INFECTIOUS DISEASE EPIDEMIOLOGY APPLIANCE (IDEA)

Terry Antony[1], Lucia Fernandez[1], Kyle Milligan[1], Rui Ponte[1], Meena Sengottuvelu[1], Tianwen Chu[2], Frank D'Ippolito[2], Michael S. Atkins[2]

[1]University of Maryland, College Park, MD; [2]FedCentric Technologies, LLC., College Park, MD

Introduction

The flavivirus genus consists of 73 mosquito and tick borne viruses that pose a considerable threat to public health. A deeper understanding of the molecular evolution of flaviviruses is needed to guide public health decisions and to prevent flavivirus epidemics, such as the 2015-2016 Zika virus epidemic in the Americas. Comparative genomic analysis can lead to many discoveries regarding molecular evolution and epidemiology, but computational power remains a limiting factor due to the size of the data and the complexity of the algorithms. Here we present the Infectious Disease Epidemiology Appliance or IDEA, a tool currently in development at FedCentric Technologies for the rapid analysis and visualization of over 6,200 genomic and 93,000 proteomic flavivirus sequences. The long term goal with IDEA is to create a robust epidemiological tool than can be used to quickly identify, monitor, and contain any infectious disease outbreak wherever it occurs, from nosocomial infections in a local hospital to remote hot zones anywhere in the world.

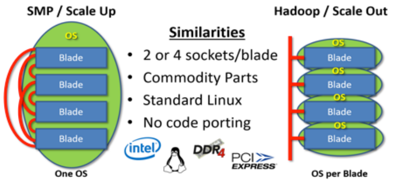
Methods

Hardware

FedCentric Labs SGI UV2000 system

- x86/Linux Operating System
- Scales up to 4092 cores
- Scales up to 64TB RAM (all data in memory)
- Very low latency, high performance

Figure 1: Scale-Up vs. Scale-Out Architecture



Similarities

- 2 or 4 sockets/blade
- Commodity Parts
- Standard Linux
- No code porting

Scale Up

- One large server
- Scales up by adding memory and CPUs to the same system
- Blades are linked with high speed internal bus

Scale Out

- Cluster of small servers
- Scales out by adding more physical servers
- Servers are linked through external network connection

Software

Apache Spark

- A big data processing framework
- Driver node manages parallel operations carried out by executor nodes
- Implements graphs through its GraphFrame data structure

Shiny

- An RStudio package used for web application development

Methods (continued)

Data

- Whole genome and protein sequences from the NIAID Virus Pathogen Database and Analysis Resource (ViPR)

Graph Model

- The graph consists of virus sample nodes linked together with distance calculation edges
- Each sample has 14 protein sequences, 1 polyprotein sequence, and 1 nucleotide sequence
- 6,201 nodes and 307,569,600 edges

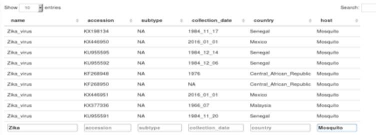
User Interface

- Edit distance comparison between protein and nucleotide sequences
- Phylogenetic tree builder with maximum likelihood
- Multiple Sequence Alignment using MAFFT (Multiple Alignment using Fast Fourier Transform)
- Choropleth map view

Results

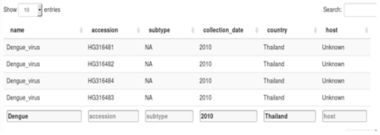
Sequence Comparison

- All-to-All comparison of protein and nucleotide sequences
- Distance value is calculated with edit distance corrected with Jukes-Cantor



name	accession	subtype	collection_date	country	host
Zika_virus	KX198134	NA	1984_11_17	Senegal	Mosquito
Zika_virus	KX446950	NA	2016_01_01	Mexico	Mosquito
Zika_virus	KU955595	NA	1984_12_14	Senegal	Mosquito
Zika_virus	KU955592	NA	1984_12_06	Senegal	Mosquito
Zika_virus	KF268948	NA	1976	Central_African_Republic	Mosquito
Zika_virus	KF268950	NA	NA	Central_African_Republic	Mosquito
Zika_virus	KX446951	NA	2016_01_01	Mexico	Mosquito
Zika_virus	KX377336	NA	1966_07	Malaysia	Mosquito
Zika_virus	KU955591	NA	1984_11_20	Senegal	Mosquito
Zika	accession	subtype	collection_date	country	Mosquito

Table 1: Searchable database allows easy strain lookup and querying



name	accession	subtype	collection_date	country	host
Dengue_virus	HG316481	NA	2010	Thailand	Unknown
Dengue_virus	HG316482	NA	2010	Thailand	Unknown
Dengue_virus	HG316484	NA	2010	Thailand	Unknown
Dengue_virus	HG316483	NA	2010	Thailand	Unknown
Dengue	accession	subtype	2010	Thailand	host

Table 2: Researchers can search data with multiple filters at the same time

Choropleth Map View

- Displays the density of samples of each country
- Provides the user with a timeline slider to view the spread over time

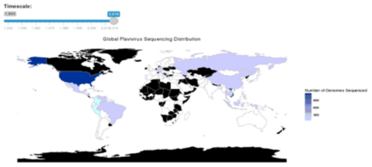

Timescale:
Global Flavivirus Sequencing Distribution
Number of Genomes Sequenced
900
600
300

Figure 2: Map visualization tool allows users to track the spread of a pathogen

Results (continued)

Phylogenetic Trees:

- Trees constructed using the Maximum Likelihood Algorithm
- Maximum Likelihood uses statistical techniques to infer a probability distribution for the given data
- The likelihood of the tree is proportional to the probability of the tree given by the distribution
- Generalised Time Reversible was used as the substitution model

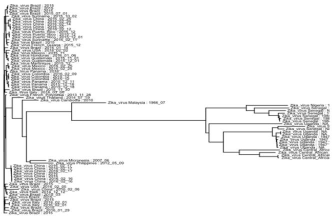


Figure 3: Dendrogram of Zika Virus genomes

- The final trees were validated using Bootstrapping, which is a resampling method
- Intuitive way to visualize the evolution of an organism over time
- Two genomic sequences at similar locations in the tree are expected to have evolved from the same common ancestor

Conclusions

IDEA will be an invaluable tool to virologists, epidemiologists, and public health officials. The ability to accurately determine the origin of a pathogenic strain will allow for preventative interventions and quarantines when they are needed most. This technology can be rapidly expanded to other pathogenic organisms as well, such as Ebola, HIV, malaria, influenza, and bacterial infections. An application like this paired with an assembler and a sequencer will allow researchers to quickly identify and track strains of pathogens permitting better intelligence for interventions and quarantines to stop and contain outbreaks.

References

Daep, C. A., Munoz-Jordan, J. L., & Eugenin, E. A. (2014). Flaviruses an expanding threat in public health: focus on Dengue, West Nile, and Japanese encephalitis virus, J. 539-560.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 39: 783.doi:10.2307/2408678. ISSN 0014-3820

Galtier, N. & Guoy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–79.

Hougland, J. (2015). How-to: Prepare Your Apache Hadoop Cluster for PySpark. http://blog.cloudera.com/blog/2015/09/how-to-prepare-your-apache-hadoop-cluster-for-pyspark-jobs/

Acknowledgements

The results published here are in whole or in part based upon data generated by the Virus Pathogen Database and Analysis Resource (ViPR) maintained by the National Institute of Allergy and Infectious Diseases (NIAID).