Johns Hopkins University

Computational Genomics

EN 600.639

Augmenting Phylogenetic Classification of Metagenomic Reads

Authors:

Shuya Chu,

Stephen Cristiano,

Bing He,

Zhou Ye

Instructor:

Dr. Ben Langmead

December 6, 2013

1 Abstract

In recent years, metagenomics has gained increased interest in the scientific community. With relevance to public health, agriculture, biofuels, and marine biology [REFERENCE], it is important to understand the taxonomic composition of metagenomic samples and its functional impact it their respective ecosystems. In this project, we use simulated data to explore a variety of methods to taxonomically classify metagenomic reads to their correct position in a phylogenetic reference set. RESULTS

2 Introduction

The recent expansion of next generation sequencing technologies have greatly benefited our ability to study microbial communities [Reference]. Metagenomics, defined in 2005 as "The application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species" [Chen, Pacther (2005)], has seen a surge of research efforts in recent years in diverse fields including public health, agriculture, biofuels, and marine biology, and efforts such as the Human Microbiome Project generating large quantities of metagenomic data. Despite this, computational methods are still needed for analyzing and making sense of these data in an accurate but efficient matter.

One important challenge is, given a set of metagenomic reads, to accurately classify the taxonomic composition of the reads. For example, a metagenomic human gut sample will contain many different microbial species. We wish to be able to classify each read to the correct phylogenetic location in a reference set and determine the presence and abundance of different microbial species in the sample. Adding to the challenge, many of the species will be closely related (belonging to the same genus, family etc) and in a microbial sample, many of the reads make come from *de novo* genomes, *ie* from species that have never been sequenced before and unknown to the scientific community.

In this project, we propose, implement, and compare three methods for the classification of

metagenomic reads. In particular, we provide a Bloom filter approach, an approach based on Interpolated Markov Models, and an approach based using Bowtie to perform local alignment to reference genomes and classify using a novel algorithm. While similar methods exist, we also consider a number of ways in which we can improve upon these.

To perform our analyses, we obtain reference genomes from RefSeq, a large collection of genomes maintained by the National Center for Biotechnology Institute (NCBI) [Pruitt et al (2009)]

3 Related work

A large body of work has already been done on the phylogenetic classification of metagenomic reads. It's important to note that many of the early methods are only accurate for large sequence reads and fail at correctly classifying the short ($\sim$ 100 bp) reads generated by next generation sequencers [Brady, Salzberg (2009)]. We briefly review a few of the more recent metagenomic classification methods.

3.1 PhymmBL

PhymmBL [Brady, Salzberg (2009)] uses an Interpolated Markov Model to phylogenetically classify genomic reads to genomes in RefSeq, a large collection of genomes. It is capable of performing classification at different levels of the phylogentic tree (species, genus, family etc). In addition to classification using an IMM, PhymmBL also uses BLAST for local alignment to classify. By applying a linear function of the scores from the IMM and BLAST, the authors report greater classification accuracy than either the IMM or BLAST alone. Though very accurate, PhymmBL suffers from extremely slow running time, due to using BLAST for local alignment.

3.2 FACS

Fast Accurate Classification of Sequences (FACS) [CITATION] transforms the reference genomes to Bloom filters and query Bloom filters for exact matches. Bloom filters are an compact hash-based data structure which are used to quickly determine whether an element is a member of a set or not. In a trade-off for compactness and speed of look-up, Bloom filters come with a risk of giving false positives which must be controlled. FACS is implemented in PERL and achieves high speed, but is limited to using Bloom filters $< 312MB$ and does not have a scoring system optimized for classification. In a comparison of classification software, other authors have reported an inability to reproduce the results reported in the FACS paper [Bazinet, Cummings (2012)].

3.3 MetaPhlAn

MetaPhlAn (Metagenomic Phylogenetic Analyis) [HUTTENHOWER] is a popular tool for metagenomic classification.

3.4 Bowtie

Bowtie [Langmead et al (2010)] utilizes a Burrow-Wheeler Transform and an FM index to perform extremely fast alignment of short DNA reads to a reference genome. By performing local alignment of the reads to each of the reference genomes from a phylogenetic tree, and using an appropriate scoring system, Bowtie can be adapted for classification of metagenomic reads.

4 Methods and software

We obtain our reference genomes from RefSeq, obtained from the NCBI. Due to the vast size of RefSeq, we initially considered only the phylum *Proteobacterium*, but even that exceeded our memory limits. For this reason, we decided to use comparisons of the methods for

our reference set the genomes from *Escherichia coli* (strain K-12 substain. DH10B) and [**BDELLO GENOME**]. To simulate our data, we use the software MetaSim [Ricther et al (2008)] which will generate metagenomic reads into a FASTA file given a set of reference genomes.

As the main purpose of the Bowtie method was to develop an algorithm that will not have align to every genome in a reference set, we also considered 15 genomes from *Proteobacterium* but the same simulated reads. However, the comparisons with the other methods were still conducted only on the two previously mentioned reference genomes.

4.1 Bloom filter

Our first method was to create a Bloom filter for each reference genome and query the reads to them.

4.2 Improvement with Bowtie2

Bowtie2 is an ultrafast aligner using FM index. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters. Since our simulate testing data is around 100-200 long, using Bowtie2 will lead to fast aligning.

The idea for improvement is by skipping some unnecessary alignment between the test data and reference, identifying the species for test data will be much more fast. Like what we did in the Boyer-Moore method in class, we use results from comparisons to skip future alignments that definitely won't match. To be more specific, if we got a really low score when aligning the test data with one species, it means the sample is definitely not from this species, thus we can skip the whole species from the same genus. For example, if we know this sample definitely does not belong to *Canis*, instead of checking whether it's *Canis lupus*, we can check whether it's *Cuon alpinus* from the genus of *Cuon*.

First we construct a tree structure to describe the taxonomy relation between the reference in python. For each node, node.data represent directory path of this reference, parent and

children is directory path for its **XXXX**.

Then by running Bowtie2, we will have scores which describe the similarity between the sample data and reference. If scores is really low, in other words, we have enough confidence to say the sample is not from this species, we skip the whole genus this reference species belongs to. If scores is not that bad, we will go through all the species in this genus to determine which species it has most similarity with.

Threshold here is very important. We didn't come up with a fixed expression to calculate it by the due date. We extract 30 species from 20 genus from the reference set and set the threshold after observing the result from those alignments. Also, we used the average score from the scores of 1000 test sequence in the .fna file.

In order to run the code, please change the value of 'projectpath' to the actual path that 'AAA.zip' you extract to. And also, make sure inside the 'projectpath', there is a /data/index folder and a /bowtie2-2.1.0 folder.

5 Results

After running `bowtie_with_skipping.py` and `without_skipping`, we can tell with **X species in X genus, skipping method is X times faster**.

6 Conclusion

In our project, we implemented and compared three methods for the phylogenetic classification of metagenomic reads. Under our simulated data,

For future work, we would ideally run our methods on reference genomes from the full RefSeq and using simulated reads from a larger clade in the tree. While we aimed to do so for our project, we found that we found that the limitations on cluster we were using were too stringent and we had to vastly scale down the scale of our reference set. In particular,

we believe that the IMM method will improve as we scale up the data and the number of reference genomes. We also would like to scale our comparisons to real data instead of simulated reads, especially a dataset that has been well studied and we have some idea of what the true classifications should be. It's also important to compare our methods to the existing methods. While we believe we sufficiently improved upon existing methods, we were unable to get existing software running on the cluster, aside from FACS.

Other things we would like to test is the ability of our methods to detect *de novo* reads. A real metagenomics sample will have reads from many bacteria which have never been discovered, which we would like our methods to classify to the correct genus, but identify as being from an unknown genome. One easy way we can explore this is by simulating reads from a reference set of genomes, and then classify against the reference set multiple times using a cross validation procedure with one reference genome left out from each classification.

In summary, we identify three methods for phylogenetic classification for metagenomics. While our initial results look very promising, we need to scale up the size of our simulations and set of reference genomes to fully quantify the accuracy, sensitivity, and specificity of these methods. We remain hopeful that our methods will perform well on a more realistic simulation and can be applied to real metagenomics data.

7 References

Chen K, Pachter L (2005). Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLoS Comput Biol*, 1(2): e24. doi:10.1371/journal.pcbi.0010024

Brady A, Salzberg S (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* 6, 673 - 676 (2009)

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* 2009 Jan; 37 (Database issue):D32-36.

Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs *BMC Bioinformatics* 2012, 13:92

Stranneheim H, Kller M, Allander T, Andersson B, Arvestad L, Lundeberg J Classification of DNA sequences using Bloom filters *Bioinformatics* (2010) 26 (13): 1595-1600.

Langmead B, Trapnell C, Pop M, Salzberg S Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* (2010) 10:R25

Richter DC, Ott F, Auch AF, Schmid R, Huson DH MetaSimA Sequencing Simulator for Genomics and Metagenomics. PLoS ONE (2008) 3(10): e3373. doi:10.1371/journal.pone.0003373