ProPhyle: a phylogeny-based metagenomic classifier using Burrows-Wheeler Transform

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

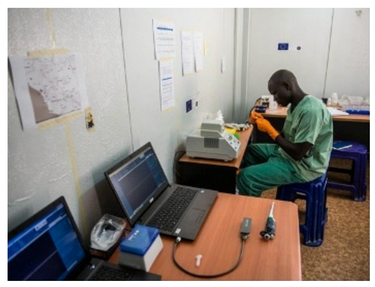
Karel Břinda, Kamil Salikhov, Simone Pignotti, Gregory Kucherov

UNIVERSITÉ PARIS-EST

Background

Metagenomics is a powerful approach to study genetic content of environmental samples and it has been strongly promoted by NGS technologies. The aim of **metagenomic classification** is to assign each sequence of the metagenome to a corresponding taxonomic unit, or to classify it as "novel".

In **point-of-care sequencing** and **disease surveillance projects** (e.g., [5]) using mobile sequencing technologies such as Oxford Nanopore, researchers are often limited to data processing on laptops with limited RAM and a slow Internet connection.



A mobile sequencing laboratory [5]

Kraken [2], the most popular tool for metagenomic classification, is very fast but suffers from high memory requirements and an inaccurate indexing structure. As a consequence, it may not be applicable in point-of-care sequencing projects.

Objectives

Our goal is to overcome two main Kraken's limits to make the classification **suitable for point-of-care sequencing**.

1. **Small memory footprint.** Whereas Kraken can be used on well-equipped clusters only, we aim at laptops with 16 GB RAM.
2. **Expressive index.** As Kraken stores only the lowest common ancestor (LCA) for every *k*-mer, the resulting classification can be inaccurate when many *k*-mers are shared between multiple genomes. This problem appears especially with phylogenetic trees for a single species. Therefore, our objective is to store a list of associated nodes for every *k*-mer.

ProPhyle

We developed ProPhyle [1], a metagenomic classifier based on BWT-index and *k*-mer propagation (see ↗) with the following features:

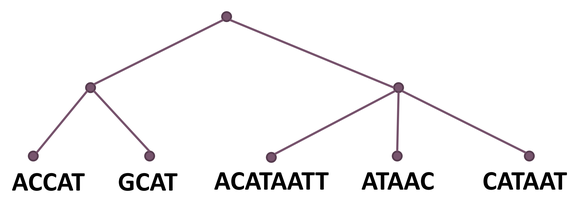
- **Small memory footprint** both during index construction and querying (see ↘).
- **Lossless representation** of *k*-mers in the index.
- **Support for multiple similarity measures for classification:** hit count, normalized hit count (corresponding to the Jaccard index), coverage and normalized coverage.
- **Support for standard formats.** Input trees are provided in **Newick/NHX** and assignments to the tree are reported in **SAM.** Therefore, it is easy to combine ProPhyle with other bioinformatics tools (e.g., SamTools).

References

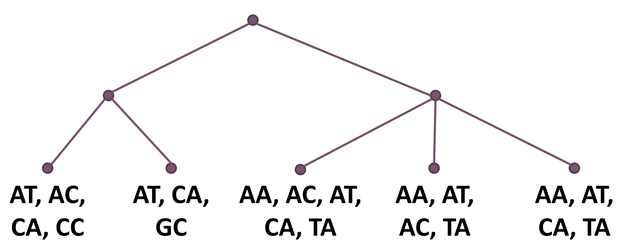
1. Břinda, K., Salikhov, K., Pignotti, S., & Kucherov, G. **ProPhyle: accurate and resource-frugal phylogeny-based metagenomic classification.** To appear.
2. Wood, D. E., & Salzberg, S. L. (2014). **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biology*, *15*(3), R46.
3. Li, H., & Durbin, R. (2009). **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics*, *25*(14), 1754–1760.
4. Ferragina, P., & Manzini, G. (2000). **Opportunistic data structures with applications.** In *Proceedings 41st Annual Symposium on Foundations of Computer Science* (pp. 390–398). IEEE Comput. Soc.
5. Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). **Real-time, portable genome sequencing for Ebola surveillance.** *Nature*, *530*(7589), 228–232.

Compressed *k*-mer index using propagation and BWT-index

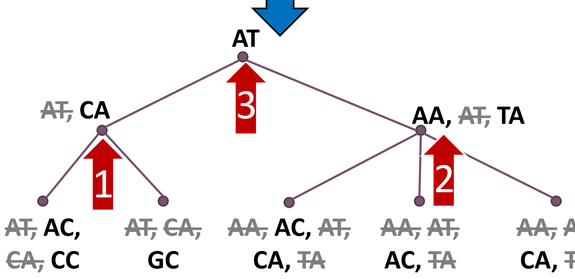
1. Initial tree



2. Sets of canonical *k*-mers

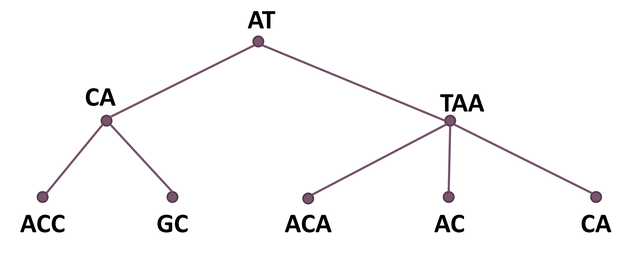


3. *k*-mer propagation



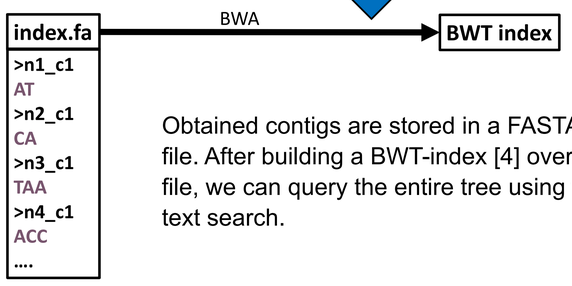
When a *k*-mer is present in all children of some node, it is **moved to the parent**. As a sequence of **local** modifications of the tree, such a propagation is memory-efficient, especially compared to Kraken (*k*-mer counting by JellyFish is a global operation).

4. Contig assembly



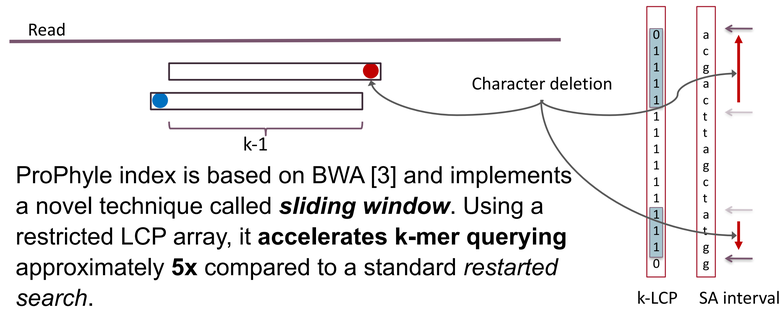
Contigs are assembled by a greedy enumeration of disjoint paths in the de-Bruijn graphs corresponding to individual nodes.

5. BWT-index construction



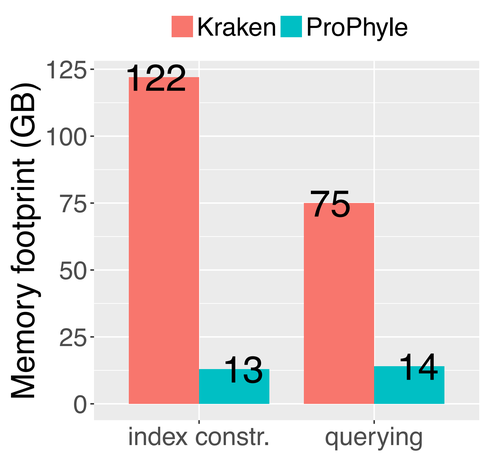
Obtained contigs are stored in a FASTA file. After building a BWT-index [4] over this file, we can query the entire tree using full-text search.

6. Querying using a sliding window

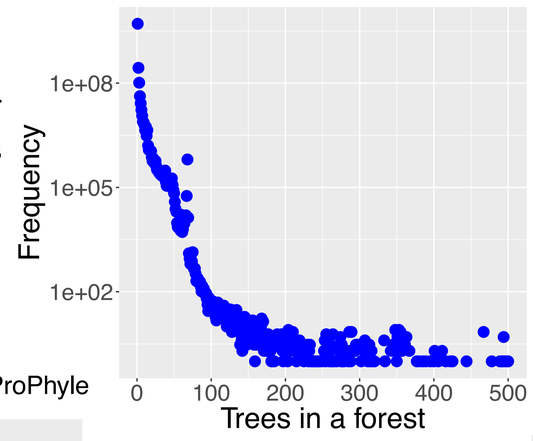


ProPhyle index is based on BWA [3] and implements a novel technique called ***sliding window***. Using a restricted LCP array, it **accelerates k-mer querying** approximately **5x** compared to a standard *restarted search*.

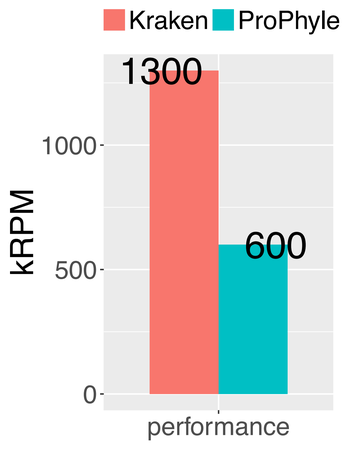
Experiment – RefSeq bacterial database (2,787 genomes, k=31)



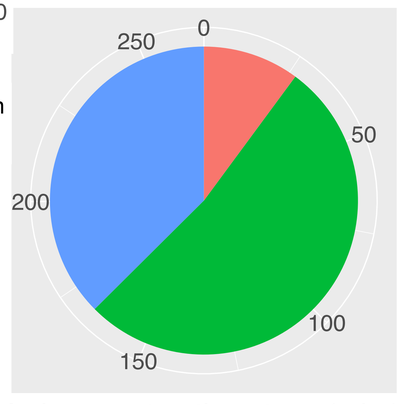
ProPhyle has a very low memory footprint. **It can be used even on laptops with 16 GB RAM.**



Unlike Kraken, ProPhyle stores a list of all genomes in which a given *k*-mer occurs. Kraken can represent accurately only the *k*-mers that occur within an entire clade, i.e., only those ones associated with the leftmost point of this graph.



ProPhyle index is approximately 2× slower compared to the index of Kraken. This is a consequence of more cache misses in BWT-indexes in general and higher expressiveness of the ProPhyle index.



Index construction currently takes approximately 4 hours. This is mainly due to non-parallelized BWT and SA construction steps in BWA.

Availability, installation and usage

github SOCIAL CODING

http://github.com/karel-brinda/prophyle

Read the Docs

http://prophyle.rtfd.io

BIOCONDA

```
$ conda install –c bioconda prophyle
```

python Package Index

```
$ pip install prophyle
```

1. Download a database

```
$ prophyle download bacteria
```

2. Build an index

```
$ prophyle index ~/prophyle/bacteria.nw bactidx
```

3. Classify your reads

```
$ prophyle classify bactidx reads.fq
```