Host prediction for viral metagenomes using oligonucleotide profiles

Michiyo Wellington-Oguri[1], Robert Schmieder[1], Barbara Bailey[2], Robert A. Edwards[1,3,4], Bas E. Dutilh[1,3,5]

1 Department of Computer Science; 2 Department of Mathematics; 3 Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA; 4 Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA; 5 Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Geert Grooteplein 28, 6525 GA, Nijmegen, The Netherlands

INTRODUCTION

Viral metagenomics is a method of sequencing virus-like particles isolated from an environment. It has led to the discovery of large numbers of previously-unsequenced viruses[1], whose roles in the ecosystem remain elusive.

Frequencies of **oligonucleotides**, short sequences of nucleic acids, are characteristic of a genome[2]. Thus, they are a useful tool for classifying sequence fragments[3]. **Random Forest (RF)** is a classification algorithm capable of integrating many variables[4]. It is essentially a group of decision trees, each of which is based on a subset of the training data.

We use a Random Forest to classify viral metagenomic fragments by host based on oligonucleotides frequencies.

METHODOLOGY

Data Compilation — Compiled 1,465 viruses with known host annotations and complete genomes from Genbank.

Host Group Creation — Grouped viruses based on host taxonomy, minimum 10 viruses per host group: 90 host groups.

RF Optimization — Tested the use of different subsets of oligonucleotide frequencies as RF predictors to minimize misclassification.

RF Testing — Used the RF to classify simulated and real metagenomes sequences and evaluated the precision of the predictions.

RANDOM FOREST OPTIMIZATION

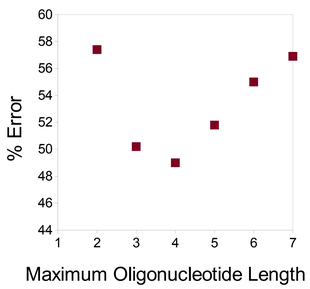

Figure 1: The classification error of the RF depends on the length of the training oligos.

The frequencies of short oligonucleotides were calculated from the combined forward and reverse strand of each viral genome. RFs based on incremental subsets of these predictors were trained and tested (Figure 1). Oligonucleotides of 1 through 4 bases in length produced the lowest error rate (49%). This is highly accurate compared to the expected error rate of 98.9% if classification were random.

RANDOM FOREST TESTING

The optimal RF was tested using simulated metagenomic sequences with varying **lengths** and **error rates**. Sequences with lengths similar to assembled contigs (>500 nt) can be classified accurately; longer sequences (>1000 nt) can also be classified with high precision (Figure 2).

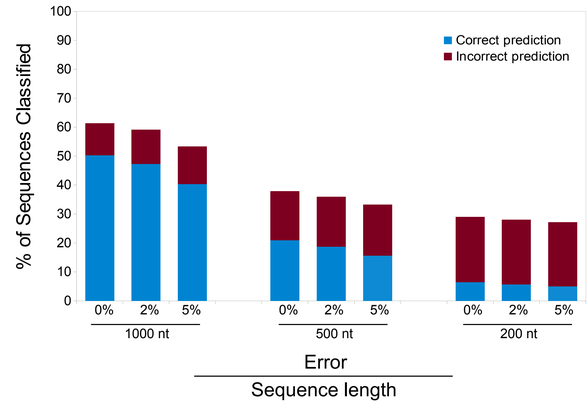

Figure 2: Performance of RF trained on oligonucleotides of length 1-4 nt on simulated metagenomic reads of varying lengths and error rates.

The confusion matrix is presented as a “heat map” where the hosts are ordered by taxonomy, so misclassifications near the diagonal show classification into related host groups (Figure 3). A larger training set may allow us to create host groups that better correspond to phylogenetic boundaries. Additionally, training with more genomes may allow additional characteristics of these host groups to be identified.

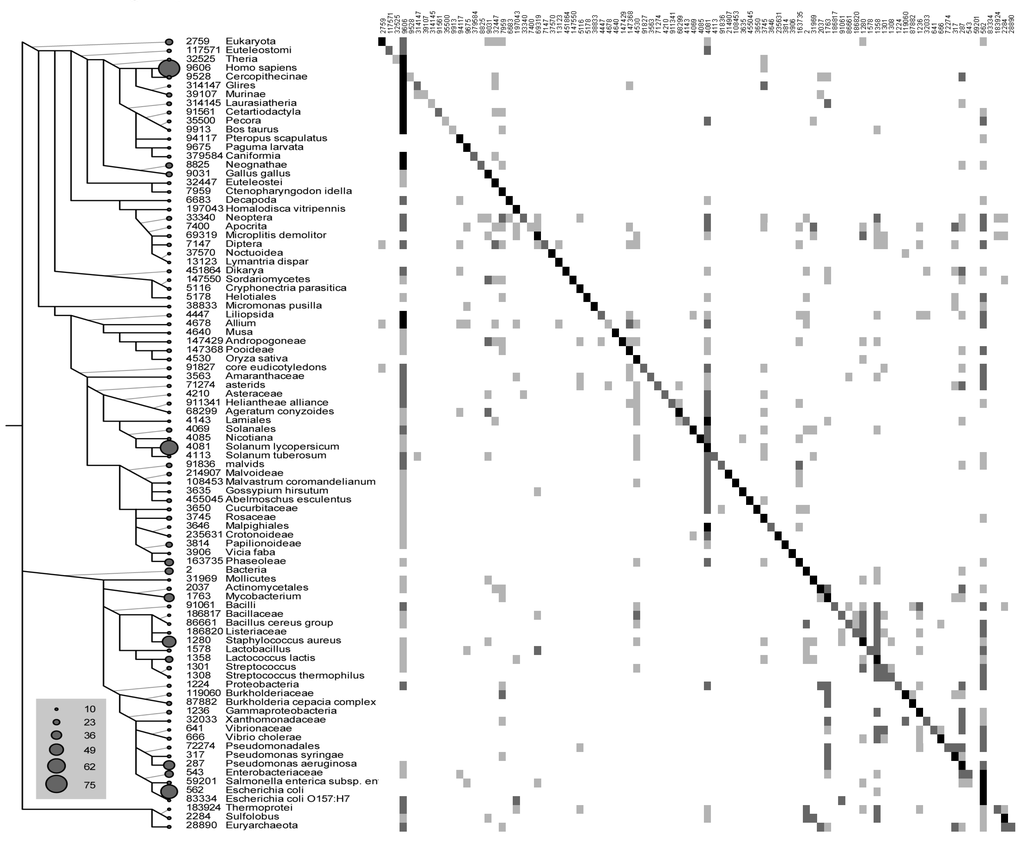

Figure 3: Predictions for reads from a simulated metagenome (length=1000, 0% error) by the optimal RF. Each cell in the matrix shows the fraction of sequence reads from that row predicted to belong to that column.

HUMAN GUT VIROMES

We predicted hosts for contigs assembled from 12 viral gut metagenomes[5] and converted back to reads to estimate abundances. For each sample, 22-48% of the reads were classified, most as bacteriophages that infect, for example, known gut bacteria (Figure 4).

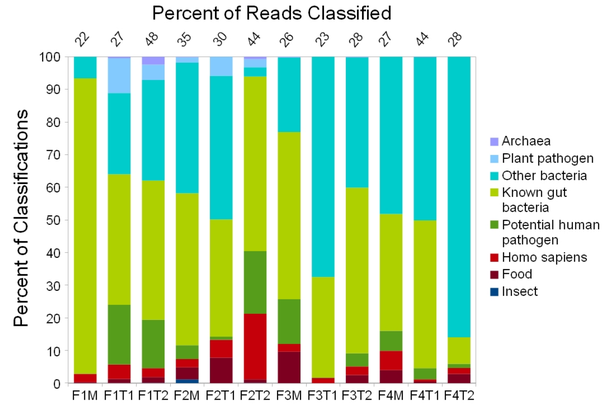

Figure 4: Summary of hosts predicted in each gut sample. Samples were taken from mothers and twins daughters of four families[5].

CONCLUSIONS

We develop a tool to predict the host of a virus from genomic sequence fragments.

Assembled metagenomic contigs can be classified accurately.

The tool is used to predict host groups for viral gut metagenomic sequences.

REFERENCES

[1] Mokili,J.L. et al. (2012) Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, 2: 1-15.
[2] Deschavanne,P. et al. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.*, 16, 1391-1399.
[3] McHardy et al. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4: 63-72.
[4] Breiman,L. (2001) Random Forests. *Mach. Learn.*, 45, 5–32
[5] Reyes,A. et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, **466**, 334-8.

SAN DIEGO STATE UNIVERSITY