Ecologically informed viral sequence detection in aquatic metagenomes using machine learning

THE UNIVERSITY OF ARIZONA®

Introduction

Modern 'omics allows for the exploration of in situ relationships between phages and their hosts, to provide new insights into the impact of viral populations on numerous biological systems. Retrieving viral sequences from metagenomes is critical to bacterial understanding host-viruses interactions in a given ecosystem. Homology to known viral genes is a primary method to retrieve viral from complex metagenomes. contigs However these approaches limit the discovery of novel viral sequences with no similarity to previously known viruses.

Recently, VirFinder, a new tool to detect viral sequences in bacterial metagenomes using a machine learning method, was released [Ren et al. 2017]. This method distinguishes viral from bacterial sequences based on their kmer signatures, rather than through homology based searches to viral genes.



Current Limitations: VirFinder relies on a model trained on known viral ar genomes from the RefSeq database, that shows a bias toward the detection abundant viral groups in reference databases [Ren et al. 2017].



Models trained on RefSeq genomes show

Fig2. AUC ROC of 3 independantly trained logistic classifiers trained on RefSeq bacterial and viral (DNA virus) genomes broken down to a size A. The models performances are evaluated using RefSeg genomes broken down to 5000bp and not used for the training step and grouped by B. Evaluation Bacterial examples are provided by RefSeq bacterial genomes borken down to 5000bp and not used for the training step. Evaluation provided by contigs from JGI img/vr broken down to 5000bp and grouped by isolation environment type (annotation from GOLD).

Viromes represent a large collection of viral sequences that are unbiased methods and cover a wide variety of ecosystems. These sequences ar interesting source of information about viral k-mer signatures.

We present a scalable computational framework to train machine learning on curated ecosystem-specific metagenomic contigs. This approach aims to viral sequence detection even in ecosystems less studied, where a smaller amount of phages have been previously isolated and sequenced, and provide the user with ecosystem specific predictions. Finally, our approach takes into account the possibility of eukaryotic contamination that is fundamental in various environments.

Alise J. Ponsero and Bonnie L. Hurwitz Department of Biosystems Engineering, University of Arizona

Creating marine-specific models

As a proof of concept, we provide the results obtained for a marine specific model, trained using the Tara Ocean assemblies dataset [Sunagawa et al. 2015]. Viromes and microbiomes always carry small amount of sequence contamination. However, in order to ensure good performance for the models, it is necessary to provide a clean training set.



for 3 independently trained logistic classifiers, trained on Tata Ocean contigs (3000bp), after different data cleaning steps The different cleaning steps are described in the pipeline schema

Better resiliance to contaminations

VirFinder training set is void of any Eukaryotic training examples. The tool is thus naïve to sequences of Eukaryotic origin, which can lead to an increased false positive rate in certain ecosystems were sequences from micro-eukaryotes are naturally present in metagenomes. Including protists and micro-eukaryotes genomes in a RefSeq trained model is limited by the low number of sequences currently available for those organisms.

Training classifiers using metagenomes allows the model to take into account microeukaryotes sequences, when they are present naturally in the ecosystems.



Fig5. AUC ROC of 3 independently trained logistic classifiers trained on RefSeq bacterial and viral (DNA virus) genomes broken down to a size of 5000bp or on Tara Ocean viromes and microbiomes (contigs size of 5000bp) The models performances are evaluated using RefSeq bacterial, viral and eukaryotic genomes isolated in a marine anvironment and broken down to 5000pb. These genomes were not used for the training step.

Similarity search
Kmer matches
Kmer signature
genomes assembly inces from metagenomes
nd bacterial of the most
ingevaluationbacterial mes 2018)RefSeq bacterial genomes (May 2018)DNA nomes 2018)RefSeq DNA phages (May 2018)
g evaluation cterial RefSeq bacterial genomes 18) (May 2018) NA RefSeq DNA mes phages 18) (May 2018)
e of 5000bp. viral host groups. tion viral examples are
d by cultivation are a vast and
models directly ensure reliable



	Training Tara	Training RefSeq	evaluation
ial Ies	Tara Ocean assemblies (0.2-1.3-3µm)	RefSeq bacterial genomes (May 2018)	RefSeq marine bacterial genomes (May 2018)
les	Tara Ocean assemblies (<0.2µm)	RefSeqDNA virus genomes (May 2018)	RefSeq marine viral genomes (May 2018)
otic Ies	Tara Ocean assemblies (0.2-1.3-3µm)		RefSeq marine protist and algae genomes (May 2018)

An ensemble method to improve detection of rare events

Even when using state-of-the-art tools, the detection of rare events often leads to a low precision : TP/(TP+FP). This can be easily explained using the Baye's theorem.

		expected			
		viral	non-viral	total	
prodictod	viral	211	189	400	
predicted	non-viral	39	4561	4600	
total		250	4750	5000	

Fig 7 : confusion matrix for a logistic classifier trained on Tara Ocean assemblies and evaluated on RefSeq genomes grinded at 5000pb (250 viral examples and 4740 bacterial examples)

It is possible to improve the precision obtained in such conditions using an ensemble approach [Ali and Pazzani, 1996]. The ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A simplest but powerful ensemble is the Bootstrap Aggregation technique (Bagging).





Fig8. TPR and precision for models trained on Tara Ocean assemblies. The evaluation examples are RefSeq genomes from prokaryotes and DNA viruses isolated in marine environment and broken down to 5000bp

Conclusions

- curated metagenomes
- ecosystem

References

- 173-202.
- viral sequences from assembled metagenomic data, Microbiome, 5, p. 69

 $P(A|B) = \frac{P(B|A)P(A)}{P(A)}$

Eq 1 : Baye's theorem formula

Viral sequences in microbiomes constitute often less than 10% of the total sequences. Thus the use of a tool like VirFinder will lead to more false positives in most conditions.

• We present a framework allowing the training of ecosystem-specific models directly on • This approach allows a better resilience toward eukaryotic sequences in the studied • Training simple ecosystem-specific models allows for easily implemention of an ensemble method, that improves the detection of viral contigs in metagenomes.

• Ali, M., & Pazzani, M.J. (1996). Error reduction through learning multiple descriptions. Machine Learning, 24(3),

• Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying • Sunagawa S. et al (2015) Structure and function of the global ocean microbiome, Science, 348 (6237)