GaLSIC Colloquium
Guest Seminar

“Species boundaries, tree topology metrics, and the quest for comprehensive phylogenomics”

Dr. Russell Neches

Specially Appointed Researcher

Kyoto University

Institute for Chemical Research,

Bioinformatics Center

Monday 19th May 2025　16:35 ~ 18.05

Engineering Building　E207

For any query, please message Achille Zappa Ph.D. (zappa@soka.ac.jp)

Dr. Russell Neches

Kyoto University
Institute for Chemical Research,
Bioinformatics Center

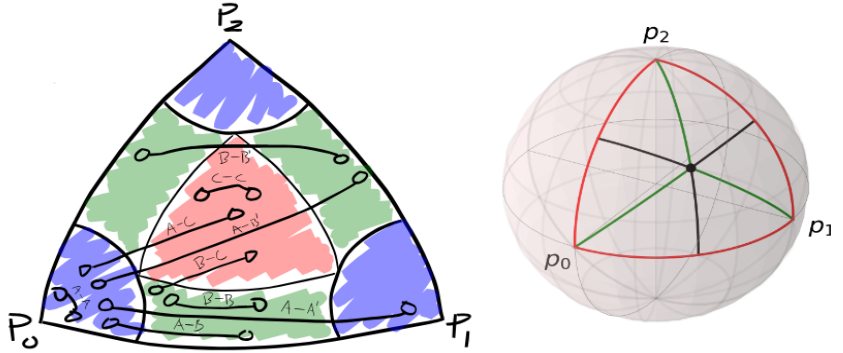
Research

- Interpreting major evolutionary events from gene phylogenies remains challenging, as results are highly sensitive to initial assumptions. However, if assumptions can be treated as explicit parameters, a systematic approach allows us to examine their influence on the results and on the method itself. While the most exciting questions in evolution tend to be found near the root of the tree of life, it is not necessarily the ideal place to test new approaches.

Bio

- Russell is a theoretical biologist studying the evolution and ecology of viruses and microbes by leveraging graph theory, machine learning and High-Performance Computing.

- After earning a B.S. in physics from Northeastern University and a Ph.D. in microbiology from UC Davis, he went on to a postdoc at the Joint Genome Institute at Lawrence Berkeley National Laboratory before joining Kyoto University as a JSPS postdoctoral fellow.



(a) Schematic view (b) Spherical embedding

Figure 1: A schematic view of quartet frequency vectors mapped to the $(+,+,+)$ oct of the unit sphere (panel (a)). The unit vectors $\widehat{p}_0$, $\widehat{p}_1$ and $\widehat{p}_2$ correspond to the frequenc of each of the three possible topologies for each quartet in an ensemble of trees. Provic that the same ordering is used for each ensemble, the order of the topologies is arbitr for any given quartet. For simplicity, the relationships are illustrated in order from high to lowest, so that the highest frequency topology in the first ensemble always falls n $\widehat{p}_0$. The domain may be divided into three types of zones representing qualitativ different frequency states. Quartets whose topology frequencies are dominated by a sir topology fall near the poles, shaded in blue. Quartets whose topology frequencies dominated by two topologies fall near the edges, shaded in green. Quartets whose topol frequencies are roughly equal fall near the center of the octant, shaded in red. Am two ensembles, there are nine unique types of relationships. For quantitative treatmen is more straightforward to evaluate these relationships with respect to the "dead poi (panel (b)). Here, the domain boundary (red), the arc of a single dominant freque (green) and the arc of a mixture of two frequencies (black) are represented as arcs. T "dead point" is the back dot in the center.

Species boundaries, tree topology metrics, and the quest for comprehensive phylogenomics

The rarefaction of events as one looks deeper into the tree of life is a boon for explanatory power, but it's bad for statistics. If we want to develop and test new methods, it is perhaps better to focus our attention at the other extreme, the species formation process. The distribution of average nucleotide identity (ANI) among genomes exhibits a very striking feature, where genomes belonging to different species almost always have an ANI above about 95%, while genomes belonging to different species almost always have a ANI below about 85%.

There are very, very few pairs of genomes with an ANI between 85% and 95%. This "ANI gap" phenomenon is extremely robust, and holds true for Bacteria, Archaea, viruses, and (with some caveats) Eukaryota, and does not seem to be a technical, sampling or statistical artifact. While this phenomenon offers little insight into the mechanisms that drive species formation in itself, it provides a ground truth for methods that model the species formation process.

One of the most useful (and thus widespread) assumptions in modern evolutionary biology is that there exists a handful of "marker" or "core" genes whose phylogenies trace the history of the lineage itself. This powerful concept has yielded extraordinary results over several decades, but because it was born more from technical necessity than from biological reality, it also leads to unavoidable problems. Most distressingly, one can draw upon the best available data and the most well-tested methods and yet arrive at mutually exclusive inferences simply by making different (wholly reasonable) choices about which genes are "core genes," and which are not. Molecular phylogenetics matured during the era in which DNA sequences were precious, and thus carefully curated.

This is no longer even remotely the case today. In principle, the choice of core genes is now a parameter that can be tested. Or, with the right mathematical treatment, perhaps the core genome concept could be set aside entirely in favor of a comprehensive phylogenomic approach to evolutionary inference. In this talk, I will discuss a proposed mathematical treatment that might help achieve this, as well as a project using the species boundary of Giant Viruses as the proving grounds for the approach.