BMC Bioinformatics

POSTER PRESENTATION Open Access

Alignment-free methods for metagenomic profiling

Shanshan Gao, Diem-Trang Pham, Vinhthuy Phan*

From 14th Annual UT-KBRIN Bioinformatics Summit 2015
Buchanan, TN, USA. 20-22 March 2015

Background

The primary goal of metagenomic studies is to analyze and evaluate the rich microbial communities present in all natural environments. The construction and utilization of a large index required by alignment-based methods for thousands of microbial genomes can be computationally prohibitive. To avoid this computational cost, we investigated three different variations of an alignment-free method for profiling abundances of microbial communities.

Materials and methods

The main idea of the method is reformulate the problem of determining abundance of microbial genomes as finding optimal solutions of linear equations that satisfy specific constraints. A set of genomic markers for the entire set of genomes is represented by a matrix F, where $\mathbf{F_{ij}}$ represents the frequency of marker **i** in genome **j**. The occurrence vector **b** represents the frequencies of markers in reads. We would like to find an optimal solution **x**, the abundance vector in which $\mathbf{x_j}$ represents the abundance of genome **j**. To find the abundance vector x, we solve the linear equation $\mathbf{Fx = b}$. The methods to choose F and b are the key factor to find the optimal value of x. We introduced a concept of *genome specific marker* (GSM), which is a kmer that occurs in only one genome and no other. We exhaustively determine such markers from the entire dataset and represent the frequencies of these markers in the matrix F. Given a set of reads from a metagenomic dataset, we compute the frequency of GSM as b. Then, three variations can be formulated, respectively, as a linear programming problem (LP), a least-square approximation problem (L2), and an L1-approximation problem.

Results

So far, our investigation on two data sets consisting of 100 and 1105 microbial genomes showed that the linear programming formulation (LP) yielded the best prediction of abundances of microbial genomes. This result was consistent across different levels of abundances. The LP variant also achieved better results across the board compared to a popular metagenomic profiler, FOCUS[1], which was found to be superior to other methods.

Conclusions

In the future, we need to investigate deeper into the matrix F which consists not only of the GSM, but also the kmers that occur in more than one genome.

Published: 23 October 2015

Reference

1. Silva GGZ, *et al*: "FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares.". *PeerJ* 2014, **2**:e425.

doi:10.1186/1471-2105-16-S15-P4
Cite this article as: Gao *et al.*: Alignment-free methods for metagenomic profiling. *BMC Bioinformatics* 2015 **16**(Suppl 15):P4.

* Correspondence: vphan@memphis.edu
Department of Computer Science, University of Memphis, Memphis, TN 38152, USA