Predicting Viral Host based on Metagenomic Data

Jeffrey Jeyachandren, October 2020

Domain Background

Viruses have many types of hosts. They can infect plants, animals, bacteria, and fungus. Modern high-throughput DNA profiling techniques have led to an abundance of virus discoveries[1]. However, for most of these discoveries, the virus host remains unclassified. Classification of a virus' host is important for many reasons. It can tell us about the evolutionary relationship between viruses, between viral hosts, and most likely candidates for new hosts. Thus, research into the viral genome has been an important area of healthcare research and computational genomics.

Studying the viral genome involves looking at raw virus genome sequences. The most common approach to describing a virus is by evaluating their sequence and phylogenetic similarities with known viruses[1]. This can be computationally expensive, since a full genomic characterization would require looking at all possible k-length subsequences (k-mers) in the viral DNA. Additionally, viruses have an extremely high rate of mutation that is compounded by host-specific interactions occurring at the genomic level[2]. Looking at DNA sequences also includes all possible manifestations of amino acids in the virus' cellular-level machinery, which includes the DNA, RNA, and protein (or proteome). To address complexities in raw sequencing data, meta-genomics has emerged as a way of looking at abstracted features of the DNA. Using abstracted features from raw sequences, we can make predictions about genomic data.

Recently, researchers have had success with using meta-genomic data as an accurate predictor for a viral host[2]. Incorporating features from k-mer compositions and predicted protein domains, scientists were able to greatly improve our ability to classify a virus' host based on simple DNA abstractions. Inspired by this, I would like to evaluate other meta-genomic features such as GC%, CDS, and genome size. These features are defined as follows:

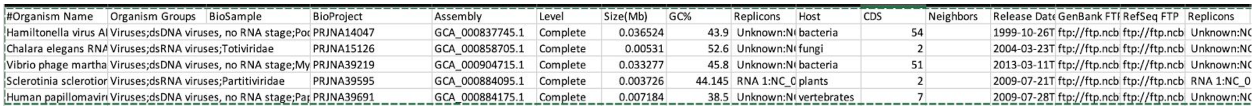
- **GC%:** the percent of guanine or cytosine in DNA. It is highly varied among different types of organisms. Molecularly, it provides stability to the DNA strand and prevents protein denaturing.
- **CDS:** A coding sequence in the DNA. A count of CDS would refer to the number of unique segments in the DNA that start with a starting codon and end with a stopping codon.
- **Genome size:** the size of the sequence measured in mega base pairs (Mb, or 1,000,000 base pairs).

Problem Statement

Currently, only 5% of the viral genomes in the IMG/VR databases are labeled with an associated host[3]. This data is often unavailable since viral DNA samples are taken from the environment and not directly from the hosts. To resolve this, we need to computationally predict and assign viruses to their host to improve the quality of this data set. Unfortunately, most techniques involving virus host prediction are limited to correlations with the host genomes and analysis of k-mers. K-mer analysis is especially problematic since viruses are subject to rapid mutation[3]. Thus, we must look to the field of meta-genomics to help make predictions about viral hosts.

Dataset and Inputs

The dataset and inputs for this project come from the Kaggle dataset, "Genome Information for Sequenced Organisms"[4]. Specifically, we will be using the virus.csv file that is provided. The file contains information about 7363 known viruses. For each virus, the dataset has information about it's size (Mb), GC%, replicons, host, CDS, and more. The model will be using the Size, GC%, and CDS as features to target predictions on the host. Here is a sample of the first five entries from the data:



#Organism Name	Organism Groups	BioSample	BioProject	Assembly	Level	Size(Mb)	GC%	Replicons	Host	CDS	Neighbors	Release Date	GenBank FTI	RefSeq FTP	Replicons
Hamiltonella virus Al	Viruses;dsDNA viruses, no RNA stage;Poc		PRJNA14047	GCA_000837745.1	Complete	0.036524	43.9	Unknown:N(	bacteria	54		1999-10-26T	ftp://ftp.ncb	ftp://ftp.ncb	Unknown:NC
Chalara elegans RNA	Viruses;dsRNA viruses;Totiviridae		PRJNA15126	GCA_000858705.1	Complete	0.00531	52.6	Unknown:N(	fungi	2		2004-03-23T	ftp://ftp.ncb	ftp://ftp.ncb	Unknown:NC
Vibrio phage martha	Viruses;dsDNA viruses, no RNA stage;My		PRJNA39219	GCA_000904715.1	Complete	0.033277	45.8	Unknown:N(	bacteria	51		2013-03-11T	ftp://ftp.ncb	ftp://ftp.ncb	Unknown:NC
Sclerotinia sclerotior	Viruses;dsRNA viruses;Partitiviridae		PRJNA39595	GCA_000884095.1	Complete	0.003726	44.145	RNA 1:NC_0	plants	2		2009-07-21T	ftp://ftp.ncb	ftp://ftp.ncb	RNA 1:NC_0
Human papillomavir	Viruses;dsDNA viruses, no RNA stage;Pap		PRJNA39691	GCA_000884175.1	Complete	0.007184	38.5	Unknown:N(	vertebrates	7		2009-07-28T	ftp://ftp.ncb	ftp://ftp.ncb	Unknown:NC

From our target (host), the distribution of the classes is as follows:

algae	34
archaea	4
bacteria	3787
fungi	55
human	1
invertebrates	107
invertebrates, plants	11
plants	306
protozoa	34

vertebrates	550
vertebrates, human	2440
vertebrates, invertebrates	17
vertebrates, invertebrates, human	8
(blank)	
Grand Total	**7354**

Solutions Statement

The proposed solution to the problem is to evaluate the predictive potential of meta-genomic data. The meta-genomic data that will be used includes viral genome size, GC%, and CDS count. These features will be used to predict the viral host. They will then be compared to benchmarks that have been established in the scientific literature that use a different set of metagenomic features (k-mer composition and protein domains)[3].

Benchmark Model

Researchers have most recently used SVM to predict the viral host based on metagenomic features[3]. The accuracy of the model is evaluated using the area under the ROC curve, or the AUC. To remain consistent with the scientific literature, I will be using the same benchmarks for my model. In previous research, AUC scores above 0.5 were indicative of a predictive signal in the feature set. A plot of true positive rates and false positive rates will also be used, with a target of 75% for both. This is a benchmark that has been established in the recent scientific literature.

Evaluation Metrics

Since the ML model will be built using SVMs, the evaluation metric will be the AUC. An AUC score of 0.5 or more will be indicative of a predictive signal in the feature set. An AUC score of 0.75 or more will be considered on par with the work that has been done in the scientific literature. Additionally, a metric of true positive vs. false positive rates will be used to further evaluate the accuracy of the model.

Project Design

The project will proceed as follows:

1. The dataset will be modified to create a balanced binary dataset for each host type. One hot encoding will be used to separate the host column into distinct columns for each host type (bacteria, invertebrate, plant, vertebrate, etc.).
2. The Size and CDS columns will be normalized.
    a. The dataset will be split into training and test sets, with a split of 0.8 and 0.2.
3. If this split is too small, or if there is an imbalance in the data with respect to known viral hosts, cross validation will be used during the training.
4. SVM algorithms will be used via the SciKit-Learn python library. The model will be trained on the training set (and also via cross validation if necessary).
5. The model will be evaluated by taking the AUC score.
6. The model will be further evaluated by looking at the true positive and false positive rates.

Jeffrey Jeyachandren

[1] Raj, A., Dewar, M., Palacios, G., Rabadan, R., & Wiggins, C. H. (2011). Identifying hosts of families of viruses: a machine learning approach. PloS one, 6(12), e27631. https://doi.org/10.1371/journal.pone.0027631

[2] Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 6.3, Viruses: Structure, Function, and Uses. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21523/

[3] Lodish H, Berk A, Zipursky SL, et al. Molecular Cell Biology. 4th edition. New York: W. H. Freeman; 2000. Section 6.3, Viruses: Structure, Function, and Uses. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21523/

[4]

https://www.kaggle.com/camnugent/genome-information-for-sequenced-organisms?select=viruses.csv
/