Accuracy of Morphology-based Phylogenetic Fossil Placement under Maximum Likelihood

Simon A. Berger and Alexandros Stamatakis
Dept. of Computer Science, Technische Universität München, Boltzmannstr. 3, 85748 Garching b. München, Germany
Email: simon.berger@in.tum.de, stamatak@in.tum.de

***Abstract*—The capability to conduct Maximum Likelihood based phylogenetic (evolutionary) analyses on datasets that contain both morphological, as well as molecular data partitions with programs such as RAxML, gives rise to new methodological questions. As we demonstrate on 5 real world datasets that comprise morphological as well as DNA data the trees inferred by separately using the morphological or molecular data partitions are highly incongruent. Since in typical current-day phylogenomic alignments, there is significantly more molecular than morphological data available, and hence the final tree shape in a concatenated analysis is dominated by molecular data, the question arises how morphological data can be used within this context. One important application lies in the phylogenetic placement of fossil taxa (for which only morphological data is available) into a fixed, given molecular or otherwise well-established reference tree. By using real and simulated datasets we conduct the first assessment of placement accuracy for fossil taxa under the Maximum Likelihood criterion. We demonstrate that, despite conflicting phylogenetic signals from the morphological and molecular partitions, the Maximum Likelihood criterion is powerful enough to yield accurate fossil placements. Moreover, we develop and make available a new morphological site weight calibration algorithm that yields an average improvement of fossil placement accuracy of 20% on more than 2,500 simulated datasets and of 25% on the 5 real-world datasets that all contain highly conflicting phylogenetic signal.**

***Index Terms*—Phylogenetic inference; Morphological Data; Fossil Placement; RAxML**

I. Introduction

The on-going extension of analysis capabilities for partitioned (phylogenomic) datasets in programs for Maximum Likelihood (ML [1]) based tree inference allows to address novel methodological questions. Recently, we have integrated additional ML substitution models for binary morphological data [2] into RAxML [3] (version 7.2.6, available at: http://wwwkramer.in.tum.de/exelixis/software.html). The new version of RAxML allows for the analysis of super-matrices (also called total evidence approach or multi-gene/phylogenomic alignments) that contain a mix of data-types, i.e., an input alignment may consist of concatenated morphological, DNA, and protein (amino acid) sequence partitions that represent the organisms under study.

An example for such a phylogenomic alignment with 4 present-day organisms (the great apes for instance) and one fossil taxon (e.g., some common extinct ancestor of the human and the chimpanzee) that entails a binary/morphological data partition with 6 sites (columns/morphological characters) and a DNA data partition with 24 sites is given below. As already mentioned, there will usually not be any molecular data available for the fossil taxa under study, hence the molecular sequence part of the fossil is filled with gaps, which are treated as undetermined characters in all standard ML-based and Bayesian implementations, and thereby do not influence the likelihood computations.

```
Fossil     001101------------------------
Human      000111A-GGCATATCCCATACAAAGGTTA
Chimp      000100ATGGCACACCCAACGCAAGGGTGA
Gorilla    001111ATGGCCAACCACTCCCAAAAGTCA
Orangutang 111011CGGGCACATGCAGCGCAA-A-T-A
```

Within this context we analyze the potential applications of morphological (throughout this paper we exclusively use binary characters, i.e., we do not consider multi-state morphological characters) data for gaining novel evolutionary insights. For this purpose we use 5 real-world partitioned input datasets that contain both morphological and molecular (DNA) partitions as well as more than 2,500 simulated morphological datasets. It is important to emphasize, that at present it is hard to use a larger number of real-world datasets for the purposes of our study, because they are not readily available in standard tree and alignment repositories such as TreeBase [4].

A general problem with morphological data within the phylogenomic context, is that only a few morphological character sites (typically 50–500 alignment columns, see Table I) are available compared to a constantly growing number of molecular character sites (typically 1,000 to tens of thousands in current phylogenomic studies, see, e.g., [5]). Thus, the overall per site log likelihood contribution of the morphological sites will be very small and therefore only have a negligible impact on the shape of the overall tree topology that is inferred based on the concatenated morphological and molecular dataset. In addition, there can be a significant incongruence between best-known ML trees (ML for phylogenetic trees is NP-hard [6]) obtained from individual tree searches on *either* the morphological *or* the molecular partitions of the input dataset. Given that, tree shapes are largely dominated by the molecular part of the input datasets because of the masses of molecular data that have now become available, the question arises what the potential use of those comparatively few (a couple of hundred compared to tens of thousands) morphological columns might be, since they will mostly add some insignificant noise to the signal of broadly sampled phylogenomic datasets. Currently, there exist two application scenarios that make use of a given

"true" reference tree, which we assume to be the molecular tree in this paper, though this assumption can evidently be challenged. As a reference tree one may also consider using a well-established species tree from the literature or for instance use the NCBI (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/) taxonomy to obtain a "true" reference tree.

Scenario I: Given a reference tree and a morphological data matrix one may use this matrix for the phylogenetic placement of a fossil taxon for which no molecular data exists. This essentially means that we impose a well-established reference topology onto the morphological character matrix and then try to insert (place) the fossil by computing the best-scoring insertion position under Maximum Likelihood in the reference tree.

Scenario II: One may also be interested in inferring the ancestral states on a fixed and potentially dated reference topology in order to determine at which point of time in the past (on which branch) a transition between, e.g., green eye color and blue eye color occured.

A problem that is, as shown in this paper, inherent to both use cases for morphological data is that of incongruence, i.e., conflict of phylogenetic signal, between the morphological tree and the (molecular) reference tree. Here, we address issues pertaining to *Scenario I*, i.e., we assess the accuracy of fossil placement for morphological data with an incongruent tree signal. While our computational experiments show, that placement accuracy under ML is already relatively good (above 85%) and robust against noise, despite conflicting signals in the data, we also devise a new statistical method that further improves fossil placement accuracy by approximately 20-25% on average.

The method which we denote as morphological weight calibration method, can infer weights for morphological alignment sites in such a way, that sites which are congruent to the reference tree obtain a higher weight than incongruent sites, such that they contribute more to the overall likelihood during the phylogenetic placement process. The above methods have been implemented in the current version 7.2.6 of RAxML which is freely available as open-source code at http://wwwkramer.in.tum.de/exelixis/software.html.

To the best of our knowledge, this work represents the first systematic assessment of fossil placement accuracy and introduces the first statistical procedure for morphological site weight calibration under Maximum Likelihood. In addition, we provide the first complete Maximum Likelihood-based framework, including a placement and a weight calibration method, for fossil placement as open-source code.

The remainder of this paper is organized as follows: In Section II we briefly cover related work. Thereafter (Section III), we outline our fossil placement algorithm and in Section IV our statistical weight calibration method. In Section V we describe the experimental setup and datasets used. In the following Section VI we describe experimental results on simulated and real data, and also discuss the placements of real fossils in a biological context. We conclude in Section VII.

II. Related Work

The assignment of weights to morphological sites, which we henceforth denote as weight calibration problem, has previously mainly been addressed within the framework of correct value range treatment for quantitative versus qualitative traits [7], [8], i.e., not with the goal to reduce incongruence, but with better biological modeling in mind. Those methods are primarily used in phylogenetic analyses under Maximum Parsimony (MP [9]), where each morphologic trait (character) needs to have the same relative weight. For MP, weight calibration is used to eliminate unequal weightings that may arise from different value ranges on multi-state morphological characters. Because nothing is known about the relative informativeness of transformations on different characters, equal weighting should be assumed a priori [8], [10]. As pointed out by J.J. Wiens [7] this issue has generally received little attention, despite its importance and biological relevance.

In contrast to the above, we investigate (i) to which extent incongruent signal in the morphological and molecular data partitions can bias placement accuracy and (ii) if morphological site weight calibration can be used to filter out morphological sites that are highly congruent to the reference tree.

In a recent paper J.J. Wiens [11], addresses the question if the addition of molecular data (instead of using morphological data alone) can improve the phylogenetic position/placement of fossils (for which molecular data is not available) in trees, by exclusively using simulated datasets and Bayesian as well as Maximum Parsimony methods. While he finds that the usage of molecular data in addition to morphological data can increase accuracy, or will at least not affect accuracy in the worst case, he does not address the effects of incongruent signal in the morphological and molecular partitions on placement accuracy. Our approach is different in that we assume, that the molecular tree is the reference tree and that there may be a significant amount of incongruence in the trees favored by the molecular and morphological partitions. We also demonstrate this incongruence on real datasets.

III. Fossil Placement Algorithm

Initially, we require a method to place our fossil(s) into a given molecular or otherwise well-established reference tree by exclusively using the morphological part of data for which fossil data is available. An example is provided in Figure 1, where we intend to place a fossil into a reference tree with 4 current-day organisms, once again using the example of the great apes. For the sake of simplicity we will only consider the case were we need to place a single fossil into the tree; the placement procedure for more than one fossil is analogous.

The input for the fossil placement algorithm in RAxML consists of the reference tree t_{ref} that comprises the n morphological sequences (4 in our example) of present-day species. The input alignment contains the n present-day sequences as well as the fossil sequence(s) we intend to place into the tree. As already mentioned we assume that t_{ref} has been obtained via a thorough ML analysis of the

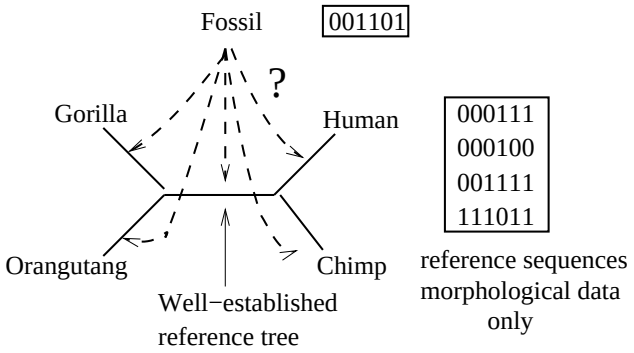

Fig. 1. Example for a phylogenetic fossil placement problem.

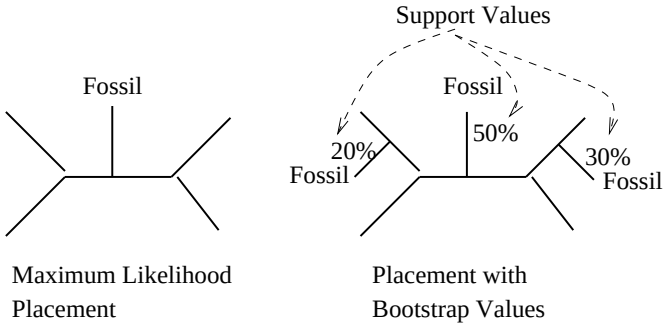

Fig. 2. Example output of the phylogenetic fossil placement procedure without and with Bootstrap support values.

corresponding molecular sequence data or by using a well-established species tree, e.g., as obtained from the literature or the NCBI taxonomy database. Initially, the algorithm will read t_{ref} and the alignment and mark all sequences (in this case only the one fossil sequence) of the alignment that are *not* contained in t_{ref} as a query sequence(s). Thereafter, the ML model parameters and branch lengths of t_{ref} will be optimized.

After this step, the actual placement algorithm is invoked which will successively insert (and remove again) the fossil sequence into all $2n-3$ branches of t_{ref} and compute the respective ML score (insertion score). The insertion score will then be stored in a list that keeps track of the insertion scores of the fossil into all $2n-3$ branches (5 branches in the example) of the reference tree. The output of this procedure for evolutionary placement is then simply the input reference tree, extended by an assignment of the fossil to the respective best-scoring insertion branch in t_{ref} as outlined in Figure 2. Our placement algorithm can also conduct a phylogenetic bootstrapping procedure [12], i.e., repeat the computation of the best insertion score for the fossil under slight alterations of the input data. This allows for assessing uncertainty in the placement of the fossil by including several potential insertion positions into the reference tree and assigning respective bootstrap support values to each potential placement. The respective output of the bootstrapping procedure is also depicted in Figure 2. Finally, we can also invoke the above placement algorithm (with and without the Bootstrapping option) using an explicit weight vector to specify per column (per alignment site) integer weights. This option is important for using weight calibration results (see Section IV).

Finally, we also require measures to quantify the placement accuracy of the fossil. Therefore we will assume that a "true" position, i.e., the true insertion branch of the fossil, is known. We can than measure the distance between the true insertion position and the calculated insertion position. To quantify placement accuracy, we use two distance measures based on the topology and branch lengths of the reference tree.

To quantify the distance between the true position and the calculated position of the fossil we use the following measures: The "Node Distance" (ND), is the unweighted path length in the reference tree between the true and the calculated insertion branche. This corresponds to the number of nodes located on the path that connects the two insertion branches (see Figure 3a) and represents an absolute distance measure. The second measure is the sum of branch lengths on the path connecting the calculated with the true insertion branch. This measure includes 50% of the branch length of the insertion-branch and 50% of the length of the "true" original branch (see Figure 3a). For comparability between different trees and in order to obtain a relative measure of placement accuracy, we normalize the branch path length by dividing it through the maximum tree diameter (see Figure 3b). The maximum tree diameter is the branch path of maximum length between two taxa in the reference tree. This distance measure is henceforth denoted as "Branch Distance Normalized" (BDN%).

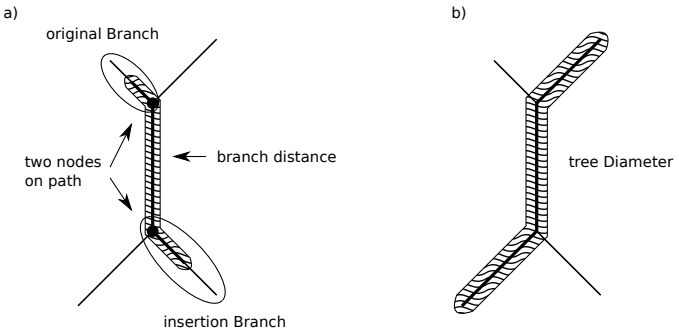

Fig. 3. Example tree with two branches (true and calculated insertion branch) highlighted. There are two nodes on the path, so the node distance is 2. The branch distance corresponds to the length of the connecting path, where of the two end branches only half of the branch length is used. (b) Tree diameter which is used to normalize the branch distance

When the placement algorithm is used with bootstrapping, more than one potential insertion branch can be proposed for a fossil (see Figure 2), which means that we need to appropriately adapt our distance measures to incorporate Bootstrap support values. For a bootstrap run with N_{bs} bootstrap replicates, the output of the algorithm contains a set of $i = 1...N$, where $N \leq N_{bs}$, insertion positions for the fossil with bootstrap values S_i. Using this information we derive a

set of ND or BDN distances D_i to the correct branch for each alternative Bootstrap placement i. We use the D_i to represent the bootstrap placement information as a single quantity for the fossil placement accuracy by defining the Weighted Root Mean Squared Distance (WRMSD), D_{wrms} as follows:

$$D_{wrms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{S_i}{N_{bs}}D_i)^2} \quad (1)$$

IV. Weight Calibration Algorithm

As already mentioned, there may be significant incongruence between the phylogenetic signal, i.e., the trees that are favored by the morphological and molecular partitions of the data. As such, one of the key questions is to which extent this incongruency affects the placement accuracy of fossils into a tree derived from molecular data, and if there exist mechanisms to efficiently determine which sites of the morphological data partition are congruent to the molecular reference tree.

Therefore, we need to devise a criterion to determine *those* sites from a—in our case—morphological data partition that are highly congruent to a given reference tree. Ideally, we would like to calibrate the weights, i.e., execute the fossil placement algorithm described in the preceding Section with a weight vector that enhances the signal of morphological sites that are congruent to the reference tree. This weighting scheme should idelly increase fossil placement accuracy and decrease the impact of noise caused by incongruent morphological character sites.

In order to achieve this, we have designed a randomized statistical procedure that works as follows. Initially, it reads in the reference tree t_{ref} and the morphological alignment and optimizes ML model parameters on this tree, without changing the tree topology. Once the model parameters have been optimized we store the per-site log likelihood values on the reference tree in an appropriate vector $\vec{L}_{ref}$ of length m, where m is the number of sites (columns) in the morphological data.

Thereafter, we generate a set of $n = 100$ random trees, $r_1, ..., r_{100}$. For each random tree r_i, where $i = 1...100$, we re-optimize ML model parameters again in order to compute the per-site log likelihood scores $\vec{L}_{r_i}$ for random tree r_i.

Once the per-site log likelihood scores $\vec{L}_{ref}$ on the reference tree and $\vec{L}_{r_i}$ on the 100 random trees have been computed, we can then determine the degree of congruence between a specific site j, where $j = 1...m$, and the reference tree, by counting in how many random trees r_i the per-site log likelihood of site j is worse than the per-site log likelihood $\vec{L}(j)_{ref}$ in the reference tree t_{ref}. For each site j we compute a weight vector entry $\vec{W}(j) = \sum_{i=1}^{n}\delta_{j,i}$ where $\delta_{ref,i}$ is defined as follows:

$$\delta_{ref,i} = \begin{cases} 1 & if \quad \vec{L}(j)_{ref} > \vec{L}(j)_{r_i} \\ 0 & else \end{cases} \quad (2)$$

name	# taxa	# mol sites	# morph sites
D1	35	2,006	117
D2	23	16,662	414
D3	32	1,713	381
D4	81	3,675	213
D5	18	266	35

TABLE I
Overview of test datasets.

The above definition means that sites that are highly incongruent with t_{ref} will have low weights close to 0, while sites that have weights close to 100 are highly congruent to the reference tree. The rationale behind the above approach is that a site that is highly congruent to the reference tree will score worse on random trees, while a site that is highly incongruent will score better or at least not worse on most random trees. The above weight vector $\vec{W}$ can be used directly as input to a placement analysis of a fossil. The weight vector $\vec{W}$ can also be used to derive a binary weight vector $\vec{W}_{bin}$ in which we set all elements with $\vec{W}(j) \geq 95$ to 1 and all elements j with $\vec{W}(j) < 95$ to 0 (using a typical cutoff at 5%). This allows us to more radically filter out incongruent sites. When comparing the per-site log likelihoods we do not explicitly use a method for determining if values are significantly different from each other, but rather compare site-wise log likelihoods directly, since those effects will be averaged out by the random re-sampling procedure. Finally, our tests indicate (results not shown) that the computation of 100 random trees is sufficient to infer stable weight values. Our experimental results on simulated data clearly show that the above approach is able to discriminate well between congruent and incongruent sites and thereby justify this approach.

V. Experimental Setup

A. Real-World Test Datasets

We used 5 real-world test datasets that contain morphological as well as molecular DNA data. The datasets are labelled as D1 through D5 for ease of reference. Table I provides the number of taxa and number of molecular as well as morphological sites for all input datasets we used. The real-world datasets can be downloaded at http://wwwkramer.in.tum.de/exelixis/morphologyDatasets.tar.bz2.

Dataset D1 [13] contains 35 taxa of walnut trees (*Juglandaceae*). The original alignment also contained an additional 5 fossils. D2 [14] comprises 23 Marsupial sequences (the original dataset also contained 10 fossils). D3 [15] contains 32 taxa of Amphibians (*Caudates*). D4 [16] contains 81 taxa of tree-frogs (*Hylidae*). Finally, D5 [17] contains 18 taxa that span a wider variety of species than the other datasets, ranging from the chicken to the homo sapiens.

It is important to emphasize, that despite our efforts to collect more combined morphological/molecular real world datasets, a call for such datasets via the RAxML mailing list, as well as a thorough search in the TreeBase database we were not able to gather more real-world datasets. Therefore, we

also generated simulated datasets as outlined in the following Section.

B. Simulated Datasets

An initial literature search revealed that, currently, there are no freely available programs for generating simulated morphological datasets available. Therefore, we contacted J.J. Wiens, who kindly made available to us the C code for generating simulated datasets that was used in [11]. We completely re-implemented and extended the original C program in Java. The program can now read in two distinct trees, for instance, one that is congruent to a reference topology and a random tree that is incongruent to the reference topology. This allows for generating simulated morphological datasets that entail two partitions with conflicting phylogenetic signal. In addition, the simulation program can generate morphological partitions of variable length, e.g., a partition of 300 sites that are incongruent to the reference tree and a partition with 100 sites that is congruent to the reference tree. Moreover, the simulation program allows for generation of an artificial fossil sequence, that is located at the innermost branch (the most distant position from current-day species) of the tree on which the data is being generated. While a fossil in general must not necessarily be located at the innermost branch of a tree (see [13]), this setup ensures that the placement problem as such is more difficult, since the closest current-day relatives of the fossil are located as far away as possible. In our simulated data generation tool, the artificial fossil is thus automatically placed onto the branch that has the longest branch-based path length to the nearest tips (leaves) at either end of the branch where the fossil is located.

C. Computation of RF-Distances

In order to assess incongruence between trees obtained from morphological and molecular data partitions in Section VI we need to compute the topological distances between trees.

The standard Robinson-Foulds [18] distance between two trees is defined as the number of non-trivial bipartitions (splits into taxon label sets induced by the inner branches of a tree) that are contained in one of the two trees but not in both. The RF distance is typically reported as relative distance, i.e., the count of distinct bipartitions divided by $2(n-3)$ where n is the number of organisms and $n-3$ the number of inner branches (branches not leading to tips/leaves). The number $2(n-3)$ hence represents the worst case for RF, i.e., the two trees under comparison do not share any bipartitions. In addition to the RF distance, one can also define the Weighted RF (WRF) distance that takes into account the Bootstrap support values on the branches. If there are incongruent bipartitions in the tree that have low support, e.g., 10%, they will contribute a total of 0.2 to the WRF distance, while they would contribute 2 to the RF score. Therefore, the WRF distance provides a better notion of whether trees disagree in strongly (important) or weakly (unimportant) supported bipartitions. The WRF distance also better resembles the way in which Biologists usually assess there results. In our experiments we used the respective RF and WRF options as implemented in RAxML.

dataset	RF(morph,mol)	WRF(morph,mol)
D1	59%	39%
D2	60%	37%
D3	62%	47%
D4	82%	45%
D5	80%	42%

TABLE II
INCONGRUENCE BETWEEN MORPHOLOGICAL AND MOLECULAR TREES & AVERAGE BS SUPPORT INDUCED BY MORPHOLOGICAL AND MOLECULAR PARTITIONS.

VI. RESULTS

A. Incongruence of Morphological and Molecular Data

Initially, we assessed the (in)congruence between the morphological and the molecular data partitions in our real world datasets to substantiate our claim that morphological and molecular partitions typically exhibit incongruent signal.

For this, we split up each real-world data set into the morphological and molecular partitions and conducted thorough ML analyses as follows: For the morphological and the molecular datasets we seperately conducted 100 bootstrap analyses and 50 ML searches for the best-scoring ML tree under the Γ model of rate heterogeneity [19] using RAxML.

We then used the corresponding RAxML option to draw Bootstrap support values on the respective best-scoring out of 50 ML trees. The RF and WRF distances between the respective best-scoring morphological and molecular trees with Bootstrap support values were then computed in order to determine incongruence between the data partitions (see Table II).

The values provided in Table II clearly show that the molecular and morphological trees are highly incongruent based on the RF and WRF distances. RF distances exceed 50% and WRF distances oscillate around 40% which means that several highly supported bipartitions of the molecular tree are not recovered by the morphological tree.

In order to assess the stand-alone topological stability of the morphological data partitions we conducted an additional 100 ML searches per dataset (on the morphological partitions only). We then computed the maximum RF distance and the mean RF distances within those ML tree sets based on all pairwise RF distances between the resulting 100 ML trees (in this case we do not include WRF distances, as the 100 ML trees have been calculated without bootstrap support). The average and maximum distances in those tree sets as shown in Table III provide a good notion for the general topological instability of the morphological partitions. Except for datasets D3 and D4 the mean RF distance largely exceeds 10% and the maximum RF is larger than 50% in most cases, i.e., ML trees for the same dataset only share 25% of non-trivial bipartitions. Given that the datasets are relatively small with respect to the number of taxa and that the RAxML search algorithm

has been shown to be very efficient in recovering the best-known tree [3] we conclude that there is a *significant* lack of signal with respect to tree reconstruction in the real-world morphological datasets under study.

dataset	max RF	mean RF	dataset	max RF	mean RF
D1	68.75%	21.70%	D2	25.00%	06.27%
D3	58.62%	25.56%	D4	64.10%	32.59%
D5	73.33%	33.05%			

TABLE III
PAIRWISE MEAN AND MAXIMUM RF VALUES FOR SETS OF 100 ML TREES.

This initial set of experiments underlines two major claims: *Firstly*, that morphological data partitions can yield significantly different trees than molecular data partitions and *secondly*, that morphological data partitions can suffer from a significant lack of or a weak phylogenetic signal and it is therefore difficult to use them for de novo phylogenetic inference. Based on this observation we focus on assessing the usage of morphological data partitions for the placement of fossils in the following computational experiments.

B. Morphological Weight Calibration

Initially we assessed if our statistical method for determining congruent and incongruent sites works on simulated datasets. For this purpose we generated a simulated dataset based upon the real molecular tree for dataset D3 and generated one incongruent partition with 200 morphological sites and one partition congruent to the molecular tree that also comprises 200 morphological sites. We then executed our algorithm on this dataset and plotted the inferred weights over the number of sites in the simulated alignment. As Figure 4 clearly shows we are able to distinguish between incongruent (first half of the alignment, low values) and congruent (second part of the alignment, high weight values) sites.

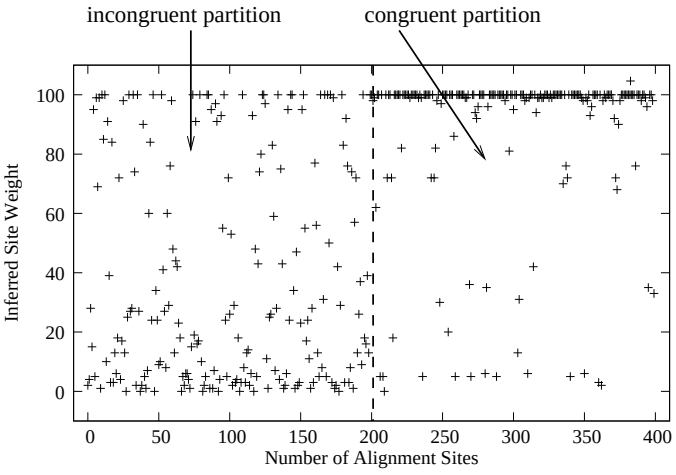

Fig. 4. Weight assignments for congruent and incongruent (with respect to a reference tree) data partitions of a simulated morphological dataset.

Results for larger datasets in terms of the number of organisms included and distinct input trees are analogous (results not shown).

C. Fossil Placement Accuracy on Simulated Datasets

To test the weight calibration algorithm on simulated datasets, we generated simulated datasets as follows: Based upon each of the 5 real molecular trees we generated 100 simulated datasets per real tree, by using different random seeds for every simulated alignment and a different random tree for sets of 10 simulated alignments in order to generate morphological sites that are incongruent to the molecular reference trees. For each set of those 100 simulation runs per reference tree we also generated morphological alignments of variable length, i.e., datasets containing 100 congruent as well as 100, 200, 300, 400, and 500 incongruent sites derived from the random tree. Thus, for each real input tree we generated a total of 500 simulated alignments. The rationale for this setup is to test up to which extent the degree of random noise in the alignment affects placement accuracy. For ease of reference we denote these simulated datasets as SX_YYY_ZZZ where SX denotes the molecular tree from datasets D1-D5 that was used to generate the congruent data partition, YYY the number of congruent sites, and ZZZ the number of incongruent sites.

For each simulated dataset size, we placed the fossil (generated as described in Section V-B) into the reference tree using our phylogenetic placement algorithm (see Section III) with Bootstrapping in order to avoid any random effects that may be caused by placement runs without Bootstrapping. We executed those placement runs for the unweighted case (all morphological sites included, without weight calibration; denoted as **UNW**), the case with integer weights (using the calibrated weights from $\vec{W}$; denoted as **INT**), using only the incongruent data partition (SX_ZZZ; denoted as **BAD**), and only the congruent data partition (SX_YYY; denoted as **GOOD**). The accuracy was then measured using Equation 1 to compute the node-based absolute accuracy. Relative branch-based accuracy results were analogous (data not shown). Fossil placement accuracy was averaged over the 100 simulated datasets for every dataset size.

The results of the simulated dataset experiments are provided in Table IV. Except for datasets S4, the approach with calibrated site weights using integer values, clearly improves placement accuracy by 25% to over 50%. In some cases (datasets S1 and S3) it even outperforms the fossil placement accuracy achieved by exclusively using the congruent data partition. Overall the weight calibration approach improves the average placement accuracy on all datasets (including dataset S4) from 4.72 to 3.84, i.e., achieves an accuracy improvement of 20%. Overall, there is a clear tendency for placement accuracy to decrease with an increased amount of incongruent sites. The bad overall performance on dataset S4—note that the congruent morphological partition does not achieve a significantly better placement accuracy than the combined partitions—may be attributed to artefacts generated by the simulated data generation process. In addition, the molecular tree shape of S4 is particularly difficult, because it has a large number of relatively short inner branches and long branches leading to the leaves. As such, the simulated fossil

name	UNW	INT	BAD	GOOD
S1_100_100	1.37	0.00	5.02	0.66
S1_100_200	1.82	0.40	5.04	0.66
S1_100_300	2.01	0.44	5.16	0.66
S1_100_400	2.82	1.13	4.82	0.66
S1_100_500	2.93	1.11	5.05	0.66
S2_100_100	2.50	1.38	5.25	0.83
S2_100_200	3.29	1.64	5.71	0.83
S2_100_300	3.95	2.47	5.34	0.83
S2_100_400	3.87	2.57	5.14	0.83
S2_100_500	4.24	3.07	5.62	0.83
S3_100_100	1.36	0.44	7.00	0.95
S3_100_200	2.08	0.82	6.48	0.95
S3_100_300	2.44	0.85	6.66	0.95
S3_100_400	3.01	1.28	6.26	0.95
S3_100_500	3.85	1.46	6.76	0.95
S4_100_100	12.02	14.95	12.26	12.40
S4_100_200	11.25	13.93	11.59	12.40
S4_100_300	11.47	12.58	11.80	12.40
S4_100_400	12.08	12.36	11.49	12.40
S4_100_500	11.22	11.29	12.10	12.40
S5_100_100	2.04	1.32	5.55	0.92
S5_100_200	3.74	2.35	5.37	0.92
S5_100_300	3.92	2.74	5.66	0.92
S5_100_400	4.63	3.61	5.65	0.92
S5_100_500	4.14	3.28	5.54	0.92

TABLE IV
ABSOLUTE AVERAGE NODE-DISTANCE BASED ACCURACY FOR FOSSIL PLACEMENTS ON SIMULATED DATASETS FOR UNWEIGHTED, INTEGER-WEIGHTED SITES AS WELL AS INCONGRUENT AND CONGRUENT DATA PARTITIONS.

that is placed at the innermost branch of the tree will be hard to place accurately, because of the short internal branches. This also explains the better performance of the D4 reference tree on real data, because in this case we use current-day organisms that are mostly attached to long branches to assess placement accuracy. In addition to this, a congruent data partition length of 100 may not be sufficient to compute an accurate placement because the dataset has significantly more organisms than all other tested datasets. For 100 simulated datasets S4_400_400, i.e., with 400 congruent sites, the placement accuracy of the congruent partition increased to 5.30 and that of the integer weighted placement to 8.55. However, a further increase of the congruent sites to a length of 800 did not yield further significant improvements in placement accuracy.

D. Placement Accuracy on Real Datasets

The overall placement accuracy on real datasets was assessed in a different way than on simulated data. While some datasets include fossil data, unlike as for the simulated datasets, we do not know the true phylogenetic position of these fossils. To this end, we decided to base our analysis only on the current-day species for which molecular data is available. We assume that the true position of these taxa is the position in the respective molecular reference topology. In order to thoroughly test placement accuracy, from every real world molecular tree, we removed one organism at a time and then re-inserted it using only the morphological data

name	UNW	BIN	INT	100
D1	1.26	0.93	1.05	1.05
D2	0.99	0.86	0.75	0.78
D3	1.32	1.14	0.75	0.99
D4	3.51	3.02	2.06	2.34
D5	3.13	3.34	2.20	2.43

TABLE V
ABSOLUTE NODE-DISTANCE BASED ACCURACY FOR FOSSIL PLACEMENTS ON REAL-WORLD DATASETS FOR ALTERNATIVE SITE WEIGHTING SCHEMES.

via the placement algorithm. On dataset D3 for instance, we conducted 32 placement runs for each of the 32 species. Once again, we used phylogenetic placement with bootstrapping and extracted the average placement accuracy using the Weighted Root Mean Squared Distance (see Equation 1). We conducted placement runs for 4 different weighting schemes: unweighted (denoted as **UNW**), binary weights (denoted as **BIN**), integer weights (denoted as **INT**), and all weights set to 100 (denoted as **100**). In Table V we provide the absolute accuracy in terms of average placement node distance for all analyzed weighting schemes and all real-world datasets. In Table VI we provide the relative accuracy in terms of average placement branch distance. The data presented clearly show, that the approach using our weight calibration mechanism with integer weights yields the best results in terms of accuracy. An interesting observation is that the approach where a weight of 100 is assigned to every site performs better than the binary weighting scheme. This can be attributed to the application of the Bootstrap procedure and a too strict cutoff of 5% used for generating the binary weight vector. While our morphological calibration mechanism works well, assigning weights of 100 to each site assures that sites that contain congruent signal will with high probability be included in the bootstrap replicates, while this probability is low for binary weights or the unweighted placement that comprise all sites at most once as opposed to 100 times. As the relatively good accuracies obtained for the unweighted case indicate, Maximum Likelihood is able to filter out noise, i.e., incongruent signal, for placing fossils. However, the standard Bootstrap procedure may occasionally not include some congruent sites in the bootstrap replicates, which can bias the stability of the placement results. The probability for not sampling congruent sites is relatively large, because the morphological data partitions have comparably few sites.

Our placement algorithm using calibrated integer weights yields placements that are approximately 25% better in terms of node distance than the unweighted standard approach and a relative average distance improvement (over all datasets) of 25%. Thus, despite the partially highly incongruent phylogenetic signal between morphological and molecular data partitions, we are able to accurately place fossils in well-established reference trees. Even using the unweighted approach one can achieve better than 85% accuracy in the worst case.

name	UNW	BIN	INT	100
D1	3.9%	3.4%	3.2%	3.1%
D2	4.6%	4.5%	3.6%	3.5%
D3	9.6%	7.8%	5.4%	8.3%
D4	11.0%	9.9%	7.6%	8.6%
D5	14.2%	14.2%	12.7%	13.0%

TABLE VI
RELATIVE BRANCH-DISTANCE BASED ACCURACY FOR FOSSIL PLACEMENTS ON REAL-WORLD DATASETS FOR ALTERNATIVE SITE WEIGHTING SCHEMES.

E. Placement Accuracy of Real Fossils: Two Case Studies

We also used morphological data for placing the real fossil taxa that were included in the original biological analyses of dataset D1 [13] and D2 [14]. Those fossil taxa had previously been placed and analyzed using different placement approaches in the aforementioned studies [13], [14]. While a detailed biological analysis of the placements obtained via the approach we present here is outside the scope of this paper, we briefly address placements results using morphological weight calibration and discuss potential interpretations.

Figure 5 depicts the placements of the Juglandaceae fossils. The name labels of the fossil taxa in Figure 5 are preceded by the word QUERY and we have appended the bootstrap support for the insertion branch at the end of the name label. The placements of the individual fossils is partly comparable to the findings in [13].

The Polyptera, Palaeoplatycarya, and Platycarya fossils are in a clade (subtree) with Carya and Juglans which corresponds to empirical biological expectations, but are not located at the root of the subtree containing all Juglans and Carya. Also the Cruciptera fossil is placed with the Juglans, rather than being located at the root of the subtree containing all Juglans. The largest difference to the study presented by Manos et al. is that the Paleooremunnea fossil is placed at the root of the subtree containing the Oreomunnea in your tree rather than being located at the root of the subtree comprising all Juglans and Carya. However, the placement of the Paleooremunnea fossil is know to be problematic (see [13], p. 425). While in [13] its phylogenetic placement varies significantly, depending on the method used, we obtain an assignment with 100% Bootstrap support for this fossil. Overall, the fossil placements are biologically reasonable and could give rise to new biological hypotheses (D. Soltis, personal communication).

Figure 6 shows the placement of the Marsupial fossils from [14]. The placement of the Djarthia fossil is particularly interesting, as it seems to confirm the original placement as a member of Australidelphia, but outside the subtree comprising extant Australasian marsupials (see [14] Figure 3A-B). Note that, in contrast to the original studies [13], [14] which include the fossil sequences into a de novo tree inference, we placed them individually into a previously generated molecular tree using morphological data alone. This one-by-one placement of the fossil sequences generates the multi-furcating (non-binary) trees, in which more than one fossil taxon can be placed into the same branch of the molecular reference tree.

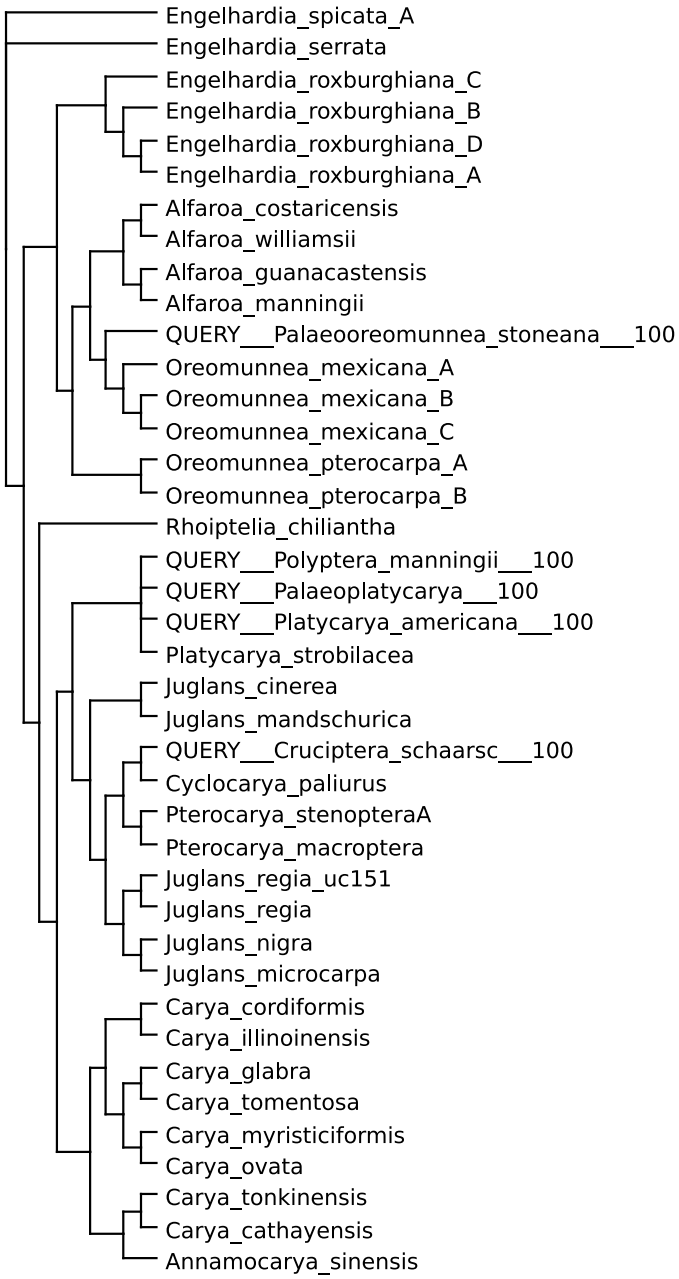

Fig. 5. Fossil placement in dataset D1. The number behind the query sequence names denotes the bootstrap support of the placement.

VII. CONCLUSION

We have conducted the first assessment of fossil placement accuracy using morphological data under the Maximum Likelihood criterion on simulated and real-world datasets based on a well-established reference tree. In addition, we have developed a statistical weight calibration mechanism that is able to identify morphological sites, that exhibit a phylogenetic signal which is congruent to that of the reference tree. By using according calibrated integer weights we can improve upon the

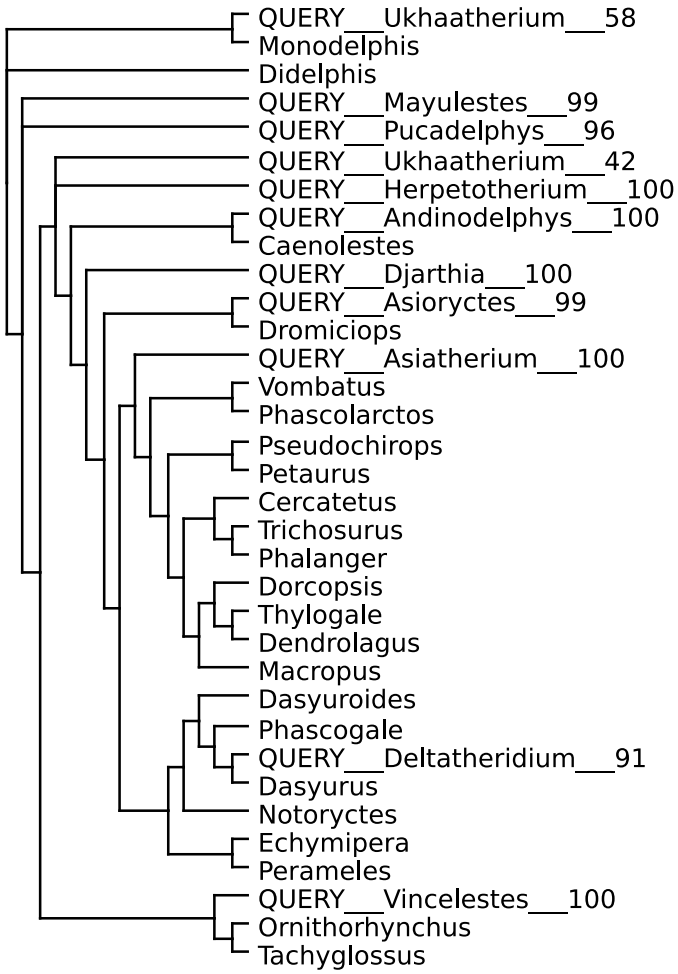

Fig. 6. Fossil placement in dataset D2. Placements with BS support $< 10\%$ have been removed.

absolute and relative placement accuracy by 20% on simulated datasets and by 25% on real-world datasets.

Moreover, we find, that despite the partially high incongruence between ML trees from the morphological and molecular data partitions, the achieved accuracy under Maximum Likelihood is sufficient for reliably placing fossils. Two biological case studies with real fossil taxa reveal that we can obtain reasonable biological results using the weight calibration and fossil placement algorithms.

The statistical weight calibration procedure as well as the phylogenetic placement algorithm have already been integrated into RAxML which is a freely available and widely used tool for phylogenetic inference.

Future work will cover a more detailed analysis and a potential refinement of the statistical weight calibration procedure to. A method for computing incongruence among sites may also be valuable for analyses of broad phylogenomic molecular datasets.

VIII. Acknowledgements

We would like to thank J.J. Wiens for making available the simulated data generation code to us and for valuable comments on an earlier version of this paper. We would also like to thank Pam and Doug Soltis, Guido Grimm and Steven Manchester for providing help on the interpretation of the real fossil placement results.

References

[1] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *J. Mol. Evol.*, vol. 17, pp. 368–376, 1981.

[2] P. Lewis, "A likelihood approach to estimating phylogeny from discrete morphological character data," *Systematic Biology*, vol. 50, no. 6, pp. 913–925, 2001.

[3] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006.

[4] M. Sanderson, M. Donoghue, W. Piel, and T. Eriksson, "TreeBASE: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life," *American Journal of Botany*, vol. 81, no. 6, p. 183, 1994.

[5] C. Dunn, A. Hejnol, D. Matus, K. Pang, W. Browne, S. Smith, E. Seaver, G. Rouse, M. Obst, G. Edgecombe, M. Sorensen, S. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. Kristensen, W. Wheeler, M. Martindale, and G. Giribet, "Broad phylogenomic sampling improves resolution of the animal tree of life," *Nature*, vol. 452, no. 7188, pp. 745–749, 2008.

[6] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics*, vol. 21, no. 1, pp. 97–106, 2005.

[7] J. Wiens, "Character analysis in morphological phylogenetics: problems and solutions," *Systematic Biology*, vol. 50, no. 5, pp. 689–699, 2001.

[8] K. Thiele, "The Holy Grail of the Perfect Character: The Cladistic Treatment of Morphometric Data," *Cladistics*, vol. 9, pp. 275–304, 1993.

[9] W. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, no. 3760, pp. 279–284, 1967.

[10] J. Farris, "Phenetics in camouflage," *Cladistics*, vol. 6, no. 1, pp. 91–100, 1990.

[11] J. Wiens, "Paleontology, Genomics, and Combined-Data Phylogenetics: Can Molecular Data Improve Phylogeny Estimation for Fossil Taxa?" *Systematic Biology*, vol. 58, no. 1, p. 87, 2009.

[12] J. Felsenstein, "Confidence Limits on Phylogenies: An Approach Using the Bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[13] P. S. Manos, P. S. Soltis, D. E. Soltis, S. R. Manchester, S.-H. Oh, C. D. Bell, D. L. Dilcher, and D. E. Stone, "Phylogeny of extant and fossil juglandaceae inferred from the integration of molecular and morphological data sets," *Systematic Biology*, vol. 56, no. 3, pp. 412–430, 2007. [Online]. Available: http://dx.doi.org/10.1080/10635150701408523

[14] R. M. Beck, H. Godthelp, V. Weisbecker, M. Archer, and S. J. Hand, "Australia's oldest marsupial fossils and their biogeographical implications," *PLoS ONE*, vol. 3, no. 3, pp. e1858+, March 2008. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0001858

[15] J. Wiens, R. Bonett, and P. Chippindale, "Ontogeny discombobulates phylogeny: Paedomorphosis and higher-level salamander relationships," *Systematic Biology*, vol. 54, no. 1, pp. 91–110, February 2005. [Online]. Available: http://dx.doi.org/10.1080/10635150590906037

[16] J. Wiens, J. Fetzner, C. Parkinson, and T. Reeder, "Hylid frog phylogeny and sampling strategies for speciose clades," *Systematic Biology*, vol. 54, no. 5, pp. 719–748, October 2005. [Online]. Available: http://dx.doi.org/10.1080/10635150500234625

[17] S. Struckmann, M. J. Arauzo-Bravo, H. R. Schoeler, R. A. Reinbold, and G. Fuellen, "Rexspecies - a tool for the analysis of the evolution of generegulation across species," *BMC Evolutionary Biology*, vol. 8, pp. 111+, April 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2148-8-111

[18] D. Robinson and L. Foulds, "Comparison of phylogenetic trees," *Math. Biosci*, vol. 53, no. 1-2, pp. 131–147, 1981.

[19] Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites," *J. Mol. Evol.*, vol. 39, pp. 306–314, 1994.