Freie Universität Berlin

Fachbereich Mathematik und Informatik

An analysis of maximum parsimony algorithms to predict parasitism in Eukaryota

using a large multifurcated phylogenetic synthesis tree

Submitted on:
3 April 2018

Lydia Buntrock
E-Mail: info@irallia.de

Supervisor:
Prof. Dr. Bernhard Y. Renard
&
Prof. Dr. rer. nat. Emanuel Heitlinger

Abstract

Parasitism can be defined as an interaction between species in which one of the interaction partners, the parasite, lives in or on the other, the host. The parasite draws food from its host and harms it in the process. According to some estimates, above 50% of all eukaryotes are parasites. Nevertheless, it is difficult to obtain information whether a particular taxon is a parasite computationally making it difficult to query large sets of taxa.

Here we test in how far it is possible to use the open tree of life (OTL), a synthesis of phylogenetic trees on a backbone taxonomy (resulting in unresolved nodes), to expand available information via phylogenetic trait prediction. We use the Global Biotic Interactions (GloBI) database to categorise 25,992 and 34,879 species as parasites and free-living, respectively and predict states for over 2.7 million (97.6%) leaf nodes without state information.

We estimate the accuracy of our maximum parsimony based predictions using cross-evaluation and simulation at roughly 80% overall, but strongly varying between clades. We describe this variation across taxa as associated with available state and topology information. We compare our results with several smaller scale studies, which used manual expert curation and conclude that computationally inferred state changes largely agree in number and placement with those. In clades in which available state information is biased (mostly towards parasites, e.g. in Nematodes) phylogenetic prediction is bound to provide results contradicting conventional wisdom.

This represents, to our knowledge, the first comprehensive computational reconstruction of the emergence of parasitism in eukaryotes. We argue that such an approach is necessary to allow further incorporation of parasitism as an important trait in species interaction databases

and in individual studies on eukaryotes e.g. in the microbiome.

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 $\approx 2.4\ \%$ 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

This thesis is about the analysis of ancestral state reconstruction algorithms for non-binary trees, applied to the currently largest phylogenetic synthesis tree of Open Tree of Life (OTL) [1], with the application of prediciton of parasitism.

1.1 Motivation

The aim of this thesis is the application of maximum parsiomony algorithms to non-binary trees and very large datasets. In particular, the example find the 'Origins of Parasitism' throughout the Eukaryota Tree of Life.
For these large phylogenetic synthesis trees, however, ancestral state reconstruction has so far only been done for Bacteria and Archaea for binary traits by Goberna and Verdú [5]. However, this differs from eukaryotes in the sense that complex traits such as parasitism depend on more than one gene.
TODO: Warum ist das relevant? Was könnte man dann tun?

1.2 Background

For about 50 years, people have been working on ancestral state reconstruction. The first paper is by Camin and Sokal, who in 1965 were working on algorithms for discrete-state data [6]. Different methods have been developed and the question is which method is the most suitable for the problem at hand: The ancestral state reconstruction for a huge multifurcated

tree with binary/two discrete states.
Royer-Carenzi et al. distinguishes two major classes of ancestral state reconstruction methods:
The first is maximum parsimony: explain the current state with the least number of state changes between the child and his ancestor.
The other class she presents describes 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 [7].

Parsimony methods are simple methods in which you can not include information such as branch length (development time of the species). This is also their main point of criticism compared to the more complex likelihood based models. Since there are no branch lengths or other information available for the present relationship tree, and thus can not take advantage of the other models, we decided in favor for parsimonious.

We will test the existing maximum parsimony algorithms Fitch [8] and Sankoff [9] for this task and estimate their predictive power. The present tree structure of OTL is not binary. A tree is multifurcated if each node has multiple ($n >= 3$) children [10].
Maximum parsimony methods are developed for phylogenies, which are usually depicted as binary trees. Parsimonious 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'.
The Sankoff method is implemented by Louca et al. for the non-binary case and is available as an R package called *Castor* [3]. In addition, we have implemented the Fitch method and adapted it for multifurcated trees.
This achieves that we can predict ancestral states and unknown states of living species for large non-binary relatives trees.

To accomplish this task, a large phylogenetic tree and information about the current species states is needed.
The biggest 'phylogenetic tree' is a synthesis of phylogenetic trees filled with a taxonomic trees given by Open Tree of Life (OTL) [1]. For the information about the current states of

the species we use the interaction database Global Biotic Interactions (GloBI) [2]. The data in GloBI are stored as interactions e.g. species A parasitize species B. From this we conclude that species A is parasitic and species B free-living.

At this point a few words to the term parasitic. There are different definitions. Since we use GloBI to classify species, we use their definition of parasitism. Again, in GloBi, Ontobee definitions are used [11]. 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."[1]. This definition includes: ecto- and endoparasites, parasitoids, kleptoparasites and pathogenes.

1.3 Structure of this work

The objectives of this work are the following points: (1) Find a suitable ancestral state reconstruction method. (2) Perform reconstructing on the Eucaryotic synthesis tree of OTL. The goal of Point 1 is to evaluate the possible methods based on a simulation of our data situation.
The one by Louca et al. implemented Sankoff algorithm is the best in our comparisons. Therefore, point 2 consists of reconstructing the ancestral states and predicting the unknown leaf states. And then perform an evaluation of the results.

[1] ontobee.org/ontology/RO?iri=http://purl.obolibrary.org/obo/RO_000244; Last checked: 22.03.2018.

2 Methods

As stated in the introduction, in this thesis, 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. A more detailed view of the workflow can be found in the appendix 5.1.

2.1 Description of data sets

Two types of data are needed for an ancestral state reconstruction: a tree and information about the states.
For the Tree Open Tree of Life (OTL) is used [1] and for the state information the Global Biotic Interaction database (GloBI) is taken [2].

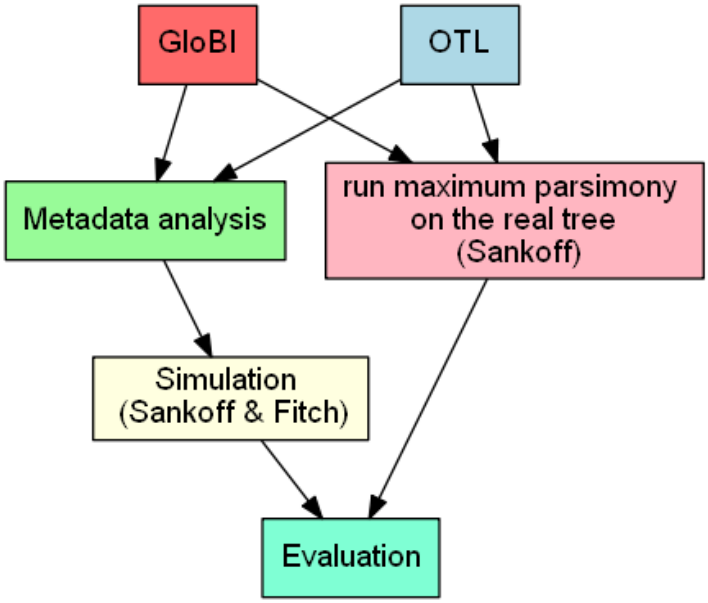


Figure 2.1: The Workflow of the resulting procedure with the following steps:
(1) Retrieve phylogenentic tree data as input for the tree (OTL) and the state information (GloBI). (2) Get metadata of these for a realisitc the simulation of the maximum parsimony algorithms (Fitch & Sankoff). (3) Build and 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) Evalute and interprete results.

2.1.1 OTL

For this 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[1]. 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) [12]
- NCBI (National Center for Biotechnology Information) [13]

[1] https://tree.opentreeoflife.org/about/synthesis-release/v9.1; Last checked: 22.03.2018.

- WORMS (World Register of Marine Species) [14]
- GBIF (Global Biodiversity Information Facility) [15]
- OTT (OpenTreeOfLife-Taxonomy) [1]

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 preffering 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 for this study, there is a section in the appendix 5.2.

2.1.2 GloBI

The most big interaction databases are offline or outdated. For example: IWDB (Interaction Web Database) [16], Webs on the Web [17], Animal Diversity Web [18] and ecoweb [19]. GloBI is including most of the known ones and is still growing actively [2]. So the question was answered rather quickly which interaction database could be used.

This database consists of entries of the form: species A (source) interacts with B (target). A number of interactions have been identified[2] , 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:

[2]https://github.com/jhpoelen/eol-globi-data/.../InteractType.java; Last checked: 22.03.2018.

- 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

Of these interactions, e.g. species A parasitize species B, the state of the species is determined, here is species A parasitic and species B free-living. The case a parasite conquers (parasitizes) another parasite yields conflicting states for the second species. This is solved by preferring parasitic.
For each species known IDs are stored in GloBI. This includes OTT (the taxonomy of OTL). All species that have stored an OTT id and have a matching interaction are formed into two lists: parasites and free-livings.

2.2 Metadata analysis

In order to generate a realistic simulation, influencing parameters are investigated.
Since the transitions are minimized in an ancestral state reconstruction, this is an important parameter to consider. On the other hand, the completeness of our input data are influencing values. Therefore, two major types are distinguished:

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 2.4).

2.2.1 Transition probabilities

As mentioned above, in an ancestral state reconstruction, transitions are minimized, with transition probabilities playing a role.
This subsection deals with these transition probabilities from free-living (hereinafter / as a formula FL) to parasitic (hereinafter P) and vice versa: $\mathcal{P}(FL \to P)$, $\mathcal{P}(P \to FL)$.
Different parasite types have different transition probabilities. It is very difficult to make a statement about these probabilities. It is generally assumed in this work that there are 40 % parasites and 60 % free-livings which is based on the estimates by Windsor [20] and $\mathcal{P}(FL \to P) = \mathcal{P}(P \to FL)$, as a result of the diversity of parasites and the lack of general determinations for this. These parameters are debated in subsection 4.2.1 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.

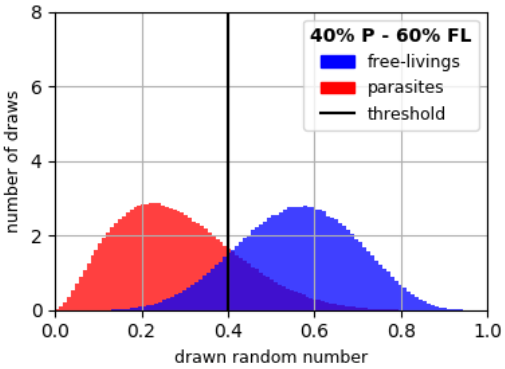


Figure 2.2: 60 % Free-living - 40 % Parasites
red: parasites, blue: free-living,
the threshold is at 0.4

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. In order to compare models of different complexity, the BIC (Bayesian Information Criterion) values were calculated in addition to the residuals. There are two different information criteria: AIC (Akaike Information Criteria) and BIC.
The advantage of the BIC is that the penalty is dependent on the sample size and is therefore advantageous for large samples.

For each node, depth, min, max and mean height were noted. Where the depth of a node is the distance (number of edges) to the root node and the height of a knot is described as the largest distance to a leaf node. In this work, a distinction is made between minimum, maximum and average distance ($\rightarrow$ min, max and mean height).
The influence in the modeling of these parameters was tested, additive as well as multiplicative.

For all these calculations, the following R functions were used: *glm()*, *anova()* and *BIC()*.

2.3 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 [7].
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 divergece times (branch length) into account. Since the OTL does not include development times of species, this is irrelevant.
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 unimportant to the problem at hand, because in the analysis of the entire Eukaryota tree only generalized models make sense.

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.
Anmerkung: Nochmal darstellen warum und selbst wenn sie fully sufficient wäre, heißt es ja

nicht, dass man sie nehmen muss. Also: motiviere! (Renard)
Felsenstein [10] discusses in his book two algorithms that generalize all previous methods (from Camin and Sokal [6], TODO: Kluge and Farris and Farris [21]): Fitch parsiomony [8] and Sankoff parsimony [9].
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 [22]

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].

2.3.1 Fitch maximum parsimony

Fitch maximum parsimony is an algorithm for rooted, binary trees and describes an ancestral state reconstruction for discrete states [8] 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 [10] 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 Algortihmus 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 informations in the leaf nodes. Each node is depicted as a set of states. There are only two states in this thesis, 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 trough the tree and fills these sets.
In each step, two sets are considered and their intersect formed. There are two cases:

i) The intersection is not empty and corresponds to the new set.

ii) The intersection is empty. $\rightarrow$ Build the union of these sets as new 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 traversion (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:

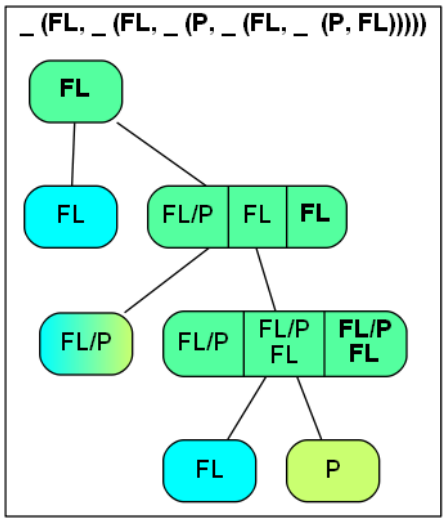


Figure 2.3: Fitch algorithm for binary trees.
The unknown leaf node is discribed with both states. Computed internal nodes (exclusive the root node) consists of three sets, where the last set is the final one (bold).
From the second internal node (seen from the root node) there are several possibilities to create the second and third set.

- 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 2.3 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.

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 2.4 shows this over all unknown node percentage. At 95 % unknown nodes and 95 % of multifurcation of the internal nodes, version 1 was 88.37 %, version 2 was 88.37 %, version 3 was 88.4 %, and version 4 was 88.39 % 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

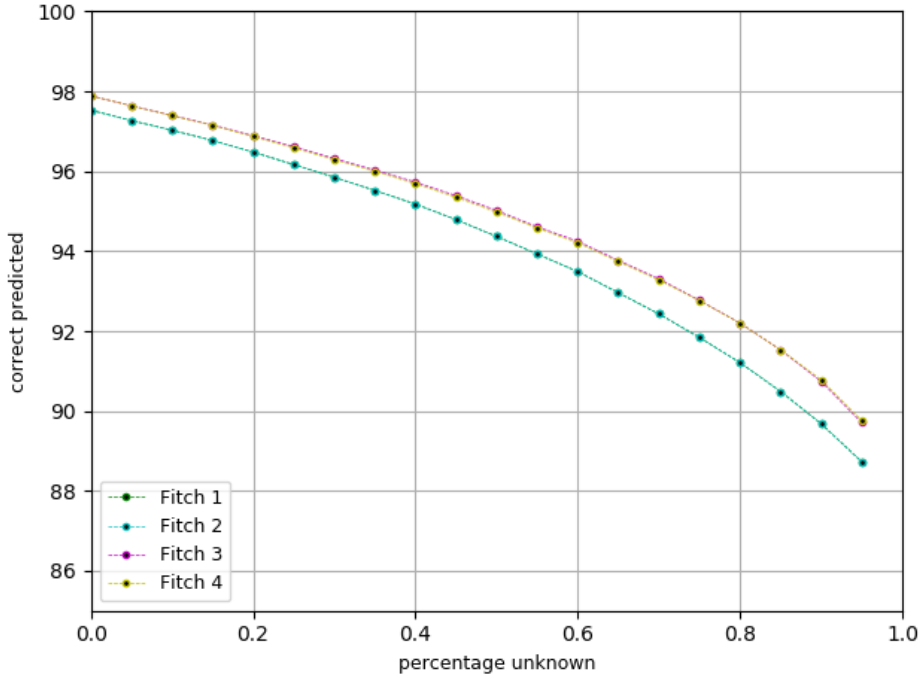


Figure 2.4: Test of Fitch Versions.

2.4 Simulation

The simulation compares these different of ancestral state reconstruction 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].

The course of a simulation is shown in Figure 2.5. The individual steps are explained in the following.

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.

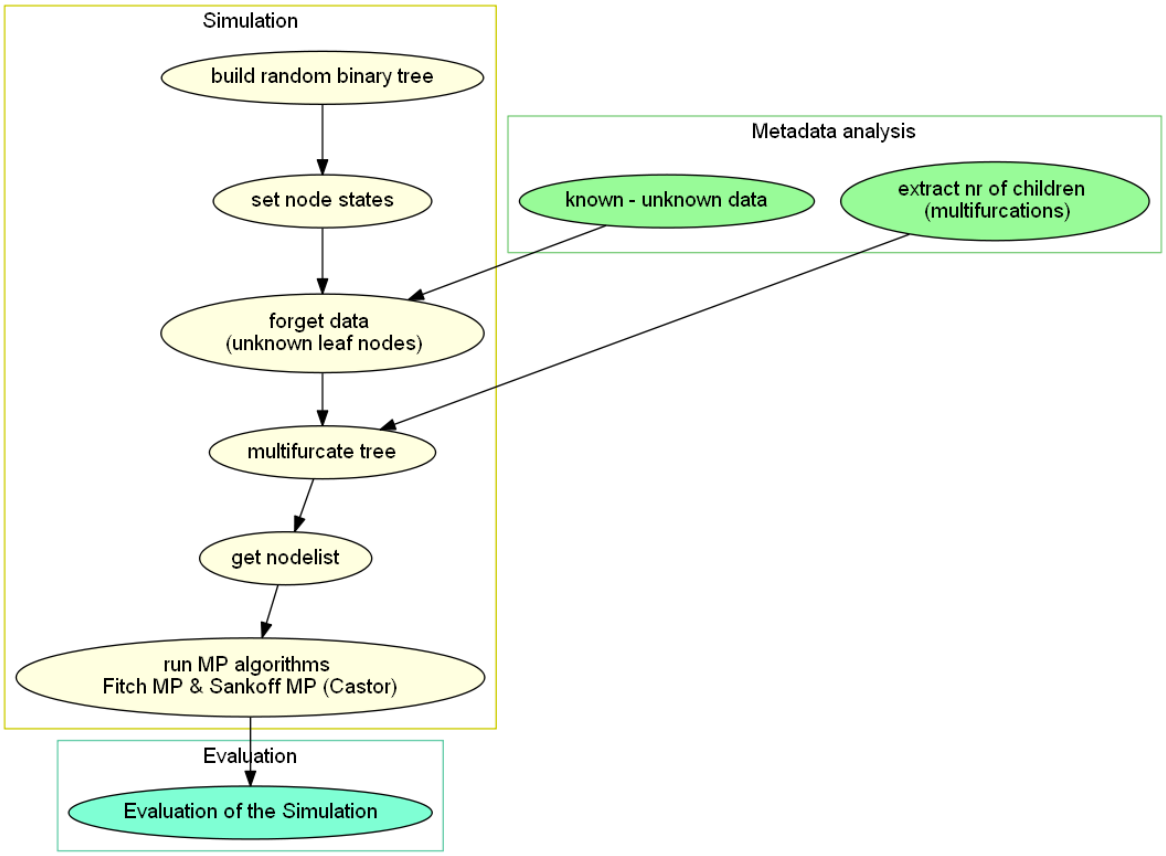


Figure 2.5: A simulation was performed to compare different maximum parsimony algorithms. The course of the simulation with influence of the metadata analysis from the real data can be seen:
(1) A phylogenetic tree is created randomly. (2) Simulate node states for all nodes. (3) 'Forget' internal states and some leaf node states. (4) 'Lose' phylogeny information. (5) Make a nodelist for the algorithm. (6) Run algorithms. (7) Evaluate results.
Points 3 and 4 are influenced by metadata of the real-data analysis.

To get a random binary tree, the Phylo package from biopython is used [24]. They offer a *randomized()* function which returns a BaseTree[3]. The credibility of this choice is examined in subsection 4.2.2 of the discussion.

The next step is to simulate states for all nodes.
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

[3]https://github.com/biopython/biopython/blob/master/Bio/Phylo/BaseTree.py

traverse from the root to the leaf nodes, always pulling out of the current distribution until the randomly drawn number is above the threshold and the new node changes state. Figure 2.6 shows a part of this simulating states.
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 2.6.

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.

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.

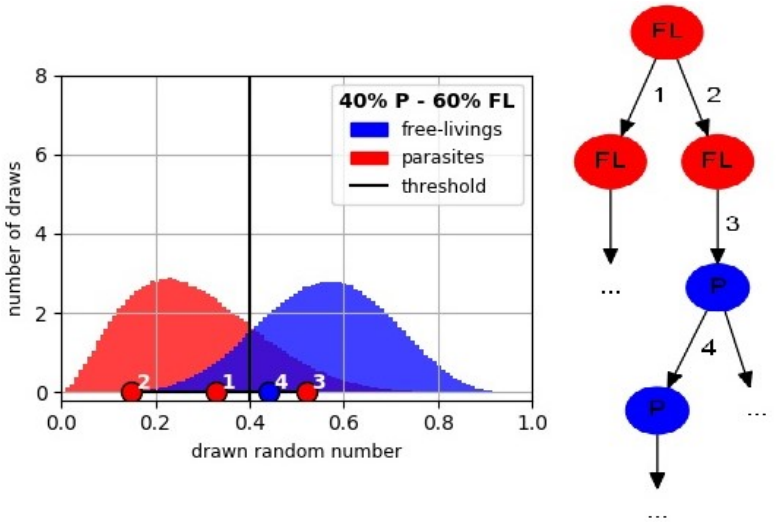


Figure 2.6: Set node states: Distribution of states (left); traversion through the tree (right).
Start with a free-living root node (FL: red).
(1) + (2) Draw random numbers for its children from the free-living distribution (red), the numbers are under the threshold → the nodes are again free-living; go on with the children.
(3) The number drawn is above the threshold. → The node state changes to parasitic (P: blue).
(4) Now draw random numbers from the parasite distribution (red) until one number lies unter the threshold. Then change back.

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

Different percentages of forgetting information are simulated, see figure 3.2 in the results.

The last step is the evaluation of the results. This is done with a simple difference calculation of the node states.
In the nodelist, the originally simulated states and the newly calculated states are stored for each node ($FL = 0$, $P = 1$). The sum of the differences of the node states gives the distance of the prediction to the original tree.

2.5 real data analysis

TODO: Evaluation section in methods?

- Import tree
- Import interactions
- run castor algorithm / and others?
- interprete results (leave-100-out cross validation)

We then evaluated the results in two ways. On the one hand, we have examined the results of some subtrees for their credibility with the help of biological background knowledge. On the other hand, we performed a leave-100-out cross validation.

2.6 Implementation

The complete code is located 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

This work deals with the ancestral state reconstruction of the entire Eukaryota relatives tree.

For this reconstruction, we first analyzed our data. This is the tree of OTL [1] and the data for the leaf nodes (free-living or parasite states) from GloBI [2] (section 3.1).
Next we compared different possible methods. We decided to take a closer look at maximum parsimony algorithms because they are best suited to the problem at hand. The reason for this can be found in section 2.3. We tested these on different simulations of the data and compared their predictive power (section 3.2).
We used the Sankoff algorithm [9] implemented by Louca et al. [3], which performs best, for the actual reconstruction on the real data.

We then evaluated the results in two ways. On the one hand, we have examined the results of some subtrees for their credibility with the help of biological background knowledge (subsection 3.3.1). On the other hand, we performed a leave-100-out cross validation (Subsection 3.3.2) and predicted approximately 98.17 % states correct.

3.1 Metadata analysis - Missing information

As previously presented in the methods (section 2.2.2), we have two types of missing information: unknown states of leaf nodes and multifurcation.

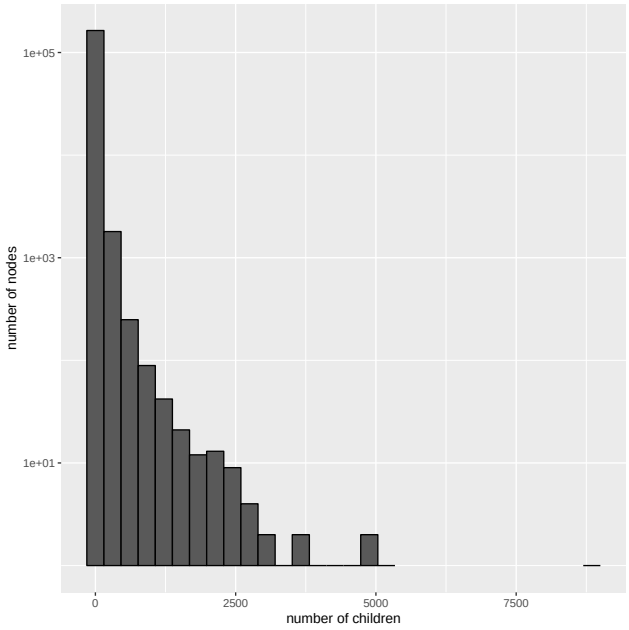
A tree is multifurcated if there are nodes that have more than two children. In the case of a phylogenetic tree, the ridge of multifurcation describes the amount of lack of information

about the topology of the tree. A complete phylogenetic tree would be binary, meaning that the number of leaf nodes equals the number of internal nodes minus one. Since we only work with a artificially constructed tree (synthesis of several phylogenetic trees), this tree is strongly multifurcated: It has 241,974 internal nodes for 2,293,463 leaf nodes $\rightarrow$ 98.16 % missing nodes.
We calculated these percentages of missing nodes and also missing state information for some subtrees and plotted them in table 3.1.

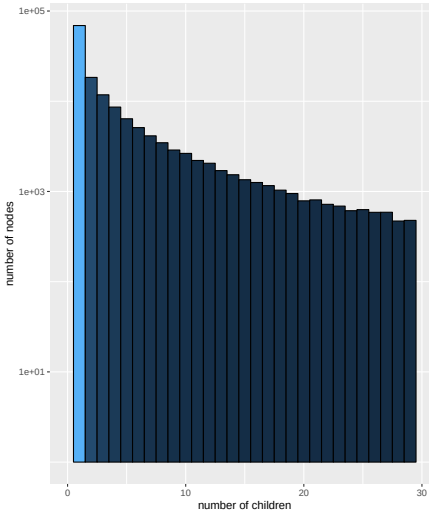
Subtree of	**Unknown States**	**Multifurcation**
Eukaryota	97.34 %	98.16 %
Metazoa	96.44 %	87.93 %
Fungi	98.87 %	96.97 %
Chloroplastida	99.14 %	89.46 %
Apicomplexa	86.26 %	87.16 %
Nematoda	89.01 %	88.59 %
Chordata	88.59 %	66.49 %
Platyhelminthes	68.73 %	80.34 %
Insecta	97.11 %	90.78 %

Table 3.1: Examination of subtrees regarding missing information.
The percentage values show the ratio of missing information of: unknown states (missing state information of leaf nodes) and multifurcation (missing internal nodes).
The subtrees are from different taxa: domain (Eukaryota), kingdom (Metazoa, Fungi, Chloroplastida), phylum (Apicomplexa, Nematoda, Chordata, Platyhelminthes) and class (Insecta).
The two by far smallest values were highlighted in green.

We also have for a first overview we collected for each 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 of every node (a node with two children is binary). That means it discribes the number of nodes which we have lost from the real (binary) phylogenetic tree.
It can be recognized that we are very far from a binary tree.



(a) Histogram with automatic binwidth.



(b) Histogram with $binwidth = 1$. Higher multifurcation than 30 has been cut off.
light blue: binary;
dark blue: multifurcation

Figure 3.1: Histograms about the multifurcation of the internal nodes of the synthesis tree. For each node, the number of children (degree -1) was collected. A node is multifurcated if it has more than two children, so we deducted two from each number of children. We have plotted this in two histograms: number of children -2 on the x-axis with log scale and the number of nodes with this amount on the y-axis.

3.1.1 Data artifacts

At this point we also found out that there are nodes with only one child node (55,700 nodes). The most of these nodes are next to a leaf node, others are deep in the tree (3,956 with height > 2). They are probably a result from the fact that taxonomic information has been incorporated into a phylogeny.
Anmerkung: probably? Da ist eine Spekulation hier? Kann man das nicht herausfinden? Spekulieren solltest du am ehesten erst in der Diskussion. (Thilo)
Some examples of these nodes are:

- 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

3.1.2 Poisson regression of the multifurcation

For a regression analysis of the multifurcation first set up a simple generalized linear model and then output the intercept of it:

```
kingdom.furc.mod <- glm(multifurc~kingdom,
                        data=inner.taxa,
                            family="poisson")
# get the intercept:
summary(kingdom.furc.mod)
```

Since the multifurcation can also be 0, the intercept describes the expected mean value and thus, if it is greater than 0, it says that we have a multifurcation: $2.62 > 0 \Rightarrow$ there is a multifurcation.

We next compared the intercept of the different kingdoms and found that

$$1.22 \text{ (Fungi intercept)} > -0.11 \text{ (Metazoa intercept)} > -0.09 \text{ (Chloroplastida intercept)}$$

that means the strength of the multifurcation is the same/has the same size sorting. TODO: kann das sein?

We investigated several models that could describe multifurcation:
In doing so, we allowed the different influence of the taxa and the heights and depths of a node to be included. From this we got 9 times 4 models of different complexity levels (see first row of table 3.2).
We first compared the models based on their residuals. These give us the error of the model. If the value is small, our data will be well modeled. These values are meaningful only for the same complexity, therefore we then calculated the BIC (Bayesian information criterion) values of the models and used them to evaluate the models.
The deviance table with the residuals can be found in the appendix (table 5.3). Table 3.2. lists the BICs.

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

Table 3.2: BIC (Bayesian information criterion) values of the multifurcation models. These models were created with the R function *glm()* and compared with the *BIC()* function. This results in the listed BIC values.

Within the every complexity class it can be seen that the mean height gives the best additional factor.

Despite higher complexity, the BIC values are getting smaller from model to model, meaning that the finest model available here is also the best one of these. Lower taxa than orders (eg family) were computationally too expensive to calculate.
The model *multifurc* $\sim$ *order* * *mean.height* turns out to be the best of our models, whereby it is possible that eg *multifurc* $\sim$ *family* * *mean.height* is better.

3.1.3 Binomial regression of the unknown state information

Next to the problem of the multifurcation of the tree is the less of interaction data we have for the species. For the ancestral state reconstruction, we need information about the states (free-living or parasite) in the leaf nodes.
The eukaryotic synthesis tree has 293,463 leaf nodes. The GloBI database has 5,346,414 interactions (at 29.01.2018). Out of this data we got 51,337 distinct free-living species and 47 332 distinct parasite species for the whole tree of life. From the Eukaryota we could determine 25,992 and 34,879 species as parasites and free-living. With 2,293,463 leaf nodes we still have about 97.34 % unknown leaf nodes. In the discussion, section 4.1, we will talk about ways to extend this data.

We also compared different models in terms of their BICs (Table: 3.3). The Residuals are not very meaningful here, since all models have different complexities. For the sake of completeness, the associated deviance tables are located with the residuals in the appendix 5.4.

Model / Taxa	Kingdom	Phylum	Class	Order
multifurc $\sim$ taxa	545799	500004	485121	484681
multifurc $\sim$ taxa + depth	544862	493808	481869	478851
multifurc $\sim$ taxa * depth	544179	489845	481494	478188

Table 3.3: BIC (Bayesian information criterion) values unknown state information models. These models were created with the R function *glm()* and compared with the *BIC()* function. This results in the listed BIC values.

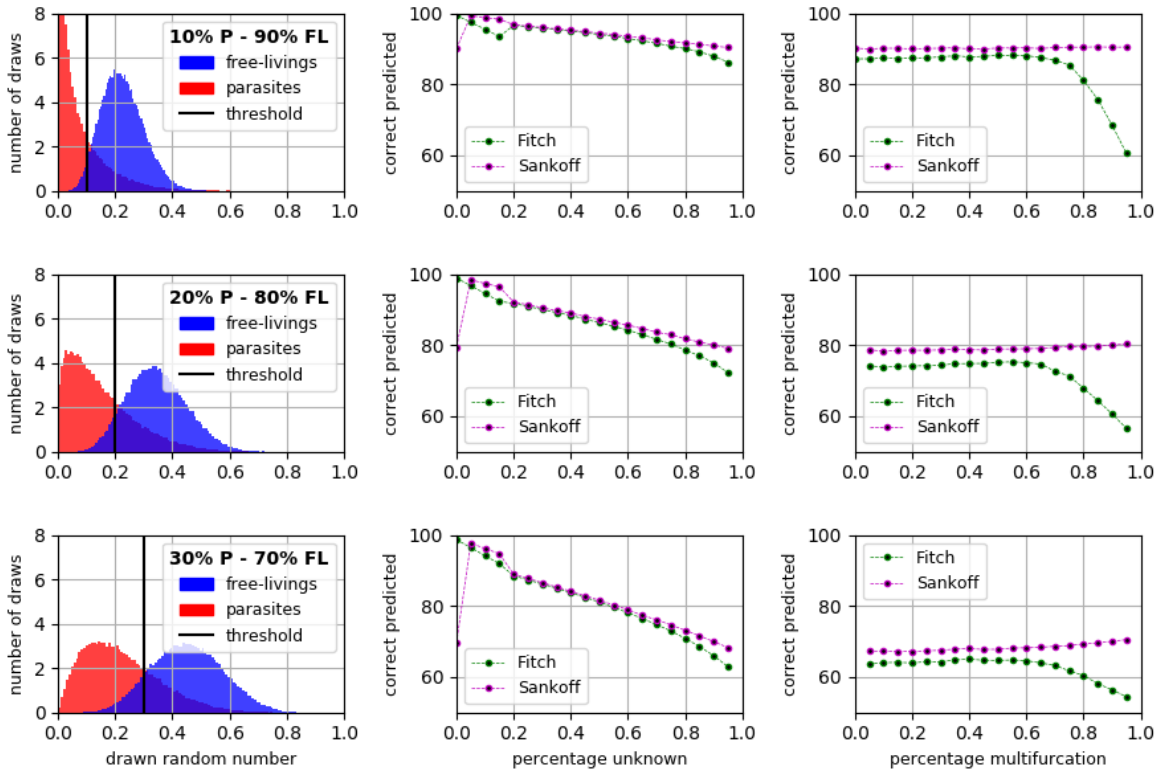
It also follows from this table that the most complex model is the best. In general, the BIC values are smaller than those of the multifurcation models. The modeling here is thus better.

Again, the calculation of finer models (eg order or family) was too expensive.

These missing data modeling results can be used to better simulate the data.

3.2 Results of simulation / Influence of different parameters

As presented, we compare two methods in our simulation to their prediction accuracy: Fitch and Sankoff.
Figure 3.2 shows the results of examining various parameters.



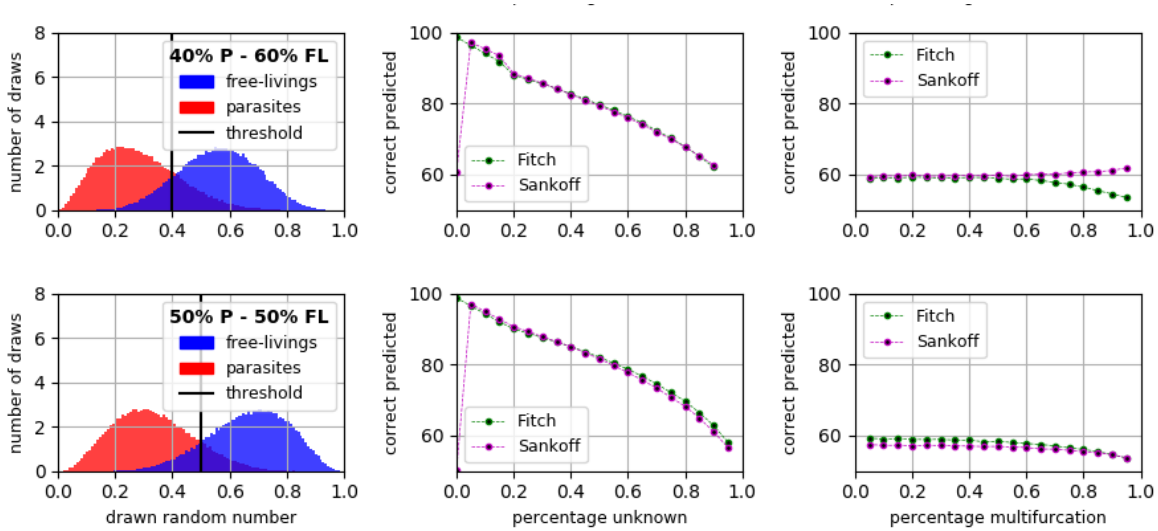


Figure 3.2: Influence of unknown data to prediction.
The first column describes the distributions of free-livings and parasites with a given threshold for the respective simulations to the right.
The middle column investigates the influence of the unknown states, the right the influence of the strength 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.

First, we tested different trasition probability distributions including threshold (first column). It can be observed that the more balanced the percentages of free-livings and parasites, the worse the prediction of the algorithms (second and third column).
It can be seen that both algortithms are always over 50 % prediction 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.
On the other hand, we examined the influence of missing internal nodes (ridge of multifurcation) and missing leaf node information (unknown leaf nodes).
Both factors have a relatively linear influence on the Sankoff method. Fitch, on the other hand, breaks significantly in his prediction from about 70% unknown leaf nodes or mutifurcation.

Since we have a lot of missing data in most subtrees (both internal nodes and state informations), the Sankoff gives a better prediction and was thus used for real data analysis.

3.3 Results of the real data analysis created with Sankoff

This section is about evaluating the prediction of real data using the Sankoff method.
It is divided into two subsections. 3.3.1: The analysis of some subtrees using biological background knowledge. 3.3.2: Presentation of the results of the leave-100-out cross validation.

3.3.1 Biological view

To analyze the results, we have selected some phyla (subtrees) to evaluate our results selectively from the biological point of view: Chordata, Nematoda, Platyhelminthes and Apicomplexa.
In Table 3.4 we compare the given states with the predicted ones.
Several factors such as the distribution of existing input data via parasitics and free-livings, faulty input data from GloBI and reinforcement of errors by multifurcation play an important role, which crystallize out in these examples.
An important factor here is that the credibility of the results. The accuracy of the input data stands and falls with the presence and the correctness of the data of GloBI. 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. This is 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 3.4: Some selected phyla (subtrees), only leaf nodes, were examined for the evolution of the amount and percentage of given data to predicted data.

In contrast to the other phyla examined, the phylogeny in the Chordata is more pronounced (less multifurcation) (see Table 3.1). This results in less variance of errors. What is reflected in the results from the table 3.4. 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).

We mapped the few parasitic species in a rough/simple 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 ate* known as broodparasite and some others.
An example of the amplification of mistakes here are the carp. There is a paper from which GloBI concludes: Grass carp (*Ctenopharyngodon idella)* has Pathogen common carp (*Cyprinus carpio*)[1]. Since there is hardly any information about free-living, it follows that all siblings are also predicted to be parasitic.

The Apicomplexa are a parasitic phylum. We found only one input organism: Stemonitis fusca as a free-living specie. In GloBI it is listed as being parasitized by Nectria candicans

[1] https://www.globalbioticinteractions.org/?interactionType=hasParasite&targetTaxon=Cyprinus%20carpio

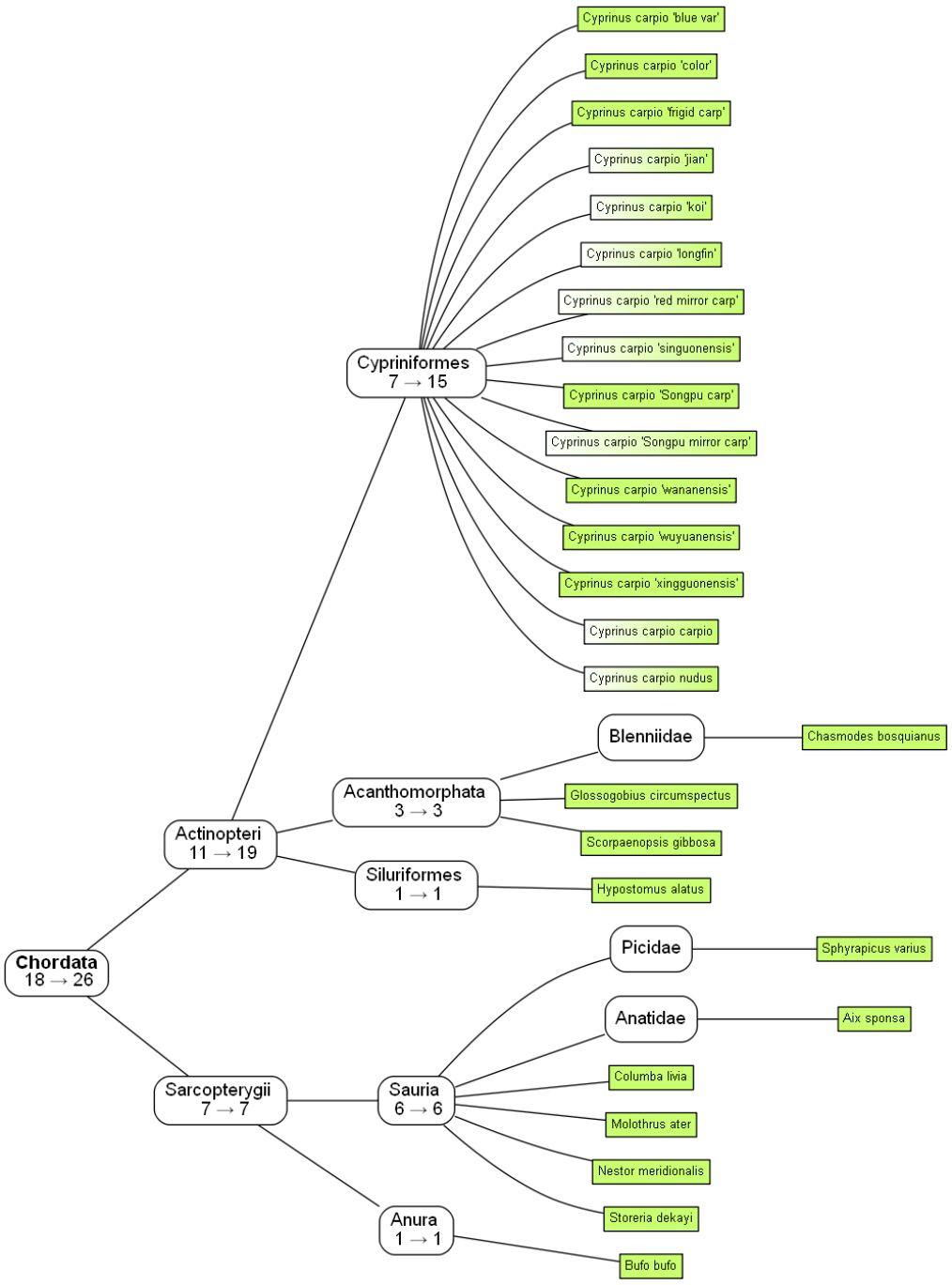


Figure 3.3: Parasites of Phylum: Chordata.
All parasite data of the chordata were mapped into a rough taxonomy (phylum, class, order, family) in order to understand their distribution and affiliation.
The internal nodes are the wanted taxa from OTL (with the addition of # input parasites to → # predicted parasites).
The leaf nodes are the input parasites (green) and the predicted parasites (white → green).

and Nectriopsis sporangiicola[2]. The algorithm has not predicted new free-livings.

Most species of Platyhelminthes (flatworms) are parasites, although there are also free-living, predatory feeding species. These are summarized in the Turbellaria, while the parasites are divided into three other classes [26]. This also corresponds to our observations. There is one class (Rhabditophora) that contains all but one single exception of free-living species of this phylum, which includes the Turbellaria.
It should be noted, however, that this classification is outdated, as it has been proven that the Turbellaria are not monophyletic. But we will not go into that here.
For the Platyhelminthes we had more state information for the leaf nodes compared to the other considered subtrees, see Table 3.1. We start with 0.1 % free-livings and predicted 0.77 % as free-living species.

With the Nematoda it looks more complicated. Large parts of Nematoda are free-living, but we found only 5.32 % of them. Blaxter et al. estimates the order of 25,000 parasites in the Nematoda [27] and speaks of at least seven independently arosed parasitism [28]. In a recent article Blaxter identifies 18 origins [27] in Nematoda.
The problem at this point, however, is: Hallan speaks of the fact that only 23,000 species were described by the Nematoda but is it assumed that one million or more species are in this phylum.[3]. TODO: Link ist nicht erreichbar! The parasites have been much more studied and thus we start with only 0.63 % (table 3.4) free-living species. Against such a shifted data situation, the algorithm is almost powerless to make correct predictions. And yet the percentage has increased to 5.34 %.

TODO: Im Folgenden folgen 3 weitere ähnliche tabellen. Einmal eine ähnliche Tabelle inklusive interner Knoten 3.5, zweimal die Übersicht über die Kingdoms Blattknoten 3.6 bzw interne Knoten 3.7. Welche nehmen wir? Rest Appendix oder ganz raus?

[2] https://www.globalbioticinteractions.org/?interactionType=parasiteOf&targetTaxon=Stemonitis%20fusca
[3] J. Hallan, unpublished; https://insects.tamu.edu/research/collection/hallan/

Phylum	# nodes	original states FL	original states P	final states 0 (FL)	final states 0.4	final states 0.5	final states 0.67	final states 0.75	final states 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 3.5: Some selected phyla (subtrees), including internal nodes, were examined for the evolution of the amount and percentage of given data to predicted data.

Kingdom	# nodes	original states FL	original states P	final states 0 (FL)	final states 0.25	final states 0.33	final states 0.4	final states 0.5	final states 0.67	final states 0.75	final states 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 3.6: All kingdoms (subtrees), only leaf nodes, were examined for the evolution of the amount and percentage of given data to predicted data.

Kingdom	# nodes	original states FL	original states P	final states 0 (FL)	final states 0.25	final states 0.33	final states 0.4	final states 0.5	final states 0.67	final states 0.75	final states 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 3.7: All kingdoms (subtrees), inkluding internal nodes, were examined for the evolution of the amount and percentage of given data to predicted data.

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. The question is, how to count

these state changes. We decided to round the values.

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

Table 3.8: Origins and losses

In Table 3.8 we can see, that we found some more origins than Weinstein and on top of that some losses.

When more closely regarding 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 also look credible. 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 [28]. 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.

3.3.2 Leave-100-out cross validation

For a further validation of our results, we carried out a leave-100-out cross validation.
In order to achieve about 15% of the 60,871 input node states with a validation, we randomly left out 100 states 100 times. To make smaller amounts up to leave-one-out had too much computational effort. TODO: warum 15%?

Of these 10,000 nodes, 9,238 were unique. From the unique nodes, we predicted approximately 98.17 % correct and thus 1.82 % wrong, with duplicate draws always having the same prediction (data not shown).

We have again dealt with the question of how this data is best modeled. The influence of the taxa (kingdom, phylum, class) and the depth of the leaf nodes was re-modeled and the BICs compared (table 3.9). Lower taxa than classes (eg order) were computationally too expensive to calculate.
The Residuals are included in the appendix (table 5.5) for the sake of completeness; here too, due to the different complexity, they do not have much meaning for the comparison of the models.

TODO: table for these numbers?

Model / Taxa	Kingdom	Phylum	Class
correct predicted $\sim$ taxa	117936	112242	111733
correct predicted $\sim$ taxa + depth	117776	111304	111273
correct predicted $\sim$ taxa * depth	117709	111262	113135

Table 3.9: BIC (Bayesian information criterion) values of cross validation prediction models.
These models were created with the R function *glm()* and compared with the *BIC()* function. This results in the listed BIC values.

The BIC values this time did not prove that the finest model is the best. Of our calculated models, *correct predicted* $\sim$ *phylum * depth* has the smallest value.

We examined the influence of the omitted data on the prediction. On average, about twice as many leaf nodes are predicted differently. The variance is very high. This means that we have a high degree of dispersion and thus a stochastic situation exists. It also describes the width of the present probability function. The standard deviation is about five times as high as the number of omitted nodes, so the variability is quintupled. Table 3.10 shows these results.

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

Table 3.10: Statistics about the leave-100-out cross validation
The distance between original and new states was calculated using the Euclidean metric. This was summed over all states / all leaf node states and all internal node states.
The lower half of the table describes the distribution of the 'lost' states between parasites (P) and free-livings (FL).

3.4 Effects of Taxa in the different models

The comparisons of the effects of the taxa can be found in Table 3.11 and showed ...

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 3.11: Effects of Taxa in models for unknown data

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

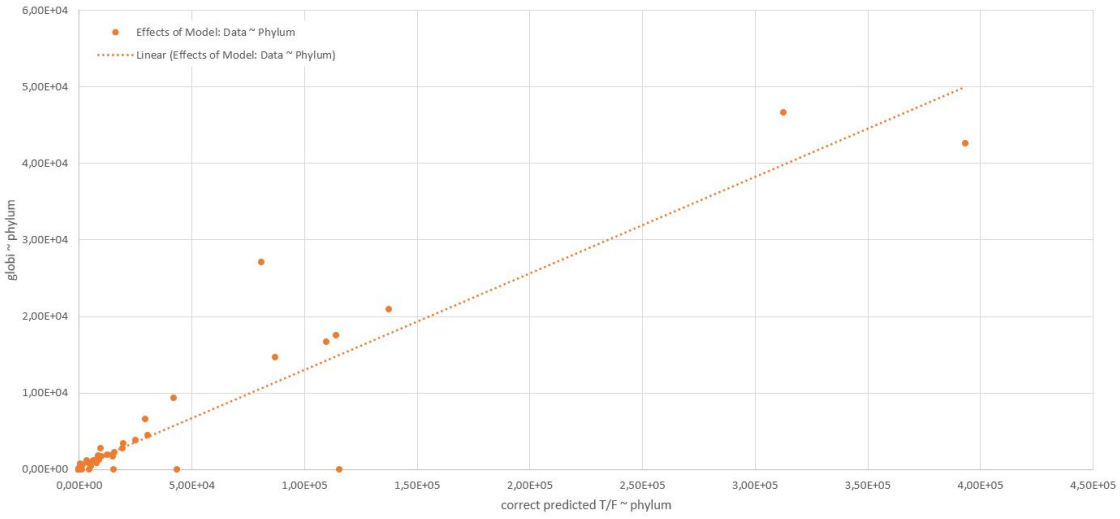


Figure 3.4: Effects of Model: Data ~ Phylum

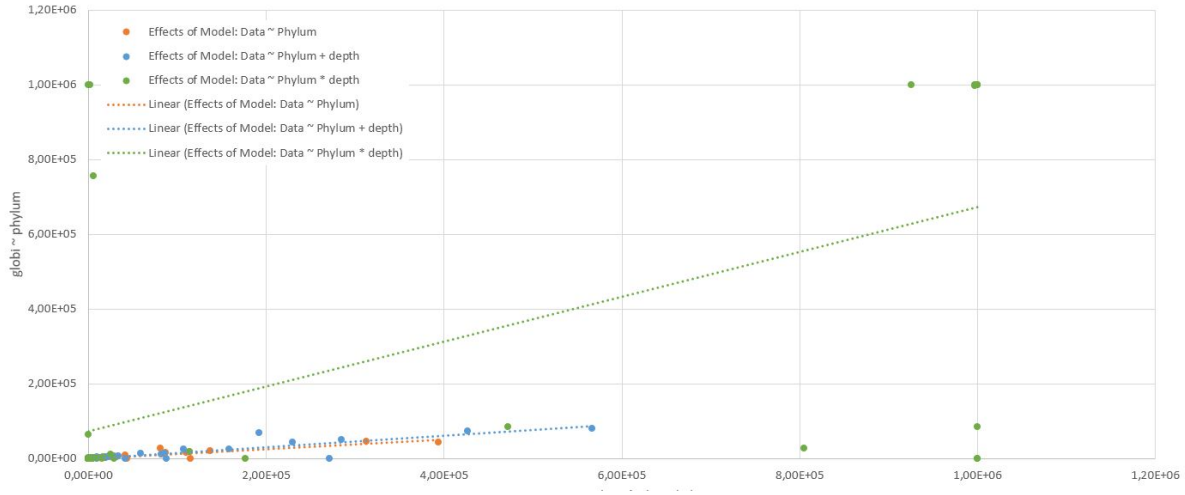


Figure 3.5: Effects of 3 Models: Data $\sim$ Phylum ($+$/* depth)

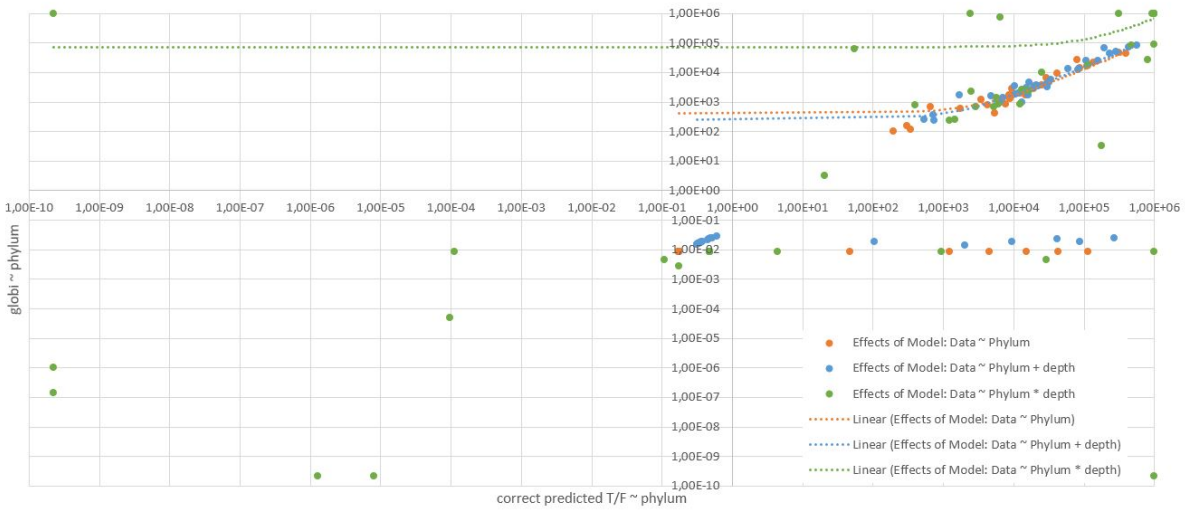


Figure 3.6: Effects of 3 Models: Data $\sim$ Phylum ($+$/* depth)
both axis in log scale

4 Discussion

This chapter deals with the question of how trustworthy our result of the ancestral state reconstruction of the Eukaryota tree is. And further on how well our simulation can simulate the real problem and thus make statements about the predictive power of the applied Sankoff algorithm.

We have pursued this question in various investigations and yet, of course, further possibilities for improvement remain. Despite these possibilities for improvement, our reconstruction gives a first good assessment of the whole tree.

TODO: Why is the study important?
TODO: How does this study relate to previous studies?

We evaluated our ancestral state reconstruction and the relational state prediction using three methods: a simulation, a selective analysis from a biological perspective and a leave-100-out cross validation.

Each theme has its own limitations.
There are various expansion options and points of criticism for the simulation, which in particular revolve around the question of how well we can simulate our real data. We dedicate this question to the subsection 4.2.
Above all, the 'biological view' gives a first impression of the credibility of the results. But he can not make any statistical statements. Here the interested researcher may have to analyze the subtrees of relevant to him for himself.
Of course, cross-validation can be expanded. With more computing power, a leave-1-out cross validation could also be performed. However, our result on this point yields a strong result with 98.17 % correct prediction that this does not seem necessary.
One last limitation also exists in the data situation itself. We have pointed out in various sections that we have a strong multifurcation and only a few data about leaf node states. In

figure 3.2 you can see how the predictive power can improve if we have more data available. We discuss this point in section 4.1.

4.1 Data situation

The used eukaryotic synthesis tree from OTL [1] has 241,974 internal nodes and 293,463 leaf nodes. In addition, we could specify 25,992 parasitic and 34,879 free-living species from GloBI [2].
This gives us a high number of missing internal nodes (high multifurcation) and a low number of node states of the leaf nodes. In table 3.1 we have written down this percentage including some subtrees.

The investigation of the taxonomy revealed that the OTL 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 [29].
In section 43 of the appendix is a list of all phyla.

Also, the global database is not fallible. We found out (section x) that there is some misinformation.
We were able to report some of these directly to Globi.

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. At this point there is thus the possibility to use more of the existing data, if one performs a matching with the other Ids.

4.2 Simulation

The aim of the simulation was to test the influence of various unknown or unsafe parameters in order to test the credibility of the prediction.
Some of these influences could not be tested: Firstly, the distribution of parasites and free-livings in the tree produces various parameters that we could not test all (see subsection 4.2.1) and, secondly, we discuss the simulation of a phylogenetic tree on which the evaluation based on certain parts (see subsection 4.2.2).

4.2.1 Transition probabilities

We assume that different parasite types have different transition probabilities. Establishing a common distribution across the eukaryotic tree is therefore difficult.
Based on the estimates of Windsor [20], we have assumed a distribution of 40% parasites to 60% free livings in this work. As a result of the diversity of parasites and the lack of generalizations, we have generally stated that $\mathcal{P}(FL \to P) = \mathcal{P}(P \to FL)$. But it is also reasonable to assume that in general $\mathcal{P}(FL \to P) > \mathcal{P}(P \to FL)$, because a reverse mutation is usually less likely. However, one would have to determine how much this difference is and thus discuss another parameter.

In the simulation, we tested different beta distributions with different thresholds and found that this combination has a considerable influence on the predictive power of the algorithms. At this point it would be possible to test other distributions with equal threshold values or different threshold values with equal distributions.
However, if you choose a strong free-living distribution and balance it with the threshold to achieve a 40:60 parasite free-living distribution, for example, this could be very computationally expensive.

With these further simulations you could find out if the issue of distributions plays a big role. Conversely, one could estimate possible distributions based on the data location in the tree. However, this is likely to be very difficult given the poor data. For this reason, we have

decided in this work to accept very general values and not to speculate much.

4.2.2 Simulation of a phylogeny

In our simulation, we start with the simulation of a phylogeny and then depict our data situation.
For this first step we use the *randomized()* function of the phylo package from biopython [24]. The question is how close is such a randomized tree to a phylogeny.

The problems that can arise are that different species develop at different rates. This means that there could be sections in the tree with many branches (for example, subtrees of unicellular species), ie with high depths and opposite sections (of very complex species, for example). It can thus be assumed that the tree is not balanced.
For more precise statements, one would have to take deeper undercurrents, which was not possible in the context of this work.

4.3 Conclusion

From the simulation, we can conclude that we predict correctly about 60% of the nodes in the present data situation (see section 3.2). The leave-100-out cross validation even showed that we predict the omitted nodes to be 98.17 % percent correct (see section 3.3.2).
This allows for the assumption that the data is grouped and not uniformly distributed and thus smaller subtrees are present in which data are to be found in the simulation with smaller multifurcation and smaller value for unknown nodes. This is also confirmed by our biological analysis of the Chordata subtree (see subsection 3.3.1).
However, this means that the ancestral states' data in the direction of root node are probably particularly unbelievable. Which makes the localization of Origins direction root node difficult. The question remains, how much this affects the estimation of the number of origins. Our comparison with the paper from Weinstein and Kuris [4] (see subsection 3.3.1), however, leaves us with optimism.

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] Goberna, M.; Verdú, M. Predicting microbial traits with phylogenies. *The Isme Journal* **2015**, *10*, 959 EP –, Original Article.

[6] Camin, J. H.; Sokal, R. R. A Method for Deducing Branching Sequences in Phylogeny. *Evolution* **1965**, *19*, 311–326.

[7] Royer-Carenzi, M.; Pontarotti, P.; Didier, G. Choosing the best ancestral character state reconstruction method. *Mathematical Biosciences* **2013**, *242*, 95 – 109.

[8] Fitch, W. M. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* **1971**, *20*, 406–416.

[9] Sankoff, D. Minimal Mutation Trees of Sequences. **1975**, *28*.

[10] Felsenstein, J. *Inferring Phylogenies*; Sinauer, 2003.

[11] Xiang, Z.; Mungall, C.; Ruttenberg, A.; He, Y. Ontobee: A linked data server and browser for ontology terms. *Neoplasia* **2011**, *833*, 279–281.

[12] ITIS, Integrated Taxonomic Information System. https://www.itis.gov.

[13] 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/.

[14] Horton, T. et al. World Register of Marine Species (WoRMS). http://www.marinespecies.org, 2018; http://www.marinespecies.org, Accessed: 2018-02-27.

[15] GBIF, Global Biodiversity Information Facility. https://www.GBIF.org.

[16] Diego Vázquez, R. N., Jeremy Goldberg Interaction Web Database (IWDB). 2003; https://www.nceas.ucsb.edu/interactionweb/.

[17] Webs on the Web (WOW): 3D visualization of ecological networks on the WWW for collaborative research and education. 2004; pp 5295 – 5295 – 9.

[18] Myers, P.; Espinosa, R.; Parr, C. S.; Jones, T.; Hammond, G. S.; Dewey, T. A. The Animal Diversity Web (online). 2018; https://animaldiversity.org.

[19] Cohen, J. E. c. Ecologists' Co-Operative Web Bank. Version 1.1. Machine-readable database of food webs. *New York: The Rockefeller University* **2010**,

[20] Windsor, D. A. Controversies in parasitology, Most of the species on Earth are parasites. *International Journal for Parasitology* **1998**, *28*, 1939–1941.

[21] Farris, J. S. Methods for Computing Wagner Trees. *Systematic Biology* **1970**, *19*, 83–92.

[22] Swofford, D. L.; Maddison, W. P. Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences* **1987**, *87*, 199 – 229.

[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] 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.

[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] Ax, P. Verwandtschaftsbeziehungen und Phylogenie der Turbellarien. Ergebnisse der Biologie. Berlin, Heidelberg, 1961; pp 1–68.

[27] BLAXTER, M.; KOUTSOVOULOS, G. The evolution of parasitism in Nematoda. *Parasitology* **2015**, *142*, S26–S39.

[28] 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.

[29] Cavalier-Smith, T. Eukaryote kingdoms: Seven or nine? *Biosystems* **1981**, *14*, 461 – 481.

5 Appendices

5.1 Methods overview

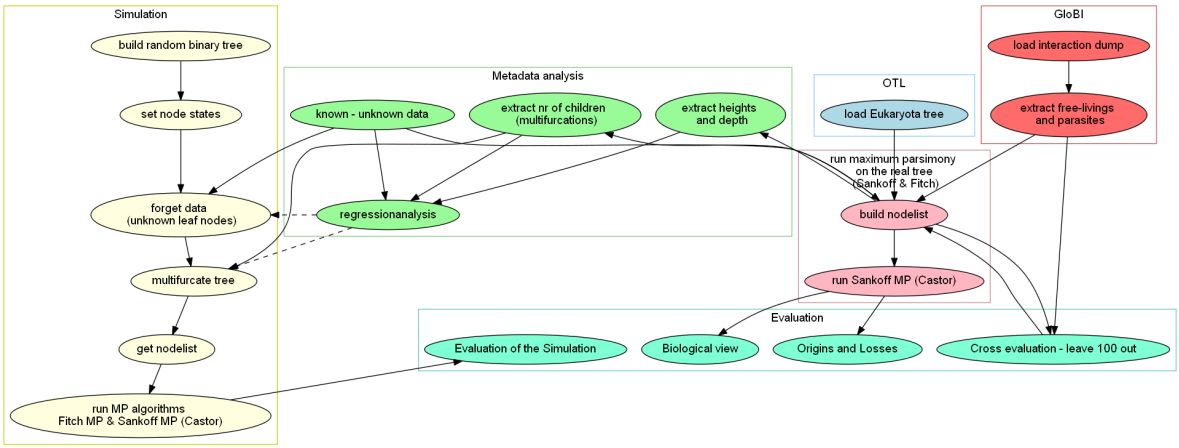


Figure 5.1: Big overview of the whole Workflow

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, Cycliophora, 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

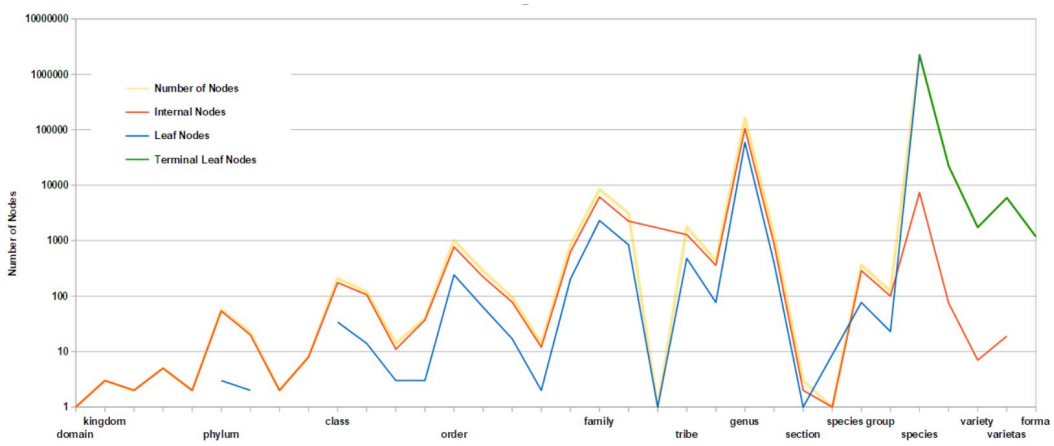


Figure 5.2: Distribution of Nodes in Rank-Cathegories

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).

5.3 Missing internal node modelling - Resiudal table

5.4 Missing leaf state modelling - Resiudal table

5.5 Cross validation prediction - Resiudal table

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: ...

Model / Taxa	Kingdom	Phylum	Class	Order
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 5.3: Residuals of multifurcation models
These models were created with the R function *glm()* and compared with the *anova()* function. This results in the listed residuals.

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

Table 5.4: Residuals of unknown information models
These models were created with the R function *glm()* and compared with the *anova()* function. This results in the listed residuals.

Model / Taxa	Kingdom	Phylum	Class
correct predicted ~ taxa	117877	111466	108878
correct predicted ~ taxa + depth	117703	110513	108403
correct predicted ~ taxa * depth	117592	109827	107994

Table 5.5: Residuals of cross validation prediction
These models were created with the R function *glm()* and compared with the *anova()* function. This results in the listed residuals.