

Mapping species boundaries among the Nucleocytoviricota using high-resolution phylogenomics
高解像度系統ゲノム解析を用いたヌクレオサイトウイルス門における種境界のマッピング

京都大学 化学研究所 数理生物情報研究領域
NECHES Russell Young (ネчес ラッセル ヤン)

研究成果概要

The Nucleocytoviricota, the phylum including the giant viruses, tends to exhibit large genomes with intricate histories of gene gain, loss and exchange, and a vast number of genes without known or predicted function. Within that history, the phylum has a complex evolutionary interrelationship with the Eukaryotes, and many of its extant species occupy key ecosystem roles. This presents unique challenges for large scale phylogenomic analysis; topology methods such as Robinson-Foulds, patristic correlation and Laplacian spectra perform well for trees with similar topologies, but saturate when applied to gene trees spanning this diverse group of viruses.

Previously, the Supercomputer System at the Institute for Chemical Research was used to construct a large dataset of 8865 gene trees were over 104,220 Nucleocytoviricota draft and reference genomes. Analysis of this dataset called for the development of a novel tree topology statistic, which we have tentatively named Quartet Ensemble Discordance (QED). A new software package called Quarimo (<https://github.com/ryneches/quarimo/>) has been developed to apply this new metric to this large and complex dataset.

QED uses the relative frequencies of quartet topologies within bootstrap ensembles to extract information from topologies with minority support within the alignments. The relative frequencies of quartets within ensembles are expressed as an intuitive, entropy-like measure for which we observe rapid numerical convergence with simulated and biologically-derived tree ensembles.

Because QED requires deep sampling of quartet space over large ensembles of trees, a high performance strategy for querying tree topologies was required. For this, Quarimo was built using minimum range queries (RMQ) of Eulerian tours implemented on compressed sparse row (CSR) offset tables constructed using a flat-packed memory layout. An in-GPU sampling kernel over quartet topologies and Steiner trees leverages hardware acceleration to perform massively parallel topology sampling. With a user-defined parameter, topological similarity at a specified divergence range may be quantified. Polytomies are supported via degenerate topology detection, and paralogs are resolved via GPU accelerated QED-maximizing combinatoric gradient descent.

Quarimo implements pluggable kernels for multiple numerical backends, including CPU parallelization via Numba's just-in-time (JIT) compiler, Nvidia CUDA via Numba CUDA's JIT compiler, and Metal Shader Language on Apple silicon via MLX. These hardware optimizations, in combination with the $O(1)$ complexity scaling of the RMQ-CSR algorithm with respect tree depth, achieves extremely high performance and favorable scaling for large datasets. We hope efficient use of latent, non-tree-like signal will widen the scale of divergence admissible to large scale phylogenomic analyses.