# On Whole Genome Phylogeny of Viruses

Levon Aslanyan
Discrete Mathematics Department,
Institute for Informatics and Automation Problems of NAS RA,
Yerevan, Armenia
e-mail: lasl@sci.am

Zaven Karalyan
Laboratory of Cell Biology and Virology,
Institute of Molecular Biology of NAS RA,
Yerevan, Armenia
Department of Medical Biology,
Yerevan State Medical University,
Yerevan Armenia
e-mail: zkaralyan@yahoo.com

***Abstract*—To infer evolution of viridae and to better define its evolutionary dynamics and genetic diversity, a genome-scale phylogenetic analysis is required. But the typical virus groups are of large approximate genome lengths of order of 200,000 nucleotides.**

**The branch lengths on the bifurcation phylogenetic tree are directly related to genetic distances accrued along those branches (clocklike evolution). The branches at the phylogenetic tree of Asfarviridae, in the similar analysis, are much shorter than those in Baculoviridae. Shorter branches in population suggest that this group has passed a shorter period of evolution and can be more closely related. Precise phylogeny requires careful mathematical design and high performance computations.**

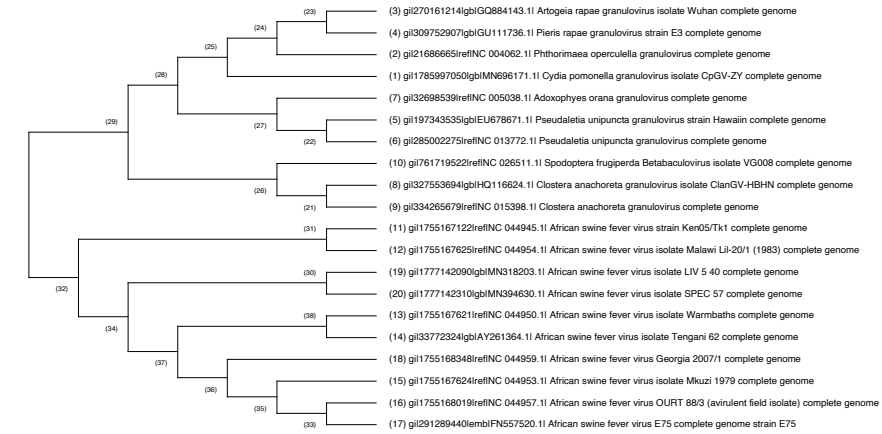
## I. Problem and Discussion

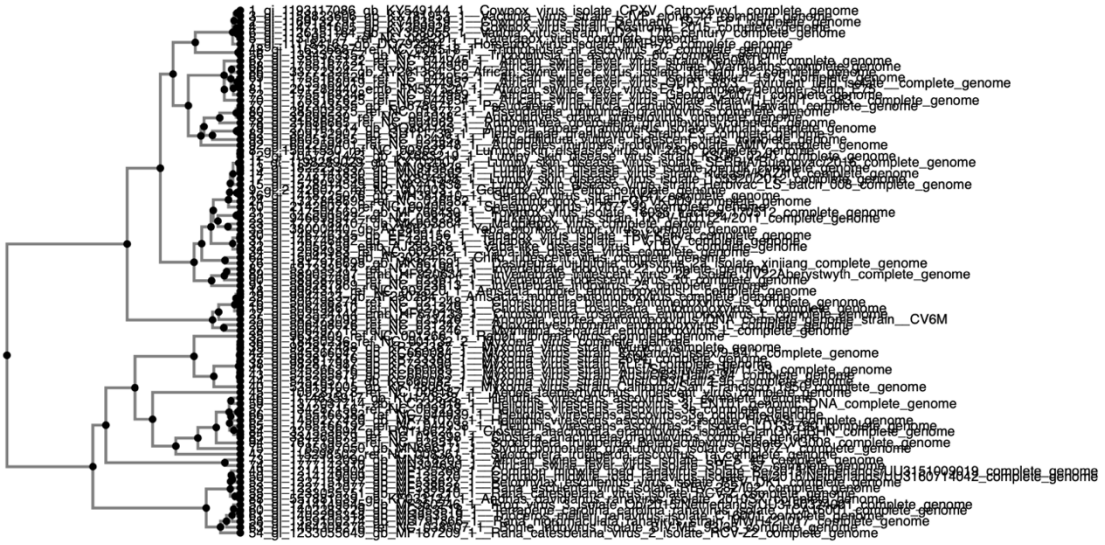
Viral genomes usually evolve at a high rate, and accumulated changes through either mutations or recombination with other strains or species, that are first fixed in the genome of successful virus isolates that give rise to genetic lineages. The relationship between biological lineages related by common descent is called 'phylogeny'. When the history of evolution is coded from ancestry, we derive a tree with the root as ancestor. It is a completely different scenario when we are given some population of viruses trying to solve the inverse problem about the evolution that generated this population, describing the way and steps of this evolution. Here it is to adopt some parametric characterization of mutations and recombination that is derived from one or another tree.

A complex mathematical apparatus has been developed for phylogeny inferences. It can determine and evaluate inter-species differences followed by phylogenic tree deduction and comparison. A reconstructed tree is an approximation of the "true" phylogeny that generally remains unknown. The phylogenetic analysis is used in applied and basic virology research, including epidemiology, diagnostics, forensic studies, phylogeography, evolutionary studies, and virus taxonomy. It can provide an evolutionary perspective on variation of any trait that can be measured for a group of viruses.

The "real" and approximate distances between the sequences are considered. The most appropriate similar/dissimilar measure between the sequences is the longest common subsequence (LCS) measure, even though the actual LCS is hardly computable. In combinatorial interpretation, LCS algorithms are regarded to the class of NP (nondeterministic polynomial) hard problems, which still require studies for finding successful polynomial algorithmic and heuristic solutions [1,2]. The known approximations to LCS are achieved through the k-mer, FFT and other combinatorial means. For $m$ sequence collection of $C_n^m$ pairwise distances can be computed by $C_n^m$ different LCS runs, or it might be approximated by an integrative multiple sequence alignment composition. Sequence Demarcation Tool, SDTv1.2 recommends the pairwise distance computations which is time consuming if tractable. The problem is in length $l$ of the whole genome. Complexity of computation and testing for recombination is increasing exponentially with the number of sequences $m$, and it is increasing linearly with the lengths $l$ of sequences examined. However, even the linearity by $l$ complicates the computations often making it practically intractable. Although the session of multiple sequence alignment (MSA) is also time-consuming, it provides acceptable approximation for $C_n^m$ pairwise distances in one run. Our strategy was to apply this approximation with testing it in "narrow places". On encountering any phylogenetic join that was subject to test, the set of sequences was narrowed by deleting the evidently correctly classified sequences and keeping the neighbourhood of the suspicious join. Then we compute the actual pairwise distances, scaling and checking the distance differences in MSA and pairwise LCS.

The evolutionary history is being inferred by the Neighbor-Joining method [3]. All ambiguous positions were removed for each sequence pair (pairwise deletion option). In a typical situation there were considered a total of 281332 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.



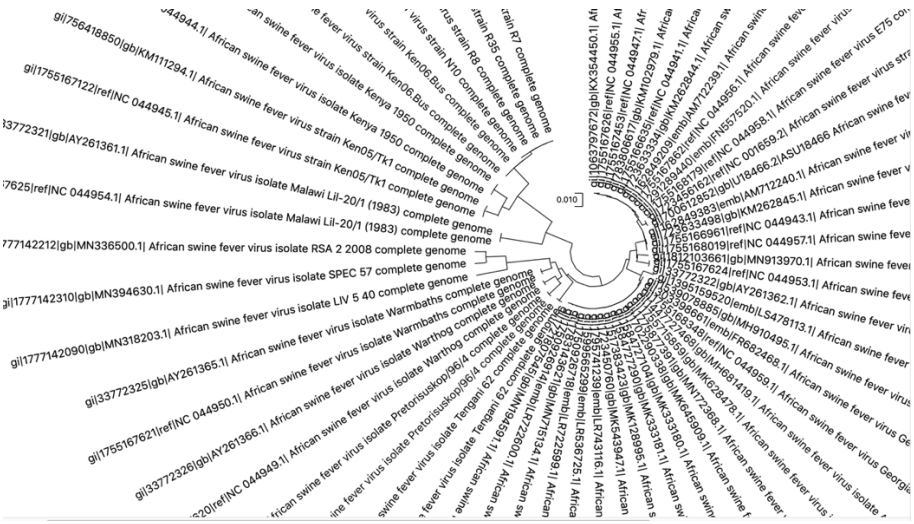



Figure 1. The phylogeny trees computed.

Computer experiments are conducted on 16-processor 16GB GENOME SERVER, a virtual computational environment at the Institute for Informatics and Automation Problems of National Academy of Sciences of the Republic of Armenia.

## REFERENCE


[1] L. Aslanyan, V. Minasyan, "LCS Algorithm with Vector-markers", *CSIT 2017, Revised Selected Papers, IEEE Xplore*, pp. 117-124, 2018.

[2] L. Aslanyan, H. Avagyan, Z. Karalyan, "Whole genome phylogeny of ASF viruses", *Veterinary World*, vol. 13(10), pp. 2118-2125, 2020.

[3] N. Saitou, M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees", *Molecular Biology and Evolution*, vol. 4, pp. 406-425, 1987.