Linear mixed models for bacterial strain identification in metagenomic datasets

Contact: t.abeel@tudelft.nl or lvandijk@broadinstitute.org

Motivation and background

The world around and within us contain many complex microbial communities. Examples are your skin, your gut, or the soil. The concept of species, however, is very loosely defined in the microbial world—all known genomes of one species can differ only a by a few percent, while other species sometimes share as less as 40% of their total genome content. There's a growing need for software tools that can analyze these microbial communities down to the strain level, that is, down to specific instances of a species, and identify any variants in their genomes that sets them apart from others. These tools will aid us in analyzing how bacteria evolve, analyzing how they spread within hospitals, identifying mixed infections in patients or how the composition of your gut (your microbiome) affects your health.

This is a challenging problem: with our current technology we only obtain short (about 150 base pair) DNA fragments from our microbial community, not their complete genomes. Algorithms are required to figure out what microbes are present in the community based on the millions of short DNA fragments in your dataset. Here we use our growing knowledge on biology and the ever-increasing database of known bacterial genomes to aid in these kinds of analyses.

A common building block used in a lot of bioinformatics algorithms is the "k-mer". Slide a window of size k over a DNA sequence and each position yields a substring of length k, a k-mer. One common application is comparison of genomes: by counting the k-mers and comparing those with the k-mer counts of other genome sequences, you get a proxy on how similar these genomes are. K-mers are computationally efficient, and often give good approximations compared to sequence alignment based methods, which are more computationally demanding.

Project description

Consider our metagenomic dataset again with millions of short DNA fragments: you could imagine the k-mer counts in this dataset as a linear mixing of a set of k-mer counts from known genomes in a database. Of course, there are a few other factors to take into account, for example sequencing error or what to do with data from species/strains not in your database. The goal of this project is to research the possibilities and design an algorithm that reports which strains from a known database are present in given metagenomic dataset together with their relative abundances.

An initial approach could be formulating a multivariate linear regression problem—for each k-mer you estimate which linear combination of strains best predicts the count in the metagenomic dataset—but additional techniques from machine learning are probably necessary to properly deal with noise and other unknown factors.

Deliverables

- Algorithm description and implementation that reports strains and their relative abundances in a metagenomic dataset
- Benchmarks comparing performance to other tools
- Report
- Nice to have: mathematical formulation

Further reading

- https://en.wikipedia.org/wiki/K-mer
- Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** (2014).
- Ahn, T. H., Chai, J. & Pan, C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* **31,** 170–177 (2015).
- Roosaare, M. *et al.* StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ* **5,** e3353 (2017).

About

Broad Institute of MIT and Harvard was launched in 2004 to improve human health by using genomics to advance our understanding of the biology and treatment of human disease, and to help lay the groundwork for a new generation of therapies. It's one of the largest sequencing centers in the world and brings together researchers in medicine, biology, chemistry, computation, engineering, and mathematics from across MIT, Harvard, and the Harvard-affiliated hospitals, along with collaborators around the world.

The Delft Bioinformatics Lab develops novel computational tools to analyze genomic data to gain new insights in biology. Our approaches heavily use techniques from pattern recognition, machine learning and deep learning to obtain insights from large biological datasets. Our applications areas cover medicine, biotechnology and plant research. We teach bioinformatics courses across several programs within TU Delft and nationally.