Bioinformatics, 35(7), 2019, 1151–1158
doi: 10.1093/bioinformatics/bty767
Advance Access Publication Date: 31 August 2018
Original Paper

Phylogenetics

Methods for automatic reference trees and multilevel phylogenetic placement

Lucas Czech [1,*], **Pierre Barbera** [1] **and Alexandros Stamatakis**[1,2]

[1]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg, Germany and [2]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on April 12, 2018; revised on July 24, 2018; editorial decision on August 28, 2018; accepted on August 30, 2018

Abstract

Motivation: In most metagenomic sequencing studies, the initial analysis step consists in assessing the evolutionary provenance of the sequences. Phylogenetic (or Evolutionary) Placement methods can be employed to determine the evolutionary position of sequences with respect to a given reference phylogeny. These placement methods do however face certain limitations: The manual selection of reference sequences is labor-intensive; the computational effort to infer reference phylogenies is substantially larger than for methods that rely on sequence similarity; the number of taxa in the reference phylogeny should be small enough to allow for visually inspecting the results.
Results: We present algorithms to overcome the above limitations. First, we introduce a method to automatically construct representative sequences from databases to infer reference phylogenies. Second, we present an approach for conducting large-scale phylogenetic placements on nested phylogenies. Third, we describe a preprocessing pipeline that allows for handling huge sequence datasets. Our experiments on empirical data show that our methods substantially accelerate the workflow and yield highly accurate placement results.
Availability and implementation: Freely available under GPLv3 at http://github.com/lczech/gappa.
Contact: lucas.czech@h-its.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput DNA sequencing technologies have revolutionized biology by transforming it into a data-driven computational discipline (Escobar-Zepeda *et al.*, 2015). Next Generation Sequencing (NGS) methods now allow for studying microbial samples directly extracted from their environment (Edwards and Holt, 2013). For each sample, these methods yield a set of short, anonymous DNA sequences, so-called reads. A typical task in such studies is to identify and classify the reads by relating them to known reference sequences, either taxonomically or phylogenetically.

Conventional methods based on sequence similarity are fast and work reasonably well if the reads are similar enough to the reference sequences, that is, if they represent species that are closely related to known species. However, they might *not* yield the most closely related species (Koski and Golding, 2001). This is particularly true for environments where available reference databases do not exhibit sufficient taxon coverage (Mahé *et al.*, 2017). As insufficient taxon coverage cannot be detected by methods that are based on sequence similarity, they can potentially bias downstream analyses.

So-called phylogenetic (or evolutionary) placement methods (Barbera *et al.*, 2018; Berger *et al.*, 2011; Matsen *et al.*, 2010) provide a more accurate means for identifying reads. Instead of relying on sequence similarity, they identify reads based on a phylogenetic tree of reference sequences. Thereby, they can incorporate information about the evolutionary history of the species under study.

In short, phylogenetic placement calculates the most probable insertion branches for a query sequence (QS) on a given reference tree (RT). For metagenomic studies, the QSs are the reads from the environmental samples and barcoding regions or marker genes are used predominantly, see below. First, the QSs are aligned against the reference alignment of the RT. For a given QS and a given branch in the RT, the QS is inserted as a new tip into the branch; the affected branch lengths are then re-optimized; the likelihood score of the tree is evaluated; and the QS is removed again from the branch. This process yields a so-called *placement* of the QS for every branch of the RT, that is, an optimized position on the branch, along with a likelihood score for the entire RT. These likelihood scores are then transformed into probabilities to quantify the uncertainty of the QS placement into the respective branch. This placement process is repeated independently for each QS on the original RT. Phylogenetic placement thus yields a mapping of each QS to all branches of the RT, along with a probability for each placement of a QS on a specific branch.

This mapping can be seen as an identification and classification of the QSs in terms of the RT, similar to a taxonomic assignment. However, phylogenetic placement also allows for more elaborate downstream analyses. Firstly, the reference tree usually offers a higher resolution than simple per-taxon abundance counts and the amount of mapped QSs per branch can be directly visualized on the RT (Mahé *et al.*, 2017). Secondly, established methods such as Edge PCA and Squash Clustering (Matsen and Evans, 2011) allow for identifying subtle differences between distinct samples, thus enabling comparative studies directly based on phylogenetic placement. Lastly, we recently proposed novel methods for visualizing and clustering phylogenetic placement data (Czech and Stamatakis, 2018), which, for example, can reveal correlations of per-sample meta-data features with sequence abundances.

A phylogenetic placement can be carried out *if* the QSs can be aligned to the reference alignment. Often, barcoding regions such as 16S or 18S are used, but there also exist studies that use different, or even a set of, maker genes (Sunagawa *et al.*, 2013). Furthermore, other types of sequences such as $_{\mathrm{mi}}$tags (Logares *et al.*, 2014) can be used. Phylogenetic placement is particularly helpful for studying new, unexplored environments, for which no closely related sequences exist in reference databases (e.g. Mahé *et al.*, 2017). However, the selection of suitable reference sequences for inferring the RT constitutes a challenge for studying such environments, as this typically is a manual process. Furthermore, conducting phylogenetic placement requires a higher computational effort with respect to the placement algorithms *per se*, but also the pre- and post-processing, than, for instance, similarity based methods. Nonetheless, existing placement algorithms are being increasingly used and cited. Due to the continuous advances in molecular sequencing, existing placement methods as well as respective pre- and post-processing tools have already reached their scalability limits.

2 Materials and methods

Here, we introduce methods to overcome the aforementioned limitations, that is, to (i) automatically obtain a high quality reference tree for phylogenetic placement, (ii) split up the placement process into two steps using smaller phylogenies and (iii) accelerate the computation of placements via appropriate data pre-processing approaches. All methods are implemented as part of our GAPPA tool, which is freely available under GPLv3 at http://github.com/lczech/gappa.

2.1 Phylogenetic automatic (reference) trees

2.1.1 Motivation

Molecular environmental sequencing studies, particularly those that aim to conduct phylogenetic placement, often rely on a set of manually selected and aligned reference sequences to infer an RT (de Vargas *et al.*, 2015; Mahé *et al.*, 2017; Tedersoo *et al.*, 2014; Thompson *et al.*, 2017). Creating and maintaining databases of such reference sequences constitutes a labor-intensive and potentially error-prone process. Moreover, this approach is impractical for highly diverse samples that comprise sequences from a plethora taxonomic clades, or samples obtained from unexplored environments. Lastly, even if a large RT is available, the visualization of placements on such an RT might be confusing and thus hard to interpret.

The reference tree (RT) used for phylogenetic placement should ideally (i) cover all major taxonomic groups that occur in the QSs, (ii) use high-quality error-free reference sequences and (iii) not be too large to allow for unambiguous visualization and interpretation. These criteria can be met for small datasets by manually selecting curated sequences from databases. For large and taxonomically diverse samples one key challenge is that sequence databases such as GREENGENES (DeSantis *et al.*, 2006), UNITE (Abarenkov *et al.*, 2010), PR2 (Guillou *et al.*, 2012), EZTAXON (Kim *et al.*, 2012), SILVA (Quast *et al.*, 2013) and RDP (Cole *et al.*, 2014) maintain reference collections of thousands to millions of taxonomically annotated sequences. Therefore, one needs to appropriately sub-sample sequences such that the RT can be inferred in reasonable time and sufficiently covers the diversity of the sample.

To this end, we present a computationally efficient approach for obtaining sequences from large databases to infer an RT. This RT is then used for conducting phylogenetic placement analyses. The input of our method is a database of aligned sequences of known species including their taxonomic labels. Our approach then identifies sets of sequences that are similar to each other based on their entropy. It subsequently reduces the sequences in these sets to a predefined number of consensus sequences. This set of sequences is the output of our method. It represents the taxonomic clades and is then used to infer the RT.

2.1.2 Sequence entropy

First, we define a measure to quantify the ensemble similarity of a set s of sequences, based on their entropy (Shannon and Weaver, 1951). Variants of sequence entropy have been used before in numerous biological and phylogenetic contexts, for example, to assess the information content of sequences (Schmitt and Herzel, 1997; Vinga, 2014) or to measure substitution saturation (Xia *et al.*, 2003). Here, we use entropy for alignment sites, that is, we define the entropy (uncertainty) H at alignment site i as

$$H_i = -\sum_c f_{c,i} \times \log_2 f_{c,i}$$

where $c \in \{A, C, G, T, -\}$ is the set of nucleotide states including gaps and $f_{c,i}$ is the frequency of character c at site i of the alignment. Including gaps (–) in the summation reduces the contribution of sites that contain a large fraction of gaps. Their contribution is weighed down as all standard phylogenetic inference tools model gaps as undetermined states, that is, they do not contribute anything to the likelihood score. The entropy is 0 for sites that only contain a single character. It increases the more different characters an alignment site contains and the more similar their frequencies are. Its maximum occurs if all characters appear with the same frequency (each of them 20%). Note that we also treat ambiguous characters as gaps. As only 0.008% of the non-gap characters in our test database (SILVA) are ambiguous, their influence is negligible. Ambiguous characters could however be incorporated by using fractional character counts.

Finally, the total entropy of a set s of aligned sequences is simply the sum over all per-site entropies: $H(s) = \sum_i H_i$. We use this entropy to quantify the ensemble similarity of a set of sequences. This can be regarded as an information content estimate of the sequences.

2.1.3 Sequence grouping

The goal of this step is to group the sequences of a database into a given target number of groups/sets, such that the groups reflect the diversity of the sequences in the database. We use the taxonomy to identify potential candidate groups of sequences that could be represented by a consensus sequence. We interpret a taxonomy as a sequence labeling, where similar sequences have related labels. Thus, a taxonomy represents a pre-classification of similar sequences that can be exploited to group them.

For a clade t of the taxonomic tree, we denote by $H(t)$ the entropy of all sequences that belong to that clade, including all sequences in its sub-clades, that is, its lower taxonomic ranks. Clades with low entropy imply that they contain highly similar sequences that can in turn be represented by a consensus sequence without sacrificing too much diversity. Inversely, clades with high entropy contain diverse sequences, implying that a consensus sequence is not likely to sufficiently capture the inherent sequence diversity. It is thus better to expand these clades and construct separate consensus sequences for their respective sub-clades. An example is shown in Figure 1. As the clade structure of a taxonomy forms a tree, this criterion can then be applied recursively, as shown in Algorithm 1.

The algorithm works as follows: We initialize a list of candidate clades with the highest-ranking clades that we want to consider. In the most general case, these can be "Archaea", "Bacteria" and "Eukaryota". We then select the most diverse candidate clade, that is, the clade t whose sequences exhibit the highest entropy $H(t)$. This clade is then expanded and we do not consider it as a potential candidate for building a consensus sequence. The high entropy clade is then removed from our list and its immediate sub-clades are added as new candidates to the list. Finally, the current count of how many candidates we have already selected is updated accordingly. By expanding clades with high entropy, we descend into the lower ranks of the taxonomy. On average, this decreases the entropy, because low ranking clades generally tend to contain more similar sequences. This process is repeated until our list contains approximately as many candidate clades as the desired target count of reference sequences, which is provided as input. As the sizes of expanded clades can vary substantially, the target count cannot always be met exactly. In our tests, the average deviation was 0.2%, as shown in Supplementary Table S1.

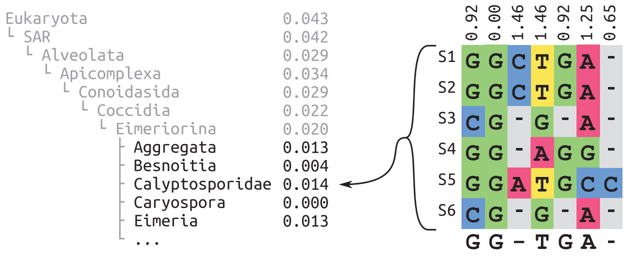

Fig. 1. Entropy and consensus sequence of a taxonomic clade. The left hand side shows the exemplary clade *Eimeriorina* in its taxonomic context, listing its super- and sub-clades with the normalized entropy of their respective sequences. The right hand side is an excerpt from the alignment of six sequences that belong to the *Calyptosporidae* sub-clade. At its top, the per-site entropies for the alignment columns are shown. At the bottom, the majority rule consensus sequence is shown, which is used to represent the sub-clade (Color version of this figure is available at *Bioinformatics* online.)

Algorithm 1 Taxonomy Expansion

```
$Candidates \leftarrow$ list of highest ranking clades
$TaxaCount \leftarrow$ size of $Candidates$
while $TaxaCount < TargetCount$ do
    $MostDiverse \leftarrow \arg\max_{t \in Candidates} H(t)$
    remove $MostDiverse$ from $Candidates$
    add sub-clades of $MostDiverse$ to $Candidates$
    $TaxaCount \leftarrow TaxaCount - 1 +$ size of $MostDiverse$
return $Candidates$
```

Given this list of clades from different taxonomic ranks, we can now compute the consensus sequences. For each clade, all sequences in that clade and its sub-clades are used to construct a consensus sequence, which represents the clade diversity and serves as the reference sequence for that clade. A simple per-site majority rule consensus (May, 1952) works well, but we also assessed alternative methods; see Supplementary Figures S2 and S3 for details. The above process yields a set of consensus reference sequences, which capture the diversity of distinct taxonomic clades.

2.1.4 Inferring a reference tree

Once we have identified the consensus sequences, which are already aligned to each other, we can use them to infer a maximum likelihood tree, which we call a *Phylogenetic Automatic (Reference) Tree* (PhAT). As each consensus sequence is associated with a taxonomic clade, the corresponding taxonomic path can be used to label the tips of the tree. Note that since clades with low entropy might not be expanded, the tip labels do not necessarily correspond to species or genera. Furthermore, the PhAT will not necessarily be congruent to the taxonomy.

A PhAT satisfies all criteria we listed: (i) all taxonomic groups occurring in the QSs can be covered by using a suitable taxonomy as input, (ii) by using consensus sequences, potential sequencing errors can be alleviated and (iii) the size of the tree can be specified by the user. However, the resolution of the trees is limited by the underlying taxonomy, see Supplementary Figures S5 and S7 for details. Thus, one needs to verify that the resulting tree is appropriate for the dataset to be placed on it. This also holds for manually selected reference sequences. Furthermore, using consensus sequences may obscure the degree of sequence diversity in sub-clades, which in turn can affect the accuracy of subsequent phylogenetic placements on that tree. Our algorithm as described here cannot fully compensate for this. We present a method to address both issues (tree resolution and obscured diversity) in the next Section.

2.2 Multilevel placement

When conducting phylogenetic placement, the computationally limiting factors are (i) the number of QSs to be placed (addressed in the next section) and (ii) the size of the RT (number of taxa) and corresponding alignment length (addressed below). Using RTs with more taxa increases the phylogenetic resolution of the placements, at the cost of increased computational effort for inferring the RT, aligning

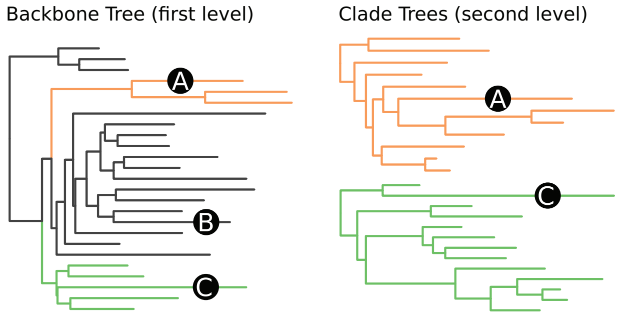

Fig. 2. Multilevel Placement. The left shows a backbone tree (BT); the right shows two clade trees (CTs). Branches in the BT that are associated with a CT are marked in the same visual style. The trees "overlap" each other, meaning that each CT is represented by multiple branches in the BT. Three sequences A, B and C are placed on the BT, which is the first level. A and C are placed on branches associated with a CT. Hence, their second level placement is conducted on the respective CT. B is placed on a branch that is not associated with any CT, and thus not used in the second level (Color version of this figure is available at *Bioinformatics* online.)

the QSs and placing the QSs. Furthermore, longer reference alignments (if appropriate data is available) are required to accurately infer large trees under the maximum likelihood criterion (Yang, 1994), thus further increasing the computational costs. Lastly, placement on large trees that comprise reference sequences with high evolutionary distances can reduce placement accuracy (Mirarab *et al.*, 2012). Thus, using a large number of reference sequences is not always desirable in practice.

To address this issue, we present an approach called *Multilevel* or *Russian Doll* Placement, which is summarized in Figure 2. Instead of working with one large RT comprising *all* taxa of interest, we use a smaller, but taxonomically broad backbone tree (BT) for pre-classifying the QSs (first level) and a set of refined clade trees (CTs) for the final, more accurate placements (second level). These CTs comprise the reference sequences that are of interest for a particular study. For example, if a study is concerned with *Apicomplexa* and *Cercozoa*, a broad *Eukaryotes* BT can be used for the first level and two respective CTs for the second level, in analogy to (Mahé *et al.*, 2017). Each CT is associated with the set of branches of a specific BT clade.

The method then works in three steps:

1. Align and place the QSs using the BT (first level).
2. For each CT, collect the QSs that are placed on the BT branches associated with the CT.
3. Align and place these QSs again, using their specific CTs (second level).

While this approach requires some additional bookkeeping, the total computational cost is reduced, because the QSs do not have to be placed on all branches of all CTs. The speed gain depends on the relative sizes of the BT *and* the CTs with respect to the size of the substantially larger (often one order of magnitude or more) comprehensive tree. For example, by splitting a tree with 10 000 taxa into a BT and 10 CTs with 1000 taxa each, the computational cost decreases by a factor of 5. Furthermore, at each level, the amount of required main memory is reduced by a factor of 10 compared to the large tree. Lastly, this method allows for fine-grained control over the clades of interest at both placement levels:

Firstly, the BT provides a means for phylogenetically informed sequence filtering – that is, to identify and remove "spurious" QSs. Sequences with low similarity to known references are often removed in environmental sequencing studies. However, using sequence similarity as a filter criterion can remove too many QSs, particularly when studying new, unexplored environments (Mahé *et al.*, 2017). By using phylogenetic placement as a filter instead, substantially more sequences can be retained for downstream analyses. Only the QSs that are placed onto the inner branches of the BT, that is, branches with no associated CT, are omitted at the second placement level.

Secondly, using specific clade trees for lower level taxonomic clades offers the phylogenetic resolution that is necessary for downstream analyses and for biological reasoning. It is, for example, possible to use manually curated "expert" trees for each clade of interest.

In this setup, the BT is only used for pre-classification and can, for example, use our PhAT method. The aforementioned issue of obscured diversity in sub-clades can be circumvented by "overlapping" the CTs with the BT. That is, a CT can be associated with several branches of the BT, so that placements on each of these BT branches are collected and placed onto the same CT. See Figure 2 and Supplementary Figure S4 for examples. We recommend to ensure that the branches of the BT that are associated with one CT are monophyletic, meaning that there is one split that separates these branches from the rest of the BT. This can be achieved by inferring the BT with a high-level constraint that maintains the monophyly of the CTs. It ensures phylogenetic consistency between the BT and the CTs, and improves the accuracy of the first placement level, as shown in Section 3.4. Lastly, it is also possible to use more than two levels, which might become necessary when working with RTs and datasets even larger than what is currently available.

2.3 Data preprocessing for phylogenetic placement

Apart from the RT size, handling the sheer number of QSs also induces computational limitations for conducting phylogenetic placements. Most metagenomic studies publish their data in unprocessed formats, which are sometimes filtered to only contain reads from certain barcoding or marker regions. Those data often contain duplicates of exactly identical sequences, both *within* and *across* samples. Identical sequences are however treated the same in phylogenetic placement algorithms and therefore induce unnecessary computational overhead. Furthermore, sample sizes, that is, the number of sequences per sample, can vary by several orders of magnitude. If the placement algorithm is parallelized over samples, this leads to an uneven load balance across compute nodes.

In order to solve these issues, that is, reduce computational cost and achieve good load balancing, one can pre-process the sequences with our GAPPA tool. First, sequences are de-duplicated across all samples and fused into chunks of equal size. The chunk size should be chosen to allow aligning and placing a chunk within wall time on the intended hardware; we recommend chunk sizes of 50 000 or larger. Our tool assigns an identifier to each unique sequence and computes a list of abundance counts for each sequence in a sample. Given an RT and its underlying alignment, the QS chunks are then aligned to the reference multiple sequence alignment, using programs such as PaPaRa (Berger and Stamatakis, 2012) or HMMALIGN (Eddy, 1998) and subsequently placed on the RT, for example by PPLACER, RAxML-EPA or EPA-NG (Barbera *et al.*, 2018; Berger *et al.*, 2011; Matsen *et al.*, 2010). The resulting per-chunk placement result files in combination with the per-sample abundance counts can then be parsed and analyzed by GAPPA to generate final per-sample placement files, containing a placement for each sequence in the original sample.

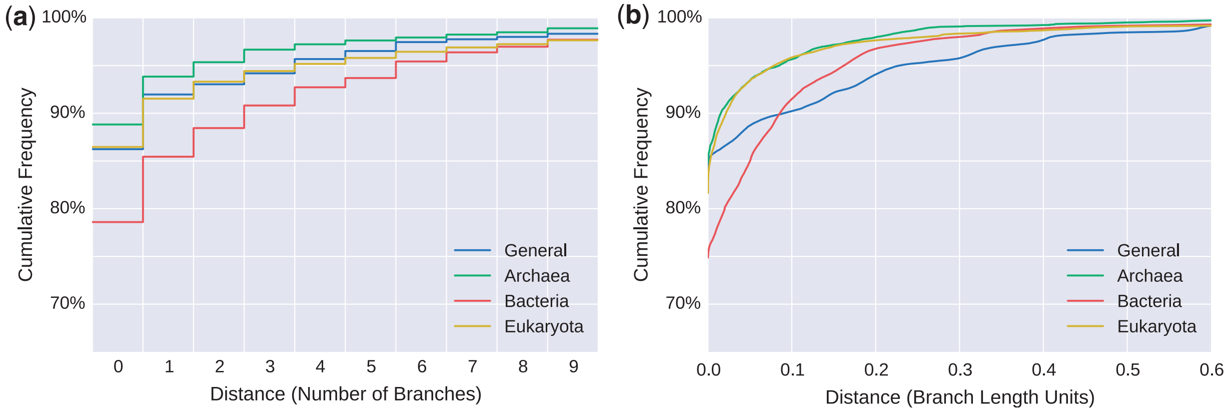

Fig. 3. Weighted distances to expected edges for unconstrained trees. We evaluated the accuracy of our PhATs by placing sequences and measuring the weighted distances to their respective expected placement branches. The Figure shows the cumulative frequencies of number of sequences versus distances, measured (a) in number of branches and (b) in branch length units. In other words, it shows how many sequences are placed within a certain radius from their expected branches. For example, in (a), more than 85% of the sequences of the *Bacteria* are placed within a radius of at most one branch from their expected branch, and in (b), more than 95% of the *Eukaryota* are within a radius of 0.1 branch length units from their expected branches (Color version of this figure is available at *Bioinformatics* online.)

The speedup induced by this preprocessing is proportional to the ratio of total versus unique sequences; the gain in parallel efficiency depends on the ratio of the smallest to the largest sample (in terms of number of sequences). This approach allows to analyze datasets that are orders of magnitude larger than in previous published studies. For example, in 2012, an analysis of Bacterial Vaginosis (BV) data placed a total of 426 612 sequences, thereof 15 060 unique, on an RT with 796 tips (Srinivasan *et al.*, 2012). Using a prototype of GAPPA, we were able to analyze a neotropical soils dataset with a total of 50 118 536 sequences, thereof 10 567 804 unique, with an RT comprising 512 taxa (Mahé *et al.*, 2017). To demonstrate the scalability of our methods for this paper, we analyzed datasets with up to 116 520 289 total sequences, thereof 63 221 538 unique, from the Human Microbiome Project (HMP) (Huttenhower *et al.*, 2012; Methé *et al.*, 2012), using RTs with up to 2059 tips. This corresponds to a computational effort that is four orders of magnitude greater than for the BV study.

3 Results

To test the Phylogenetic Automatic (Reference) Tree (PhAT) method, we used the “SSU Ref NR 99” sequences of the SILVA database (Quast *et al.*, 2013) version 123.1 and the corresponding taxonomic framework (Yilmaz *et al.*, 2014). The database contains 598 470 aligned sequences from all three domains of life, classified into 11 860 distinct taxonomic labels.

We constructed four sets of consensus sequences from the SILVA database: a *General* set (“all of life”), as well as separate sets for the domains *Archaea*, *Bacteria* and *Eukaryota*. For each set except the *Archaea*, the recursive expansion of taxonomic clades was applied to obtain approximately 2000 (*General*) and 1800 (*Bacteria*, *Eukaryota*) consensus reference sequences. This is large enough to cover the diversity well, while still being computationally feasible for the subsequent steps. The *Archaea* taxonomy in SILVA is smaller, containing 248 taxa at *Genus* level, which is the lowest level in their taxonomy. Hence, the *Archaea* tree also comprises 248 taxa. Furthermore, in the three domain-specific trees, we included sequences at the *Phylum* level of the respective two other domains, to ensure that our methods also work with outgroups. The assembly of these four datasets required in total about 30 min and 10 GB of main memory on a standard laptop computer. An overview of the tree sizes is shown in Supplementary Table S1. We then inferred constrained and unconstrained maximum likelihood trees for the consensus sequences. The constrained trees comply with the SILVA taxonomy and are used to assess how taxonomic constraints affect the phylogenetic placement and the subsequent analyses. Details are provided in Supplementary Section S1, which also discusses differences between the constrained and unconstrained trees. Details of the trees are shown in Supplementary Table S3; Supplementary Figure S4 shows the unconstrained *Bacteria* tree as an example.

In total, our setup yields eight distinct RTs for evaluation: the *General* tree, the three domain trees and the respective taxonomically constrained variants.

3.1 Accuracy

Here, we assess how using a PhAT affects phylogenetic placement accuracy. Each terminal branch of our RTs represents a consensus sequence, which is computed from *Species* level sequences that share the same taxonomic label. We evaluate an RT by placing these species sequences onto the RT: Each species sequence is expected to be placed onto the branch leading to the consensus sequence that represents this particular species sequence. For example, sequences S1-6 in Figure 1 are represented by the consensus sequence for the *Calyptosporidae* clade, which is shown below the six sequences in the Figure. They are thus expected to be placed onto the *Calyptosporidae* branch in the RT.

We placed the respective subset of the SILVA database species sequences onto each of the eight RTs. We quantify placement accuracy for a sequence by the distance to its expected placement branch. More precisely, we measured (a) the (discrete) number of branches between the actual placement and the expected branch and (b) the (continuous) distance in branch lengths units. As a sequence can have multiple placement locations, the distances, are, in fact, weighted averages incorporating the placement probabilities (likelihood weights). The results for the four unconstrained trees are shown in Figure 3; Supplementary Figure S1 depicts the results for the constrained trees. Further details are provided in Supplementary Table S3.

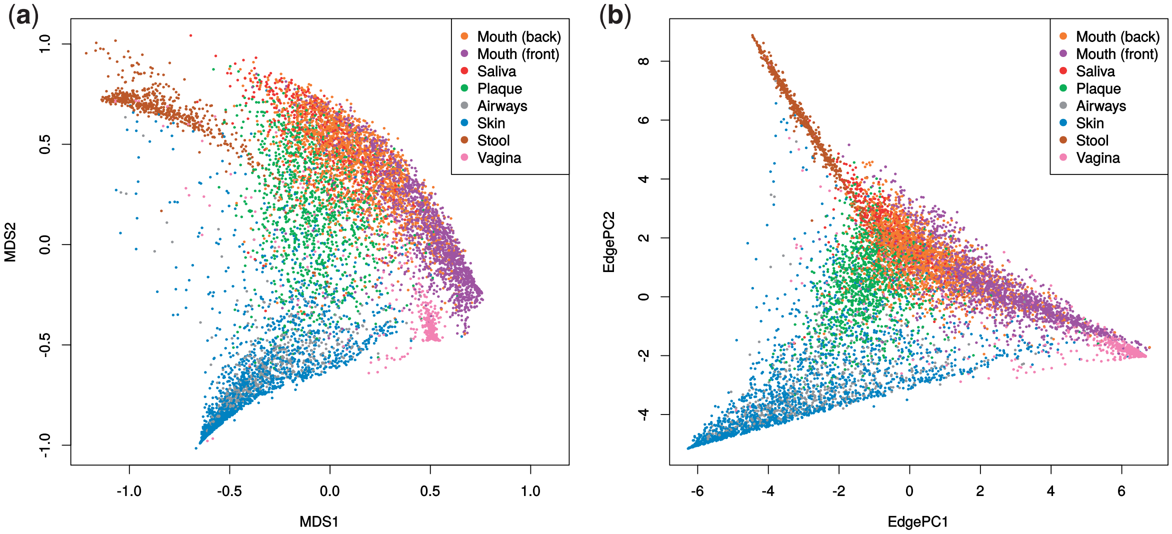

Fig. 4. Assessment of a PhAT for large dataset analyses. Here, we test the unconstrained *Bacteria* tree for placing and analyzing a large sequence dataset. For this, we used the Human Microbiome Project (HMP) (Huttenhower *et al.*, 2012; Methé *et al.*, 2012) data, and selected 9 192 samples from different body sites with a total of 117 million sequences. For details on the processing, see Supplementary Section S2. We categorized the 18 original body site labels into 8 regions for readability, see Supplementary Table S4. The sequences were placed on the tree, and subsequently analyzed with two different methods. Both subfigures show that the tree, despite only representing higher taxonomic levels, suffices to separate different body site regions from each other. (a) Visualization of the pairwise phylogenetic Kantorovich-Rubinstein (KR) distance (Matsen and Evans, 2011) between all samples. The KR distance is a generalization of the UniFrac distance (Lozupone, 2005) for phylogenetic placement data. The high-dimensional pairwise distance matrix was embedded into the plot by performing Multidimensional scaling (MDS, Borg, 2005.). (b) Edge PCA (Matsen and Evans, 2011) analysis of the samples. The grouping of body sites is again clearly visible with this method (Color version of this figure is available at *Bioinformatics* online.)

Considering the size of the trees, most sequences are placed in close vicinity to their expected branches. This is corroborated by the short average distances reported in Supplementary Table S3. Furthermore, the average expected distance between placement locations (EDPL, Matsen *et al.*, 2010) is low, indicating that the placements of a specific sequence mostly cluster in a small neighborhood of the tree. We observed that errors occur mostly in parts of the tree with short branches, which might be explained by the inability of 16S SSU sequences to properly resolve certain clades (Janda and Abbott, 2007). In addition, the placement likelihood differences are small between neighboring, short branches, such that the placement signal is fuzzy.

With 77% of the sequences placed exactly on their expected branch, the accuracy is generally lowest for the *Bacteria* tree. This might be because the *Bacteria* have the most sequences in Silva and exhibit a high diversity. In the other three trees, more than 90% of the sequences are placed at most one branch away from their respective expected branch. The constrained trees (Supplementary Figure S1) exhibit similar placement accuracy. Particularly when using Multilevel Placement with overlapping RTs, placement differences of a few branches on the first level tree are acceptable, as they do not change the second level tree on which the sequence is placed. See Section 3.4 for details.

As outlined in the method description, we represent clade diversity via majority rule consensus sequences. To assess the impact of the consensus method, we repeated the above evaluation, using two alternative consensus methods, but found little difference between the methods, see Supplementary Figure S2. Finally, we also tested an automated approach that uses actual sequences (instead of consensus sequences) from the database to represent the taxonomic clades, see Supplementary Figure S3. We found that this approach yields trees that are less accurate for phylogenetic placement.

3.2 Empirical datasets

PhATs are intended for conducting phylogenetic placement of environmental sequences. As the true evolutionary history of such sequences is unknown, we cannot repeat the previous accuracy tests on empirical environmental datasets. Instead, we assess if the PhATs yield meaningful quantitative results for typical post-analysis methods. To this end, we placed two empirical metagenomic amplicon barcoding datasets on our unconstrained *Bacteria* tree. To assess the placement results obtained from the PhATs, we performed Squash Clustering and Edge PCA (Matsen and Evans, 2011) post-analyses on the placement results, see Supplementary Section S2 for details. The results are shown in Figure 4 and Supplementary Figure S5, and reveal that the PhAT reproduces results of previous studies based on custom RTs with manually selected reference sequences. Furthermore, the PhAT is able to classify samples (e.g., body regions or healthy versus sick patients), at least to the extent that is expected from its phylogenetic resolution. That is, samples that only differ in placements at the *Species* level cannot be classified using a broad, high-level tree such as our *Bacteria* tree. In order to obtain finer taxonomic resolution, it is thus necessary either to use a PhAT that contains more taxa, or to use our multilevel approach instead (see next Section).

3.3 Taxonomic assignment and profiling

Here, we assess how PhATs perform when used for obtaining a taxonomic profile of a set of samples in conjunction with placement. We emphasize though that taxonomic assignment and profiling are neither the focus of PhATs, nor the intended standard applications of phylogenetic placement. To perform the evaluation, we used the *mouse gut* dataset of the 2nd CAMI Challenge (Bremges and McHardy, 2018; Sczyrba *et al.*, 2017) and phylogenetically placed the reads of the 16S locus (≈ 0.08% of the total data) on our constrained and unconstrained *Bacteria* trees. We then used this placement data to taxonomically assign the reads based on the underlying Silva taxonomy of the trees, in analogy to the method used by Sativa (Kozlov *et al.*, 2016). Unfortunately, the CAMI Challenge uses the NCBI taxonomy for the respective evaluation. We thus had to compute a mapping between the two taxonomies, which introduces some incongruities (Balvočiūtė and Huson, 2017). The resulting per-read assignment was then used to generate a taxonomic profile of the data.

Despite only using a small fraction of the reads and despite having to use incongruent taxonomies, our PhAT-based taxonomic profiling is in the mid-range of the tools evaluated by CAMI. Therefore, our method yields reasonable accuracy for taxonomic assignment and profiling. Note that the resolution of the assignment is limited by the taxonomy used when running the PhAT method, that is, we could not assign reads at *Species* level. Details of the process and its results are provided in Supplementary Section S4; Supplementary Figure S7 and Supplementary Table S5 show the most important evaluation results.

3.4 Sub-clades and multilevel placement

We selected five bacterial clades to evaluate PhAT accuracy on smaller clades, as well as to assess some properties of the Multilevel Placement approach. The same clades were already scrutinized in Sativa (Kozlov *et al.*, 2016). Supplementary Figure S4 shows the *Bacteria* tree with the five test clades highlighted.

First, using the sequences and taxonomies of these five clades, we built unconstrained and constrained PhATs. We then conducted the same accuracy analysis as explained before on these 10 trees. That is, we placed the Silva sequences of the five clades onto their respective PhAT and evaluated distances to expected branches. Thereby, we evaluated the accuracy of these PhATs when used as second level clade trees. The results are shown in Supplementary Figure S6. The placement accuracy is slightly worse for the clade trees than for the eight comprehensive PhATs evaluated before. This is again likely due to 16S SSU sequences being unable to properly resolve lower taxonomic levels (Janda and Abbott, 2007).

Next, using the five clades, we evaluated the accuracy of the first placement level when conducting Multilevel Placement. So far, our evaluation focused on the distance from a sequence placement to its expected placement branch. For the first placement level on a backbone tree (BT), it is however more important that a sequence is placed into the correct clade. Thus, we used the unconstrained *Bacteria* BT again and assessed how many sequences were placed in the clades shown in Supplementary Figure S4. Of the 450 313 sequences in Silva in these clades, 98.0% were placed (most likely placement) into a branch of their corresponding clade. Thus, for multilevel placement, they will be assigned to the correct second level clade tree (CT). More specifically, the *Firmicutes* perform worst, as only 94.7% of the *Firmicute* sequences are placed into the corresponding clade. This can be explained by the high amount of paraphyletic branches of this clade, cf. Supplementary Figure S6, which is a known issue (Parks *et al.*, 2018). The sequences of the other four clades we tested achieve a clade identification accuracy exceeding 99%.

As mentioned before, a high-level taxonomic constraint can improve the accuracy of placing a sequence into the correct BT clade. To show this, we inferred the *Bacteria* RT again, but used a *Phylum* level constraint that separates the five clades from each other and from the rest of the tree. All branches within the clades were resolved using maximum likelihood. The tree (not shown) is similar to the tree in Supplementary Figure S4, but all five clades are now monophyletic. Using this tree, 99.3% of the sequences were placed into the correct clade. Particularly the accuracy for *Firmicutes* improved, yielding an accuracy of 99.5%.

Overall, our experiments show that the first level placement is highly accurate, even if an extremely diverse "all bacteria" backbone tree is used. The accuracy on the second level is slightly worse when using PhATs as CTs.

4 Discussion and conclusion

We presented algorithms and software tools to facilitate and accelerate phylogenetic placement of large environmental sequencing studies.

The Phylogenetic Automatic (Reference) Tree (PhAT) method provides a means for automatically obtaining suitable reference trees by using the taxonomy of large sequence databases. Using the Silva database as a test case, we showed that it can be applied for accurately (pre-)placing environmental sequences into taxonomic clades. The method can also be used for rapid data exploration in environmental sequencing studies: A PhAT might be useful to obtain an overview of the taxa that are necessary to capture the diversity of a sequence dataset, without the substantial human effort and potential bias of manually selecting reference sequences. As we showed, PhATs can also be used to obtain taxonomic assignments and profiles for a set of samples, in conjunction with phylogenetic placement. To capture clade diversity with finer resolution, for example for a second placement level, clade-specific PhATs can be inferred. If species-level resolution is required, we recommend that the sequences are inspected by an expert, in order to confirm that the tree is appropriate for the dataset to be placed on it. Furthermore, as our automated approach inevitably suffers from errors in the database it is based on, we recommend using Sativa (Kozlov *et al.*, 2016) to identify potentially mislabeled sequences in the database. One should also keep in mind that phylogenetic placement does not necessarily provide resolution at the *Species* level (Dunthorn *et al.*, 2014).

As we show, our multilevel placement method as well as the preprocessing pipeline accelerates the placement process without sacrificing accuracy. By first placing the query sequences on a broad backbone tree (BT), novel environments with sequences of unknown evolutionary origin can be classified without having to process a large tree comprising all taxa of interest. A second placement on a set of clade trees (CTs) provides sufficient resolution for biological interpretation. Placement accuracy can be further improved by inferring the BT with a high-level constraint that separates the clades of the CTs from each other and thus ensures monophyly of these clades.

The methods presented here are implemented as part of our gappa tool, which is freely available under GPLv3 at http://github.com/lczech/gappa (see Supplementary Section S3 for an overview of the corresponding commands). All scripts and data used for this paper are available at http://github.com/lczech/placement-methods-paper.

Acknowledgements

We thank **S. Srinivasan** and **E. Matsen** for providing the Bacterial Vaginosis dataset (Srinivasan *et al.*, 2012); **M. Dunthorn**, **G. Lentendu** and **A. Kozlov** for their feedback on our methods and this manuscript.

Funding

This work was financially supported by the Klaus Tschira Stiftung gGmbH in Heidelberg, Germany.

Conflict of Interest: none declared.

References

Abarenkov,K. *et al.* (2010) The UNITE database for molecular identification of fungi–recent updates and future perspectives. *New Phytol.*, **186**, 281–285.

Balvočiūtė,M. and Huson,D.H. (2017) SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genom.*, **18**, 114.

Barbera,P. *et al.* (2018) EPA-ng: massively parallel evolutionary placement of genetic sequences. *bioRxiv*.

Berger,S. and Stamatakis,A. (2012). *PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension. Technical Report*, Institute for Theoretical Studies, Heidelberg.

Berger,S. *et al.* (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.*, **60**, 291–302.

Borg,I. and Groenen,P.J.F. (2005). *Modern Multidimensional Scaling: Theory and Applications*, 2nd edn. Springer-Verlag, New York.

Bremges,A. and McHardy,A.C. (2018) Critical assessment of metagenome interpretation enters the second round. *mSystems*, **3**.

Cole,J.R. *et al.* (2014) Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, **42**, D633.

Czech,L. and Stamatakis,A. (2018) Scalable methods for post-processing, visualizing, and analyzing phylogenetic placements. *bioRxiv*.

de Vargas,C. *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.

DeSantis,T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.

Dunthorn,M. *et al.* (2014) Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Mol. Biol. Evol.*, **31**, 993–1009.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edwards,D.J. and Holt,K.E. (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb. Inform. Exp.*, **3**, 2.

Escobar-Zepeda,A. *et al.* (2015) The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. *Front. Genet.*, **6**, 1–15.

Guillou,L. *et al.* (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.*, **41**, D597–D604.

Huttenhower,C. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Janda,J.M. and Abbott,S.L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.*, **45**, 2761–2764.

Kim,O.-S. *et al.* (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int. J. Syst. Evol. Microbiol.*, **62**, 716–721.

Koski,L.B. and Golding,G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.*, **52**, 540–542.

Kozlov,A.M. *et al.* (2016) Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Res.*, **44**, 5022–5033.

Logares,R. *et al.* (2014) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.*, **16**, 2659–2671.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Mahé,F. *et al.* (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat. Ecol. Evol.*, **1**, 91.

Matsen,F.A. and Evans,S.N. (2011) Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One*, **8**, 1–17.

Matsen,F.A. *et al.* (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, **11**, 538.

May,K.O. (1952) A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, **20**, 680–684.

Methé,B.A. *et al.* (2012) A framework for human microbiome research. *Nature*, **486**, 215–221.

Mirarab,S. *et al.* (2012) SEPP: SATé-enabled phylogenetic placement. In: *Proceedings of the Conference Pacific Symposium on Biocomputing. World Scientific*, pp. 247–258.

Parks,D.H. *et al.* (2018) A proposal for a standardized bacterial taxonomy based on genome phylogeny. *bioRxiv*.

Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

Schmitt,a. O. and Herzel,H. (1997) Estimating the entropy of DNA sequences. *J. Theor. Biol.*, **188**, 369–377.

Sczyrba,A. *et al.* (2017) Critical Assessment of Metagenome Interpretation a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.

Shannon,C.E. and Weaver,W. (1951). *The Mathematical Theory of Communication*. University of Illinois Press, Champaign, Illinois, USA.

Srinivasan,S. *et al.* (2012) Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One*, **7**, e37818.

Sunagawa,S. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods*, **10**, 1196.

Tedersoo,L. *et al.* (2014) Global diversity and geography of soil fungi. *Science*, **346**, 1256688.

Thompson,L.R. *et al.* (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, **551**, 457–463.

Vinga,S. (2014) Information theory applications for biological sequence analysis. *Brief. Bioinform.*, **15**, 376–389.

Xia,X. *et al.* (2003) An index of substitution saturation and its application. *Mol. Phylogenet. Evol.*, **26**, 1–7.

Yang,Z. (1994) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.*, **43**, 329–342.

Yilmaz,P. *et al.* (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.*, **42**, D643–D648.