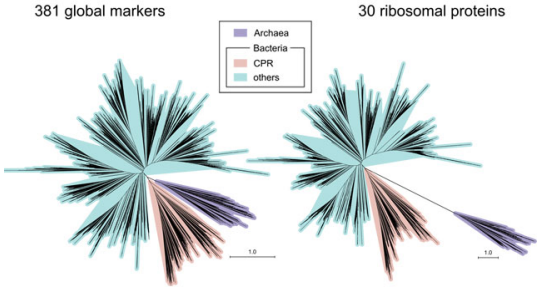
UC San Diego
UC San Diego News Center

January 17, 2020 | By Jan Zverina

UC San Diego-led Study Finds Close Evolutionary Proximity Between 'Tree of Life' Microbial Domains

SDSC's 'Comet' supercomputer used in analysis of bacteria and archaea

A comprehensive analysis of 10,575 genomes as part of a multi-national study led by researchers at UC San Diego has revealed close evolutionary proximity between the microbial domains at the base of the tree of life, the branching pattern of evolution described by Charles Darwin more than 160 years ago in his book, *On the Origin of Species*.



These two trees illustrate the study's primary finding. The evolutionary distance between Archaea and Bacteria is much less in the left tree obtained from a global set of 381 global marker genes than in the right tree obtained from only 30 genes for ribosomal proteins, similar to those used in other studies. This study provides strong evidence that trees based on more broadly selected genes better reflect genome-level evolution and a more accurate view of the tree of life. Image courtesy of Qiyun Zhu, et al.

The currently accepted tree of life consists of two microbial domains, Bacteria and Archaea, and a third domain, Eukaryota, of higher organisms whose cells have nuclei to enclose their DNA and that may have evolved from Archaea.

The study, published last month in *Nature Communications*, found much closer evolutionary proximity between Archaea and Bacteria than have most previous studies. This new result arises from the use of a comprehensive set of 381 marker genes versus a couple of dozen core genes such as ribosomal proteins typically used in previous studies, according to Qiyun Zhu, a postdoctoral scholar in the UC San Diego School of Medicine's Department of Pediatrics and lead author of the paper.

"Our work shows that insufficient or uneven sampling of genetic information, as in most previous work, results in a biased view of the tree of life, therefore limiting our ability to establish evolutionary relationships," said Zhu.

The researchers also generated time-calibrated trees, assuming a universal molecular clock and that the split between Cyanobacteria and Melainabacteria occurred about 2.5 billion years ago when the atmosphere became oxygenated. The base of these trees implies that the origin of life occurred about 4 billion years ago when 381 marker genes are considered, versus about 7 billion years ago when 30 ribosomal proteins are considered. The latter time is not credible, said researchers, since it is older than the age of the Earth, which further supports the choice of genes adopted in the study.

Rob Knight, founding director of the Center for Microbiome Innovation and Professor of Pediatrics and Computer Science & Engineering at UC San Diego, and senior author of the new study, said that its significance from a pediatric standpoint is that many diseases that strike in adulthood have their roots in the human microbiome in childhood.

“Our ability to collect DNA sequences from the human microbiome has expanded dramatically in the past 15 years, but our ability to interpret the data relies on reference databases that are highly incomplete,” said Knight. “Improving the precision of our understanding of evolutionary relationships among microbes gives us better precision in understanding how these changes occur, and how to target them to improve the microbiome in childhood to address not only microbiome-based early-life diseases, but to improve health throughout a person’s lifespan.”

Zhu further noted: “We expect that our tree with 10,575 genomes selected in a statistically even way will be a valuable resource. We have made our results publicly available in a reference database and have developed computational tools to explore it. In multiple microbiome studies currently taking place in the Knight Lab, we have already witnessed remarkable improvements by using this resource.”

Scalable Algorithm and a Powerful Supercomputer

The availability of a scalable algorithm and a powerful supercomputer were essential for carrying out the study.

Phylogenetic trees were generated using two algorithmic approaches: concatenation and summary. Summary methods, which are relatively new, combine potentially different evolutionary histories of different genes to obtain a master “species tree”. A leading summary method is ASTRAL, developed by (among others) Siavash Mirarab, Assistant Professor in Electrical and Computer Engineering at UC San Diego.

Both approaches gave similar trees, but the summary approach better resolved the basal relationships among major microbial lineages because it is inherently scalable and can use all genomic data, whereas the concatenation approach requires subsampling to be computationally feasible. To facilitate analysis of the very large amount of data in the study, Uyen Mai, a PhD student in the Mirarab Lab and co-first author of the paper, developed new methods to extend the summary approach.

Most of the computations were done on the *Comet* supercomputer of the San Diego Supercomputer Center (SDSC) at UC San Diego. Wayne Pfeiffer, Distinguished Scientist at SDSC, made more than 2,000 runs on the standard compute nodes of *Comet* to generate the gene trees, while Mai combined these trees using ASTRAL on the GPU nodes of *Comet*.

Zhu summarized: “We advanced the state-of-the-art of phylogenetic research along three dimensions: larger and more even representation of microbial life forms, more comprehensive use of whole-genome information, and improved methodology for accurate resolution of evolutionary relationships. This was made possible with the supercomputing power at SDSC.”

MEDIA CONTACT

Jan Zverina, 858-534-5111, jzverina@sdsc.edu
Heather Buschman, 858-249-0456, hbuschman@ucsd.edu