HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

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

Karel Břinda, Kamil Salikhov, Simone Pignotti, and Gregory Kucherov

UNIVERSITÉ PARIS-EST

Background

The aim of **metagenomic classification** is to assign each sequence of a metagenome to a corresponding taxonomic unit, or to classify it as "novel".

In **point-of-care sequencing** and **disease surveillance projects** (e.g., [4]) 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 [4]

Kraken [1], 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, in particular, 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, a metagenomic classifier based on BWT-index and *k*-mer propagation, with the following features:

1) Low memory requirements

- **Small memory footprint** (both during index construction and querying)
- Easy to use on **laptops**
- The resulting index can be **further compressed** for an easy transmission

2) High resolution index and classification

- **Lossless indexing** of *k*-mers
- **Deterministic behavior**
- Support for **ambiguous assignments** (e.g., in case of reads from a core genome)

3) Standard formats & flexibility

- Support for **standard formats:**
  Trees: Newick or NHX (arbitrary phylogenetic trees)
  Assignments: SAM or the format of Kraken
  Reports: the formats of Kraken, Centrifuge and MetaPhlan
- Therefore, ProPhyle can **easily replace Kraken** in existing pipelines
- Support for **multiple measures** (hit count and read coverage, possibly normalized)
- Easy to install using a single command

4) Simple user interface

Download a database *$ prophyle download bacteria*
Download standard RefSeq databases with the NCBI taxonomy.

Build an index *$ prophyle index ~/prophyle/bacteria.nw idx*
Build the index either from a downloaded database, or from a user-provided database.

Classify your reads *$ prophyle classify idx reads.fq > class.sam*
Classification of individual reads.

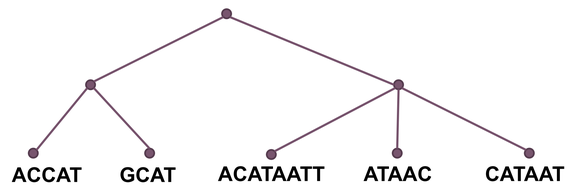
Compute abundancies *$ prophyle analyze idx class.sam exp_report*
Report summaries in various formats.

Compress index *$ prophyle compress idx*
Create an archive for an easy transmission.

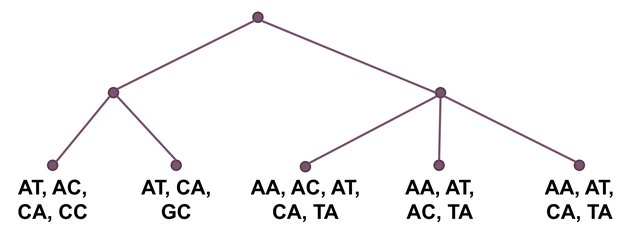
Decompress index *$ prophyle decompress idx.tar.gz ./*
Decompression after the transmission.

Methods – compressed *k*-mer index using propagation and BWT-index

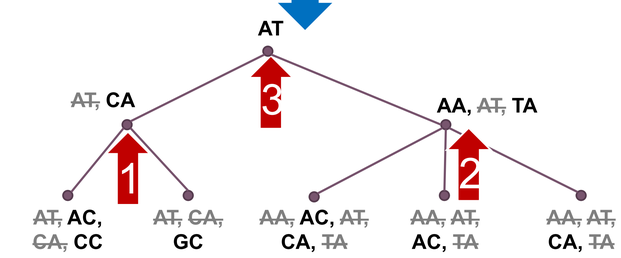
1. Initial phylogenetic tree



2. Sets of canonical *k*-mers

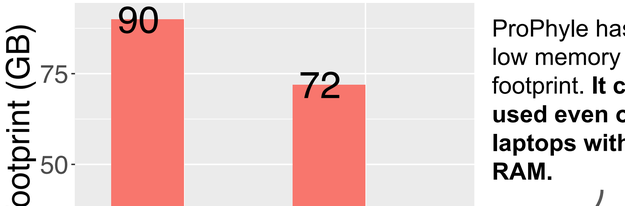


3. *k*-mer propagation



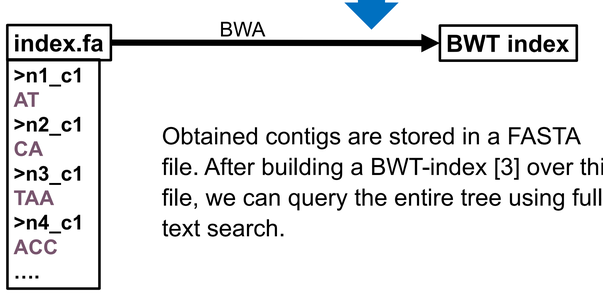
When a *k*-mer is present in all children of some node, it is **moved to the parent**. As a sequence of local modifications of the tree, such a propagation is **memory-efficient**.

4. Contig assembly



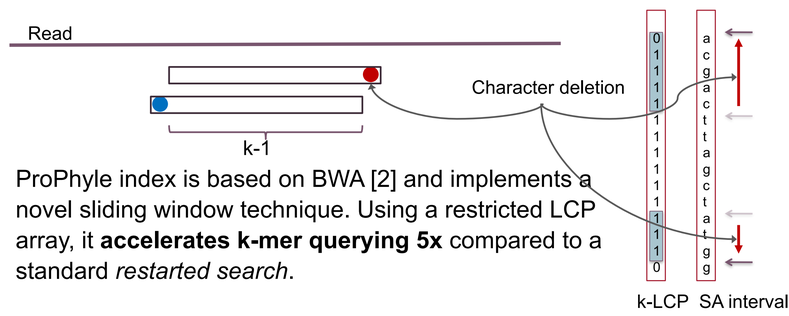
Contigs are **assembled** by a greedy enumeration of disjoint paths in the de-Bruijn graphs corresponding to individual nodes.

5. BWT-index construction



Obtained contigs are stored in a FASTA file. After building a BWT-index [3] over this file, we can query the entire tree using full-text search.

6. Querying using a sliding window

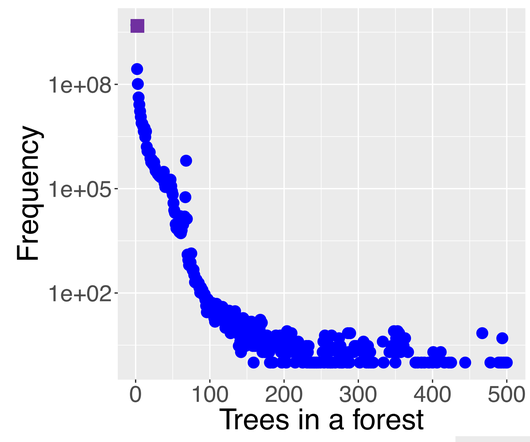


ProPhyle index is based on BWA [2] and implements a novel sliding window technique. Using a restricted LCP array, it **accelerates k-mer querying 5x** compared to a standard *restarted search*.

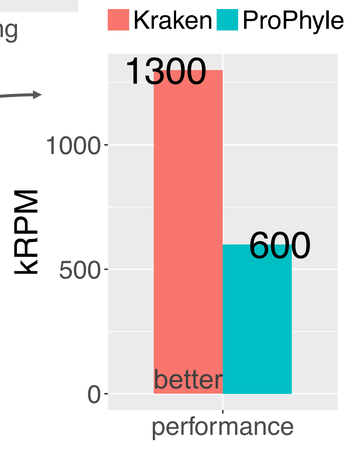
Results

RefSeq bacterial database, 2,787 genomes, k=31

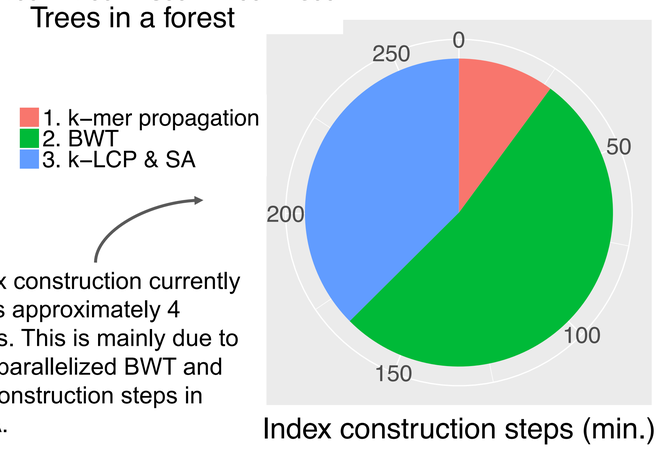
ProPhyle has a very low memory footprint. **It can be used even on laptops with 16 GB RAM.**



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 subtree, i.e., only those *k*-mers represented by ■.



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.

Index construction steps (min.)

Availability and installation

github SOCIAL CODING — http://github.com/karel-brinda/prophyle

Read the Docs — http://prophyle.rtfd.io

BIOCONDA — $ conda install prophyle

python Package Index — $ pip install prophyle

References

1. Wood, D. E., & Salzberg, S. L. (2014). **Kraken: ultrafast metagenomic sequence classification using exact alignments.** *Genome Biology*, *15*(3), R46.
2. Li, H., & Durbin, R. (2009). **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics*, *25*(14).
3. Ferragina, P., & Manzini, G. (2000). **Opportunistic data structures with applications.** In *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE Comput. Soc.
4. Quick, J., Loman, N. et al. (2016). **Real-time, portable genome sequencing for Ebola surveillance.** *Nature*, *530*(7589).