Morphology-based phylogenetic binning of the lichen genera Allographa and Graphis via molecular site weight calibration

Simon A Berger [1], Alexandros Stamatakis*[1] and Robert Lücking[2]

[1]The Exelixis Lab, Dept. of Computer Science (I12), Technische Universität München, Boltzmannstr. 3, D-85748 Garching b. München, Germany

[2]Dept. of Botany, The Field Museum, 1400 South Lake Shore Drive, Chicago, Illinois 60605-2496, USA

Email: Simon A Berger - bergers@in.tum.de; Alexandros Stamatakis*- stamatak@in.tum.de; Robert Lücking - rlucking@fieldmuseum.org;

*Corresponding author

Abstract

Background: Despite the potential shortcomings of using phenotype ("morphological") data for phylogenetic inference, there exist scenarios where only morphological data is available for systematic classification (e.g., phylogenetic placement of fossil records, analysis of large taxonomic groups for which DNA data are only available for a small number of species). Because of the frequently incongruent phylogenetic signal between morphological and molecular data partitions, we need to devise computational methods to determine morphological site patterns that are congruent with the molecular tree (which we assume to represent the "true" tree relative to any tree inferred from morphological data), to improve the accuracy of the phylogenetic classification/placement of taxa for which only morphological data exist.

Results: We developed methods for determining morphological characters that are congruent with the molecular tree (site weight calibration) and for conducting phylogenetic binning (assignment of morphological taxa to branches of the molecular reference tree) and implemented those methods in the widely used program RAxML. We applied our methods to a real world case, the taxonomy of the lichen genera *Allographa* and *Graphis*, and show that these methods can improve the assignment accuracy of morphologically defined taxa to the two genera. We also tested our methods systematically on five additional datasets that contained both morphological and molecular data.

Conclusions: We demonstrate that site weight calibration can be used to improve the systematic assignment/binning accuracy of taxa for which only morphological data exist.

Availability: The site weight calibration and binning methods are implemented in the RAxML (v. 7.2.7) open-source code available at: http://wwwkramer.in.tum.de/exelixis/software.html.

Background

Since the advent of PCR and automated sequencing, molecular approaches have largely replaced non-molecular methods for phylogeny reconstruction. Molecular data have several advantages over morphological data:

1. The number of characters is much larger, typically around 300 parsimony-informative sites for a single-gene analysis, and increasingly over 1,000 in multi-gene approaches, whereas morphological data sets rarely contain more than 100-200 characters, especially in groups poor in phenotypic features such as fungi, including the lichens. In a maximum likelihood (ML [1]) context, molecular data typically comprises 500-1,000 distinct site patterns for a single gene and significantly more than 10,000 in current phylogenomic analyses [2]. Simulation studies [3, 4] have shown that tree reconstruction accuracy increases significantly with the number of site patterns/parsimony-informative sites.

2. Molecular data have an intrinsic discrete code, whereas morphological data have to be defined as separate characters and coded as character states, which leaves room for subjectivity and coding errors due to character misinterpretation.

3. The amount of homoplasy is significantly smaller in molecular data depending on the gene, usually less than 5-10% of parsimony-informative sites are homoplastic, whereas in morphological data sets, quite frequently 50% or more of the characters exhibit homoplasy.

As a consequence, phylogenies inferred from molecular data versus those inferred from morphological data can be highly incongruent [5], and phylogenies inferred from morphological data often lack resolution and support. Thus, simply concatenating morphological and molecular data partitions to conduct a joint (also

referred to as total evidence approach) phylogenetic inference may lead to biased topologies, if the number of homoplastic morphological characters is high, compared to the number of molecular characters. Despite the potential shortcomings of morphological data and the increasing availability as well as decreasing cost of molecular data for phylogenetic inference, there exist several application scenarios where input from morphological data is indispensable. One such scenario is the reconstruction of phylogenies involving fossils (for which molecular data are not available), or the placement of fossils into given molecular reference trees [5].

Another application scenario consists of systematic classification—denoted as 'phylogenetic binning' throughout this paper—of large taxonomic groups for which molecular data are only available for a small number of species. Consider a taxon of 500 species for which 50 species have been sequenced. Assume that the result of a phylogenetic analysis of these 50 species indicates that there exist three separate genera. Logistically, sequencing the remaining 450 species to disentangle their generic status may represent a nearly impossible task, in particular if rare species or species only known from type material are involved. One would then face the problem that 50 species can be assigned to definitive lineages, whereas the status of the 450 species would remain unresolved. Alternatively, the morphological features of these species can be used to bin (assign) them via appropriate algorithms to the corresponding genus, or to additional branches of the molecular reference tree that are located between the genera, which we term 'binning no man's land', thus indicating that potential additional lineages may exist which have not yet been sequenced. The work-flow of such a basic phylogenetic binning procedure is outlined in Figure 1. The binning procedure assigns taxa for which no molecular data are available to one of the three genera in this example, without extending the molecular reference tree to a fully bifurcating tree containing 500 taxa. One may also consider an example (as is the case for the two lichen genera we studied here) of a reference tree with two genera and an outgroup (see Figure 2). In this scenario we intended to determine how many morphologically defined taxa are binned to the left and to the right of the outgroup and also, which taxa are binned into the branch leading to the outgroup, which would represent the 'binning no man's land' in this scenario.

Based on the prolegomena, molecular and morphological data can give rise to conflicting phylogenetic signals, and because of the aforementioned properties, particularly homoplasy, phylogenies based on molecular data are usually to be trusted more than those based on morphological data. Therefore, methods that help to extract and deploy the morphological signal that is congruent to the molecular phylogenetic signal can help to improve accuracy in combined molecular and morphological data analyses such as, for

instance, phylogenetic binning.

One such method for improving accuracy consists of calibrating weights for morphological characters based upon their degree of congruence with a molecular reference topology. In other words, morphological characters that are congruent with the reference tree will receive a high weight and characters that are incongruent will receive a low weight. Evidently, the weight calibration requires a dataset or a subset of taxa for which molecular *and* morphological data are available. Using the calibrated weights, additional taxa (e.g., fossil taxa), for which only morphological characters are known can then be placed into the molecular reference tree (assigned to branches of the reference tree) using maximum likelihood [5]. Fossil placement using such a weight vector can increase placement accuracy in the tree by up to 25% on real world datasets [5].

In this paper, we investigated a different application of weight calibration, which allows for phylogenetic binning of a large number of taxa for which only morphological data are available, based on a small subset with known molecular *and* morphological data. We also extended our weight calibration algorithm [5] by a parsimony calibration method. By example of the lichen genera (bins) *Allographa* and *Graphis*, we show that weight calibration can be deployed to obtain a biologically reasonable and highly supported binning of species into two genetically distant, but morphologically similar genera. Thus, weight calibration in combination with binning can be used to assign morphological taxa to distinct genera more reliably. Thereby, one can formulate hypotheses of their systematic placement based upon an objective criterion which can then be used for targeted hypothesis testing by sequencing selected species, rather than using a shotgun approach for sequencing whatever is available.

Methods

Datasets

For this study, we used the lichen genus *Graphis* (Ascomycota: Lecanoromycetes: Ostropales: Graphidaceae) as an example, which has recently been shown to comprise two separate, distantly related lineages [6] which, however, are morphologically similar. The alignment entailed a molecular data partition of mitochondrial small subunit (mtSSU) and nuclear large subunit (nuLSU) rDNA for 16 ingroup species, and a morphological data partition with 48 characters/traits for a total of 313 ingroup species. The morphological data were used in [7] for a multivariate analysis study and included thallus and fruiting body morphology and anatomy as well as secondary chemistry (Table 1). As outgroup, we used *Fissurina marginata*, which has a basal position in *Graphidaceae*. The data were arranged as follows:

1. Molecular data set of 16 ingroup taxa plus 1 outgroup taxon.

2. Morphological data set of (the same) 16 ingroup taxa plus 1 outgroup taxon.

3. Morphological data set of the remaining 297 ingroup taxa for which no molecular data are yet available.

The PHYLIP files of the concatenated molecular dataset of 17 taxa as well as the morphological data matrix of all 314 taxa are available for download at
http://wwwkramer.in.tum.de/exelixis/phylogeneticBinning.tar.bz2.
In addition to this dataset, we also used the five concatenated molecular/morphological datasets from [5] to conduct additional systematic tests. For convenience we denote these datasets as D1 through D5. Dataset D1 [8] contains 35 taxa of walnut trees (*Juglandaceae*). D2 [9] comprises 23 Marsupial sequences. D3 [10] contains 32 taxa of Amphibians (*Caudates*). D4 [11] contains 81 taxa of tree-frogs (*Hylidae*). Finally, D5 [12] contains 18 taxa that span a wider variety of species than the other datasets, ranging from the chicken to the homo sapiens.

Algorithms

To carry out phylogenetic binning, we used and combined two algorithms that have recently been implemented in the RAxML [13] open source code for phylogenetic inference under Maximum Likelihood (v. 7.2.7, available at http://wwwkramer.in.tum.de/exelixis/software.html).

Weight Calibration

The RAxML weight calibration algorithm can be used to infer weights for morphological sites according to their degree of congruence with the molecular reference tree. Essentially, congruent sites are up-weighted, while incongruent sites are down-weighted. Previous computational experiments showed that deploying calibrated weights can improve the accuracy of fossil placement on real datasets by 25% on average [5]. Weight calibration is only conducted on those taxa of the data matrix for which molecular *and* morphological data are available.
The algorithm for inferring ML-based weights by using the molecular reference tree and the morphological data partition for the taxa in the reference tree works as follows: Initially, ML model parameters are optimized on the fixed reference tree and the per-site log likelihood scores are computed and stored. Thereafter, the algorithm generates a certain number of random trees (100 replicates have proved to be

sufficient [5]) and re-computes the per site log likelihood scores on each random tree. Integer weights are then obtained by counting the number of times each site yields a worse log likelihood score on a randomized tree than on the reference tree. In other words, sites that exhibit a signal that is highly congruent to the molecular tree receive a high weight and sites that are highly incongruent receive a low weight.
For the present analysis, we also implemented an option to infer weights under parsimony in RAxML. Initially, we read in the molecular reference tree. Then, RAxML computes the parsimony score for each site. We then compute intermediate weights for each site by subtracting the actual parsimony score from the optimal possible score based upon the composition of the site. This yields high weights for incongruent sites and low weights for congruent sites. Finally, we reverse the weight vector and normalize the weights to a range between 0 and 100 by using the maximum weight in the intermediate weight vector.

Phylogenetic Binning

Given the likelihood- or parsimony-based weight vectors, the Evolutionary Placement Algorithm (EPA) implemented in RAxML [14] can be used to bin the morphological taxa into the genera (or any given lineage) of the molecular reference tree.
The EPA was initially developed for placing short reads as obtained, for instance, from 454 pyrosequencing runs into a given reference tree based on full length (e.g., 16S) sequences. Instead of computing a fully bifurcating tree topology comprising all short reads as well as the full length reference sequences, the algorithm determines the optimal ML-based insertion position (insertion branch) for each read individually, that is, it assigns each read to a branch of the reference tree topology (independently of all other reads). The EPA algorithm can also assign an individual read to an area (several branches) of the reference tree via a standard phylogenetic bootstrap procedure [15] or via likelihood weights [16, 17]. That is, the EPA provides a means to infer placement uncertainty. For details please refer to [14]. Thus, the original application scenario for the EPA is to place short reads into a given reference tree and derive, for instance, the microbial diversity of a sample by means of the distribution of reads in the tree. However, as we show here, the algorithm can also be used to bin different morphologically defined taxa (species) into the lineages (genera) of a molecular reference tree. This represents a more coarse-grain use of the EPA, because we are only interested to infer—via phylogenetic placement—to which genus (set of insertion branches) a taxon belongs. In addition, it is straight-forward to use a likelihood- or parsimony-based integer weight vector with the EPA, since using a site-weight vector is a standard RAxML option (`-a` option).

Analysis Pipeline

The overall analysis procedure consists of the following four steps:

1. Infer a reference tree topology (e.g., best-known ML tree) using the molecular data partition only.

2. Calibrate site weights (under ML or parsimony) using the reference tree and only the morphological data of the taxa for which there also exists molecular data.

3. Invoke the EPA method (optionally with BS) using the previously computed weight vector, the morphological data of all taxa, and the reference tree as input. All taxa for which only morphological data is available, that is, all taxa not contained in the molecular reference tree, will be assigned to branches of the reference tree.

4. Execute a post-analysis script that parses the EPA output files to determine the phylogenetic assignments/bins (genera or 'no man's land') of the morphological taxa.

Binning Tool

To facilitate the usage of the methods that are presented here, we have designed a flexible binning tool in JAVA. The tool reads in a reference tree and the results of the phylogenetic placement file as obtained from the EPA algorithm in RAxML. It also allows the user to specify an arbitrary number of phylogenetic bins (clades/subtrees, see Figure 1). This can be done by listing those taxa of the reference tree that form a bin in a plain text input file. In other words, the user needs to provide several taxon lists that form bins, corresponding to monophyletic lineages of the reference tree. The user can also chose if the branch to which a bin is attached shall form part of the respective bins or not. All placements into branches that do not form part of a bin are assigned to a separate 'no man's land' bin. Alternatively, as carried out for the lichen study, the user may simply only specify the outgroup name. Our JAVA tool will then automatically divide the tree into three bins as shown in Figure 2.

Experimental setup

Real data analysis

The real data set of the lichen genera was analyzed as follows: We initially reconstructed a best-known ML tree (remember that finding the optimal ML tree is an NP-hard problem [18], that is, the number of possible trees for n taxa is so large that it is not feasible to find *the* ML tree) using RAxML for the molecular data. Then, we computed parsimony and ML (using 100 random trees) weight vectors as

described above. Thereafter, we executed three EPA runs with 100 BS replicates each for the parsimony-weighted, ML-weighted, and unweighted case.

To further analyze our data, we conducted a leave-one out experiment on the 16 taxa (excluding the outgroup) for which both morphological *and* molecular data were available. In this experiment, we pruned one of the 16 ingroup taxa at a time from the reference tree and placed it back into the tree via the Evolutionary Placement Algorithm (EPA) using *exclusively* the morphological data partition. We conducted EPA runs with and without a weight vector to show that using a weight vector can improve binning accuracy. The results are summarized in Figure 3.

Systematic analysis

We also conducted a more systematic analysis of placement accuracy using the aforementioned five real-world datasets D1 through D5 [5]. Here we also assessed binning accuracy, assuming two bins (lineages), by means of leave-one-out experiments, that is, we once again pruned one taxon at a time from the molecular reference tree and then re-inserted it using the EPA on morphological data. To define lineages in the reference trees of D1-D5, we considered, all splits (bipartitions) of the reference trees with BS support higher than 75%. Thus, for all well-supported splits, we assessed how frequently every taxon on one side (the good/correct bin) of a split ended up on the other side of that split (in the "wrong" bin) when placed back into the tree using the EPA, the corresponding weight vector, and the morphological data only. We applied this test to all well-supported branches in the reference trees. This allowed for testing how many taxa would end up on the wrong side (in the wrong bin) of a well supported branch using our binning approach. An outline of this test procedure is provided in Figure 4. Note that we used the EPA with the BS option, that is, for each taxon we counted the BS support of insertions into the correct bin, the incorrect bin, or the 'no man's land' bin (the highly supported branch across which we assess binning accuracy, see Figure 4).

Results and Discussion

Results

Real data analysis

Likelihood-based site weight calibration & binning: Based on the unweighted morphological data, 252 of the 297 taxa for which no molecular data were available could be assigned to either *Graphis s.str.* or *Allographa* with strong BS support (90% or higher) and another 20 with good BS support (70% or higher; Table 3). For 23 species, support for either clade was low (less than 70%, and two further species, *Graphis saxiseda*

and *G. evirescens*, did not cluster with neither *Graphis* nor *Allographa* but with the outgroup *Fissurina* ('no man's land'). Weighting of the morphological characters based on the 16-taxon set and using maximum likelihood weight calibration changed the proportions as follows: 281 taxa could be assigned to either *Graphis s.str.* or *Allographa* with strong support and another seven species with good support, whereas seven species remained unresolved (less than 70% bootstrap support for either of the clades) and the two species *Graphis saxiseda* and *G. evirescens* received strong support for clustering with the outgroup (Table 2). In total, no change in the genus binning or support between unweighted and weighted morphological data was observed for 229 out of 297 taxa, whereas support and/or binning improved for 58 further species when weighting the morphological characters based on maximum likelihood weight calibration. In two cases, support for the corresponding clade/bin decreased from strong to good (*G. elegans* within *Graphis*) and strong to low (*G. olivacea* within *Allographa*), respectively, whereas in a further two cases, support changed from good to low (*G. inspersolongula* and *G. uruguayensis* within *Graphis*).

Seven out of 297 species showed topological conflict (assigned to different bins) between unweighted and weighted data: *G. novopalmicola* changed from *Allographa* (low support) to *Graphis* (strong support) and *G. diplocheila* from *Allographa* (low support) to *Graphis* (strong support), whereas *G. nigrocarpa*, *G. subimmersa*, and *G. subtracta* changed from *Graphis* (low support) to *Allographa* (good support). *Graphis daintriensis* changed from *Graphis* (good support) to *Allographa* (strong support) and *G. rimulosa* from *Graphis* (strong support) to *Allographa* (good support). All conflicting taxa exhibit intermediate morphologies between *Graphis* and *Allographa* and their placement based on weighted (as compared to unweighted) morphological characters is reasonable except for one species, *G. rimulosa*, which is expected to belong in *Graphis s.str.* which is reflected by the strongly supported placement into *Graphis* in the unweighted analysis. One of the two species that clustered with the outgroup, *G. saxiseda*, has been shown to belong in a different genus, that of *Carbacanthographis*, which is not related to neither *Graphis s.str.* nor *Allographa* [19], and hence its outgroup placement makes sense.

The analysis also helped to unambiguously assign two groups of *Graphis s.lat.* with uncertain systematic affinities, the *G. dussii* and the *G. subserpentina* group [7], with confidence to either *Graphis s.str.* or *Allographa*. Of the eleven species of the *G. dussii* group, nine were strongly supported within *Allographa* and one (*G. enteroleuca*) within *Graphis*, whereas a further species, *G. regularis*, remained unresolved (Table 3). None of these species has been sequenced so far. Of the 22 species of the *G. subserpentina* group, one has been sequenced and confirmed in *Graphis*, and an additional 17 taxa showed good to strong support within *Graphis*. However, four species, *G. cycasicola*, *G. nigrocarpa*, *G. subhiascens*, and

G. superans, were supported within *Allographa* (Table 3).

Parsimony-based site weight calibration & binning: For this analysis, we inferred morphological site weights using parsimony as describe above, and once again assigned taxa to bins (genera) using the likelihood-based EPA algorithm with parsimony-based weights. Compared to weighting based on ML, MP weighting showed no change in binning or support for 266 out of 297 taxa. For 19 taxa, binning and/or support improved whereas for 11 taxa, either support decreased or the binning result appeared to be less reasonable. One taxon showed significant topological conflict. Thus, MP weighting yielded slightly better overall results than ML weighting, suggesting that both methods should be used to detect consensus and conflict between binning approaches.

Comparative result visualization: In Figure 5 we summarize the results of the binning process for all 297 species using the three tested methods (unweighted binning, likelihood-based site weight calibration binning, parsimony-based site weight calibration binning). The 297 species are located along the y-axis of the heatmap-like graph. We sorted them according to the difference between the BS support values for a binning into *Graphis* and *Allographa* using the BS supports as obtained from the unweighted binning process. The species that are assigned to the outgroup are located at the bottom of the graph and sorted according to increasing BS support for falling into the outgroup bin.
The graph indicates that site-weight calibration based placement yields a clearer—in many cases unique—assignment of species to phylogenetic bins and can thus be deployed to resolve ambiguous binnings.

Detecting mis-labelled taxa: After conducting the binning analyses on the lichen genera, we realized that sequences of *Graphis cleistoblephara* were available in GenBank. This species forms part of the *G. dussii* group [7]. Our molecular data partition did not contain sequences for this group, but according to the phylogenetic binning analysis it should fall into the *Allographa* lineage. Phylogenetic analysis, however, showed that the molecular sequence for *Graphis cleistoblephara* was strongly supported within the *Graphis* lineage. This contradicts our binning approach. However, a re-examination of the specimen that was used for sequencing, revealed that it did not represent *G. cleistoblephara* in the *G. dussii* group but a superficially similar, yet unrelated species, *G. hiascens* in the *G. subserpentina* group. This revised identification thus confirmed that the binning result obtained from our study is correct, since it places the *G. subserpentina* group within *Graphis*. Apart from demonstrating the importance of correct taxonomy of sequenced vouchers, this suggests that phylogenetic binning can potentially also be used for identifying

mis-labelled GenBank sequences.

Control experiments

We conducted two types of control experiments.

Lichen control experiment: In the first experiment we executed the likelihood-based binning approach on the 16 lichen ingroup taxa for which both morphological and molecular data was available. Figure 3 shows that binning accuracy improved substantially for all ingroup taxa t1 through t16 —except for taxon 13 that was consistently placed into the wrong bin— when calibrated morphological site weights are used. In the weighted case, taxa t8, t11, t12, and t15 were unambiguously binned with 100% BS support into the correct bin.

Taxon 13 represents *Graphis japonica* in the *G. subserpentina* group and is morphologically intermediate between *Allographa* and *Graphis*. This shows that sequencing a single taxon within a larger group and then using morphological binning can improve systematic classifications. With the morphological data alone, species of the *G. subserpentina* group would cluster with *Allographa*, as shown in a previous multivariate analysis [7]. Therefore, taxon 13 bins incorrectly with *Allographa*. However, genetically the taxon forms part of the *Graphis* lineage, and including this taxon in the reference tree will then bin other species of the *G. subserpentina* group correctly within *Graphis*.

Systematic control experiment: The results in terms of accumulated BS support for correctly binned taxa in datasets D1 through D5 are provided in Table 2. For each dataset we provide the BS support for correct binnings for unweighted and likelihood-weighted binning runs with the Evolutionary Placement Algorithm (EPA). In contrast to the control experiment on the lichen dataset, the improvements achieved by using morphological site weight calibration were marginal. However, the table demonstrates that (i) using weight calibration did not have a negative impact on binning accuracy, and (ii) that binning using the EPA, be it weighted or unweighted was highly accurate, exceeding 95% BS support on average.

Discussion

The method of assessing systematic affinities of species by means of weighting morphological characters based on maximum likelihood and parsimony calibrations and phylogenetic binning offers new possibilities in taxonomic, systematic, phylogenetic, and evolutionary research. While morphological data cannot replace molecular data for inferring phylogenies and relationships between taxa, the method offers a quantitative and objective assessment of morphological data sets including a measure of confidence of

systematic placement, in contrast to ad hoc decisions that are based upon morphological characters alone. This method is especially useful when dealing with phylogenies that involve large groups of taxa for which only a small part have been sequenced or molecular data are not available due to methodological constraints, such as fossils. The methods presented here are readily available in the widely used open source tool RAxML (v. 7.2.7) and profit from the highly optimized likelihood function implementations for binary as well as multi-state morphological characters. Given a molecular reference tree of taxa for which morphological data have also been scored, phylogenetic binning can easily be computed by invoking RAxML once for site weight calibration using those taxa that form part of the reference tree, and once for binning additional morphologically defined taxa into the reference tree topology.

In the present case, the method helped to assess the taxonomic status of species hitherto placed in the collective lichen genus *Graphis*. This genus recently received a revised classification including species of *Graphidaceae* with a set of morphological characters: fruiting bodies with conspicuous margins and carbonized excipulum, unornamented paraphyses, ascospores with interlocular plates that stain violet-blue in Lugol's solution, and a rather simple chemistry [20]. According to this revised definition, the genus included over 300 accepted species [7, 19]. However, molecular phylogenetic analysis suggested that *Graphis s.lat.* represented two separate, only distantly related lineages, *Graphis s.str.* and *Allographa* (Rivas Plata et al. 2010 [6]). Since only 16 species of *Graphis s.lat.* had been sequenced for that analysis, it required to classify the remaining nearly 300 species based on morphological characters alone. This represents a challenge since the morphological differences between the two genera are fuzzy. The alternative of re-classifying only the 16 species (for which molecular data are available) and leaving the remaining nearly 300 species unclassified until sequence data becomes available does not appear to be acceptable for practical reasons. Specifically, two species groups, the *G. dussii* and the *G. subserpentina* group, appear as being intermediate between *Graphis s.str.* and *Allographa* (Rivas Plata et al. 2010 [6]), and without sequencing a larger number of species, their placement in either genus would have been provisional without the present study. With the results at hand, both groups can be confidently placed within either *Graphis s.str.* (*G. subserpentina* group) or *Allographa* (*G. dussii* group).

In addition to the general advantages of this method, it also helped to identify taxa of particular interest or doubtful taxonomic status. An unexpected result was the clustering of *Graphis saxiseda* with the outgroup *Fissurina* ('no man's land'). This species has already been recombined in *Carbacanthographis*, a genus unrelated to both *Graphis s.str.* and *Allographa* [19], but was retained in the morphological data set by accident. The analytical method identified it as not belonging neither to *Graphis s.str.* nor to *Allographa*.

In addition, several taxa that received low support for either clade in this analysis or showed conflicting topology for unweighted and weighted characters can now be restudied or sequenced in a targeted way to clarify their systematic positions. Thus, our method provides a quantitative and objective approach to formulate hypotheses about phylogenetic relationships, allowing to sequence selected taxa in a targeted manner.

Conclusions

We presented a novel method for binning/assigning morphological taxa to reference lineages by means of the EPA algorithm that was originally designed to classify short pyrosequencing reads and by means of a morphological site weight calibration method that was initially devised to improve fossil placement accuracy. Despite impressive advances in molecular sequencing technologies, we show that there exist cases where only morphological data is available for extant species and that tools are required for analyzing such datasets.

By example of a lichen dataset, we demonstrate the increased ability of our approach to correctly bin morphological taxa into the correct lineage. This observation is supported by means of additional experiments in five other real world datasets that contain morphological and molecular data partitions as well as by leave-one-out experiments on those lichen taxa for which morphological and molecular data are available.

The methods used here are freely available for download at
http://wwwkramer.in.tum.de/exelixis/software.html.

Authors contributions

SAB and AS implemented the weight calibration methods in RAxML, wrote the additional scripts and JAVA code for the analysis, and executed the inferences. RL provided the data and suggested the application of weight calibration to the problem. RL and AS wrote the manuscript.

Acknowledgements

The morphological dataset used for this study was compiled within the framework of three grants provided by the United States National Science Foundation (NSF) to The Field Museum: "Phylogeny and Taxonomy of Ostropalean Fungi" (DEB 0516116; PI Lumbsch, Co-PI Lücking); "TICOLICHEN" (DEB 0206125; PI Lücking); and "Neotropical Epiphytic Microlichens" (DEB 0715660; PI Lücking). AS and SAB

are funded under the auspices of the Emmy-Noether program by the German Science Foundation (DFG).

We would like to thank Ziheng Yang for suggesting to also use parsimony for site weight calibration.

References

1. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach**. *J. Mol. Evol.* 1981, **17**:368–376.
2. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse G, Edgecombe G, Martinez P, Baguna J, Bailly X, Jondelius U, Wiens M, Müller W, Seaver E, Wheeler W, Martindale M, Giribet G, Dunn C: **Rooting the Bilaterian Tree with Scalable Phylogenomic and Supercomputing Tools**. *Proc. R. Soc. B* 2009, **276**:4261–4270.
3. Moret BME, Roshan U, Warnow T: **Sequence-Length Requirements for Phylogenetic Methods**. In *WABI 2002* 2002:343–356.
4. Bininda-Emonds ORP, Brady SG, Sanderson MJ, Kim J: **Scaling of accuracy in extremely large phylogenetic trees.** In *Pacific Symposium on Biocomputing* 2000:547–558.
5. Berger SA, Stamatakis A: **Accuracy of Morphology-based Phylogenetic Fossil Placement under Maximum Likelihood**. In *Proceedings of 8th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-10)* 2010. [Accepted for publication].
6. Rivas Plata E, Hernandez J, Lücking R, Staiger B, Kalb K: **Molecular data confirms that Graphis is two genera, and both include Hemithecium**. *Taxon* 2010. in press.
7. Lücking R: **The taxonomy of the genus Graphis sensu Staiger (Ascomycota: Ostropales: Graphidaceae)**. *The Lichenologist* 2009, **41**(04):319–362.
8. Manos PS, Soltis PS, Soltis DE, Manchester SR, Oh SH, Bell CD, Dilcher DL, Stone DE: **Phylogeny of Extant and Fossil Juglandaceae Inferred from the Integration of Molecular and Morphological Data Sets**. *Systematic Biology* 2007, **56**(3):412–430, [http://dx.doi.org/10.1080/10635150701408523].
9. Beck RM, Godthelp H, Weisbecker V, Archer M, Hand SJ: **Australia's Oldest Marsupial Fossils and their Biogeographical Implications**. *PLoS ONE* 2008, **3**(3):e1858+, [http://dx.doi.org/10.1371/journal.pone.0001858].
10. Wiens J, Bonett R, Chippindale P: **Ontogeny Discombobulates Phylogeny: Paedomorphosis and Higher-Level Salamander Relationships**. *Systematic Biology* 2005, **54**:91–110, [http://dx.doi.org/10.1080/10635150590906037].
11. Wiens J, Fetzner J, Parkinson C, Reeder T: **Hylid Frog Phylogeny and Sampling Strategies for Speciose Clades**. *Systematic Biology* 2005, **54**(5):719–748, [http://dx.doi.org/10.1080/10635150500234625].
12. Struckmann S, Arauzo-Bravo MJ, Schoeler HR, Reinbold RA, Fuellen G: **ReXSpecies - a tool for the analysis of the evolution of generegulation across species**. *BMC Evolutionary Biology* 2008, **8**:111+, [http://dx.doi.org/10.1186/1471-2148-8-111].
13. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688–2690.
14. Berger SA, Komornik Z, Stamatakis A: **Evolutionary Placement of Short Sequence Reads on Multi-Core Architectures**. In *Proceedings of 8th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-10)* 2010. [Accepted for publication].
15. Felsenstein J: **Confidence Limits on Phylogenies: An Approach Using the Bootstrap**. *Evolution* 1985, **39**(4):783–791.
16. Strimmer K, Rambaut A: **Inferring confidence sets of possibly misspecified gene trees**. *Proceedings of the Royal Society B: Biological Sciences* 2002, **269**(1487):137–142.
17. Von Mering C, Hugenholtz P, Raes J, Tringe S, Doerks T, Jensen L, Ward N, Bork P: **Quantitative phylogenetic assessment of microbial communities in diverse environments**. *Science* 2007, **315**(5815):1126–1130.
18. Roch S: **A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood Is Hard**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2006, :92–94.

19. Lücking R, Archer A, Aptroot A: **A world-wide key to the genus Graphis (Ostropales: Graphidaceae)**. *The Lichenologist* 2009, **41**(04):363–452.

20. Staiger B: **Die Flechtenfamilie Graphidaceae. Studien in Richtung einer nat "urlichen Gliederung**. *Bibliotheca Lichenologica* 2002, **85**:1–526.

Figures

Figure 1 - Phylogenetic binning on an unrooted tree

Outline of the phylogenetic binning procedure for morphological taxa into an unrooted binary tree with three genera.

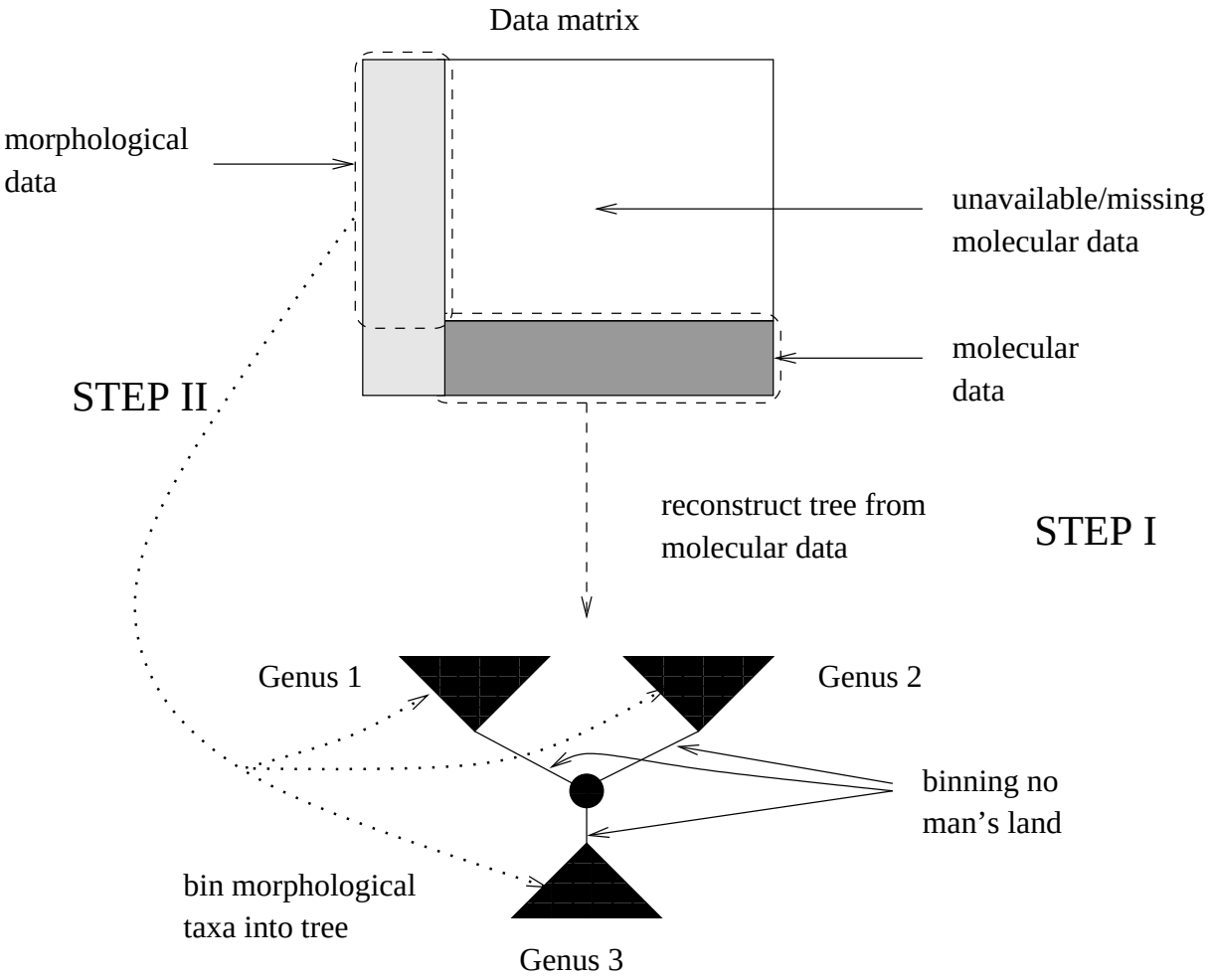


Figure 2 - Phylogenetic binning on a rooted tree

.Outline of the phylogenetic binning procedure for morphological taxa into an rooted (by an outgroup) binary tree with two genera.

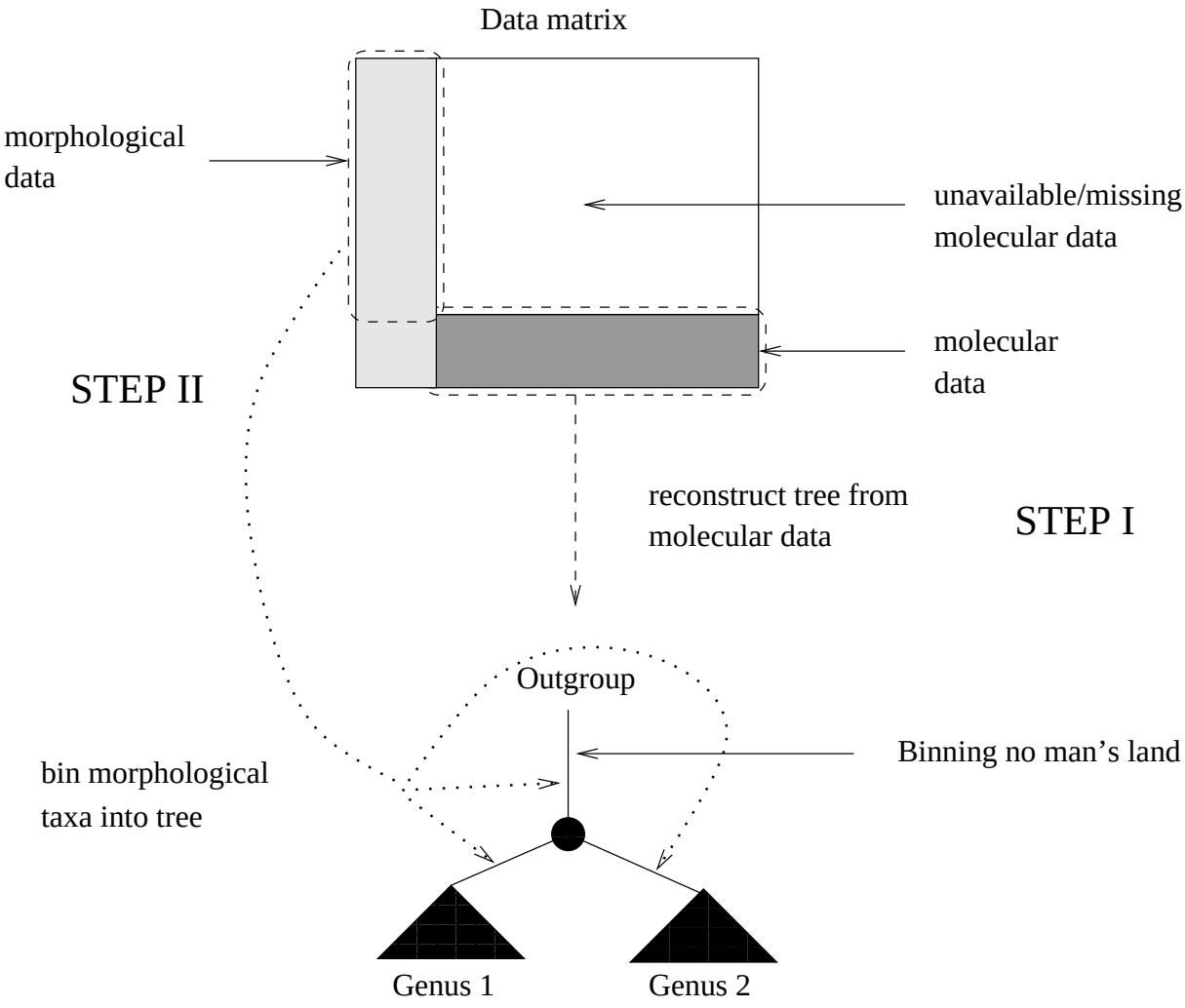


Figure 3 - BS proportions for correctly and incorrectly placed taxa in the lichen dataset

Plots of the BS support (obtained by a leave-one-out test) for correctly and incorrectly binned taxa with and without site weight calibration using taxa of the lichen dataset for which morphological *and* molecular data was available.

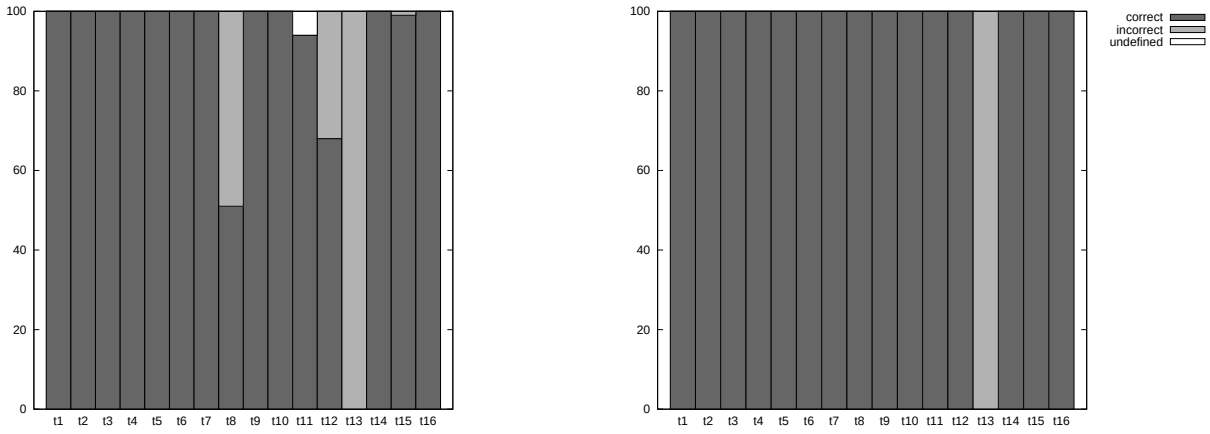


Figure 4 - Systematic test procedure

Outline of the systematic test procedure used to assess the accuracy of the binning algorithm on 5 real-world datasets.

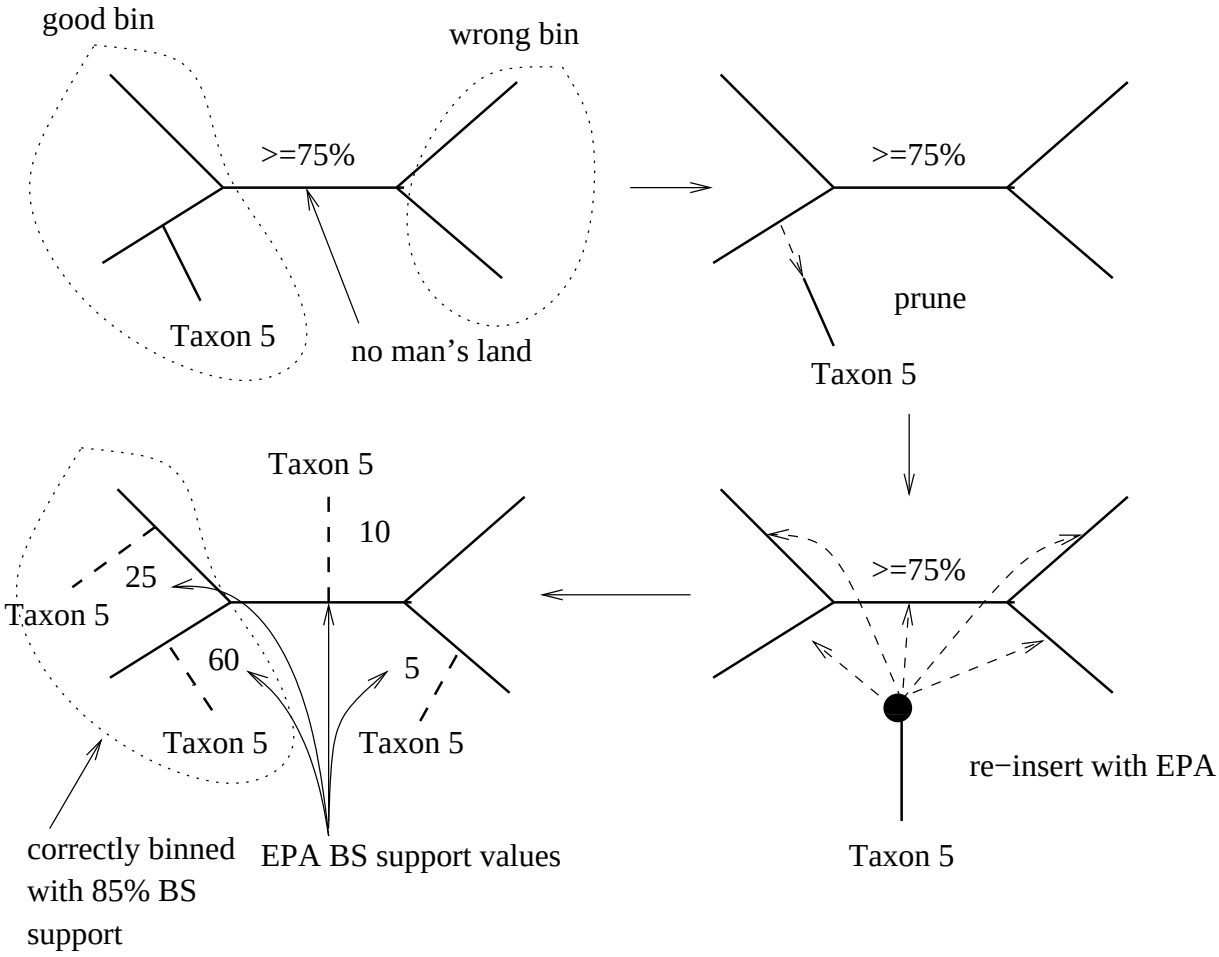
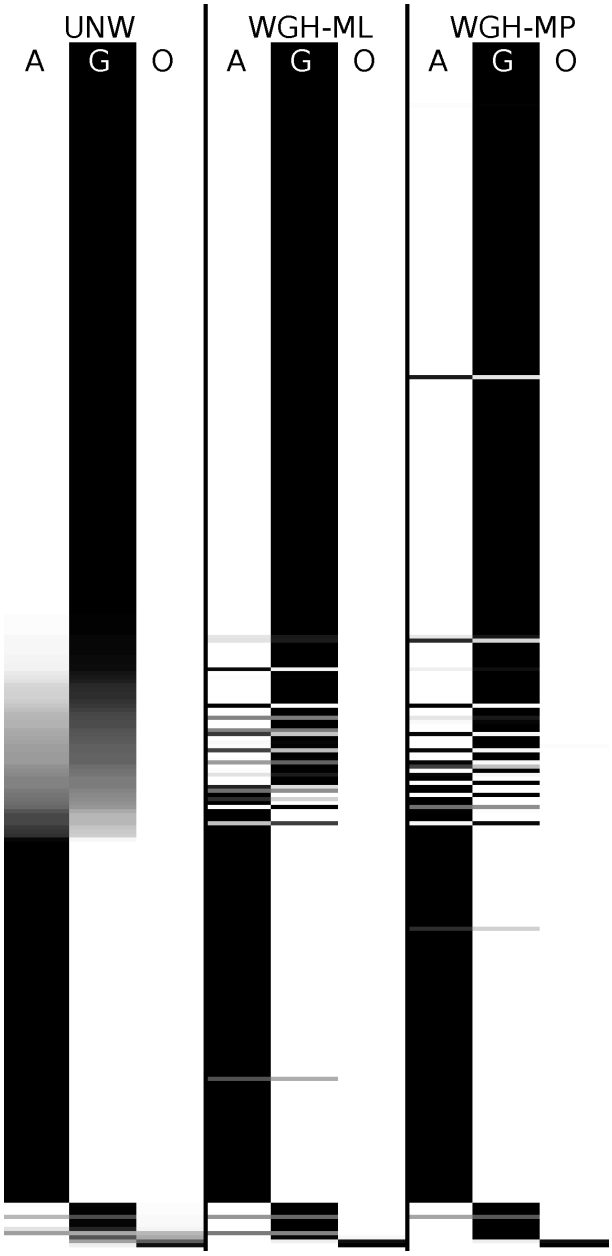


Figure 5 - Visualization of the Morphology-based Assignment of 297 lichen taxa to the three phylogenetic bins

Visualization of the morphology based assignment, including BS support, to the three bins [*Allographa (denoted by A)*, *Graphis (denoted by G)*, Outgroup (denoted by O)] using the unweighted binning method (denoted by UNW), the likelihood-weighted binning method (denoted by WGH-ML), and the parsimony-weighted method (denoted by WG-MP). Dark shaded areas indicate high support, light grey areas indicate low support.



Tables

Table 1 - Morphological characters used in the analyses and their character state definitions and coding

Character	State 1	State 2	State 3	State 4	State 5
Thallus color	olive	white-grey			
Thallus cortex	present	absent			
Thallus surface	smooth	verrucose			
Soralia	absent	present			
Isidia	absent	present			
Lirellae emergence	immersed	erumpent	prominent	sessile	
Lirellae thalline margi	compl. thick	compl. thin	lateral	basal	absent
Lateral margin thick	absent	present			
Labia sharply delimited	absent	present			
Thalline margin flaking	absent	present			
Lirellae length	0.5–2 mm	2–3 mm	3–5 mm	5–80 mm	
Lirellae width	0.1–0.3 mm	0.3–0.5 mm	0.5–1 mm	1–1.5 mm	
Length to width ratio	1–1.5	1.5–2	2–5	5–10	10–100
Lirellae branching	absent	sparse	irregular	radiate	stellate
Lirellae pseudostromata	absent	present			
Labia white cover	absent	present			
Labia pruina	absent	present			
Labia pruina yellow	absent	present			
Labia pruina orange	absent	present			
Disc exposure	absent	present			
Disc pruina	absent	present			
Disc pruina orange	absent	present			
Labia striation	entire	striate			
Excipulum carbonization	absent	apical	lateral	complete	
Hymenium inspersion A	absent	present			
Hymenium inspersion B	absent	present			
Hymenium pigment	absent	present			
Ascospores number	8/ascus	4–8/ascus	2–4/ascus	1–2/ascus	
Ascospores length	10–20 μm	20–50 μm	50–100 μm	100–200 μm	200–300 μm
Ascospores width	5–8 μm	8–15 μm	15–30 μm	30–50 μm	
Length to width ratio	1–2	2–4	4–8	8–15	
Septa transversal	3–5	5–9	9–19	19–39	
Septa longitudinal	absent	terminal	0–2/segm.	3–7/segm.	
Endospore	well-devel.	reduced			
Pigmentation	hyaline	grey-brown			
Chemistry norstictic	absent	present			
Chemistry salazinic	absent	present			
Chemistry stictic	absent	present			
Chemistry hypostictic	absent	present			
Chemistry hirtifructic	absent	present			
Chemistry protocetraric	absent	present			
Chemistry lichexanthone	absent	present			
Chemistry pigm. yellow	absent	present			
Chemistry pigm. orange	absent	present			
Chemistry isohypocrelli	absent	present			
Chemistry unknown	absent	present			

Table 2 - Binning accuracy for weighted and unweighted placement runs

Binning accuracy for runs with and without calibrated morphological site weights on 5 real-world datasets.

	UNWEIGHTED	WEIGHTED
D1	98.7476%	98.7551%
D2	99.2424%	99.2424%
D3	98.0072%	98.9187%
D4	98.6638%	98.932%
D5	96.6667%	96.3889%