Rapid and Portable Genome Classification System

Masooda Omari, Tyler Barrus, Mark Sanders, Shane Mitchell, and Dr. Sterling Thomas

Noblis Inc, Reston, VA

Introduction

Noblis development of Rapid and Portable Genome Classification System (RPGCS) is a biodefense and forensics technology that performs identification of bio-threat agents in a clinical or environmental sample. Our solution includes a technical approach that balances accuracy and speed in diagnosis with an intuitive and easy-to-use interface. It functions in a unified process by combining modern genomics techniques, statistical modeling, data analytics and machine learning algorithms. It consists of a probabilistic data structure called a Bloom Filter that allows users to insert Whole Genome Sequencing (WGS) complete genomes and k-merize raw read sets to query for organisms present with a high confidence probability. This technique allows for the rank ordering of likely organisms by confidence. The innovation of this technology is, translating big genomic data from the reference library of National Center Biotechnology Institute (NCBI) and genomic samples into compacted filters. This makes processing of a large sized data possible on a small portable device such as Raspberry Pi.

Capabilities

- Combat outbreaks of bioterrorism on the field
- Fast response to infections in hospital or clinical setting
- WGS analysis requires no advanced computational resources
- Low storage requirement
- Fast membership checking
- Low false positive probability
- Can operate on a raspberry pi
- No internet connectivity required
- Classify any type of organism
- Highly parallelizable
- Small storage for reference library

Methods Flow Chart

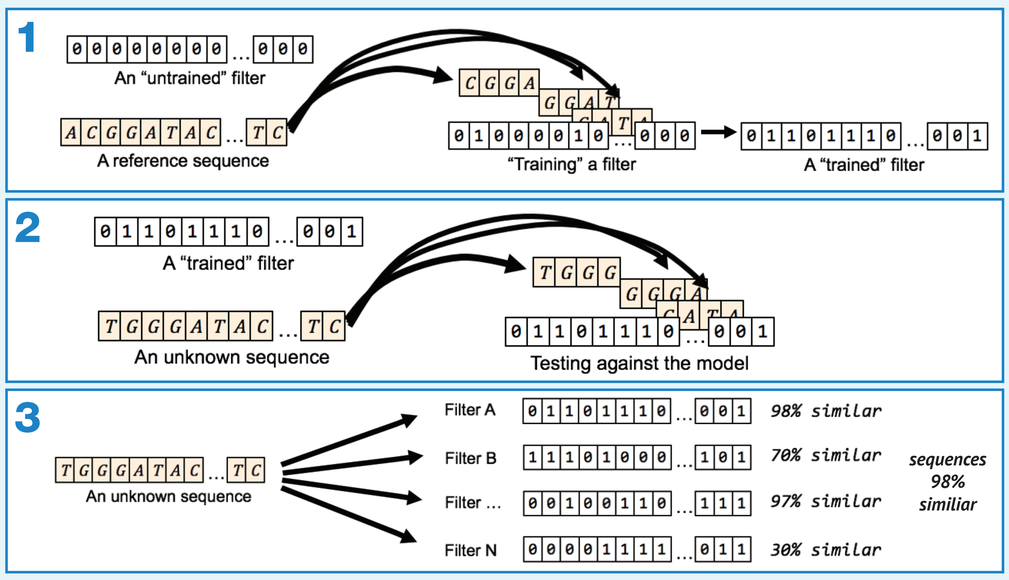


Figure 1. Steps involved in Bloom Filter data structure.

Methods

RPGCS is build based on a probabilistic data structures, such as Bloom Filters (1) that tests whether a data point is a member of a set (2) . The main goal of the technology is to filter out possible organisms in a clinical sequenced sample. This process will result in a most likely organism identity based on the matches found in the data structure. The identification can be performed on any type of organism. It is currently capable of performing bacterial and viral sequences classification.

Bloom Filters have distinct qualities that make them useful for quick identification of information, including a zero-percent false-negative rate and a settable, non-zero, false-positive rate. In other words the Bloom Filter allow users to insert data, such as WGS read data, into the data structure and query it to determine if the data is most likely present or definitely not present. This probabilistic data structure gets the zero false-negative rate by virtue of how data is inserted and queried. Upon insertion, each data point is hashed and converted into two or more bit positions. These bits are then set to true (1) within the bit array. It is possible that multiple data points overlap on one or more bits. When looking up a data point to see if it has been included, the data or this case read sets of samples to be looked up is hashed in the same manner used to hash the original data or reference genomes . All bits are checked to verify that they are all set. If any bit is not set to true, one can be confident that the data point was never inserted (see Figure 1). One of the main advantage of the Bloom Filter data structure is that it stores any form of data in hash tables and not the object itself which allows to save a high amount of disk space (4) In addition, unlike a set or traditional database, the data in a Bloom Filter index library is irretrievable once it is converted into Bloom Filters.

Reference genomes: As a first step in building the the Bloom Filters index library for all viruses , more than 7000 curated and complete genomes were downloaded from NCBI. These genomes were then k-merized and stored in the Bloom Filter index library.

Data: The next step involved processing the read sets into 17 bp k-mers of sub-sequences and determining if these sub-sequences are present in a larger genome. This list of k-mers were then queried across all Bloom Filters. K-mers that matched the species represented by the Bloom Filter was signaled as a match. For this study read sets of random viruses from distinct species were downloaded from the SRA database.

Results and Discussion

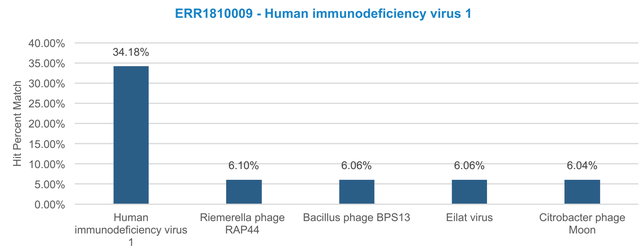
The read sets results were produced in the excel spread sheet. A rank order or probability score was assigned from each matched reference genome to that read set. The graphs below shows the top 5 complete genome hits with their respective read sets and their hit percent kmers.

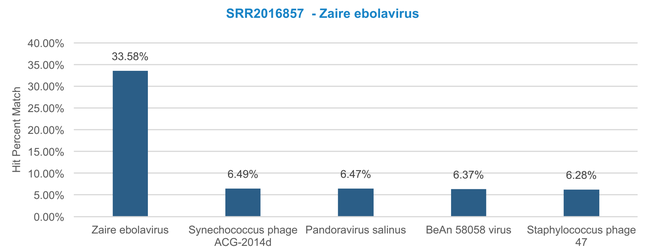
Size:

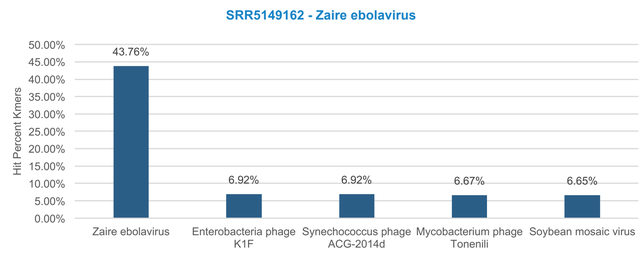
- Over 7000 complete viral genomes uncompressed were 252 MB, when processed into blooms compressed into 195 MB

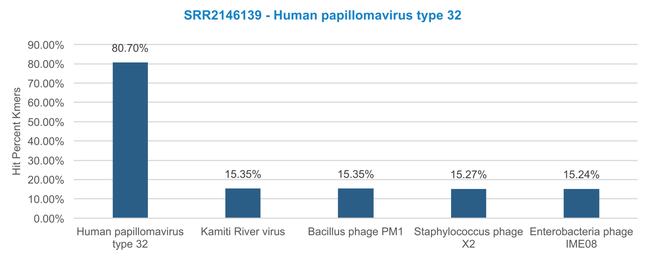
Speed:

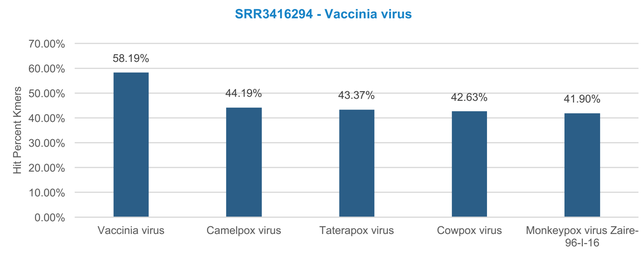
- The blooms of more than 7000 genomes were generated in 7 minutes - single threaded
- The classification of a read file of 19 MB took 5 minutes on average
- The classification of a read file of 7 GB took about 6 hours on average











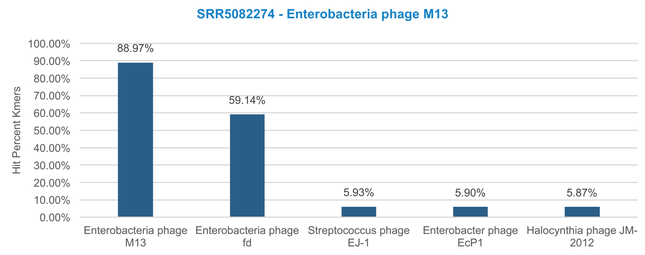


Figure 1.2. Viral genomes of interest classified with the high probability.

Future Developments

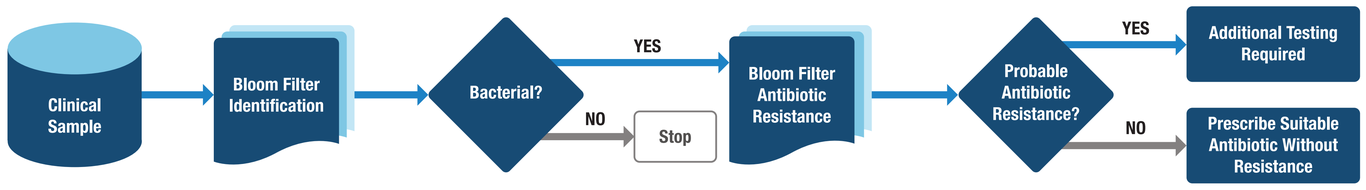


Figure 2. If the organism is deemed to be bacterial, another set of Bloom Filters will be used to determine a resistance profile.

Conclusion

RPGCS is an automated technology that accepts sequence data in fasta format as input and returns pathogen identity present in the sample. Bloom Filters data structure used in the infrastructure of this technology compactly represent large sets of information and test those sets for membership (within some computable error boundary) (3). By using a collection of Bloom Filter classifiers, we train a machine learning classification model using very large libraries of reference genomes and then rapidly classify a given sample against that model, with the model providing a ranked order of likely organism matches. Highly ranked results were accomplished that are as good as more computationally intensive approaches, with reasonable disk utilization characteristics. The analysis from a very initial development of the data structure demonstrated that Bloom Filter data structure classified the correct viral organisms based on its high probability metrics .

Future development to the algorithm will leverage upon the false positive rate, ideal kmer length, and coverage depth across the complete genome. These optimizations of the technology will expand in its capaiblities in antibiotic resistence identification, synthetic genome identification, and human dna identification in a clinical or non clinical genomic sample.

Acknowledgement

The Big Data Lab and special thanks to Tyler Barrus and Mark Sanders for playing a major role in the ideation and development of the technology.

References:

1. Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. Commun. ACM 13, 7 (July 1970), 422-426. DOI=http://dx.doi.org/10.1145/362686.362692.
2. Wikipedia, the free encyclopedia. Bloom filter. https://en.wikipedia.org/ wiki/Bloom_filter. Accessed May 4 2017.
3. Barrus T, Bloom, 2016. Github Repository, https://github.com/barrust/Bloom
4. https://www.slideshare.net/deveshmaru/bloom-filters