Estimating Species Richness from Virome Data Accounting for Variations within the Virus Population

H.M.D.K. Herath[1]*, S.L. Tang[2]

[1]*Department of Computer Engineering, University of Peradeniya, 20400, Sri Lanka*
[2]*Biodiversity Research Center, Academia Sinica, Taipei, 341400, Taiwan*
**damayanthiherath@eng.pdn.ac.lk*

Species richness is a key species diversity measure. It corresponds to the number of species in an environmental sample. Estimating species richness of a metagenome of viruses (i.e., a virome) based on the reference data is challenging because of the limited amount of sequence data of viruses available in reference databases. A limitation identified with the methods that do not rely on reference sequence data in estimating species richness while being based on the contig spectrum is the assumption of equal genome length for all the species in the sample. This work aims to formulate a mathematical model to estimate species richness from a virome considering the variability of the genome lengths of species in the sample in contrast to the mentioned methods. A model is derived for the expected contig spectrum and the parameters of the model including the species richness is estimated through optimization for the least error between expected and observed contig spectra. Genetic Algorithm is used as the optimization algorithm in parameter estimation. The optimisation procedure incorporated in the proposed approach is shown to be robust based on the results with simulated data. This work enables inference of genome lengths distribution from the metagenomic sequence data in addition to estimating the species richness and can be applied to virome originating from any environmental sample.

Keywords: Metagenomics, Phages, Species Richness, Optimisation

Acknowledgement: Financial assistance given by the University of Peradeniya - University Research Grant (Grant No. URG/2021/15/E) is acknowledged.