Evolutionary Placement of Short Reads

Methods, Applications, and Visualization

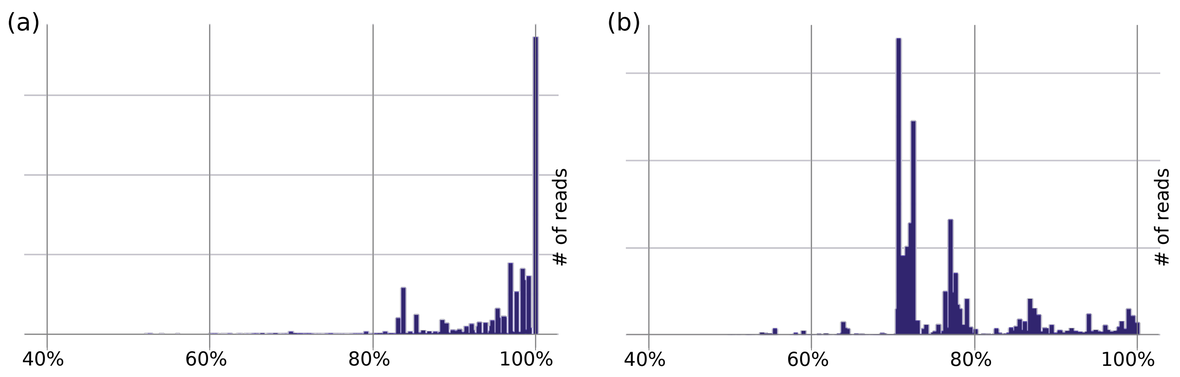
Lucas Czech[1], Simon Berger, Denis Krompaß, Jiajie Zhang, Paschalia Kapli, Pavlos Pavlidis and Alexandros Stamatakis[1,2]

[1] Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Germany
[2] Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Germany
Email: {alexandros.stamatakis, lucas.czech}@h-its.org

Background and Motivation

Metagenomic studies often need to biologically classify millions of reads. This assignment of reads to known reference sequences helps to assess the composition and diversity of microbial communities and allows for comparing them.

Traditional methods for read classification, particularly approaches based on mere sequence similarity (e.g., BLAST), however, have their limits [1]. If the sampling of the reference sequences is sparse or inappropriate, the reference might not contain sequences that are sufficiently closely related to the reads. Many environmental sequencing studies hence disregard reads that are less than 80% similar to those in reference databases.



The figure shows the similarity of sequences to a reference database for (a) marine protists and (b) tropical soil protists.
In (a), the majority of reads are sufficiently similar to the reference database. In (b), however, many reads have less than 80% similarity to the reference database.
This might result in removing too many reads and hence biasing the diversity estimate or missing new species.

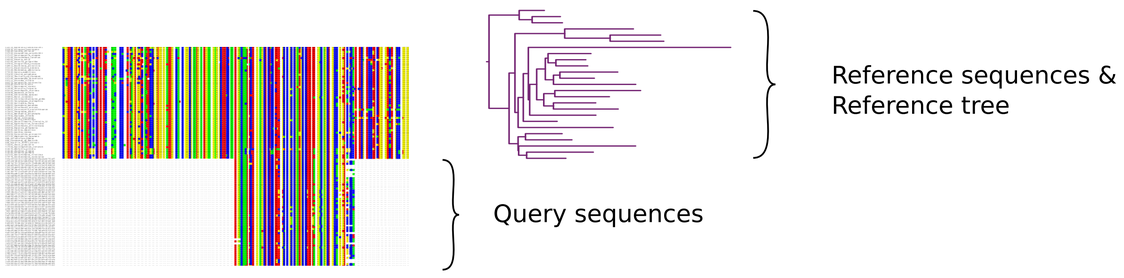
The Evolutionary Placement Algorithm offers a solution to this problem.

Evolutionary Placement Algorithm

Input Data

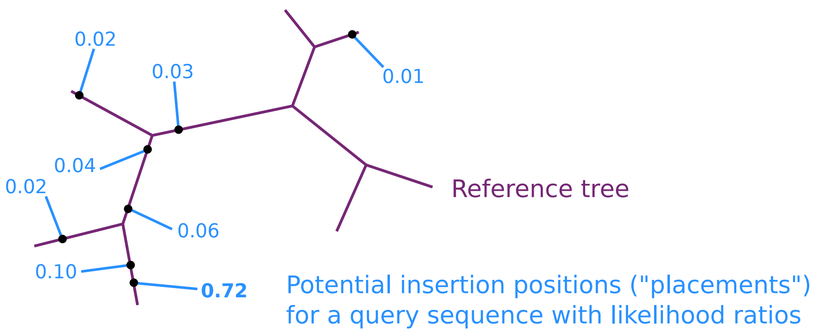
The Evolutionary Placement Algorithm (EPA) [2] takes as input:

- An alignment of reference sequences (e.g., single 16S or barcoding gene)
- A species tree (usually inferred from the reference sequences)
- Aligned query sequences (e.g., Illumina reads) [3]



Algorithm

The algorithm finds the most likely (via maximum likelihood) insertion position for every query sequence on the reference tree. The resulting assignment of a query sequence to a branch is called a "placement".



This can be done in parallel for each query and each branch and is thus highly scalable.

Advantages over BLAST

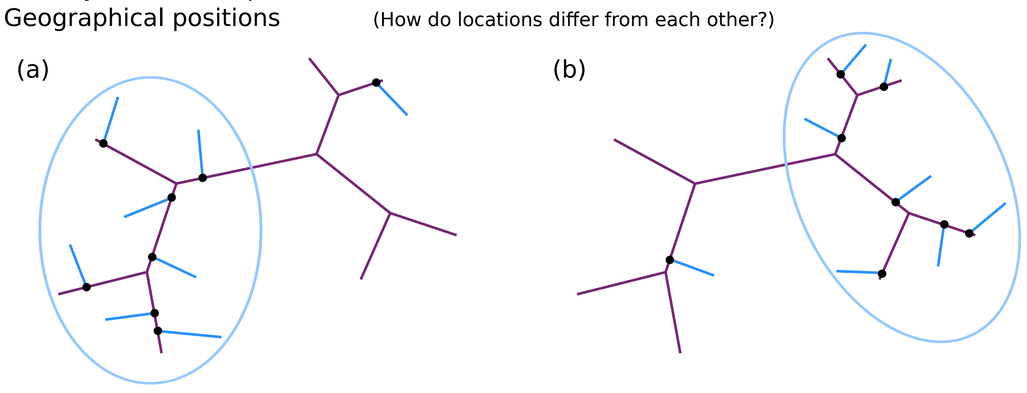
- Assigns environmental reads to an evolutionary history (assignment to all branches, not only tips)
- Best BLAST hit is not necessarily the evolutionary closest sequence [1]
- Explicit model of rate heterogeneity
- Can reveal taxon sampling problems

The output of EPA is standardized using an easy-to-read JSON-based text format [4].

Applications

Compare placements from several samples

- Time series (How does an environment change over time?)
- Healthy versus sick patients (What are the distinctive features of a disease?)
- Geographical positions (How do locations differ from each other?)

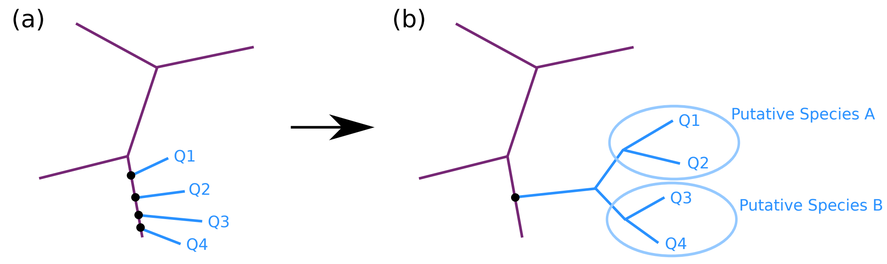


The figure shows two different distributions of placements on the same reference tree.
For comparing them, we need to quantify their difference:

- Earth Movers Distance
  Calculates the minimal amount of "work" required to transform one sample/distribution into the other.
- Edge Principal Component Analysis [5]
  Visualizes the weight differences of placements per branch (split) of the tree.

Count species in placements of a single sample

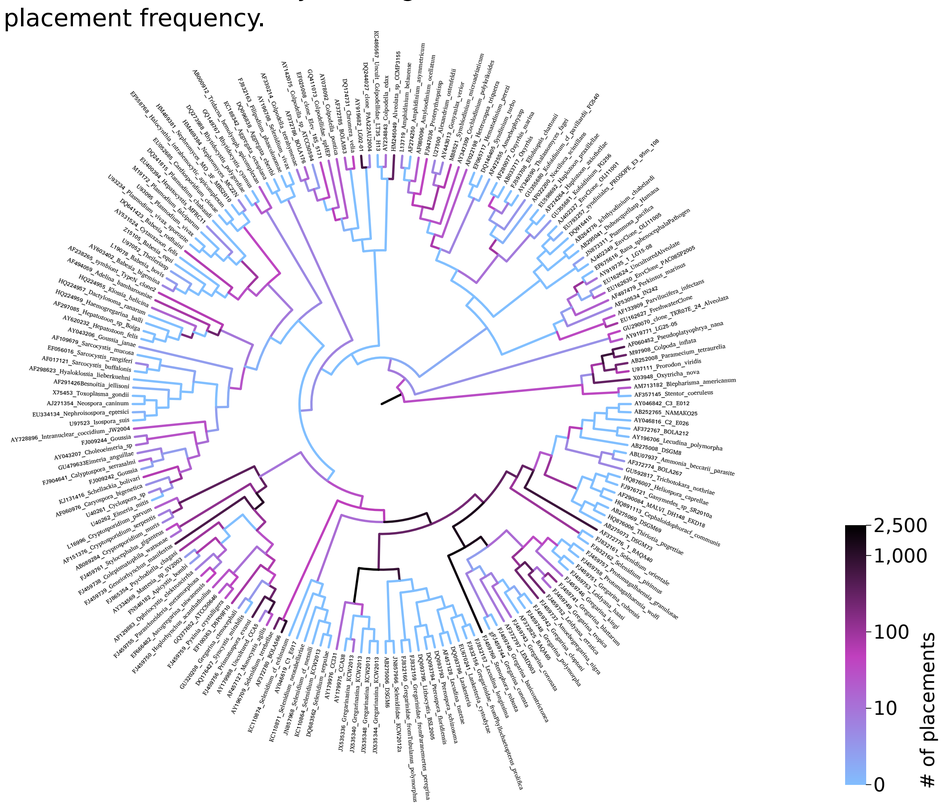
We look at the placements on each branch separately (a). Their relationship (multifurcation) is resolved using standard maximum-likelihood. The resulting tree is then rooted and can be interpreted as a new sub-tree of the reference tree (b).
Given this sub-tree, we can delimit (putative) species based on the placements [6].



This is useful for estimating the diversity of these species.

Visualization

For millions of query sequences, visualization of each placement is not helpful.
Instead, we summarize the information by coloring the branches of the reference tree according to the placement frequency.



This exemplary figure shows V4 OTU data of Apicomplexa from rainforest soils collected in Costa Rica, Panama, and Ecuador.

It is easy and straight forward to identify the "interesting" regions of the tree,
i.e., the dark regions where many query sequences are placed.

References

[1] L. B. Koski and G. B. Golding, “The closest BLAST hit is often not the nearest neighbor.,” J. Mol. Evol., vol. 52, no. 6, pp. 540–2, Jun. 2001.

[2] S. Berger, D. Krompass, and A. Stamatakis, “Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood,” Syst. Biol., vol. 60, no. 3, pp. 291–302, 2011.

[3] S. Berger and A. Stamatakis, “Aligning short reads to reference alignments and trees,” Bioinformatics, vol. 27, no. 15, pp. 2068–2075, 2011.

[4] F. A. Matsen, N. G. Hoffman, A. Gallagher, and A. Stamatakis, “A format for phylogenetic placements,” PLoS One, vol. 7, no. 2, pp. 1–4, Jan. 2012.

[5] F. A. Matsen and S. N. Evans, “Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison.,” PLoS One, vol. 8, no. 3, pp. 1–17, Jan. 2011.

[6] J. Zhang, P. Kapli, P. Pavlidis, and A. Stamatakis, “A general species delimitation method with applications to phylogenetic placements,” Bioinformatics, vol. 29, no. 22, pp. 2869–2876, 2013.

Availability and Acknowledgements

EPA is available via Standard RAxML, see github.com/stamatak/standard-RAxML

This work was financially supported by the Klaus Tschira Foundation.

The original methods and implementations were developed by S.A. Berger, D. Krompaß, J. Zhang, P. Kapli, P. Pavlidis and A. Stamatakis [2,3,6].

Further thanks go to Micah Dunthorn for providing the data used for the diagrams and visualizations.

www.exelixis-lab.org

Heidelberger Institut für Theoretische Studien

HITS

Karlsruher Institut für Technologie