FOCUS: An Alignment-Free Model To Identify Organisms In Metagenomes Using Non-Negative Least Squares

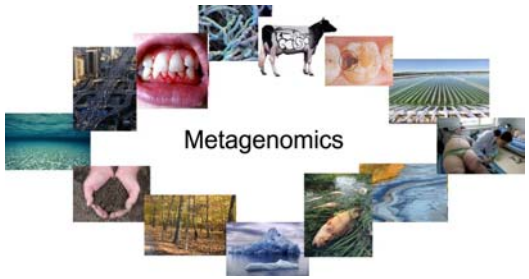
Genivaldo Gueiros Z. Silva and Robert A. Edwards

SAN DIEGO STATE UNIVERSITY

COMPUTATIONAL SCIENCE RESEARCH CENTER, SAN DIEGO STATE UNIVERSITY

Introduction

Microbes are more abundant that any other organism, and it is important to understand what those organisms are doing and who they are. In many environments more than 99% of the members of the microbial community cannot be cultured.



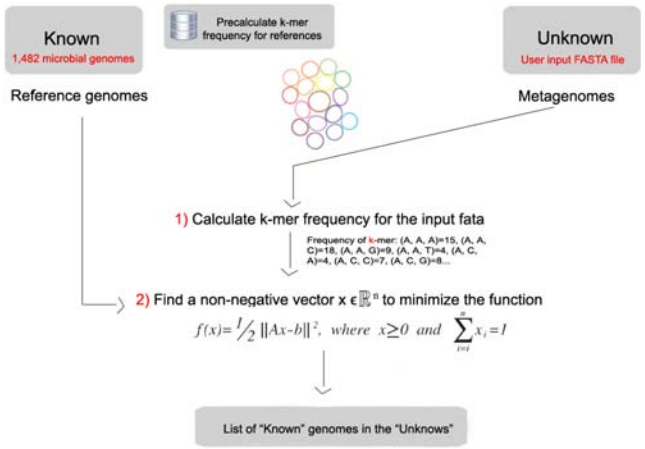
One of the major goals in metagenomics is to identify organisms present in the microbial community from a huge set of unknown DNA sequences; this profiling has valuable applications in multiple important areas of medical research such as disease diagnostics. Nevertheless, it is not a simple task, and many approaches that have been developed are slow and depend on the read length of the DNA sequences.

FOCUS

An innovative and agile composition based approach using non-negative least squares to profile and report abundant organisms present in metagenomic samples and their relative abundance. The results show that our approach accurately predicts the organisms present in microbial communities.

Methods

FOCUS is a model written in Python. The Figure below illustrates how **FOCUS** models the data.



Non-Negative Least Squares (NNLS)

NNLS is useful to solve problems, such as metagenome profiling, where there cannot be negative values for the fitted parameters.
In FOCUS, the reference dataset is represented by the matrix A. b defines the user input data, where m is the number of normalized k-mer frequencies and n is the number of reference species. The goal is to compute the set x that explains as well as possible the abundance of each species from the training set in the user input.

Results

The application was evaluated using simulated and real microbial metagenomes, and the results show that our modeling approach provides a completely different, and exceptionally fast and accurate, method for predicting the organisms that are there, though not necessarily which sequences they contribute to the sample.

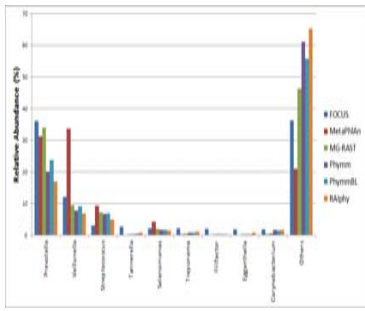


Figure 1: Genera-level taxonomy classification for the human oral cavity under disease metagenome using FOCUS, MetaPhlAn, MG-RAST, Phymn, PhymnBL, and RAlphy.

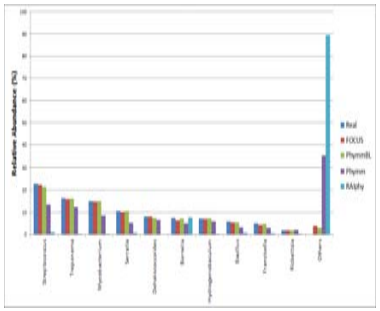


Figure 2: Genera-level taxonomy classification for the SimShort dataset using FOCUS, Phymn, PhymnBL, and RAlphy.

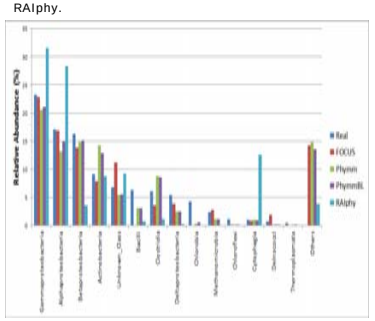


Figure 3: Class-level taxonomy classification for the SimHC (FaMeS) dataset using FOCUS, Phymn, PhymnBL, and RAlphy.

Tool	Running Time
FOCUS	20 seconds
MetaPhlAn	3 minutes
RAlphy	2 hours
Phymm	6 days
Phymmbl	6 days
MG-RAST	4 days to 1 week

Figure 4: Running time comparison for the human oral cavity under disease metagenome using FOCUS, MetaPhlAn, MG-RAST, Phymn, PhymnBL, and RAlphy.

Conclusions

FOCUS helps to identify which organisms are present in metagenomes; the algorithm presented will help biologists explore the microbes present in their samples.

Web-based version

http://edwards.sdsu.edu/**FOCUS**

Please contact me at genivaldo.gueiros@gmail.com for further information.