DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by the Mobilize Center, an NIH Big Data to Knowledge Center of Excellence

Nonprofit Org.
U.S. Postage Paid
Permit No. 28
Palo Alto, CA

Stanford University
318 Campus Drive
Clark Center Room W352
Stanford, CA 94305-5444

SeeingScience

BY KATHARINE MILLER

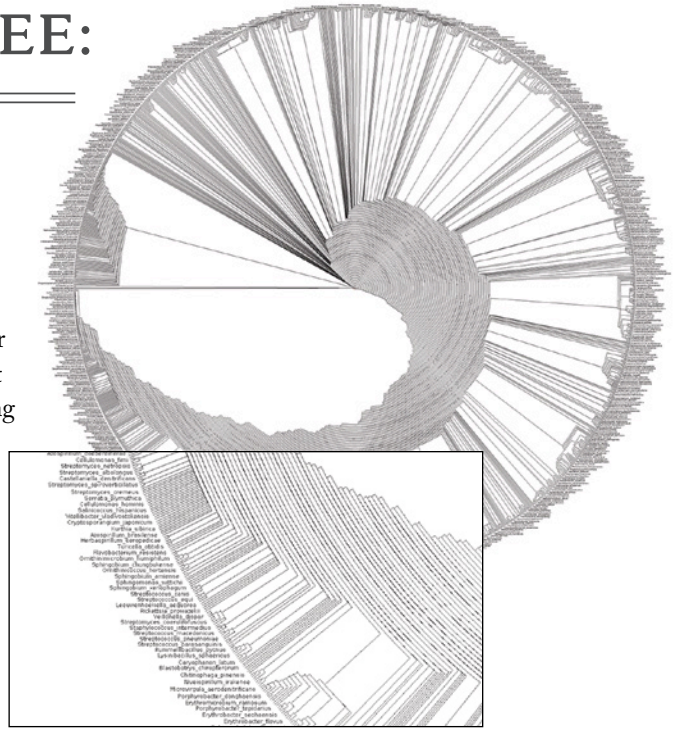
AN AUTOMATED SUPERTREE:

A Model for Extracting Literature-based Knowledge

Too much scientific knowledge is buried in published literature. Case in point: The phylogenetic relationships among microbial species are locked into numerous publications about individual species and their close relatives. And because those publications don't include machine-readable data, the information is difficult to extract. Thus, efforts to create supertrees (large trees assembled from a combination of many smaller phylogenetic trees) typically involve a handful of graduate students doing a massive cut-and-paste job—connecting trees bracket by bracket on a computer. "It's mind-numbingly tedious," says **Ross Mounce, PhD**, Open Access Grants Manager at the Arcadia Fund.

As a postdoctoral research associate at Cambridge University, Mounce set out to create a microbial supertree by using computer vision to extract information from smaller phylogenetic trees in a single journal. "We point the program at the file and it will do its best to extract phylogenetic data from the image," Mounce says. The result is not the best tree, Mounce says, but a proof-of-concept for developing a scalable, automated process. "It's a solvable problem," he says, that has been made easier in the United Kingdom by recent changes to copyright laws. As long as a researcher has legitimate access to a published piece of literature (through a university library, for example), "it's legal to do sophisticated analyses on it without asking permission of the copyright holder," Mounce says. Without that legal right, it would be nearly impossible to perform scalable syntheses of the literature.

"The future is really exciting, because if you had an ongoing reproducible pipeline, you could have a tree of life that self-updates every day," Mounce says. The same is true for any piece of scientific knowledge: "You could check back and see a self-updating synthesis of the current evidence on any topic," he says. "That's the idea really." □



Using an automated, scalable method, Mounce and his colleagues applied computer vision techniques to automatically convert phylogenetic trees from figures in a single journal (the International Journal of Systematics and Evolutionary Microbiology*) back into re-usable, computable, phylogenetic data. They then used established supertree methods to generate the tree of microbial life shown here. Reprinted from Mounce R, Murray-Rust P, Wills M, A machine-compiled microbial supertree from figure-mining thousands of papers.* Research Ideas and Outcomes *3: e13589. https://doi.org/10.3897/rio.3.e13589, (2017).*