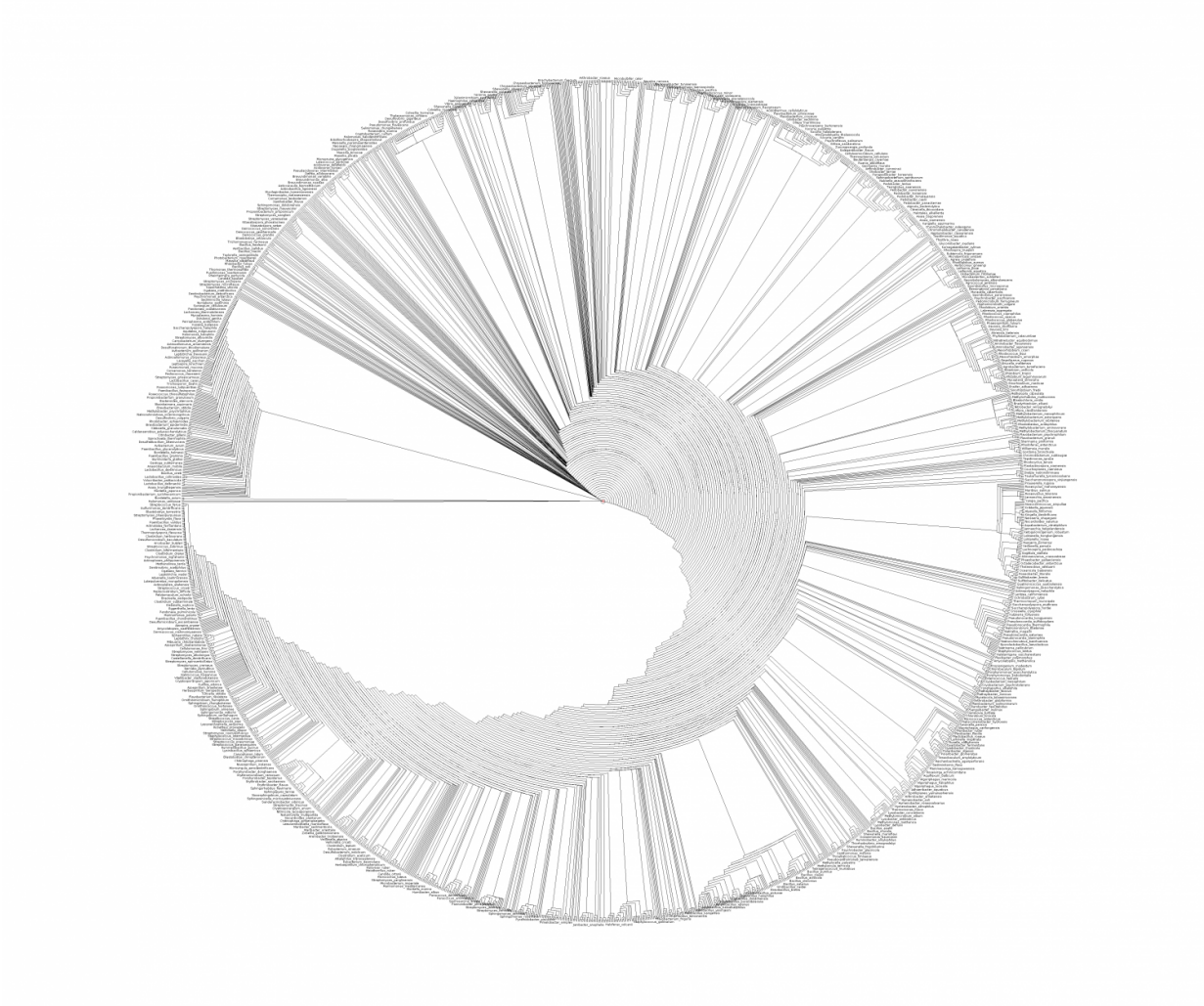
The first microbial supertree from figure-mining thousands of papers

May 9 2017



The consensus supertree produced from an analysis of 924 source trees from the journal IJSEM. Credit: Ross Mounce

While recent reports reveal the existence of more than 114,000,000 documents of published scientific literature, finding a way to improve the access to this knowledge and efficiently synthesise it becomes an increasingly pressing issue.

Seeking to address the problem through their PLUTo workflow, British scientists Ross Mounce and Peter Murray-Rust, University of Cambridge and Matthew Wills, University of Bath perform the world's first attempt at automated supertree construction using data exclusively extracted by machines from published figure images. Their results are published in the open science journal *Research Ideas and Outcomes* (RIO).

For their study, the researchers picked the *International Journal of Systematics and Evolutionary Microbiology* (*IJSEM*) - the sole repository hosting all new validly described prokaryote taxa and, therefore, an excellent choice against which to test systems for the automated and semi-automated synthesis of published phylogenies. According to the authors, IJSEM publishes a greater number of phylogenetic tree figure images a year than any other journal.

An eleven-year span of articles dating back to January, 2003 was systematically downloaded so that all image files of phylogenetic tree figures could be extracted for analysis. Computer vision techniques then allowed for the automatic conversion of the images back into re-usable, computable, phylogenetic data and used for a formal supertree synthesis of all the evidence.

During their research, the scientists had to overcome various challenges posed by copyrights formally covering almost all of the documents they needed to mine for the purpose of their work. At this point, they faced quite a paradox - while easy access and re-use of data published in scientific literature is generally supported and strongly promoted, common copyright practices make it difficult for a scientist to be

confident when incorporating previously compiled data into their own work. The authors discuss recent changes to UK copyright law that have allowed for their work to see the light of day. As a result, they provide their output as facts, and assign them to the public domain by using the CC0 waiver of Creative Commons, to enable worry-free re-use by anyone.

"We are now at the stage where no individual has the time to read even just the titles of all published papers, let alone the abstracts," comment the authors.

"We believe that machines are now essential to enable us to make sense of the stream of published science, and this paper addresses several of the key problems inherent in doing this."

"We have deliberately selected a subsection of the literature (limited to one journal) to reduce the volume, velocity and variety, concentrating primarily on validity. We ask whether high-throughput machine extraction of data from the semistructured scientific literature is possible and valuable."

More information: Ross Mounce et al, A machine-compiled microbial supertree from figure-mining thousands of papers, *Research Ideas and Outcomes* (2017). DOI: 10.3897/rio.3.e13589

Provided by Pensoft Publishers