Johns Hopkins University

Computational Genomics

EN 600.639

Classification of Metagenomic Reads

Author:

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.

4 Methods and software

To simulate our data, we

5 Results

6 Conclusion

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.