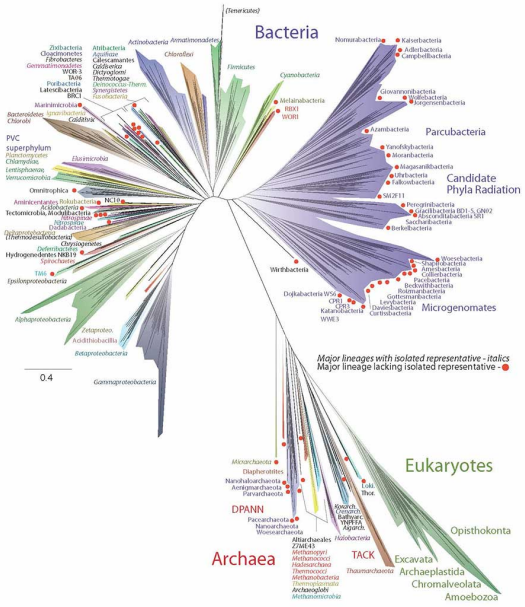
UC San Diego
UC San Diego News Center

April 25, 2016 | By Robert Sanders and Warren Froelich

# SDSC Supercomputers, CIPRES Gateway Help Define New "Tree of Life"

An outline for a new tree of life, depicting the evolution of life on this planet that included more than 1,000 new types of bacteria and Archaea lurking in the Earth's nooks and crannies, was made possible with the help of supercomputing resources and a phylogenetics "gateway" created at the San Diego Supercomputer Center (SDSC), based at the University of California San Diego.

"The CIPRES Science Gateway was critical to our work," said Laura Hug, now a biology faculty member at the University of Waterloo, Canada, and former postdoctoral fellow at UC Berkeley, where the study was conducted. "Previous attempts to infer the trees presented severe problems with run time, memory allocation and a lack of parallelized implementation of the RAxML (for Randomized Axelerated Maximum Likelihood, a popular program for phylogenetic analysis of large datasets)."




*An image of the new 'Tree of Life'. Courtesy of Laura Hug, Jill Banfield, and Nature Microbiology.*

"No run had successfully finished prior to our introduction to CIPRES," she added. CIPRES, which stands for Cyberinfrastructure for Phylogenetic Research, is a web-based portal, or gateway, that allows researchers to explore evolutionary relationships between species using supercomputers provided by the National Science Foundation's XSEDE (eXtreme Science and Engineering Discovery Environment) project.

The new tree, published online April 11 in the new journal *Nature Microbiology* and widely publicized, reinforced once again that the life we see around us – plants, animals, humans and other so-called eukaryotes – represents a tiny percentage of the world's biodiversity.

“The tree of life is one of the most important organizing principles in biology,” said Jill Banfield, a UC Berkeley professor of earth and planetary science, policy, and management, and the study’s principal investigator. “The new depiction will be of use not only to biologists who study microbial ecology, but also biochemists searching for novel genes and researchers studying evolution and earth history.”

Charles Darwin first sketched a tree of life in 1837 as he sought ways to show how animals and plants are related to one another. The idea took root in the 19$^{th}$ century, with the tips of the twigs representing life on Earth today, while the branches connecting them to the trunk implied evolutionary relationships among these creatures. A branch that divides into two twigs near the tips of the tree implies that these organisms have a recent common ancestor, while a forking branch close to the trunk implies an evolutionary split in the distant past.

Since then, researchers have tried to build on Darwin’s initial sketch, gradually adding twigs and branches to the original, and then in 1977 a large trunk called Archaea—microbes that live in extreme environments such as hot springs and oxygen-free wetlands. This new familial category represented a third great trunk of the tree of life which includes eukaryotes, including animals, plants, fungi and protozoans; familiar bacteria like Escherchia coli; and the Archaea.

The revolution in DNA sequencing during the past couple of decades has led to an explosion of data offering yet more complete descriptions of the genetic relationships among species. Researchers began sequencing whole communities or organisms at once and picking out the individual groups based on their genes alone. This metagenomic sequencing revealed whole new groups of bacteria and Archaea, many of them from extreme environments, such as the toxic puddles in abandoned mines, the dirt under toxic waste sites, and the human gut. Some of these had been detected before, but nothing was known about them because they wouldn’t survive when isolated in a lab dish.

For the new paper, Banfield and Hug, along with more than a dozen other researchers who have sequenced new microbial species, gathered 1,011 previously unpublished genomes to add to already known genome sequences of organisms representing the major groups of life on Earth.

## Counting on SDSC’s *Comet* and *Gordon* supercomputers

Access to supercomputers was a key part of this study, helping researchers to investigate relationships by comparing DNA sequences information between species. This type of analysis is becoming more powerful as the number of DNA sequences available is increasing rapidly,

with new, larger data sets requiring higher levels of computational power.

For their supercomputer resources, the researchers reached out to the CIPRES gateway, which initially allowed them access to SDSC's *Gordon*, the first high-performance supercomputer to use massive amounts of flash-based SSD (solid state drive) storage. The final trees were generated by SDSC's latest system, *Comet,* a petascale supercomputer designed to transform advanced scientific computing by expanding access and capacity among traditional as well as non-traditional research domains. The two jobs ran for a total of about five days, using 48 cores.

"The CIPRES Gateway allows scientists to conduct their research in significantly shorter times without having to understand how to operate supercomputers," said Mark Miller, principal investigator of the CIPRES gateway and an SDSC researcher.

Added Hug: "I spent over a month attempting to conduct these jobs on other servers with no success – the jobs always failed prior to finishing. Also, Mark (Miller) with customer assistance (from CIPRES) was invaluable in troubleshooting our analyses on CIPRES."

Their investigation, representing the total diversity among all sequenced genomes, produced a tree with branches dominated by bacteria, especially by uncultivated bacteria. A second view of the tree grouped organisms by their evolutionary distance from one another rather than current taxonomic definitions, making clear that about one-third of all biodiversity comes from bacteria, one-third from uncultivated bacteria and a bit less than one-third from Archaea and Eukaryotes.

"The two main take-home points I see in this tree are the prominence of major lineages that have no cultivable representatives, and the great diversity in the bacterial domain, most importantly, the prominence of candidate phyla radiation," Banfield said. "The candidate phyla radiation has as much diversity within it as the rest of the bacteria combined."

---

MEDIA CONTACT

**Jan Zverina**, 858-534-5111, jzverina@sdsc.edu
Robert Sanders, 510-643-6998 rlsanders@berkeley.edu