximum parsimony in phylogeny refers to favoring the tree that needs the least evolutionary change to explain the observed data.

In our case, it is about the change of states 'is free-living' or 'is parasitic'.

2 Methods

As initiated, in this paper, a maximum parsimony algorithm is applied to the whole tree of life to obtain an ancestral state reconstruction of free-living versus parasite states.

So far, these reconstructions have been made mainly on binary trees with better data availability. Therefore, a simulation is first performed to evaluate existing algorithms and decide how they may be adapted to our given problem. This is to perform the ancestral state reconstruction for a multifurcated (non-binary) tree using binary states.

Accordingly, in addition to the necessary data sets (GloBI, OTL), the chosen algorithm and the evaluation of its results, this chapter also deals with the previously performed simulation and the evaluation of the various algorithms and their parameters.

Figure 2.1 briefly outlines these relationships.

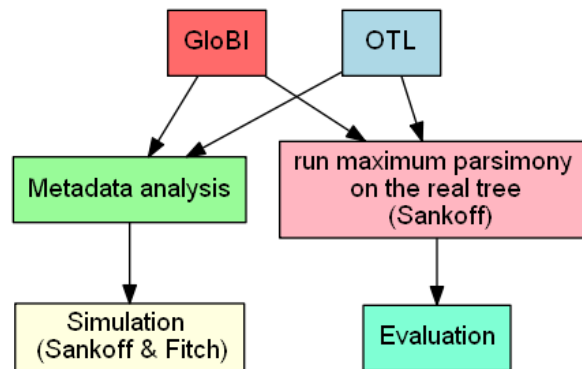


Figure 2.1: Workflow

The resulting procedure is as follows:

- (1) Retrieve phylogenetic tree data as input for the tree (OTL) and the state information (GloBI).
- (2) Get metadata of these for a realistic simulation of the maximum parsimony algorithms (Fitch & Sankoff).
- (3) Build and run the simulation.
- (4) Evaluation of parameters for the simulation and the ancestral state reconstruction of the real tree.
- (5) Evaluate the accuracy of developed algorithms and choose the best.
- (6) Run the resulted algorithm on the original data.
- (7) Evaluate and interpret results.

TODO: In den einzelnen Abschnitten Teile von dem großen Workflow verwenden?: 5.1

2.1 Description of data sets

For **our** research two types of data are required: a tree and information about the states.

For the tree **we** decided to use Open Tree of Life (OTL) [1].

For the state information, **we** decided to use the Global biotic interaction database (GloBI) [2].

2.1.1 OTL

For **our** project a large database for phylogenetic trees and also for a taxonomic tree is needed. Since an ancestral state reconstruction algorithm is applied to the phylogenetic tree, and for the assessment and other properties the taxonomy provides much more information. OTL gives a synthesis of phylogenetic trees (currently 819 trees) and a taxonomic tree. OTL also includes the large phylogenetic database TreeBASE [1].

TODO: Das steht auf der Website nicht in dem Paper...

For phylogenetic data, there are at least five big data collections, namely:

- ITIS (Integrated Taxonomic Information System) [10]
- NCBI (National Center for Biotechnology Information) [11]
- WORMS (World Register of Marine Species) [12]
- GBIF (Global Biodiversity Information Facility) [13]
- OTT (OpenTreeOfLife-Taxonomy) [1]

TODO: In die discussion?:

ITIS is only a small set of 100 % confirmed and named species. GBIF is not composed with the help of phylogeny, the same is valid for the NCBI taxonomy. The WORMS taxonomy is a way too small dataset of mostly marine species.

Here the taxonomy from OTL is used because it is including most of the known taxonomies and is synthesised by preferring taxonomies that match with available phylogenetic data. Furthermore the team from OTL preferre a maximum number of species [1]. This is resulting in somekind of hybrid between taxonomy and phylogeny. *Anmerkung: Wie genau ist das ein Hybrid? Genauer beschreiben, was Du damit meinst... (Thilo)*

A closer look is being made to some of the features of the synthesis tree. On the one hand the distribution of the taxa and on the other the distribution of the nodes on the taxa. Since this is not directly relevant to **our** study, there is a section in the appendix 5.2.

2.1.2 GloBI

TODO: Marius: "There aren't many big active interaction databases out there, most of them are offline or outdated. For example: IWDB (Interaction Web Database) [14], Webs on the Web [15], Animal Diversity Web [16] and ecoweb [17]. GloBI is including most of the known ones and is still growing actively [2]. So the question which interaction database could be used was answered rather quickly."

This database consists of entries of the form: species A (source) interacts with B (target). A number of interactions have been identified¹, including those that the species source or target has become a parasite or a free-living species from the biological perspective. These are the following:

- free-living source: preysOn, eats, flowersVisitedBy, hasPathogen, pollinatedBy, hasParasite, hostOf
- free-living target: preyedUponBy, parasiteOf, visitsFlowersOf, pathogenOf, hasHost
- parasite source: parasiteOf, pathogenOf
- parasite target: hasParasite, hasPathogen

¹<https://github.com/jhpoelen/eol-globi-data/.../InteractType.java>

Two lists are formed: parasites and free-livings, and the source or targets of an interaction are added to these.

2.2 Metadata analysis

In order to generate a realistic simulation, influencing parameters are investigated.

There are two major types of parameters:

i) Biological parameters (A result of the evolutionary process.):

- transition probabilities

ii) Distribution of the loss of information:

- Loss of topology (\rightarrow mutlifurcations)
- Unknown information about states of some leaf nodes

The influence of these parameters are tested on our result using our simulation (Section 2.3).

2.2.1 Transition probabilities

This subsection deals with the transition probabilities from free-living (hereinafter / as a formula FL) to parasitic (hereinafter P) and vice versa: $\mathcal{P}(FL \rightarrow P)$, $\mathcal{P}(P \rightarrow FL)$.

Different parasite types have different transition probabilities. It is very difficult to make a statement about these probabilities. In general, **we** assume that there are 40 % parasites and 60 % free-livings which is based on the estimates by Windsor [18] and $\mathcal{P}(FL \rightarrow P) > \mathcal{P}(P \rightarrow FL)$, because a reverse mutation is usually less likely.

TODO: This is discussed in section x of the discussion.

For the maximum parsimony analysis of the real data, all transition probabilities are equated. However, the used castor package [3] offers the possibility to enter different transition probabilities.

In the simulation two beta distributions have been chosen and a threshold that indicates the change between states.

Different thresholds with different beta distributions are simulated, with different distributions of parasites and free-livings:

- 50 % P to 50 % FL,
- 40 % P to 60 % FL,
- 30 % P to 70 % FL and
- 20 % P to 80 % FL

(TODO: see results simulation, ref...).

Figure 2.2 shows one example of these.

TODO: Plot neu erstellen: Achsenbeschriftung, threshold

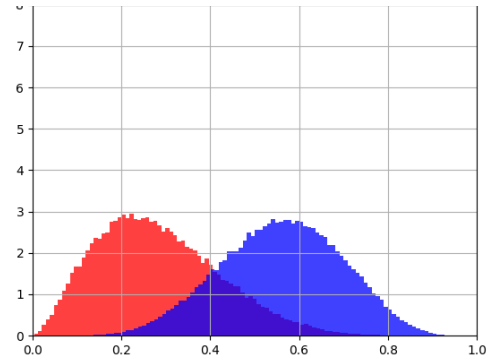


Figure 2.2: 60 % Free-living - 40 % Parasites
red: parasites, blue: free-living,
the threshold is at 0.4

2.2.2 Missing information

A binary tree with n leaf nodes has $n - 1$ internal nodes. The present Eukaryota tree of OTL has 2,293,463 leaf nodes and only 41,974 internal nodes, that is:

$$100 - \frac{100}{(2293463 - 1) \times 41974} \approx 98.16\%$$

missing internal nodes. This means that there is a lack of information about the underlying phylogeny. Instead of a binary tree this tree is highly multifurcated.

For the present Eukaryota tree with 2,293,463 leaf nodes, 34,869 free-livings and 25,962 parasites are found, which are

$$100 - \frac{100}{2293463 \times (34860 + 25962)} \approx 97.34\%$$

unknown states of leaf nodes.

In the simulation, the influence of the multifurcation and missing data in leaf nodes on the predictive accuracy of the ancestral state reconstruction algorithms is tested.

For the real data, generalized linear models are compared with poisson respectively binomial regression according to their residuals and BICs. **Anmerkung: Erklären, warum und wozu das gemacht wird. Warum BIC und nicht AIC? Erklären, was beide bedeuten, wäre sinnvoll. (Thilo)**

2.3 Simulation

There are various possibilities of ancestral state reconstruction. The simulation compares different algorithms with each other.

First different implementations of the Fitch maximum parsimony are compared and then the best of them is compared with the implementation of the sankoff algorithm of the castor package [3].

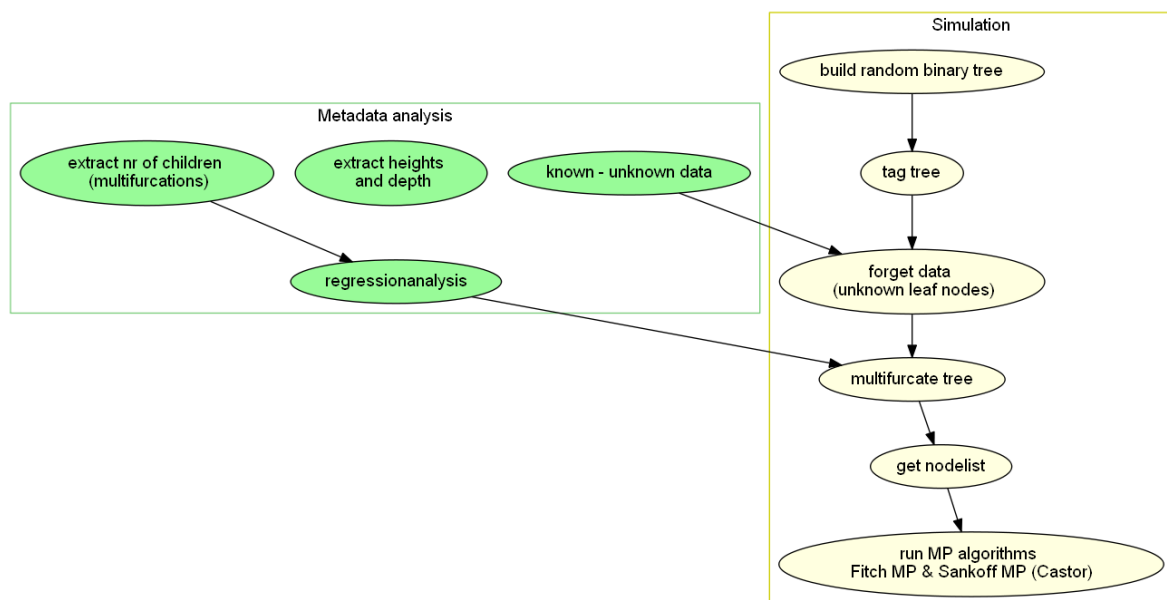


Figure 2.3: Course of the simulation with influence of the metadata analysis from the real data.

TODO: Figure 2.3 shows the course of the simulation. The individual steps are explained in the following subsections: ...

TODO: Evaluation - compare trees (distances)

2.3.1 random binary tree

A tree is needed to do a simulation of ancestral state reconstruction. It had to be decided whether to take the real tree or simulate a tree. In this simulation, trees are created randomly, as one can replicate a complete binary phylogentic tree. Thus, there is also the possibility to simulate the multifurcation.

To get a random binary tree, the Phylo package from biopython is used [19]. They offer a randomized function which returns a BaseTree².

TODO: ref in die discussion über die randomized function? Diskutieren wir das?

2.3.2 simulating states and transitions between them

The next step is to simulate the states of the nodes using the transitions. Again, fully known states are simulated and then everything is 'forgotten' but a few in the leaf nodes so that **you** can later compare the reconstruction with the origin.

The root node is defined as ancestor of all subsequent species and in this case, determined to be free-living. Therefore, a beta distribution for free-living is used at the beginning. Now traverse from the root to the leaf nodes, always pulling out of the current distribution until **you** get above the threshold and the new node changes state. *Anmerkung: Für den Tree/Graph Traversal würde sich eine erklärende Abbildung anbieten. Nur Text macht es sehr schwer verständlich an der Stelle. (Thilo)*

To ensure that the parameter of the binomial distribution is restricted to the [0,1] interval, it is modeled with a beta distribution as in Figure 2.2.

²<https://github.com/biopython/biopython/blob/master/Bio/Phylo/BaseTree.py>

After traversing through the tree, each state is saved in a nodelist associated with the node ID which is the OTT from OTL.

Here begins the simulation of the lost information. This is on the one hand the states and on the other the topology of the tree. Some splits of nodes are unknown with which the tree is multifurcated (explained in the following section [TODO: pageref](#)).

In the real tree, there is usually only information about species living today → leaf nodes. And beyond only a small percentage of these. All information about the states of the internal node and one leaf node is 'forgotten' and stored in another column to the node.

Different percentages of forgetting the information are simulated, as [you](#) can read in the [TODO: section ... from the results](#) .

2.3.3 simulating loss of information of the tree topology

As previously explained, some divisions in the tree are not known, so the real tree is not binary. This multifurcation was simulated by an equally distributed percentage of forgotten internal nodes.

2.4 Ancestral state reconstruction methods

As presented in the introduction, there are some methods for ancestral state reconstruction. For this purpose, various studied methods and their advantages and disadvantages are compared below.

Royer-Carenzi et al. distinguishes two major classes of ancestral state reconstruction methods:

The first is to explain the current state with the least number of state changes between an ancestor and his child, this is called parsimonious.

The other class she presents involves modeling the character evolution as a stochastic process

and using the likelihoods to compute the possible ancestral character states. This is generally done with a continuous time Markov model [20].

TODO: Pasqualin et al. unterscheiden noch eine weitere Methode: stochastic mapping...

One of the major disadvantages of parsimony methods is that, unlike likelihood approaches, they can not take divergence times (branch length) into account. Since there are no development times of species in our case, **you** can ignore this.

Another problem pointed out by Royer-Carenzi is that parsimony approaches are either based on predefined parameters (generalized parsimony) or on strong and often controversial assumptions, like irreversibility of transitions for Dollo parsimony. Again, this problem is irrelevant to the problem at hand, because **you** can only work with generalized models in the analysis of the entire Eukaryota tree.

Parsimony-based methods are used in this work, since they are fully sufficient for the presented use case here. Following the principle of the simpler model first.

Felsenstein [9] discusses in his book two algorithms that generalize all previous methods (from Camin and Sokal [5], **TODO: Kluge and Farris** and Farris [21]): Fitch parsimony [6] and Sankoff parsimony [22].

Anmerkung: Unter Farris war auch noch der Begriff Wagner trees in Gebrauch, als Verallgemeinerung der parsimonious trees von Camin und Sokal. (Lydia)

TODO: Wagner-parsimony [7]

Thus, the methods used in this work are those of Fitch and Sankoff. For Fitch, the algorithm has been extended from binary to multifurcated trees. For the Sankoff algorithm, Louca and Doebeli have presented an implementation for non-binary trees published in an R package named *castor* [3].

2.4.1 Fitch maximum parsimony

Fitch maximum parsimony is an algorithm for rooted, binary trees and describes an ancestral state reconstruction for discrete states [6] by minimizing transitions between states.

Note, the original Fitch algorithm has the sole purpose of minimizing the number of transitions and not reconstructing the ancestral nodes. Felsenstein [9] describes a simple extension for the reconstruction. Cunningham et al. [23] have refined these. **TODO: Wir haben mit ein paar kleinen Änderungen optimiert... und schließlich auf multifurcated angepasst... TODO: eigentlich ist Cunningham 'nur' eine kritische Neubewertung. Sie beziehen ihren Algorithmus auf Swofford und Maddison...**

To understand the differences to the multifurcated case, the algorithm for the binary case is briefly explained and referred to the extension.

Input: A rooted, binary tree, with state information in the leaf nodes. Each node is depicted as a set of states. There are only two states in this paper, free-living (FL) and parasitic (P). Internal nodes have three sets, which are empty at the beginning, excluding the root node, it has only one. Leaf nodes have their state as a set (eg {FL} or {P}), unknown leaf nodes the union of all possible states ({FL, P}).

The algorithm traverses three times through the tree and fills these sets.

In each step, two sets are considered and their intersection formed. There are two cases:

- i) The intersection is not empty and corresponds to the new set.
- ii) The intersection is empty. → Build the union of these sets as new set.

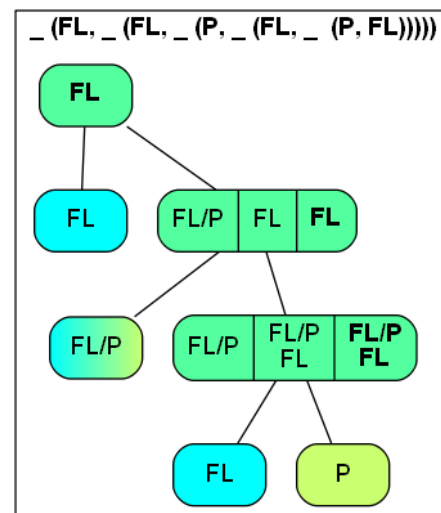


Figure 2.4: Fitch algorithm for binary trees.

The unknown leaf node is described with both states. Computed internal nodes (exclusive the root node) consists of three sets, where the last set is the final one (bold).

From the second internal node (seen from the root node) there are several possibilities to create the second and third set.

First traverse from the leaf nodes to the root / move down the tree / postorder tree traversal. Each internal node is formed from its child nodes, where at the beginning the only information lies.

Second traverse from the root node to the leafs. Each internal node is formed from its father node and its sibling node.

Last traversal (direction does not matter): Build the final state for every node. It is formed from the sets of previous traversals.

(The original Fitch algorithm was designed to minimize transitions without predicting actual states of internal nodes, so it was just the first traversal.)

The extension to the non-binary case is quite obvious, but holds some opportunities. In this case, more than two children may be present for the first traversal, but the incision or union may also be formed over more than two sets. Also in the second traversing, there may be several sibling nodes. However, there are several possibilities here that were all tested and compared in the simulation. Some of these options are already available in the binary case:

- The father node has (except for the root node) two state sets, because he came through the up-traversing previously. Are both sets used or only the first traversing?
- Since there are several siblings, do **you** first of all make the cut or union, or directly in the whole with the father node?

The first point already has an effect on the binary case. Figure 2.4 shows both possibilities of the three sets.

Cunningham uses only the first state set of the father node [23].

From these two points four different versions of Fitch were formed:

- i) Fitch 1: First state set of father node; intersection/union of siblings first.
- ii) Fitch 2: First state set of father node; intersection/union of siblings together with father node.
- iii) Fitch 3: Both state sets of father node; intersection/union of siblings first.
- iv) Fitch 4: Both state sets of father node; intersection/union of siblings together with father node sets.

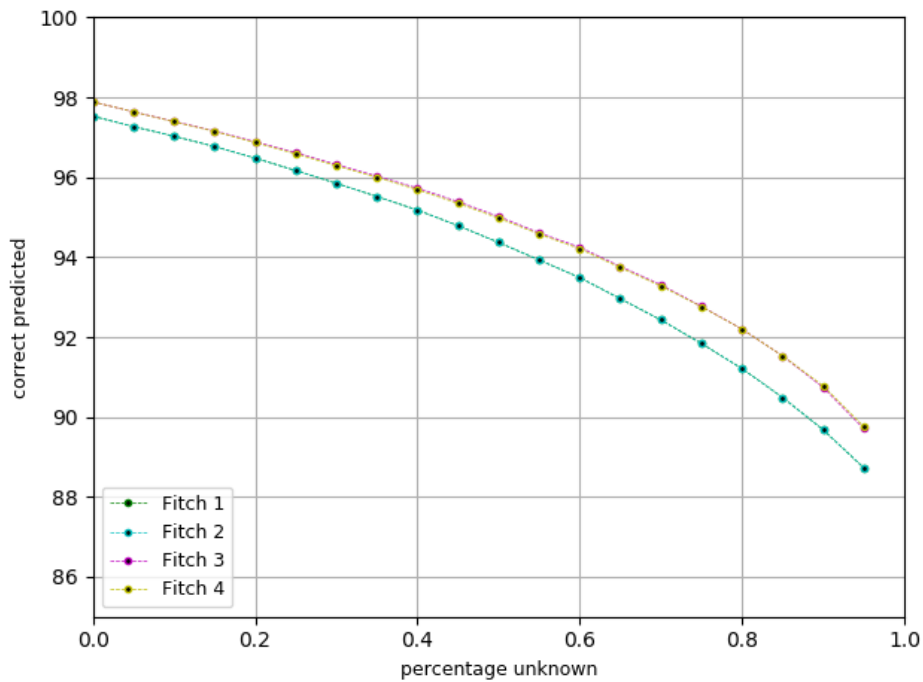


Figure 2.5: Test of Fitch Versions.

These four versions were tested in the simulation with 100 trees and 10000 leaf nodes and a distribution of 60 % FL to 40 % P. Figure 2.5 shows this over all unknown node percentage. At 90 % unknown nodes and 90 % of multifurcation of the internal nodes, version 1 was 89.26 %, version 2 was 89.26 %, version 3 was 89.35 %, and version 4 was 89.31 % correct. Therefore, only version 3 was used for all further simulations.

Sankoff

Maximum parsimony algorithm from Sankoff implemented in the R package castor [3].

TODO: transition probabilities: all equal

2.5 Implementation

You can find the full code on GitHub: github.com/Irallia/IZW-HU-Parasites.

Most of the code was written in Python. The analyzes and the use of the Castor package in R. There are some shell scripts to execute whole workflows.

3 Results

A big point in this chapter is the result of examining the input data. How is the situation? What influence does that have on our actual result? What can we do about it? Our simulation gave us some results to this.

Otherwise, this chapter is mainly about the actual reconstruction of the states. This means, on one hand investigation of origins and losses of the inner nodes and on the other, the prediction of unknown states of leaf nodes.

3.1 Metadata analysis

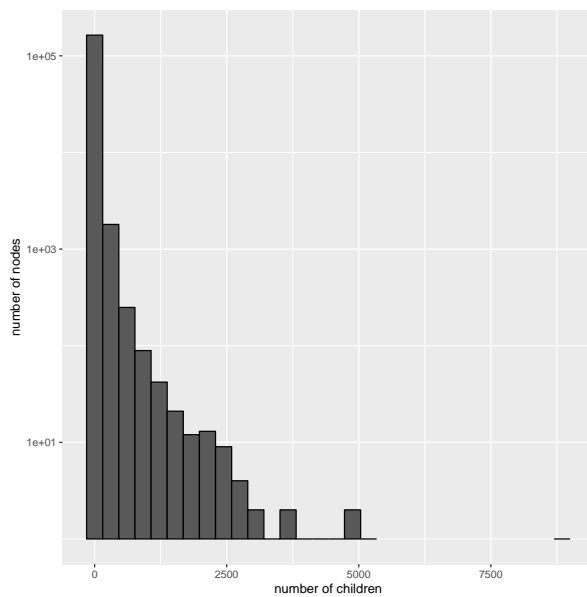
3.1.1 Missing information

As previously presented, we have two types of missing information: unknown states of leaf nodes and multifurcation.

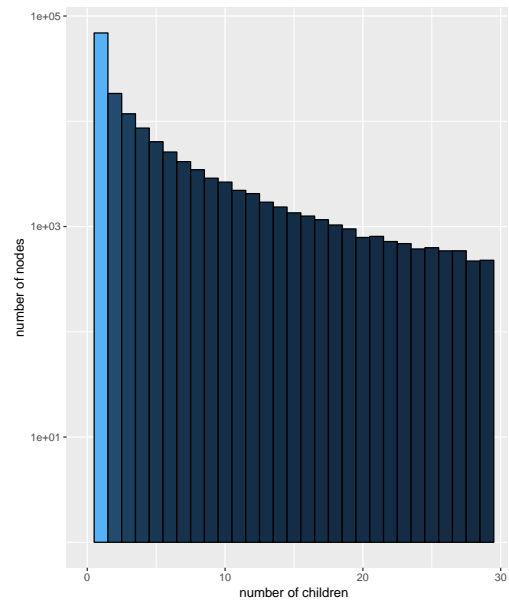
We examined the ridge of multifurcation of the tree. A complete phylogenetic tree would be binary, which means the number of leaf nodes is closely to the number of internal nodes. But since we only work with a synthesis tree, this tree is multifurcated: we have 241 974 internal nodes and 2 293 463 leaf nodes.

For a first overview we collected for every node its number of children (degree -1), and plotted this in two histograms, see figure 3.1.

The multifurcation affects only the internal nodes. We collected the number of children -2



(a) Histogram with automatic binwidth.



(b) Histogram with $binwidth = 1$.
light blue: binary; dark blue: multi-function

Figure 3.1: Histograms about the multifurcation of the internal nodes of the synthesis tree.

of every node (a node with two children is binary). That means it describes the number of nodes which we have lost from the real (binary) phylogenetic tree.

As **you** can see, we are very far from a binary tree.

Some subtrees have been examined, these have the following percentages of missing information: See table 3.1.

Data artifacts

At this point we also found out that there are some nodes with only one child node (55700 nodes).

These are both, the most nodes are right in front of a leaf, as well as some nodes are deep in the tree (3956 with height > 2). They are probably a result from the fact that taxonomic information has been incorporated into a phylogeny.

Some examples:

Domain / Kingdom / Phylum / Class	% unknown states (missing state information of leaf nodes)	% multifurcation (missing internal nodes)
Eukaryota	97.34 %	98.16 %
Metazoa	96.44 %	87.93 %
Fungi	98.87 %	96.97 %
Chloroplastida	99.14 %	89.46 %
Apicomplexa	86.26 %	87.16 %
Nematoda	89.01 %	88.59 %
Chordata	88.59 %	66.49 %
Platyhelminthes	68.73 %	80.34 %
Insecta	97.11 %	90.78 %

Table 3.1: Examination of subtrees regarding missing information

- Nephroselmidophyceae: (class)
<https://tree.opentreeoflife.org/opentree/argus/ottol@1038762>
- Phrynocrinidae: (family)
<https://tree.opentreeoflife.org/opentree/argus/ottol@3647979>
- Elaeocarpus sylvestris:
<https://tree.opentreeoflife.org/opentree/argus/opentree9.1@ott166969>

Taxa

The investigation of the taxonomy revealed that our tree has three kingdoms: Chloroplastida, Metazoa, Fungi, 53 phyla, 195 classes and 924 orders.

Since the analysis of the tree is not part of this work, it should be mentioned here that, according to recent findings, this is not complete and we lack some taxa in every rank. For example, Cavalier-Smith says that one distinguishes between seven and nine kingdoms [24]. In section 46 of the appendix **you** can find a list of all phyla.

Poisson regression of the multifurcation

The intercept is $2.821 > 0 \Rightarrow$ there is a multifunction. (Intercept: Stärke der Multifurcation)
 Comparing the different kingdoms, we find that multifunctionality is greater in Fungi than in Chloroplastida than in Metazoa:

$$4.0999(\text{FungiIntercept}) > -0.9132(\text{ChloroplastidaIntercept}) > -1.4320(\text{MetazoaIntercept})$$

Wir haben außerdem drei komplexitätsstärken von Modellen verglichen bezüglich der höhe und tiefe des Baums mit dem folgenden Deviance Table:

Model / Taxa	Kingdom	Phylum	Class	Order	Family
multifurc ~ taxa	7774454	7435700	7337241	7076068	
multifurc ~ taxa + depth	7752303	7431609	7334754	7027578	
multifurc ~ taxa + max.height	7730196	7375889	7275856	7005424	
multifurc ~ taxa + min.height	7472500	7233486	7144686	6890703	
multifurc ~ taxa + mean.height	7304402	7128318	7055313	6815271	
multifurc ~ taxa * depth	7714881	7335396	7250759	6843004	
multifurc ~ taxa * max.height	7692980	7311241	7187504	6795823	
multifurc ~ taxa * min.height	7442387	7177002	7094933	6795099	
multifurc ~ taxa * mean.height	7247309	7020258	6965794	6665565	

Table 3.2: Residuals...

* Residuals: Fehler - wieviele Werte sind nicht gut modelliert. (umso kleiner umso besser - grün)

Interpretation: Die Multifurkation ist sehr ungleich verteilt. Daher ist die vorhersage umso genauer umso kleinere Subtrees wir betrachten. ...

Because of the difference in the complexity of the models, we compared their BICs:

Model / Taxa	Kingdom	Phylum	Class	Order	Family
multifurc ~ taxa	8273333	7937828	7842157	7644249	
multifurc ~ taxa	8257680	7922207	7826490	7574154	
multifurc ~ taxa + depth	8273318	7934322	7839364	7539999	
multifurc ~ taxa + max.height	7993515	7749121	7661817	7416211	
multifurc ~ taxa + min.height	8251211	7875521	7778327	7516883	
multifurc ~ taxa + mean.height	7825417	7644249	7572474	7340741	
multifurc ~ taxa * depth	8235932	7836755	7757688	7383808	
multifurc ~ taxa * max.height	7963438	7693555	7614820	7335338	
multifurc ~ taxa * min.height	8214030	7808940	7690618	7336627	
multifurc ~ taxa * mean.height	7768360	7536296	7484953	7206369	

Table 3.3: BIC...

Binomial regression of the unknown state information

Next to the problem of the multifurcation of the tree is the less of data we have for the species. For the ancestral state reconstruction, we need information in the leaf nodes.

The eukaryotic synthesis tree has 293 463 leaf nodes. The GloBI database has 5 346 414 interactions (at this timepoint). Out of this data we got 51 337 distinct free-living species and 47 332 distinct parasite species → unknown nodes 2194794 ($\approx 95.7\%$).

We found also 57,352 (not distinct) source species and 809,993 (not distinct) target species without OTT ids. Since we currently use only OTT ids, we could not use this information.

TODO: With this only $\approx 4.3\%$ information in our leaf nodes are ...

We also compared different models in terms of their BICs (Table: 3.5). The Residuals are not very meaningful here, since all models have different dimensions.

Model / Taxa	Kingdom	Phylum	Class
multifurc ~ taxa	545740	499227	482265
multifurc ~ taxa + depth	544789	493017	478998
multifurc ~ taxa * depth	544062	488366	476382

Table 3.4: Residuals of unknown information

TODO: Taxa like order or family were too expensive to calculate...

Model / Taxa	Kingdom	Phylum	Class	Order
multifurc \sim taxa	545799	500004	485121	XXXXX
multifurc \sim taxa + depth	544862	493808	481869	478851
multifurc \sim taxa * depth	544179	489845	481494	478188

Table 3.5: BICs of unknown information

3.1.2 Results of simulation / Influence of different parameters

As presented, we compare two methods in our simulation to their prediction accuracy: Fitch and Sankoff.

We examine different parameters. In figure 3.2 is an overview of the results.

The first column describes the distributions of free-livings and parasites with a given threshold for the respective simulations to the right.

The middle column investigates the influence of the unknown states, the right the influence of the strength of the multifurcation.

The y-axes indicate the percentage of correctly predicted states (including known states).

On the x-axis the percentage of forgotten states or missing internal nodes.

Each point corresponds to the average of one hundred simulations, each with 10,000 leaf nodes.

For the middle column we set the strength of the multifurcations to 0.95% similar to the real data and in the right column the amount of the unknowns to 0.95% also similar to the real data.

As **you** can see both algorithms are always over 50% and therefore better than guessing. Moreover, they are usually close to each other, with Sankoff always makes better predictions except for equally distributed states as Fitch.

You can see that the distribution of states has a strong influence on the prediction. The more evenly distributed, the harder it is to predict.

If more than 60% of internal nodes are missing, Fitch breaks significantly in his prediction compared to Sankoff.

In the end, Sankoff is in most cases the more accurate algorithm and was therefore used for our prediction of the real data.

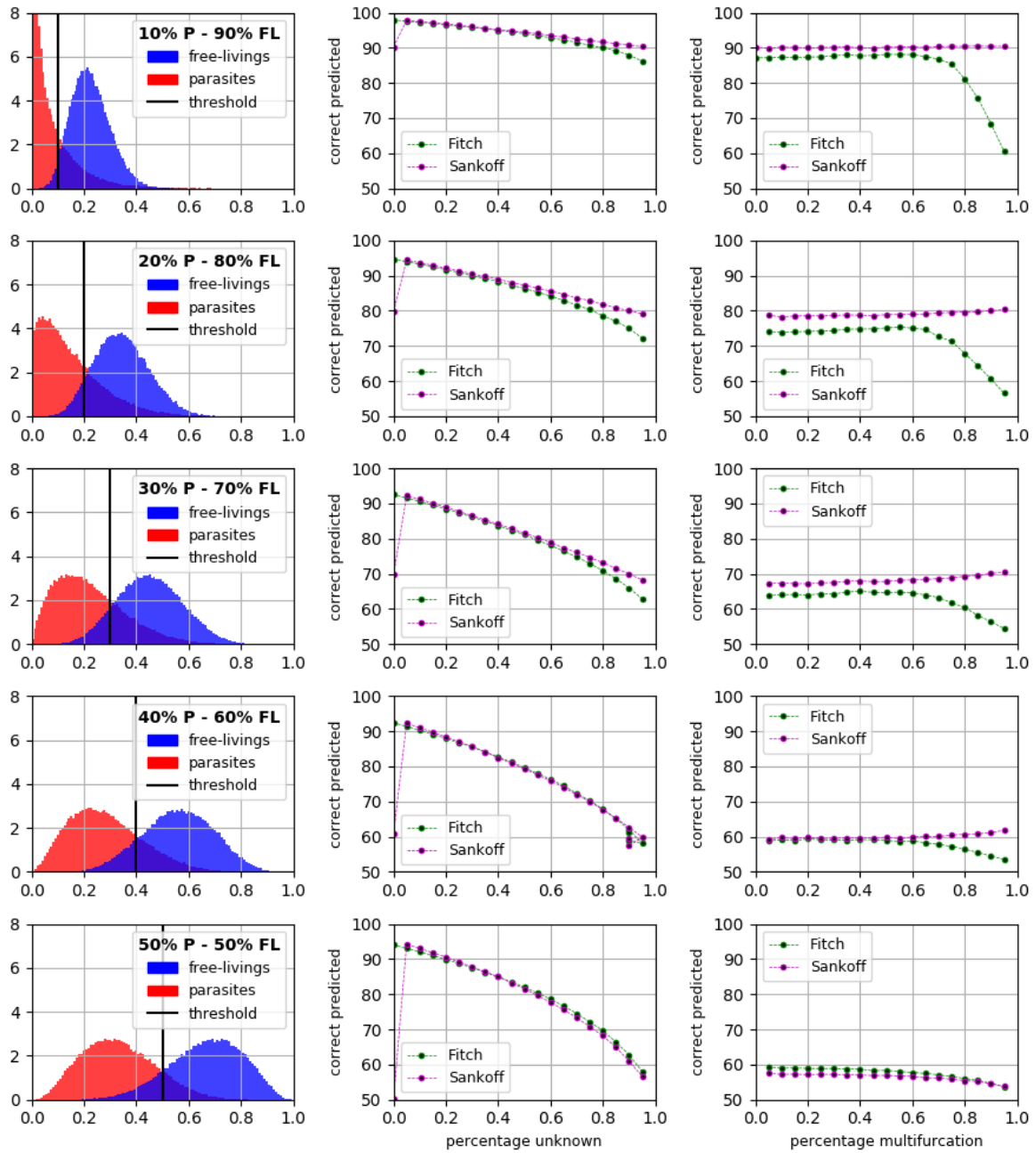


Figure 3.2: Influence of unknown data to prediction

3.2 Results of the real data analysis created with Sankoff

3.2.1 Biological view

To analyze the results, we have selected some phyla (sub-trees) to evaluate our results selectively from the biological point of view: Chordata, Nematoda, Platyhelminthes and Apicomplexa.

In Table 3.6 we compare the given states with the predicted ones.

There are several factors that play a role here, and in part, crystallize in these examples.

An important factor here is that the credibility of the results with the accuracy of the input data, i. the data of GloBI, stands and falls. Errors of incorrect input data can be amplified by incorrect prediction of unknown species and can be reversed in order to improve the data situation of GloBI.

Since we look at such large trees we can not expect to know all the parasites, so we look at individual positives. Positive in the sense that the majority have the opposite state.

Phylum	# nodes	original states		final states			
		FL	P	0 (FL)	0.3	0.5	1 (P)
Chordata	91785	10451 99.83 %	18 0.49 %	91759 99.97 %	0	0	26 0.03 %
Nematoda	30127	21 0.63 %	3289 99.37 %	1604 5.32 %	142 0.47 %	1196 3.97 %	27185 90.23 %
Platyhelminthes	22683	7 0.1 %	7086 99.9 %	175 0.77 %	0	0	22508 99.23 %
Apicomplexa	1863	1 0.39 %	255 99.61 %	1 0.05 %	0	0	1862 99.95 %

Table 3.6: Phylum (leaf nodes)

In contrast to the other phyla examined, the phylogeny is more pronounced (less multifurcation) (see Table ??). This results in a weaker spread of errors. This is reflected in the results from the table 3.6. There are 18 parasites as input data and only 8 more are predicted. The chordata mostly consist of free-living species, so this seems believable. We started with

99.83 % species and predict 99.97 % species as free-living (including already known nodes).
 TODO: A few punctual tests give us the following:

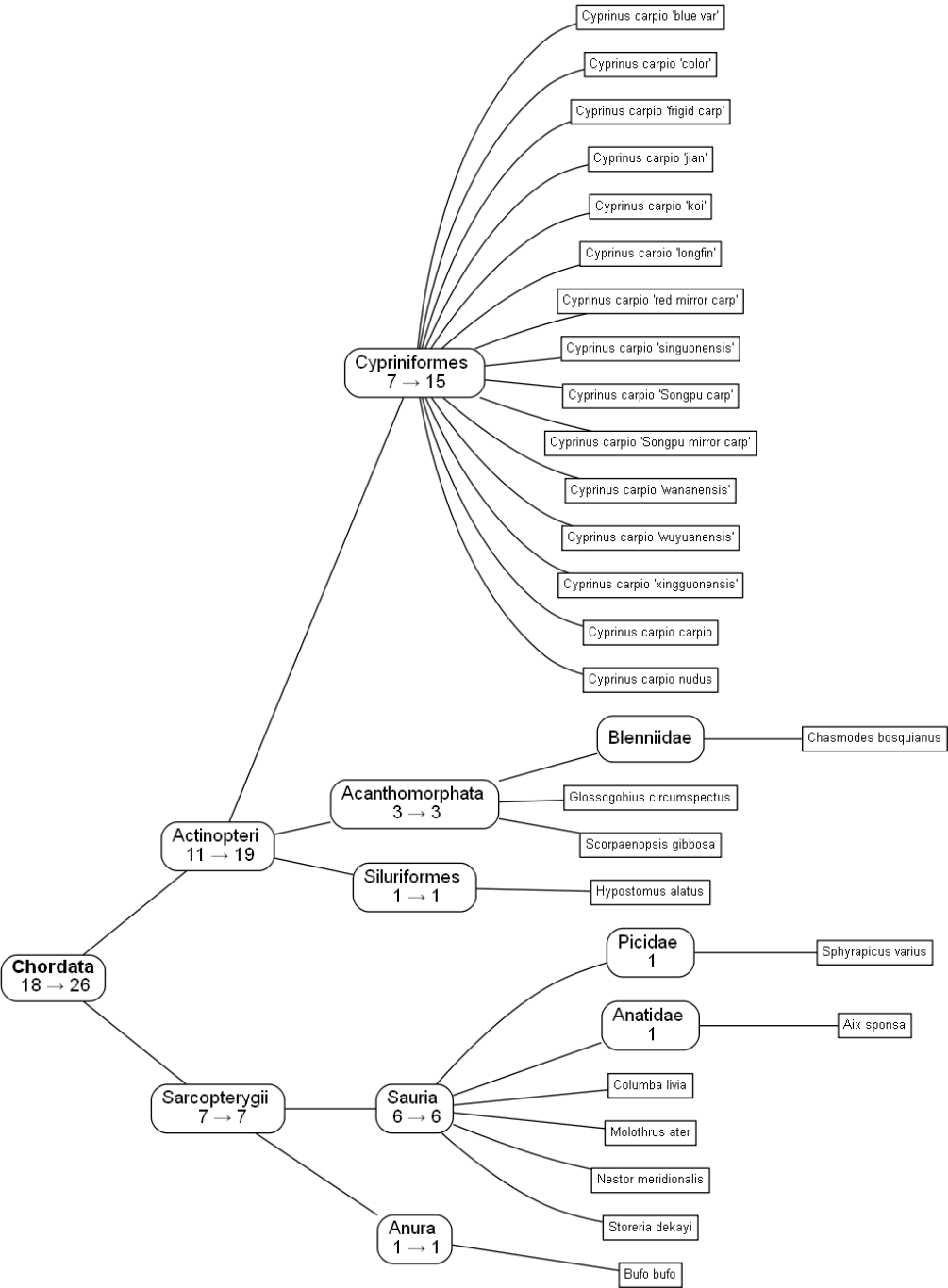


Figure 3.3: Parasites of Phylum: Chordata

We mapped the few parasitic species in a rough taxonomy (see Figure 3.3):

Known parasitic birds belong to the order Sauria. Here we know from Rothschild here there are breeding parasites, like the cuckoo and clepto-parasites as the skuas [25]. We got 6 input parasites from GloBI and there are no predictions: A woodpecker - *Sphyrapicus varius* and a duck - *Aix sponsa*, a cow bird - *Molothrus ater* known as broodparasite and some others.

The Apicomplexa are a parasitic phylum. We found only one input organism: *Stemonitis fusca* as a free-living species. In GloBI it is listed as being parasitized by *Nectria candicans* and *Nectriopsis sporangiicola*¹. The algorithm has not predicted new free-livings.

For the Platyhelminthes the literature says that there are mostly all Platyhelminthes parasites TODO: ref einfügen. But at the end we predicted 0.1 % as free-living. The class Seriata is the reason for the most of free-livings in this phylum. These are partly free-living flatworms, so the prediction looks right.

TODO: Wiki:

*The Seriata are an order of turbellarian flatworms.[1][2]

They are found in both freshwater and marine environments, and also include a number of species found in damp terrestrial conditions. Most are free-living, but the group includes the genus *Bdelloura*, which lives comensally on the gills of horseshoe crabs. Seriatans are distinguished from other related groups by the presence of a folded pharynx and of a number of diverticula arising from the intestine. The intestine itself may be either simple or branched.[3]

Note:Platyhelminthes had less missing information than starting point, see Table 3.1.

With the Nematoda it looks very different. In the Nematoda its much worse. Large parts of Nematoda are free-living, but we found only 5.32 % of them. Blaxter et al. speaks of at least seven independently arose parasitism [26]. In a recent article Blaxter identifies 18 origins [27] in Nematoda.

The problem at this point, however, is obvious: The parasites have been much more studied and thus we start with only 0.63 % free-living species. Against such a shifted

¹<https://www.globalbioticinteractions.org/?interactionType=parasiteOf&targetTaxon=Stemonitis%20fusca>

data situation, the algorithm is almost powerless. And yet the percentage has increased.

TODO: The evolution of parasitism in Nematoda:

"while only approximately 23 000 species have been described (J. Hallan, unpublished; <https://insects.tamu.edu/research/collection/hallan/>), the true species-level diversity may be 1 million or more (Lambshhead, 1993)."

" Estimates of the number of species of parasitic nematode per host suggest that there may be of the order of 25 000 nematode parasites just of vertebrates, most of which remain undescribed (Dobson et al. 2008)"

"a large proportion of nematode species may be parasites." [27]

TODO: Im Folgenden folgen 3 weitere ähnliche tabellen. Einmal eine ähnliche Tabelle inklusive interner Knoten 3.7, zweimal die Übersicht über die Kingdoms Blattknoten 3.8 bzw interne Knoten 3.9. Welche nehmen wir? Rest Appendix oder ganz raus?

Phylum	# nodes	original states		final states					
		FL	P	0 (FL)	0.4	0.5	0.67	0.75	1 (P)
Chordata	122546	10451	18	122473	0	0	0	0	73
Nematoda	33564	21	3289	846	0	1133	0	0	31585
Platyhelminthes	27142	7	7086	1010	0	175	0	0	25957
Apicomplexa	2102	1	255	1	0	0	0	0	2101

Table 3.7: Phylum (inkl internal nodes)

Kingdom	# nodes	original states		final states							
		FL	P	0 (FL)	0.25	0.33	0.4	0.5	0.67	0.75	1 (P)
none	75446	45	529	13426	220	24082	0	7792	5302	0	24493
Fungi	31457	577	2983	38520	0	0	0	5723	0	0	266463
Chloroplastida	416478	3519	77	410795	0	0	0	4182	0	0	1501
Metazoa	1491012	30758	22373	1328135	0	0	930	25535	4423	1665	130324

Table 3.8: Kingdom (leaf nodes)

Kingdom	# nodes	original states		final states							
		FL	P	0 (FL)	0.25	0.33	0.4	0.5	0.67	0.75	1 (P)
none	84456	45	529	15035	243	25910	0	8764	6183	0	28140
Fungi	324105	577	2983	39088	0	0	0	5858	0	0	274803
Chloroplastida	460457	3519	77	454211	0	0	0	4688	0	0	1558
Metazoa	1670956	30758	22373	1485749	0	0	1313	29002	5102	1957	147833

Table 3.9: Kingdom (inkl internal nodes)

3.2.2 Origins and Losses

Weinstein and Kuris have been searching for origins of parasitism in Animalia [4]. They identified 223 parasitic origins: 223 in Metazoa \supset 143 in Arthropoda \supset 87 in Insecta.

This has led us to count the origins and losses of parasitism in our investigation as well.

We count only one origin / loss in a parent node with different children's nodes.

Here we have encountered a problem: The Castor algorithm gives us probabilities for states. That means there are also nodes with state like 0.3 or 0.5. So how do **you** count? Our solution was, to round these values. We have to say that we round 0.5 to 0.

In Table 3.10 we can see, that we found some more origins than Weinstein and on top of that some losses.

Lets have a look at the same phyla as in the section before: Chordata, Nematoda, Platyhelminthes and Apicomplexa.

Chordata are full of free-living species and so we see only a few origins of parasitism. The root and mostly all species are predicted as free-living.

In Apicomplexa and the Platyhelminthes are looking fine too. Our algorithm gives us only one loss of parasitism in Apicomplexa and five in the Platyhelminthes. They are both from the root over mostly all species predicted as parasites.

Nematoda is again full of problems. The rootnode is predicted as a parasite and so we have more losses of parasitism for the less information of free-living species in this phylum. The rest is parasitic

As we have already mentioned Blaxter et al. found at least seven origins of parasitism [26].

Domain / Kingdom / Phylum / Class	# internal nodes	# leaf	Rootnode state	without # origins (FL -> P)	and with rounding # losses (P -> FL)
Eukaryota	241974	2293463	1.0 P	415 462	363 369
Metazoa	179944	1491012	0.5	294 321	123 129
Fungi	9534	314571	0.5	80 97	222 222
Chloroplastida	43486	412434	0.0 FL	40 42	2 2
Arthropoda	120479	1198981	0.0 FL	260 281	102 108
Apicomplexa	239	1863	1.0 P	0 0	1 1
Nematoda	3437	30127	1.0 P	0 2	11 11
Chordata	30761	91785	0.0 FL	12 12	1 1
Platyhelminthes	4459	22683	1.0 P	0 0	5 5
Insecta	91256	989572	0.0 FL	234 245	77 77

Table 3.10: Origins and losses

If we assume that the root node of Nematoda is free-living, then some losses would have to turn around and become Origins. So it could be that we end up in a similar size as Blaxter.

```

# possible tags: 0, 0.333, 0.4, 0.5, 0.667, 0.75, 1
# rounded to:   0  0      0   0   1     1     1
if node_state != father_state:
    if father_state == 0:
        origins += 1          # FL -> P
        new_found = True
    else:
        losses += 1           # P -> FL

```

```
new_found = True
```

TODO: without rounding change else: to elif *father_state* == 1

3.2.3 Cross evaluation - leave 100 out

We ran the castor algorithm 100 times with leaving 100 randomized free-living or parasitic species out of the input data to see how stable our result is. Of these 10,000 nodes, 9,238 were unique. Of that, we predicted 9060 ($\approx 98.17\%$) correctly and 169 ($\approx 1.82\%$) wrongly, with duplicate draws always having the same prediction.

What is the best way to model this data? We again tested the influence of the taxa and the depth of leaf nodes and calculated the BICs (Table: 3.12).

Model / Taxa	Kingdom	Phylum	Class	Order
correct predicted \sim taxa	XXXXX	XXXXX	XXXXX	XXXXX
correct predicted \sim taxa + depth	117703	XXXXX	XXXXX	XXXXX
correct predicted \sim taxa * depth	117592	XXXXX	XXXXX	XXXXX

Table 3.11: Residuals of cross validation prediction

Model / Taxa	Kingdom	Phylum	Class	Order
correct predicted \sim taxa	117936	112242	XXXXX	XXXXX
correct predicted \sim taxa + depth	117776	XXXXX	XXXXX	XXXXX
correct predicted \sim taxa * depth	XXXXX	XXXXX	XXXXX	XXXXX

Table 3.12: BICs of cross validation prediction

Residuals:

- correct predicted \sim 1: 120325
- correct predicted \sim kingdom: 117877
- correct predicted \sim phylum: 111466

	min	max	mean	variance (σ^2)	σ
all	0	3587.70	224.96	313650.61	560.05
distance leaf nodes	0	3021.12	208.69	248103.38	498.10
internal nodes	0	566.58	16.28	4927.95	70.20
changed tag	0	0	0	0	0
lost all tags	100	100	100	0	0
FL tags	44	66	57.25	19.50	4.42
P tags	34	56	42.75	19.50	4.42

Table 3.13: Statistics to Cross validation

What could happen by removing a parasite or free-living of the list?

- It could be a specie, which don't exist in the tree leaf nodes. -> no effect
- It could be a specie, which exists in both lists. -> If it was a parasite, it is now free-living, because we prefer parasites. Otherwise we have no effect again. (1053 are possible)
- Normal case: We loose information, because its a specie in our tree and we change it to a leave node with no information.

Influence on the rest of the data:

3.3 Effects of Taxa in the different models

The comparisons of the effects of the taxa can be found in Table 3.15 and showed ...

globi ~ taxa * depth: NOTE: kingdom is not a high-order term in the model

	min	max	mean	variance
distance all	1	2	1.33	0.33
	0	3587.70	217.94	273760.68
leaf nodes	1	2	1.33	0.33
	0	3021.12	202.57	209274.86
internal nodes	0	0	0.00	0.00
	0	566.58	15.37	5684.00
lost FL tags	44	49	46.67	6.33
	51	66	57.95	13.47
P tags	51	56	53.33	6.33
	34	49	42.05	13.47

Table 3.14: Statistics to Cross validation

- Less free-livings - 3 examples
- Less parasites - 64 examples

Taxa	Model / Effects	min	max	mean	median
Kingdom	globi ~ taxa	0,01	0,04	0,02	0,01
	globi ~ taxa + depth	0,01	0,03	0,02	0,01
	globi ~ taxa * depth	0,00	0,03	0,01	0,01
Phylum	globi ~ taxa	0,17	393501,80	32208,79	6149,78
	globi ~ taxa + depth	0,32	567010,90	55149,54	10783,18
	globi ~ taxa * depth	0,00	1000000,00	225375,33	2511,58

Table 3.15: Effects of Taxa in models for unknown data

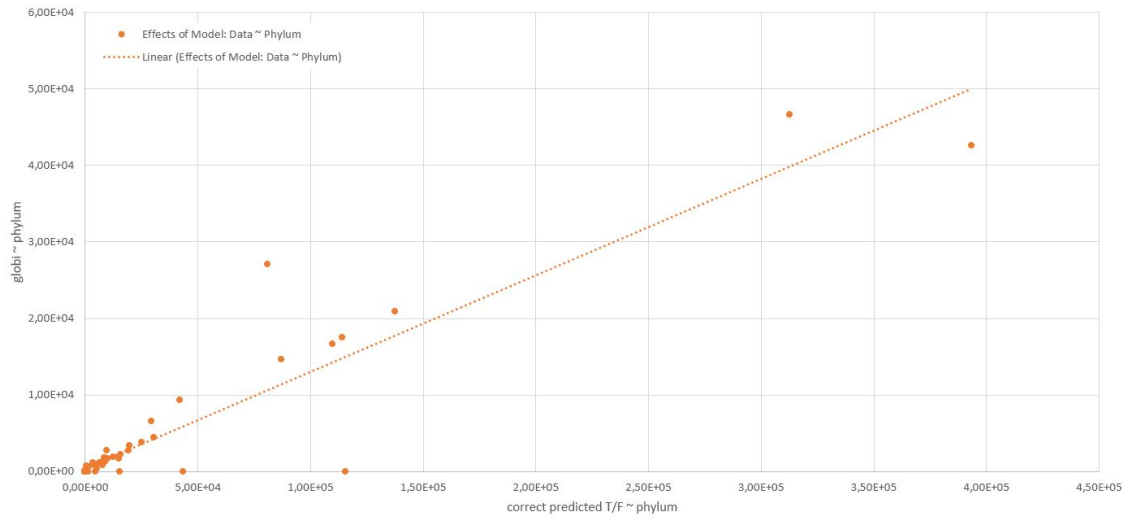


Figure 3.4: Effects of Model: Data ~ Phylum

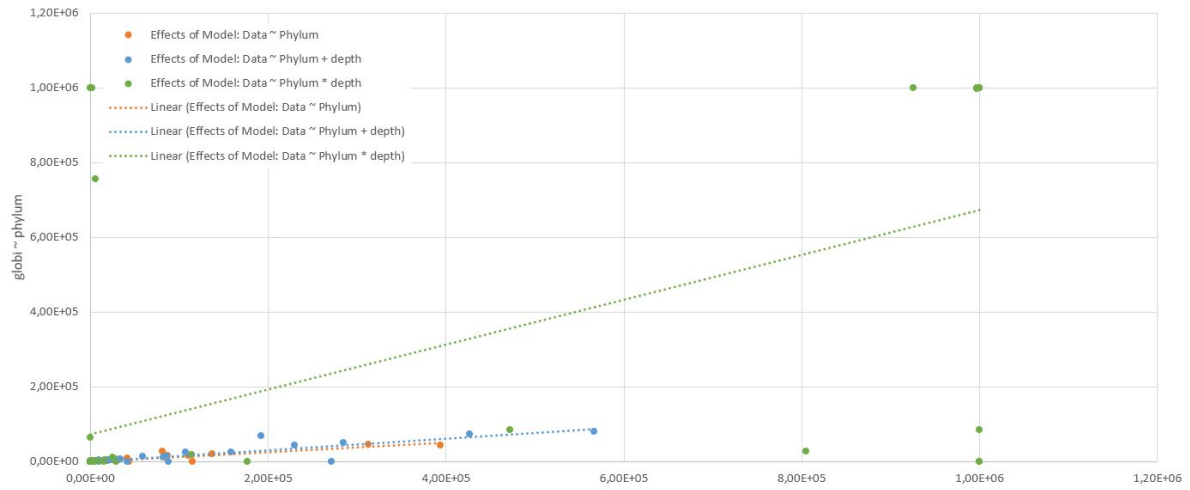


Figure 3.5: Effects of 3 Models: Data ~ Phylum (+/* depth)

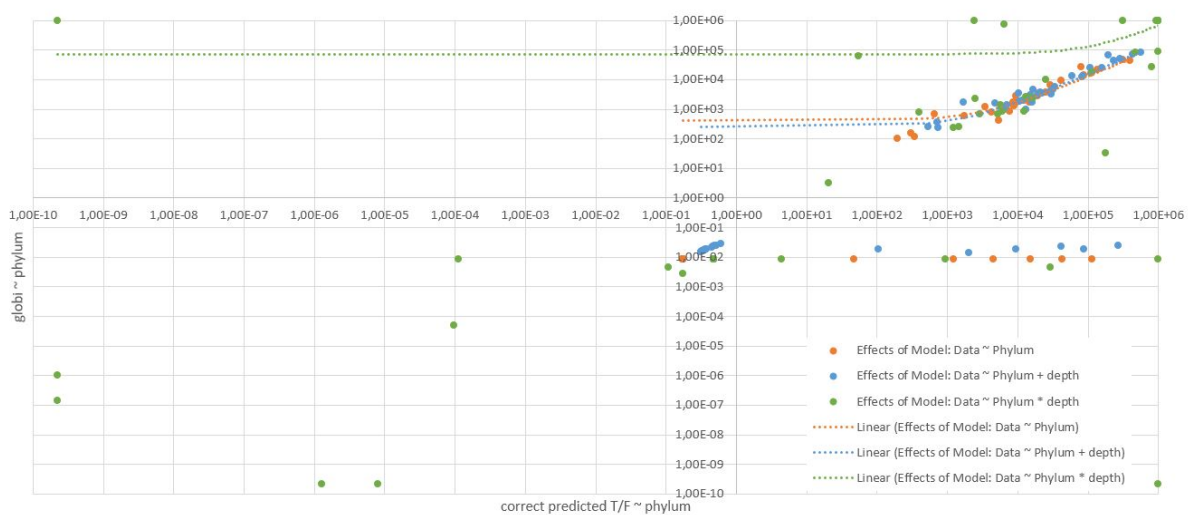


Figure 3.6: Effects of 3 Models: Data \sim Phylum (+/* depth)
both axis in log scale

4 Discussion

Fehlerquote der Daten an sich?

Wie gut ist unsere Datenlage? 3 mio knoten, 1.8 named species (leaf nodes), 200.000 leaf nodes mit Information.

Welche Teile des Baumes sind gut, an welchen muss noch viel geforscht werden.

Wieviele Origins haben wir gefunden, was bedeutet diese Zahl?

Es gibt noch ungenutzte information in GloBI.

Wir können GloBI verbessern ... Wir haben GloBI auf die gefundenen Abweichungen aufmerksam gemacht... Fehler werden durch die prediction eventuell verstärkt...

4.1 Simulation

For fixed calculations we assume a distribution of 60% free-living to 40% parasites, 95% missing data and a multifurcation rate of 95%. How well does our simulation approach the real data situation?

There are some points to discuss:

- How close is the randomized binary tree to a true phylogeny?
- The distribution of parasites to free-ranging species is a pure assumption. From different sources about 40% parasites are estimated, are these also beta-distributed?
- How well do the transition probabilities match parasitism? Depending on the type of parasitism, this will certainly look different. In addition, it can be assumed that the probability for losses is lower than for origins.

- The equal distribution of the multifurcation is a blank assumption. It is very likely to be higher towards the root node because there is less information back in time. On the other hand, there are many studies that have studied various species without considering phylogeny. Which also creates a high number of children of a node directly at the leaf nodes.

In order to be able to observe these problems, we have programmed our simulation according to the parameters. This allowed us to estimate the influence of some of these parameters.

Bibliography

- [1] Hinchliff, C. E. et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* **2015**, *112*, 12764–12769.
- [2] Poelen, J. H.; Simons, J. D.; Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* **2014**, *24*, 148 – 159.
- [3] Louca, S.; Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **2017**, btx701.
- [4] Weinstein, S. B.; Kuris, A. M. Independent origins of parasitism in Animalia. *Biology Letters* **2016**, *12*.
- [5] Camin, J. H.; Sokal, R. R. A Method for Deducing Branching Sequences in Phylogeny. *Evolution* **1965**, *19*, 311–326.
- [6] Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* **1971**, *20*, 406–416.
- [7] Swofford, D. L.; Maddison, W. P. Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences* **1987**, *87*, 199 – 229.
- [8] Goberna, M.; Verdú, M. Predicting microbial traits with phylogenies. *The Isme Journal* **2015**, *10*, 959 EP –, Original Article.
- [9] Felsenstein, J. *Inferring Phylogenies*; Sinauer, 2003.
- [10] ITIS, Integrated Taxonomic Information System. <https://www.itis.gov>.
- [11] Bethesda (MD): National Library of Medicine (US), N. C. f. B. I. National Center for Biotechnology Information (NCBI). 1988; <https://www.ncbi.nlm.nih.gov/>.

- [12] Horton, T. et al. World Register of Marine Species (WoRMS). <http://www.marinespecies.org>, 2018; <http://www.marinespecies.org>, Accessed: 2018-02-27.
- [13] GBIF, Global Biodiversity Information Facility. <https://www.GBIF.org>.
- [14] Diego Vázquez, R. N., Jeremy Goldberg Interaction Web Database (IWDB). 2003; <https://www.nceas.ucsb.edu/interactionweb/>.
- [15] Webs on the Web (WOW): 3D visualization of ecological networks on the WWW for collaborative research and education. 2004; pp 5295 – 5295 – 9.
- [16] Myers, P.; Espinosa, R.; Parr, C. S.; Jones, T.; Hammond, G. S.; Dewey, T. A. The Animal Diversity Web (online). 2018; <https://animaldiversity.org>.
- [17] Cohen, J. E. c. Ecologists' Co-Operative Web Bank. Version 1.1. Machine-readable database of food webs. *New York: The Rockefeller University* **2010**,
- [18] Windsor, D. A. Controversies in parasitology, Most of the species on Earth are parasites. *International Journal for Parasitology* **1998**, *28*, 1939–1941.
- [19] Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.
- [20] Royer-Carenzi, M.; Pontarotti, P.; Didier, G. Choosing the best ancestral character state reconstruction method. *Mathematical Biosciences* **2013**, *242*, 95 – 109.
- [21] Farris, J. S. Methods for Computing Wagner Trees. *Systematic Biology* **1970**, *19*, 83–92.
- [22] Sankoff, D. Minimal Mutation Trees of Sequences. **1975**, *28*.
- [23] Cunningham, C. W.; Omland, K. E.; Oakley, T. H. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution* **1998**, *13*, 361 – 366.
- [24] Cavalier-Smith, T. Eukaryote kingdoms: Seven or nine? *Biosystems* **1981**, *14*, 461 – 481.

- [25] Rothschild, M.; Clay, T. *Fleas, Flukes & Cuckoos; a Study of Bird Parasites*; New York, Macmillan, 1957; p 368, <https://www.biodiversitylibrary.org/bibliography/6413>.
- [26] L. Blaxter, M.; De Ley, P.; Garey, J.; Liu, L. X.; Scheldeman, P.; Vierstraete, A.; R. Vanfleteren, J.; Mackey, L.; Dorris, M.; Frisse, L.; Vida, J.; Thomas, W. A molecular evolutionary framework for the phylum Nematoda. **1998**, 392, 71–5.
- [27] BLAXTER, M.; KOUTSOVOULOS, G. The evolution of parasitism in Nematoda. *Parasitology* **2015**, 142, S26–S39.

5 Appendices

5.1 Methods overview

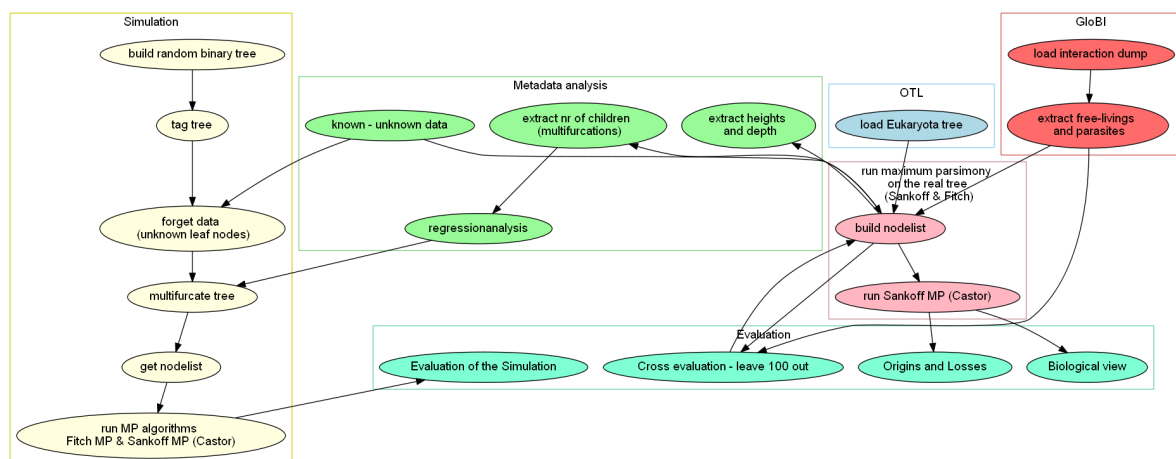


Figure 5.1: Big overview of the whole Workflow

5.2 OTL analysis

5.2.1 List of all phyla

Phyla (53):

Acanthocephala, Amoebozoa, Apicomplexa, Arthropoda, Ascomycota, Bacillariophyta, Basidiomycota, Brachiopoda, Bryozoa, Chaetognatha, Chlorophyta, Chordata, Chromerida,

Chytridiomycota, Ciliophora, Cnidaria, Colponemidia, Ctenophora, Cyclophora, Echinodermata, Entoprocta, Entorrhizomycota, Euglenida, Foraminifera, Gastrotricha, Glomeromycota, Gnathostomulida, Haplosporida, Haptophyta, Hemichordata, Kinorhyncha, Loricifera, Microsporidia, Mollusca, Myzostomida, Nematoda, Nematomorpha, Nemertea, Onychophora, Orthonectida, Phaeophyceae, Picozoa, Placozoa, Platyhelminthes, Porifera, Priapulida, Rhodophyta, Rhombozoa, Rotifera, Streptophyta, Tardigrada, Xanthophyceae
Wobei von Streptophyta -> Anthocerotophyta, Marchantiophyta, Bryophyta, Tracheophyta als Phylum im Phylum gefunden und nicht einbezogen wurden und Magnoliophyta als Phylum in Tracheophyta ebenfalls nicht.

Distribution of Taxa

- In the tree we can distinguish 28 different Taxa with the OTL taxonomic tree.
- The most of them are hardly represented. The major taxonomic groups are: ...
- Here **you** can see some characteristics of the Multifurcation of the tree.

In a phylogeny, the taxonomic division of the tree is far too coarse, meaning that there

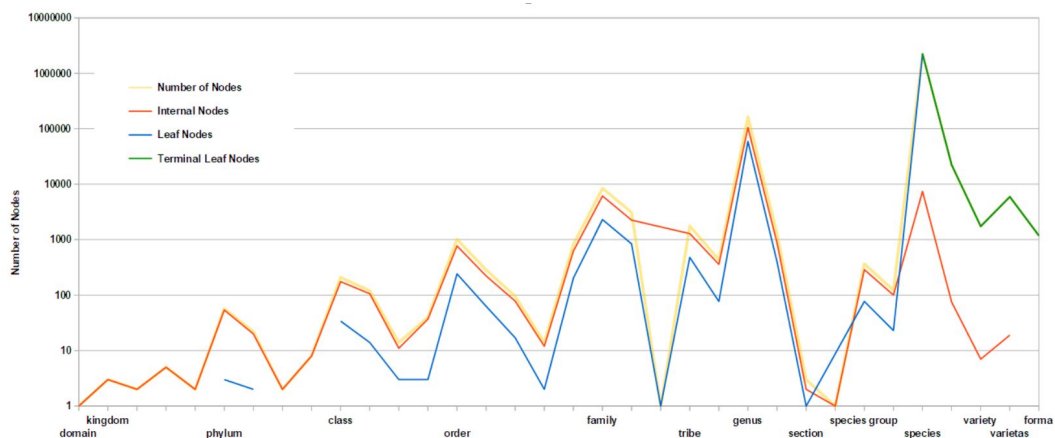


Figure 5.2: Distribution of Nodes in Rank-Categories

should be more subtaxa or 'unranked' nodes. But the closer we get to the root, the more the pure taxonomic tree is reflected. If the tree were binary, the taxa would have to double. But the multipliers for some are much bigger and for others much smaller, which **you** can

see in in figure 5.2.

... (see Table 5.2)

extended leaf nodes (real leaf nodes)

Distribution of data in the taxa

Mithilfe des taxonomischen Baums von OTL haben wir die Knoten ihren Kingdoms, Phyla und Classes zugeteilt (see Table 5.2).

Taxa	Number of Nodes	Internal Nodes	Leaf Nodes	Terminal Leaf Nodes
domain	1	1		
kingdom	3	3		
subkingdom	2	2		
infrakingdom	5	5		
superphylum	2	2		
phylum	57	54	3	
subphylum	22	20	2	
infraphylum	2	2		
superclass	8	8		
class	209	175	34	
subclass	120	106	14	
infraclass	14	11	3	
superorder	40	37	3	
order	1014	772	242	
suborder	285	222	63	
infraorder	95	78	17	
parvorder	14	12	2	
superfamily	829	626	203	
family	8449	6143	2306	
subfamily	3090	2250	840	
supertribe	1	0	1	
tribe	1764	1285	479	
subtribe	435	359	77	
genus	164656	105452	59204	
subgenus	1266	869	397	
section	3	2	1	
subsection	1	1	0	
species group	365	288	77	
species subgroup	123	100	23	
species	2247251	7423	2239828	2228993
subspecies	22437	75	22362	22239
variety	1755	7	1748	1726
varietas	5970	19	5951	5909
forma	1181		1181	1181
no rank	954	719	235	7
no rank - terminal	37452		37452	37452
(no entry)	40099	40099		

Table 5.1: TODO: ...

Kingdom (3)	Number of Nodes	Phylum (25)	Number of Nodes	max max height
Metazoa	1 465 207	Arthropoda	1 170 539	54
		Chordata	106 650	74
		Mollusca	80 022	22
		Platyhelminthes	27 141	9
		Nematoda	24 564	23
		Cnidaria	14 878	36
		Porifera	11 737	26
		Echinodermata	10 654	14
		Bryozoa	8 631	11
		Rotifera	3 093	7
		Nemertea	1 793	7
		Tardigrada	1 654	7
		Acanthocephala	1 596	6
		Brachiopoda	1 055	9
		Nematomorpha	633	7
		Chaetognatha	360	7
		Hemichordata	196	5
		Cycliophora	11	5
Fungi	254 871	Ascomycota	157 045	19
		Basidiomycota	92 931	18
		Microsporidia	1 949	6
		Glomeromycota	1 490	6
		Chytridiomycota	1 456	6
Chloroplastida	121 239	Streptophyta	120 731	49
		Chlorophyta	508	6

Table 5.2: TODO: ...