Taxonomic Read Assignment with *Read-SpaM*

Matthias Blanke[1] and Burkhard Morgenstern[1]

[1] University of Goettingen, Inst. of Microbiology and Genetics, Dep. of Bioinformatics, Germany (http://gobics.de)

Phylogenetic placement is the problem of finding an optimal position for a query taxon in an existing phylogenetic tree of reference taxa. Existing methods are e.g. *EPA(-ng)* [2], *pplacer* [6] and *RAPPAS* [5]. A drawback of these approaches is that they rely on alignments of reference sequences. An approach that does not depend on aligned reference sequences is *APPLES* [1]. This program is using estimated distances betweeen the reference and the query sequences.

An important application of phylogenetic placement is *taxonomic read assignment*, for example in metagenomics. Herein, we apply the program *Read-SpaM* [3] to this task. *Read-SpaM* is an adaption of the program *Filtered Spaced Word Matches (FSWM)* [4] to unassembled reads. Distances between sequences are estimated based on spaced-word matches, i.e. local gap-free alignments of a fixed length with matching nucleotides at certain positions, specified by a pre-defined binary pattern of *match* and *don't-care positions*.

To estimate the phylogenetic distance between two sequences, *Read-SpaM* considers all spaced-word matches with scores above some threshold and uses the number of mismatches at the *don't-care positions* to estimate the number of mismatches per position in an (unknown) alignment of the input sequences. The usual *Jukes-Cantor* correction is applied to estimate the distance between the input sequences, i.e. the average number of substitutions per sequence position since they have evolved from their last common ancestor.

We evaluated three different approaches to assign a read r to a position in a reference tree, based on distance values calculated by *Read-SpaM*: (A) r is assigned to the query sequence with the smallest distance to r, (B) r is assigned to the *lowest common ancestor* of the two reference sequences with the smallest distance to r, and (C) *APPLES* is applied to the distances calculated by *Read-SpaM*. As a comparison, we used *APPLES* with *Jukes-Cantor* corrected *Hamming* distances that the program uses by default, *RAPPAS* and *EPA-ng*.

To evaluate all six approaches, we used a set of mitochondrial genomes from 25 closely related fish taxa with known phylogeny. We removed one genome at a time from this set and generated simulated query reads from this genome using *ART*. The six approaches were then applied to place the query reads back onto the tree. We measured the accuracy of the methods by counting the average number of nodes between the original (correct) position of the query reads in the tree and the positions to which they were assigned by the evaluated methods.

The results are shown in Fig. 1. Our approaches (B) and (C) are more accurate than the three existing methods, while (A) is comparable to them. The time and memory consumption of our approaches is comparable to the competing methods. An advantage of our approach is that we do not need an alignment

of the reference sequences, and *Read-SpaM* could even be applied to unassembled reference sequences [3]. We will evaluate our approach more systematically, e.g. on further data sets from the *AFproject* study [7].

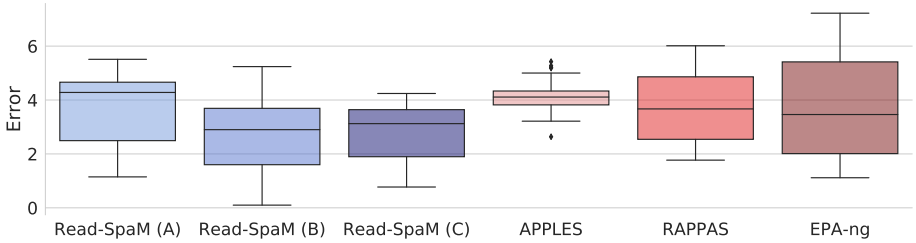

Fig. 1: Error in taxonomic read assignment, as measured by the average number of nodes between assigned and correct positions in the reference tree. For *Read-SpaM*, reads were assigned (A) to the closest reference sequence, (B) to the lowest common ancestor of the two closest reference sequences, (C) *APPLES* was applied to the *Read-SpaM* distances.

References

1. Balaban, M., Sarmashghi, S., Mirarab, S.: APPLES: Fast distance-based phylogenetic placement. Systematic Biology (in press)
2. Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., Stamatakis, A.: Epa-ng: Massively parallel evolutionary placement of genetic sequences. Systematic biology **68**, 365–369 (2019)
3. Lau, A.K., Leimeister, C.A., Morgenstern, B.: *Read-SpaM*: assembly-free and alignment-free comparison of bacterial genomes. bioRxiv **doi:10.1101/550632** (2019)
4. Leimeister, C.A., Sohrabi-Jahromi, S., Morgenstern, B.: Fast and accurate phylogeny reconstruction using filtered spaced-word matches. Bioinformatics **33**, 971–979 (2017)
5. Linard, B., Swenson, K., Pardi, F.: Rapid alignment-free phylogenetic identification of metagenomic sequences. Bioinformatics **btz068** (2019)
6. Matsen, F.A., Kodner, R.B., Armbrust, E.V.: pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics **11**, 538 (2010)
7. Zielezinski, A., Girgis, H.Z., Bernard, G., Leimeister, C.A., Tang, K., Dencker, T., Lau, A.K., Röhling, S., Choi, J., Waterman, M.S., Comin, M., Kim, S.H., Vinga, S., Almeida, J.S., Chan, C.X., James, B., Sun, F., Morgenstern, B., Karlowski, W.M.: Benchmarking of alignment-free sequence comparison methods. Genome Biology **20**, 144 (2019)