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

Supervisors:
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 estimates, above 40% of all Eukaryota are parasites. Nevertheless, it is computationally difficult to obtain information whether a particular taxon is a parasite 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 25,962 and 34,860 species as parasites and free-living, respectively and predict states for over $\sim$2.3 million (97.34%) leaf nodes without state information.

We estimate the accuracy of our maximum parsimony based predictions using cross-validation and simulation at 60-80% overall, while strongly varying between clades. The cross-validation results in an accuracy of 98.17% which is explained by the fact that the data is not uniformly distributed. 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 Eukaryota. 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 Eukaryota e.g. in the microbiome.

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] to predict parasitism in Eukaryota.

For about 50 years, people have been working on ancestral state reconstruction, the inference of evolution that leads to the given data. One of the first papers was written by Camin and Sokal, who were working on algorithms for discrete-state data in 1965 [2]. 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 non-binary tree with two discrete states.

Royer-Carenzi et al. distinguish 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 its ancestor. The other class they present describes the modeling of character evolution as a stochastic process and uses the likelihoods to compute the possible ancestral character states. This is generally done with a continuous time Markov model [3].

One of the major disadvantages of parsimony methods is that, in contrast to likelihood approaches, they can not take divergence times (branch length) into account. Since the OTL does not include development times of species, this can not be considered here. 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 (Dollo parsimony). This problem is unimportant to the present task because in the analysis of the entire Eukaryota tree only generalized models are meaningful. The comparison of methods gives us the maximum parsimony method as the simplest and fastest method that meets all our requirements.

Felsenstein [4] discusses in his book two parsimony algorithms that generalize previous methods (from Camin and Sokal [2], Farris [5] and others): Fitch parsimony [6] and Sankoff parsimony [7]. Therefore, these are the methods used in this work. For Fitch, the algorithm has been extended from binary to non-binary trees. For the Sankoff algorithm, Louca and Doebeli have presented an implementation for non-binary trees published in an R package named *castor* [8].

For an ancestral reconstruction on the phylogeny of Eukaryota, we use Open Tree of Life (OTL) [1]. Phylogeny describes the evolution of species, while the taxonomy is a classification according to certain criteria in so-called taxa. OTL is a comprehensive, dynamic and digitally available tree of life constructed from published phylogenetic trees along with a backbone of taxonomic data. Besides this, Hinchliff et al. also offer an OpenTreeOfLife-Taxonomy (OTT) with the help of which we identify the individual nodes. It follows that the biggest 'phylogenetic tree' is this synthesis of phylogenetic trees filled with a taxonomic trees given by OTL.

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 [9]. However, this differs from Eukaryota in the sense that complex traits such as parasitism depend on more than one gene.

The present tree structure of OTL is not binary but multifurcated, meaning that each node has multiple ($n > 2$) children or in other words its degree, number of adjacent nodes, is greater than 3 [4]. Parsimonious in phylogeny refers to favoring the tree that needs the least evolutionary change to explain the observed data. Maximum parsimony methods have been developed for phylogenies, which are usually depicted as binary trees. Therefore, the selected parsimony methods are not directly applicable, as they were specifically applied to much smaller subtrees, where all splits are known. We will extend and test the existing maximum parsimony algorithms of Fitch [6] and Sankoff [7] for this task and estimate their predictive power.

Note, the original Fitch algorithm has the sole purpose of minimizing the number of transitions and not the reconstruction of the ancestral nodes. Felsenstein [4] describes a simple extension for the reconstruction. In this work, the algorithm extended to reconstruction is adapted to multifurcated trees, based on the critical reevaluation of this extension by Cunningham et al. [10].

To accomplish this task, information about the states of the current species is needed in addition to the phylogenetic tree. Most of the largest interaction databases (e. g. IWDB (Interaction Web Database) [11], Webs on the Web [12], Animal Diversity Web [13] and ecoweb [14]) are offline or outdated. We use the interaction database Global Biotic Interactions (GloBI) [15] because it is including most of the known databases and is still growing actively [15]. The data in GloBI is stored as interactions e.g. species A parasitizes species B. We conclude that species A is parasitic and species B free-living. From this data, we could specify $\sim$2.3 million leaf nodes 34,860 as free-living and 25,962 as parasitic.

There are many different ways to define parasitism. Since we use GloBI to classify species, we use their definition of parasitism. In GloBi, Ontobee definitions are used [16]. 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 pathogens.

The objectives of this work are the following points: (1) Find a suitable ancestral state reconstruction method. (2) Accomplish reconstruction on the Eukaryota synthesis tree of OTL. The goal of point 1 is to evaluate the possible methods based on a simulation of our data situation. The Sankoff algorithm implemented by Louca et al. is the best in our comparisons. Therefore, point 2 consists of reconstructing the ancestral states, predicting the unknown leaf states with the aid of this algorithm and 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 Aims

The aim of this thesis is the application of maximum parsimony algorithms to non-binary trees and very large data sets to predict unknown leaf node states.
We discuss different ancestral state reconstruction methods and test the most appropriate ones on our data. Hereby, we want to find out how far it is possible to use the Open Tree of Life (OTL) to expand available information via phylogenetic trait prediction. As a basis for known states, we use the Global Biotic Interactions (GloBI) database, which categorises 34,860 and 25,962 species as free-living and parasites. This data is used to perform an ancestral state reconstruction and predict states for over $\sim$2.3 million (97.34%) leaf nodes without state information.

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 is an influencing value. Therefore, two major types are distinguished:

i) Biological parameters (a result of the evolutionary process):
   - State distributions and transition probabilities

ii) Distribution of missing information:
   - Lack of information on topology ($\rightarrow$ multifurcations)
   - Lack of information of states of some leaf nodes

3 Methods

In this thesis, a maximum parsimony algorithm is applied to the Eukaryota tree to obtain an ancestral state reconstruction of free-living versus parasite states.
This chapter is divided into the following sections: the description of the data sources (section 3.1) used and the analysis of these data as a preparation for the simulation (section 3.2). A description of the analyzed methods for the ancestral state reconstruction (section 3.3) and then an explanation of how they fit the problem at hand (section 3.4). And at the end, the real data analysis with the Sankoff method (section 3.5).
Figure 3.1 briefly outlines these relationships. A more detailed view of the workflow can be found in the appendix 7.1.

3.1 Data sources

Two types of data are needed for an ancestral state reconstruction: a tree and information about the states.

For the subsequent analyzes on the Eukaryote tree the database of Open Tree of Life (OTL) is used (downloaded on 16.02.2018) [1]. This database gives a synthesis of phylogenetic trees (currently 819 trees) and a taxonomic tree[1]. OTL also includes the large phylogenetic database TreeBASE [1].

Furthermore the Open Tree Taxonomy (OTT) from OTL is used because it includes most of the known taxonomies and is synthesised by preferring taxonomies that match with available phylogenetic data. The team from OTL prefer a maximum number of species [1], this results in the synthesis between taxonomy and phylogeny.

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

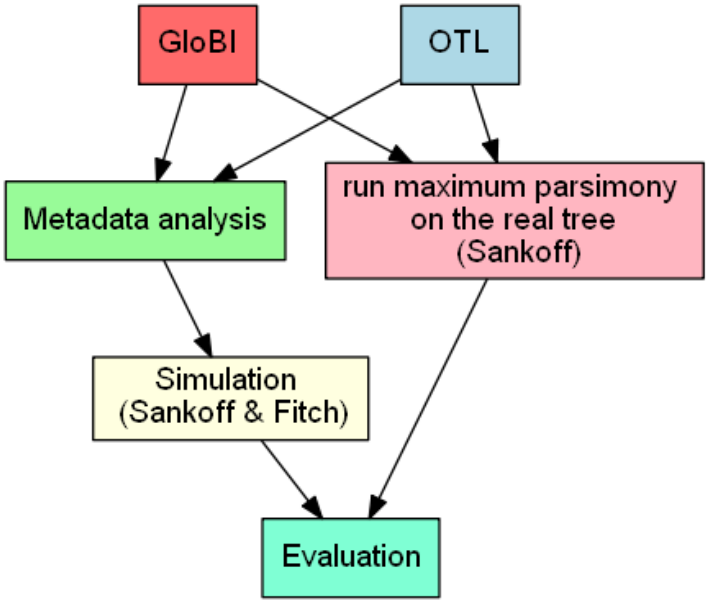


Figure 3.1: The Workflow of the resulting procedure with the following steps:
(1) Retrieve phylogenentic tree data as input for the tree (OTL) and the state information (GloBI). (2) Get metadata of these for a realistic simulation of the maximum parsimony algorithms (Fitch & Sankoff). (3) Build and run 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 resulting algorithm on the original data. (7) Evaluate and interpret the results.

For the state information the Global Biotic Interactions database (GloBI) is taken [15] (downloaded on 29.01.2018).
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 indicating whether the species source or target has become a parasite or a free-living species from the biological perspective. They 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

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

- 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. In the case that a parasite conquers (parasitizes) another parasite yields conflicting states for the second species. This is solved by preferring parasitic.
For each species known identifiers are stored in GloBI. This includes OTT (the taxonomy of OTL). All species that have stored an OTT identifier and have a matching interaction are divided into two lists: parasites and free-livings.

3.2 Metadata analysis

Based on the input 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 are 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 all these calculations, the following R functions are used: *glm()*, *anova()* and *BIC()*.

For each node, depth, min, max and mean height are noted. The node depth is calculated as the distance (number of edges) to the root node. The node height is calculated in three different ways: min, max and mean height refers to the (mininmal, maximal, mean) distance to a leaf node.
The influence in the modeling of these parameters is tested, additively as well as multiplicatively to the models.

3.3 Ancestral state reconstruction methods

The methods used to reconstruct the ancestral states are Fitch parsimony [6] and Sankoff parsimony [7], which are explained in the following subsections.

3.3.1 Fitch maximum parsimony

Based on the work of Cunningham et al. [10] the Fitch method [6] is implemented and extended by us for the multifurcated tree case. 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 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 (e.g. {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 is 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: each internal node is formed from its child nodes, where the only information lies at the beginning.
Second traverse from the root node to the leafs: each internal node is formed from its parent 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 is designed to minimize transitions without predicting actual states of internal nodes, so it is 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 intersection 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 are all tested and compared in the simulation. Some of these options are already available in the binary case:

- The parent 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?
- If there are several siblings, is the cut or union made first of these, or is everything made directly including the parent node.

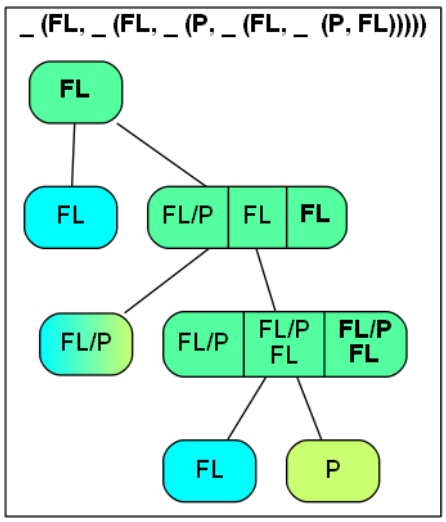


Figure 3.2: Fitch algorithm for binary trees.
The unknown leaf node is described with both states. Computed internal nodes (exclusive the root node) consists of three sets, where the last set is the final one (bold).
From the second internal node (seen from the root node) there are several possibilities to create the second and third set.

The first point already has an effect on the binary case. Figure 3.2 shows both possibilities of the three sets.
Cunningham uses only the first state set of the parent node [10].
From these two points four different versions of Fitch are formed:

i) Fitch 1: First state set of parent node; intersection/union of siblings first.

ii) Fitch 2: First state set of parent node; intersection/union of siblings together with parent node.

iii) Fitch 3: Both state sets of parent node; intersection/union of siblings first.

iv) Fitch 4: Both state sets of parent node; intersection/union of siblings together with parent node sets.

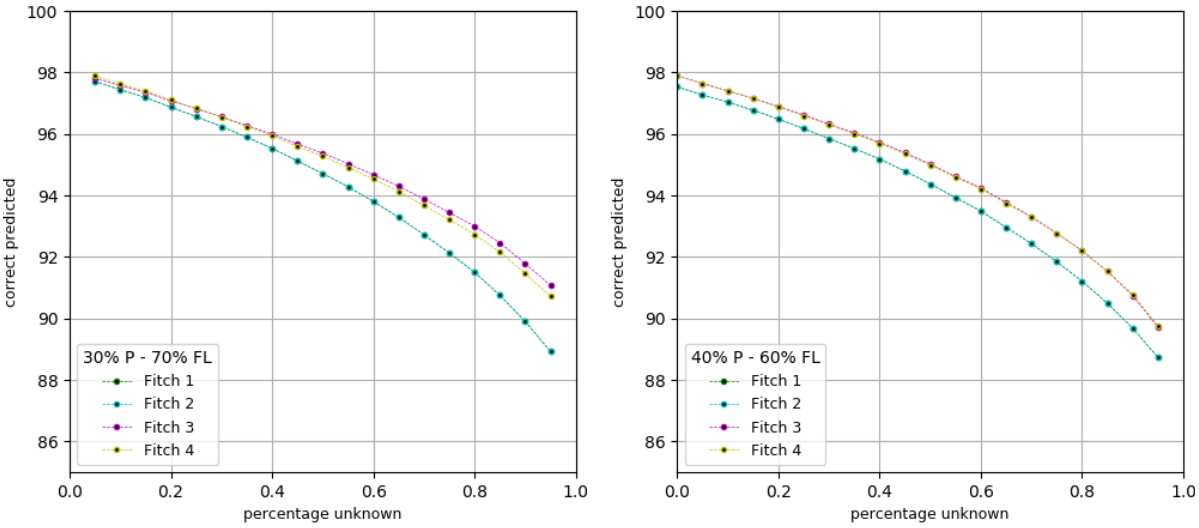


Figure 3.3: Test of Fitch Versions. TODO: beschriften!!

These four versions are tested in the simulation with 100 trees and 10000 leaf nodes and distributions of 70% FL to 30% P and 60% FL to 40% P. Figure 3.3 shows this over all unknown node percentage.
At 95% unknown nodes and 95% of multifurcation of the internal nodes, version 1 is 88.37%, version 2 is 88.37%, version 3 is 88.4%, and version 4 is 88.39% correct. Therefore, only version 3 is used for all further simulations.

3.3.2 Sankoff maximum parsimony

Maximum parsimony algorithm from Sankoff implemented in the R package *castor* [8]. From this the function $hsp\_max\_parsimony()$ is used with default settings including $transition\_costs = "all\_equal"$.

3.4 Simulation

The simulation compares these different 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 [8].

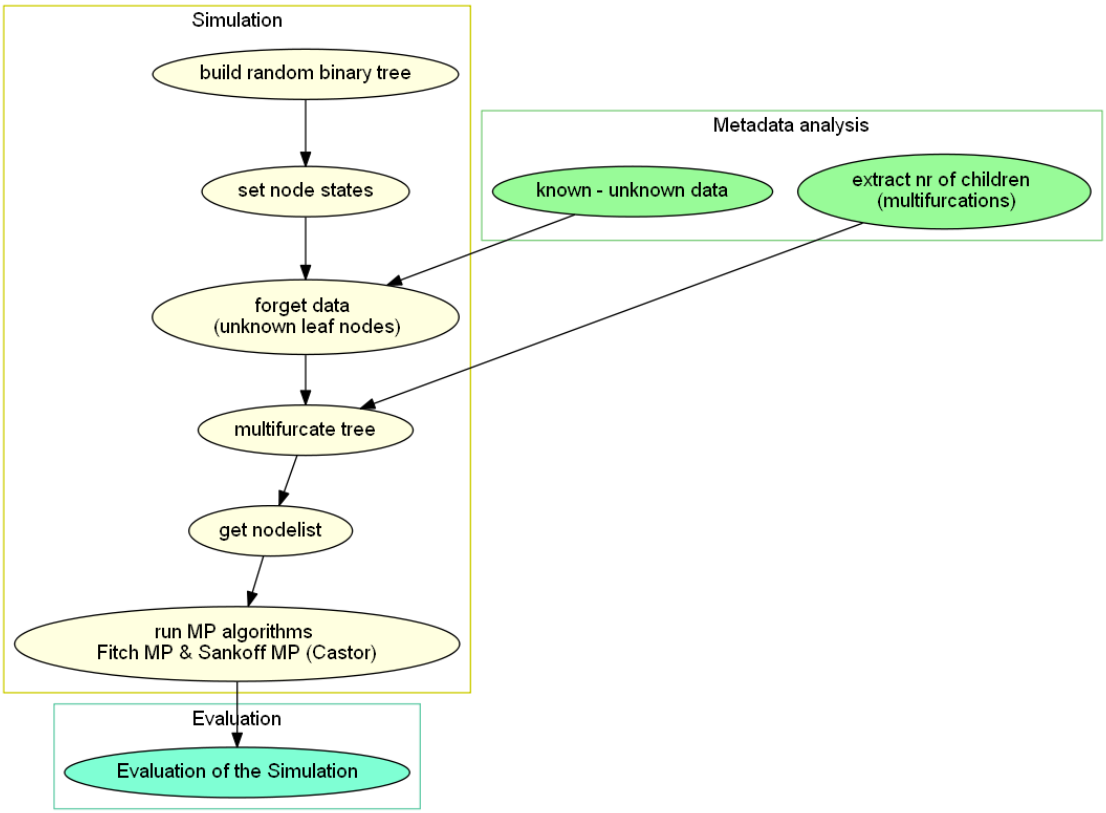


Figure 3.4: A simulation is performed to compare different maximum parsimony algorithms. The course of the simulation with influence of the metadata analysis from the real data can be seen:
(1) A phylogenetic tree is created randomly. (2) Simulate node states for all nodes. (3) 'Forget' internal states 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.

The course of a simulation is shown in Figure 3.4. The individual steps are explained in the following.

A tree is needed to perform a simulation of ancestral state reconstruction. It has 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 [17]. They offer a *randomized()* function which returns a BaseTree[3].

The next step is to simulate states for all nodes.
Here the influence of the biological parameters as transition probabilities and distributions of states is included.
Since there is no statement about general transition probabilities, these are all set the same:
$\mathcal{P}(FL \to P) = \mathcal{P}(P \to FL)$.

For the distributions of states different ratios of parasites (P) to free-livings (FL) are simulated with the help of beta distributions and a given threshold:

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

Procedure: The root node is defined as ancestor of all subsequent species and in this case, determined to be free-living. Therefore, the 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

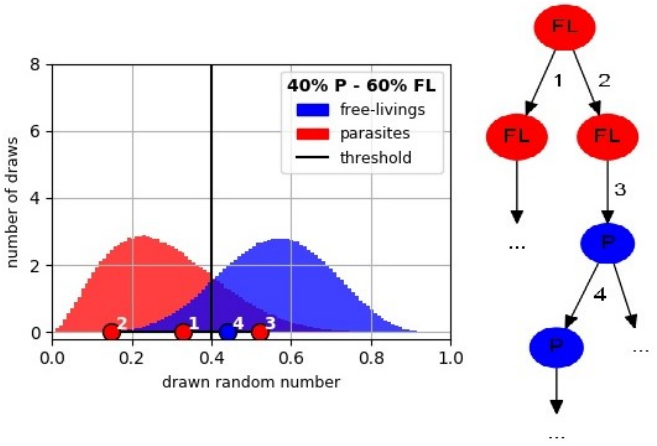


Figure 3.5: Set node states: Distribution of states (left); traversion through the tree (right). Start with a free-living root node (FL: red).
(1) + (2) Draw random numbers for its children from the free-living distribution (red), the numbers are under the threshold → the nodes are again free-living; 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 under the threshold. Then change back.

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

distribution until the randomly drawn number is above the threshold and the new node changes state. Figure 3.5 shows a part of this simulating states and the associated distributions.

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

Here begins the simulation of the lack of information, such as the lack of information on topology (multifurcation) and of states of some leaf nodes.

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

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 removed information are simulated.

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.

3.5 Real data analysis

For the evaluation of real data results, some statistics are collected and a leave-100-out cross-validation is performed. For this purpose 100 randomly distributed 100 states are left out.

The statistics are collected over the entire Eukaryota tree and over some subtrees of different taxa. There are known states besides predicted states, the predicted root node state specified

and origins and losses are counted by parasitism.
Since Sankoff predicts probabilities for states, we have rounded them to be able to count transitions from free-living (0) to parasitic (1) and vice versa.
In each case, a maximum of one changing transition from parent node to child node is counted for the origins (FL $\rightarrow$ P) or losses (P $\rightarrow$ FL).

For the leave-100-out cross-validation, analogous to the missing leaf node states, generalized linear models with binomial regression are compared according to their residuals, based on the "true" / "false" prediction.
Again, in order to compare models of different complexity, the BIC (Bayesian Information Criterion) values are calculated in addition to the residuals.
For all these calculations, the following R functions are used: *glm()*, *anova()* and *BIC()*.

3.6 Implementation

The complete code is located on GitHub: github.com/Irallia/IZW-HU-Parasites.
Most of the code is written in Python, the analyzes and the use of the *castor* package in R. Some shell scripts are used to execute whole workflows.

4 Results

In the first part of this work, we were able to provide the proof of concept for the Sankoff algorithm to perform an ancestral state reconstruction of the present Eukaryota tree. We will therefore discuss in the following how significant the result of this reconstruction is.

To do so, we first look at the input data we have: the use of Open Tree of Life for the tree gives us the greatest approximation to a phylogeny of the Eukaryota. Since such a large phylogeny has not been completely dissected, this is only a synthesis of phylogenetic trees with a taxonomy as a backbone. As a result, the tree is far from binary, we found a multifurcation of 89.45 %. However, this is highly variable and therefore the reconstruction is still very good on wide parts of the tree.
On the other hand, we used the Global Biotic Interactions (GloBI) database to get the state information for the leaf nodes. However, we were able to extract only 2.66% of the node information from GloBI. However, we found out that there are more data in GloBI that could be taken away with some work.

Next, let's consider the limitations of the simulation that led to the choice of the Sankoff algorithm. Our own main point of criticism is that we did not simulate the transition probabilities and estimate an unrealistically high number of transitions (origins and losses). The Sankoff algorithm is very well scored despite this increased difficulty (more transitions means more complex predictions) and its proven feasibility.

Finally, we analyze the validation of our ancestral state reconstruction results. We will do a selective analysis from a biological perspective and discuss the leave-100-out cross-validation. Above all, the 'biological view' gives a first impression of the credibility of the results. But it can not make any statistical statements. Here the interested researcher may have to analyze the subtrees of being relevant to him for himself.

Of course, cross-validation can be expanded. With more computing power, a leave-1-out cross-validation could be performed. However, our result on this point is 98.17% correct.

TODO: old:

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 node states (free-living or parasite states) from GloBI [15] (section 4.1), where we have determined that the entire Eukaryota tree is 89.45% multifurcated (missing internal nodes) and that we do not have any state information for 97.34% nodes.
In addition, we tested different models for these two parameters and found out that multifurcation can best be modeled with the taxa 'order' multiplied by the mean height of the knots. Also, the missing states could be modeled with the taxa 'order' and in this case multiplied by the depth of the nodes.

Next we compared different possible methods. We decided to take a closer look at maximum parsimony algorithms because they are fast and best suited to the problem at hand. We tested these on different simulations of the data and compared their predictive power (section 4.2). We compared them with the result that the Sankoff algorithm performs significantly better than Fitch. The multifurcation rate has little impact on Sankoff, but Fitch's prediction significantly breaks down at more than 60% missing internal nodes. However, the amount of missing node state information has a linear impact on both, as does the ratio of free-living to parasitic (the more balanced, the worse is the prediction).

As a result, we used the Sankoff algorithm [7] implemented by Louca et al. [8], which performed best, for the actual reconstruction on the real data.
We evaluated these on the one hand by analysing some statistics (subsection 4.3.1) and on the other hand by performing a leave-100-out cross-validation. There we predicted approximately 98.17% of the omitted states correct (subsection 4.3.2).

4.1 Availability of internal nodes and state information

As previously presented, 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. 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 241,974 internal nodes, that is:

$$100 - \frac{100}{(2293463 - 1) \times 241974} \approx 89.45\%$$

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.

Subtree of	Unknown States	Multifurcation
Eukaryota	97.34%	89.45%
Chloroplastida	99.14%	89.46%
Fungi	98.87%	96.97%
Metazoa	96.44%	87.93%
Apicomplexa	86.26%	87.16%
Arthropoda	97.49%	89.95%
Chordata	88.59%	66.49%
Nematoda	89.01%	88.59%
Platyhelminthes	68.73%	80.34%
Insecta	97.11%	90.78%

Table 4.1: Examination of subtrees regarding missing information.
The percentage values show the proportion 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 are highlighted in green.

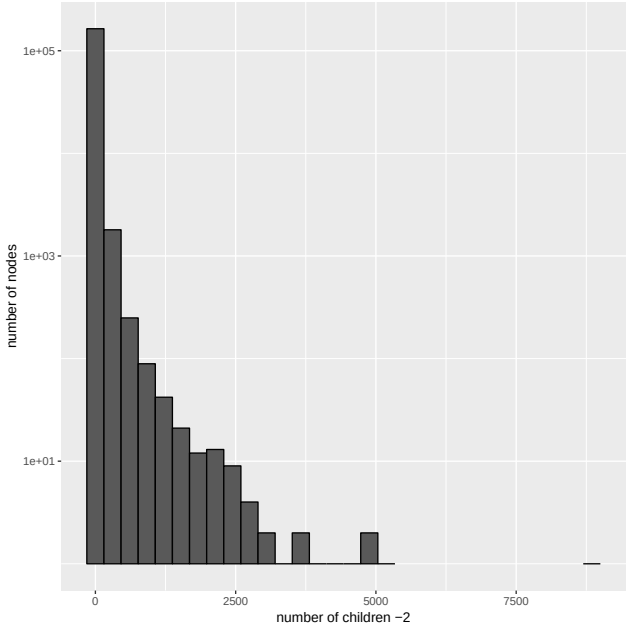
For the missing state information the present Eukaryota tree with 2,293,463 leaf nodes,

34,869 free-livings and 25,962 parasites are found. This gives

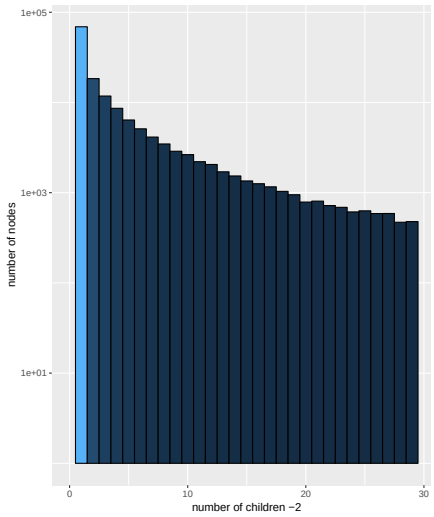
$$100 - \frac{100}{2293463 \times (34860 + 25962)} \approx 97.34\%$$

unknown states of leaf nodes.

We calculated these percentages of missing nodes and also missing state information for some subtrees and plotted them in table 4.1.



(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 4.1: Histograms showing the multifurcation of the internal nodes of the synthesis tree. For each node, the number of children (degree -1) is collected. A node is multifurcated if it has more than two children, so we deducted two from each number of children. The two histograms show: number of children -2 on the x-axis with log scale and the number of nodes with this amount on the y-axis.

For a first overview, we collected for each node its number of children (degree -1), and plotted this in two histograms, see figure 4.1.

The multifurcation affects only the internal nodes. We collected the number of children -2 of these nodes (because 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.
It can be recognized that we are very far from a binary tree.

4.1.1 Comparison of different models for multifurcation

For the regression analysis of the multifurcation we set up several generalized linear 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 (first row) and the associated BIC values (table 4.2).

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

Table 4.2: BIC (Bayesian information criterion) values of the multifurcation models. These models are created with the R function *glm()* and compared with the *BIC()* function. This results in the listed BIC values, presented in a heatmap colorization.

Within 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 (e.g. family) are computationally too expensive to calculate.
The model *multifurcation* $\sim$ *order * mean.height* turns out to be the best of our models, whereby it is possible that e.g. *multifurcation* $\sim$ *family * mean.height* is better.

4.1.2 Comparison of different models for missing state information

Next to the problem of the multifurcation of the tree is the little interaction data that 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 Eukaryota 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 free-living species and 47,332 parasite species for the whole tree of life. From the Eukaryota we could determine 25,962 and 34,860 species as parasites and free-living. With 2,293,463 leaf nodes we still have about 97.34% unknown leaf nodes.

We also compared different models in terms of their BICs (table: 4.3).

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

Table 4.3: BIC (Bayesian information criterion) values unknown state information models. These models are created with the R function *glm()* and compared with the *BIC()* function. This results in the listed BIC values, presented in a heatmap colorization.

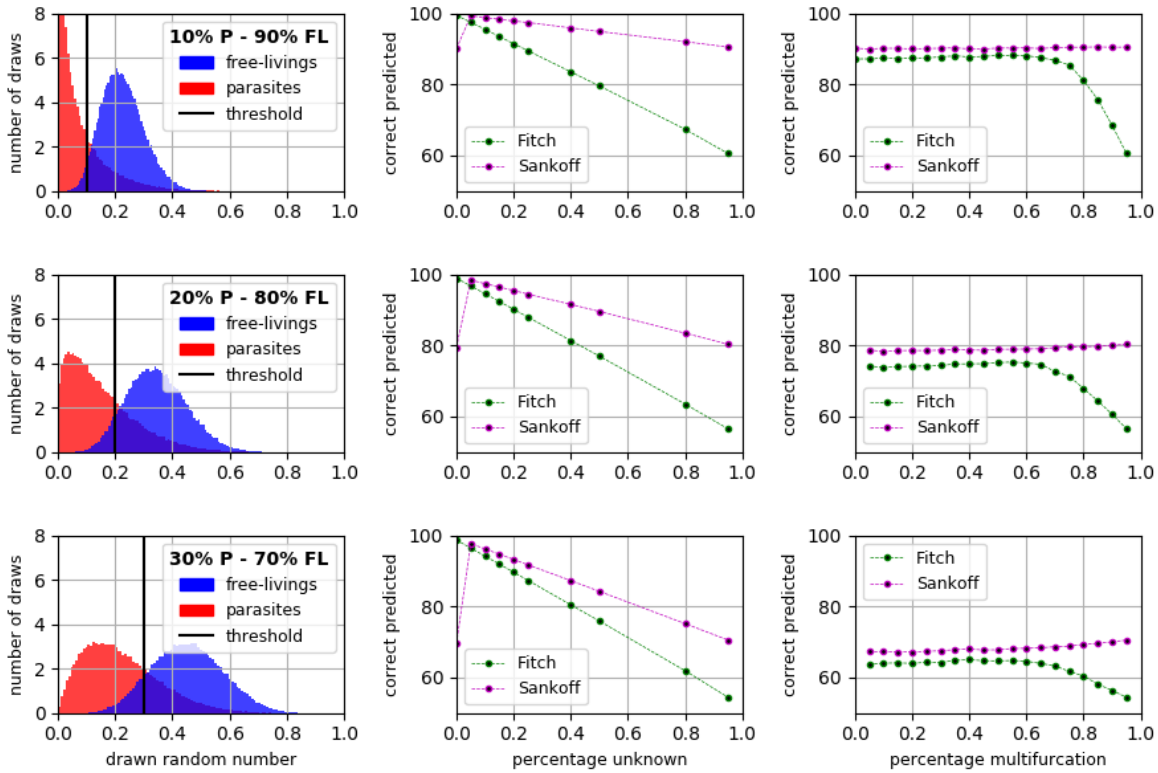
It also follows from this table that the most complex model is the best. In general, the BIC values are smaller than those of the multifurcation models. The modeling here is thus better. Again, the calculation of finer models (e.g. family) was too expensive.

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

4.2 Influence of different parameters on the prediction

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

First, we tested different distributions of parasites to free-livings 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 the predictive power of Sankoff is always greater or equal to the percentage of free-living distribution, and therefore more accurate than guessing.
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.
Figure 4.2 shows the results of examining various parameters.



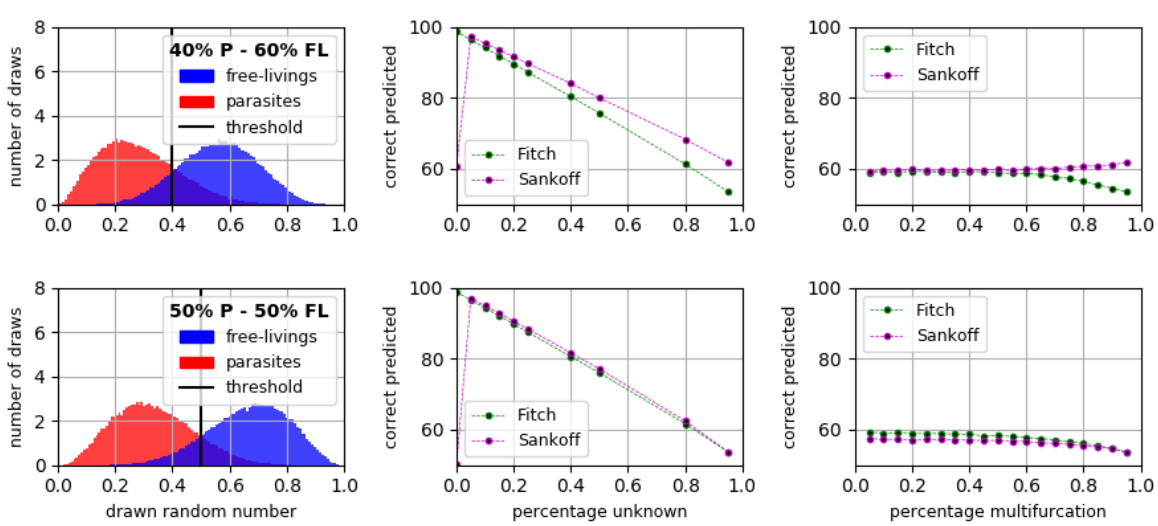


Figure 4.2: Influence of unknown data to prediction.
The first column describes the distributions of free-livings and parasites with a given threshold for the respective simulations to the right. Note, these distributions are chosen to form the specified proportions with the threshold, ignoring the number of state changes.
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.

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

4.3 Results of the real data analysis created with Sankoff

This section is about evaluating the prediction of real data using the Sankoff method and is divided into two subsections.
The analysis of some statistics on the predicted states, with the result that we predict a total of significantly more free-livings than parasites despite almost 50:50 distributed input states (subsection 4.3.1) and the presentation of the leave-100-out cross-validation results (subsection 4.3.2) This has revealed that we predicted approximately 98.17% correct with a small variance.

4.3.1 Statistics on predicted states

In the entire Eukaryota tree we start with 57.31% free-living species and 42.69% parasites. The prediction yields 80.98% free-livings 0.31% undefinable and 18.71% parasites. In addition, there are 462 origins of parasitism and 369 losses distributed throughout the tree, with the root node predicted to be parasitic.
We have predicted significantly more free-livings than parasites, but parasitism has prevailed to the root node.

In table 4.4 we compare different subtrees of different taxa on these prediction results.

Similar to the Eukaryota it is with the Metazoa subtree, which accounts for most of the Eukaryota (179,944 internal and 1,491,012 leaf nodes). However, here the root node is no longer calculated as parasitic but still as indefinable.
Strongly parasitic subtrees are for example the Apicomplexa, the Nematoda and the Platyhelminthes. For all three, the prediction is at a similar proportion between input values of the states and predicted states, e.g. In the Apicomplexa with 99.61% known parasitic states, 99.95% parasites were predicted. The same applies to the chordata, only that these mainly consist of free-living species.

↓ Taxa	Subtree	number of internal nodes	number of leaf nodes	known states FL	known states P	final states [0, 0.5) FL	final states 0.5	final states (0.5, 1] P	Root node state	#origins (with rounding) FL → P	#losses (with rounding) P → FL
Domain	Eukaryota	241974	2293463	34860	25962	2053212	7775	474450	1	462	369
				57.31%	42.69%	80.98%	0.31%	18.71%	P		
Kingdom	Chloroplastida	43486	416478	3519	77	410795	4182	1501	0.5	97	222
				97.86%	2.14%	98.63%	1.00%	0.36%			
	Fungi	9534	31457	577	2983	38520	5723	266463	0	42	2
				16.21%	83.79%	12.40%	1.84%	85.76%	FL		
	Metazoa	179944	1491012	30758	22373	1329065	25535	136412	0.5	321	129
				57.89%	42.11%	89.14%	1.71%	9.15%			
Phylum	Apicomplexa (Chloroplastida)	239	1863	1	255	1	0	1862	1	0	1
				0.39%	99.61%	0.05%	0%	99.95%	P		
	Arthropoda (Metazoa)	120479	1198981	18912	11141	1100822	22478	76064	0	281	108
				62.93%	37.07%	91.78%	1.87%	6.34%	FL		
	Chordata (Metazoa)	30761	91785	10451	18	91759	0	26	0	12	1
				99.83%	0.49%	99.97%	0%	0.03%	FL		
	Nematoda (Metazoa)	3437	30127	21	3289	1746	1196	27185	1	2	11
				0.63%	99.37%	5.79%	3.97%	90.23%	P		
	Platyhelminthes (Metazoa)	4459	22683	7	7086	175	0	22508	0	0	5
				0.10%	99.9%	0.77%	0%	99.23%	P		
Class	Insecta (Arthropoda)	91256	989572	17841	10734	1022747	976	57105	0	245	77
				62.44%	37.56%	94.63%	0.09%	5.28%	FL		

Table 4.4: Results of some selected taxa (subtrees).
(1 + 2) subtree taxa and name; (3 + 4) number of internal nodes and leaf nodes; (5 + 6) number or percentage of known states; (7 - 9) number or percentage of predicted states; (10) predicted state of the root node of the subtree; (11 + 12) origins and losses with rounded states.

Weinstein and Kuris have been searching for origins of parasitism in Animalia [18]. They identified 223 parasitic origins: 223 in Metazoa $\supset$ 143 in Arthropoda $\supset$ 87 in Insecta.
In the last columns of the table 4.4 we can see, that we found some more origins than Weinstein: 321 in Metazoa $\supset$ 281 in Arthropoda $\supset$ 245 in Insecta and on top of that some losses: 129 in Metazoa $\supset$ 108 in Arthropoda $\supset$ 77 in Insecta. The origins and losses calculated by Sankoff are thus closer to the leaves in the tree.

4.3.2 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.Omit smaller amounts of states up to leave-one-out had too much computational effort.

Of these 10,000 nodes, 9,238 are unique. From the unique nodes, we predicted approximately 98.17% correct and thus 1.82% wrong, with duplicate draws always having the same prediction. The associated variance is 0.017, which means that no large amounts of input states have to be omitted.

We have again considered how this data can best be modeled to compare whether there is a lot of variability in the taxa and whether the depth is an influencing factor over our prediction. So, the influence of the taxa (kingdom, phylum, class) and the depth of the leaf nodes is re-modeled and the BICs compared (table 4.5). Lower taxa than classes (e.g. order) are computationally too expensive to calculate.

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

Table 4.5: BIC (Bayesian information criterion) values of cross-validation prediction models. These models are created with the R function *glm()* and compared with the *BIC()* function. This results in the listed BIC values, presented in a heatmap colorization.

The BIC values this time did not prove that the finest model is the best. Of our calculated models, *correctly 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. Table 4.6 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

Table 4.6: Statistics about the leave-100-out cross-validation
The distance between original and new states is calculated using the Euclidean metric. This is summed over all states, all leaf node states and all internal node states.

5 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.

In the first part of this work, we were able to provide the proof of concept for the Sankoff algorithm to perform an ancestral state reconstruction of the present Eukaryota tree. In the following, we will first discuss the data situation and the simulation that led to the selection of the Sankoff algorithm, and then discuss how significant the result of this reconstruction is.

We evaluated our ancestral state reconstruction and the relational state prediction using a leave-100-out cross-validation. In addition, in this chapter, we will take a closer look at some subtrees and discuss their credibility from a biological perspective.

5.1 Data situation

The used Eukaryota synthesis tree from OTL [1] has 241,974 internal nodes and 293,463 leaf nodes. In addition, we could specify 25,962 parasitic and 34,860 free-living species from GloBI [15].
This gives us a high number of missing internal nodes (high multifurcation) and a low

number of node states of the leaf nodes in the entire tree. We suspect that this lack of information is not equally distributed in both cases. We conclude this from various observations.

For one thing, there were a few subtrees of higher ranks whose information share was much higher. In the Chordata there is only 66.49% multifurcation and in the Platyhelminthes only 68.73% missing state information.
Furthermore, we know that OTL is a synthesis of 819 phylogenetic trees. This suggests that just these parts of the tree are very well resolved and the remainder consists only of the rough taxonomy. For a review of this underlying taxonomy, we have compiled a list of all taxa and plottet them in relation to their distribution (see appendix section 7.2). There the number of nodes in the tree corresponding to a taxonomic rank is tabulated.

From GloBI we got very little information about leaf node states compared to their number in the Eukaryota tree. Nevertheless, the prediction of leave-100-out cross-validation was very good. This also leads to the conclusion that the information is very clustered.

As a last point, we would like to cite the modeling of the missing information. The fact that this is getting better, the lower ordered taxa (smaller subtrees) we are looking at, also suggests a high variability of the data.

We would like to mention a few other aspects of the GloBI database here.
Of course, such large data sets are flawed. We found some misinformation and were able to report some of these directly to GloBI.
We also found that there is some unused information in GloBI. We therefore found some more source species and target species without OTT identifiers. Since we currently use only OTT identifiers, 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 identifiers.

5.2 Simulation

The aim of the simulation is to test the influence of various unknown or uncertain parameters on some selected algorithms in order to test the credibility of the prediction of these for the given problem.

We assume that different parasite types have different transition probabilities. It is therefore difficult to establish a common distribution across the Eukaryota tree.
Based on the estimates of Windsor [19], 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 found that this combination has a significant influence on the predictive power of the algorithms.
We have adapted the combination of distributions and threshold to a high probability of achieving the given proportions on initial attempts. However, this means that we prefer high transition probabilities for both transitions and thus the simulated trees have a large number of transitions. It can be assumed that this may not correspond to the real data and thus means a limitation of our simulation. Further we assume that with less number of state changes, the predictive power of both algorithms would be even better, thus this limitation is probably not very restrictive.
At this point it would be possible to test other distributions with equal threshold values. this would be more computationally expensive, but would result in fewer transitions.

The used *castor* package [8] offers the possibility to enter different transition probabilities, so with information about these probabilities, you might be able to improve the prediction here.

5.3 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.

Several factors such as the distribution of existing input data via parasites and free-livings, faulty input data from GloBI and reinforcement of errors by multifurcation play an important role which becomes evident in these examples.

The accuracy of our results 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. They are positive in the sense that the majority have the opposite state.

In contrast to the other phyla examined, the phylogeny in the Chordata is more pronounced (less multifurcation). This leads to a lower variance of errors, which is reflected in the results. 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).
For a detailed view we have created and described a tree with all known and predicted parasites (see appendix 7.3).

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.

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 [20]. 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.

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

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 detail here.
For the Platyhelminthes we have more state information for the leaf nodes compared to the other considered subtrees. 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 and Koutsovoulos estimate the order of 25,000 parasites in the Nematoda and and speak of at 18 independently arosen parasites in Nematoda [21].
We assume that the parasites have been much more studied and thus we start with only 0.63% 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%.
The root node of the Nematoda is predicted as a parasite and so we got 11 losses of parasitism and only 2 origins in this phylum. 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 with a similar size as Blaxter and Koutsovoulos [21].

5.4 Conclusion

From the simulation, we can conclude that we predict correctly about 60% of the nodes in the present data situation. The leave-100-out cross-validation even showed that we predict the omitted nodes to be 98.17% percent correct.
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 statistical and biological analysis of the considered subtrees: Chordata, Nematoda, Platyhelminthes and Apicomplexa.
However, this means that the ancestral states' data in the direction of root node are probably particularly unbelievable. This makes the localization of origins in the direction of the 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 [18] an with Blaxter and

Koutsovoulos [21] showed similar numbers of origins in some subtrees, which leaves us with optimism.

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

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

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

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

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

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

[8] Louca, S.; Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **2017**, btx701.

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

[10] Cunningham, C. W.; Omland, K. E.; Oakley, T. H. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution* **1998**, *13*, 361 – 366.

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

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

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

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

[15] 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.

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

[17] 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.

[18] Weinstein, S. B.; Kuris, A. M. Independent origins of parasitism in Animalia. *Biology Letters* **2016**, *12*.

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

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

[21] Blaxter, M.; Koutsovoulos, G. The evolution of parasitism in Nematoda. *Parasitology* **2015**, *142*, S26–S39.

[22] Rothschild, M.; Clay, T. *Fleas, Flukes & Cuckoos; a Study of Bird Parasites*; New York,Macmillan,, 1957; p 368, https://www.biodiversitylibrary.org/bibliography/6413.

7 Appendix

7.1 Methods overview

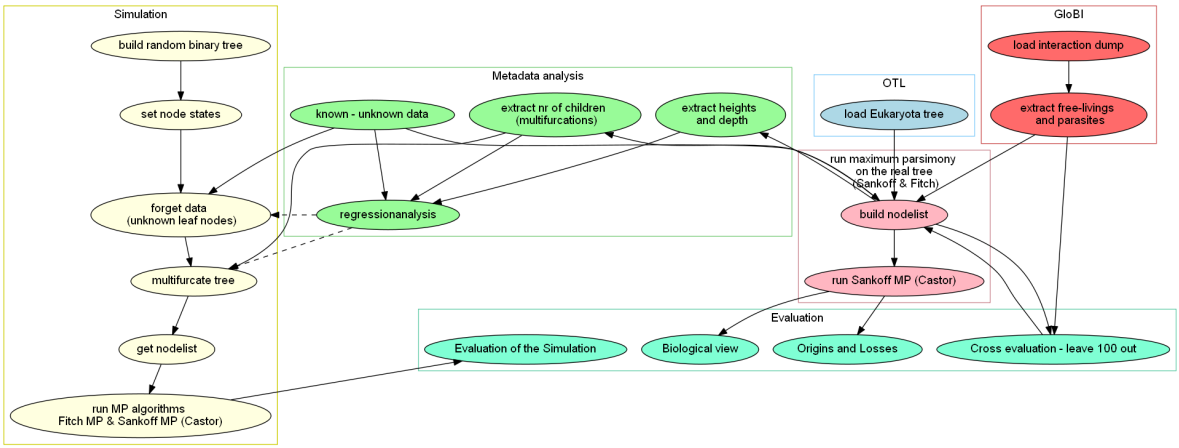


Figure 7.1: Big overview of the whole Workflow

7.2 OTL analysis

7.2.1 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.

7.2.2 Distribution of Taxa

In the tree we can distinguish 28 different Taxa in the OTL taxonomic tree. We have all Taxa sorted by rank into a table and counted their occurrences as internal nodes and leaf nodes (see table 7.1) and plotted them as a diagram (see figure 7.2).

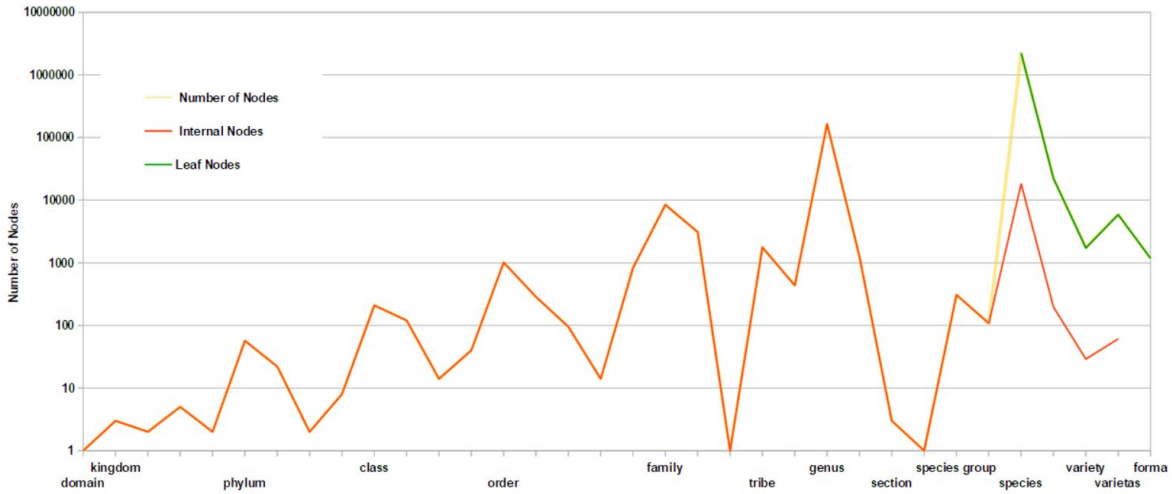


Figure 7.2: Distribution of Nodes in Rank-Cathegories

7.3 Parasites in Chordata

We mapped the few parasitic species in a rough taxonomy (see Figure 7.3):
Known parasitic birds belong to the order Sauria. Here we know from Rothschild here that are breeding parasites, like the cuckoo and clepto-parasites as the skuas [22]. 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.

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

Taxa	Number of Nodes	Internal Nodes	Leaf Nodes
domain	1	1	-
kingdom	3	3	-
subkingdom	2	2	-
infrakingdom	5	5	-
superphylum	2	2	-
phylum	57	57	-
subphylum	22	22	-
infraphylum	2	2	-
superclass	8	8	-
class	209	209	-
subclass	120	120	-
infraclass	14	14	-
superorder	40	40	-
order	1013	1013	-
suborder	285	285	-
infraorder	95	95	-
parvorder	14	14	-
superfamily	829	829	-
family	8442	8442	-
subfamily	3090	3090	-
supertribe	1	1	-
tribe	1764	1764	-
subtribe	435	435	-
genus	164592	164592	-
subgenus	1266	1266	-
section	3	3	-
subsection	1	1	-
species group	308	308	-
species subgroup	108	108	-
species	2243241	18197	2225044
subspecies	22378	196	22182
variety	1755	29	1726
varietas	5935	61	5874
forma	1179		1179
no rank	951	944	7
no rank - terminal	37451	-	37451
(no entry)	39816	39816	-

Table 7.1: Distribution of Nodes in Rank-Cathegories

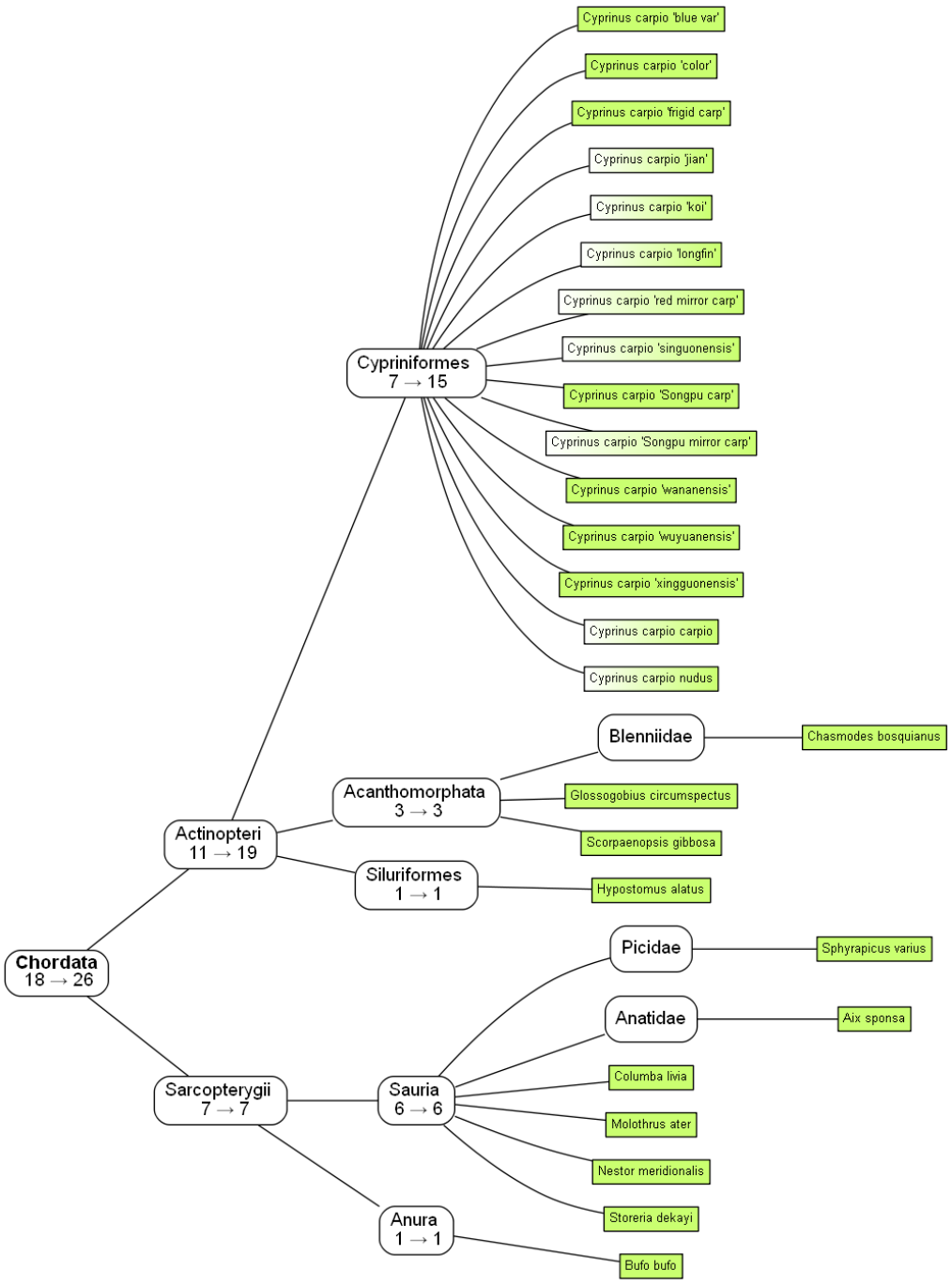


Figure 7.3: Parasites of Phylum: Chordata.
All parasite data of the chordata are mapped into a rough taxonomy (phylum, class, order, family) in order to understand their distribution and affiliation.
The internal nodes are the wanted taxa from OTL (with the addition of # input parasites to → # predicted parasites).
The leaf nodes are the input parasites (green) and the predicted parasites (white → green).