

Scientists Encounter Holes in Tree of Life, Push for Better Data Storage

University of Florida



GAINESVILLE, FL - When it comes to public access, the tree of life has holes. A new study co-authored by University of Florida researchers shows about 70 percent of published genetic sequence comparisons are not publicly accessible, leaving researchers worldwide unable to get to critical data they may need to tackle a host of problems ranging from climate change to disease control.

Scientists are using the genetic data to construct the largest open-access tree of life as part of the National Science Foundation's \$5.6-million Assembling, Visualizing and Analyzing the Tree of Life project. Understanding organismal relationships is increasingly valuable for tracking the origin and spread of emerging diseases, creating agricultural and pharmaceutical products, studying climate change, controlling invasive species and establishing plans for conservation and ecosystem restoration.

The study, appearing September 4, 2013, in *PLoS Biology* describes a significant challenge for the project, which is expected to produce an initial draft tree by the end of the year. It highlights the need for developing more effective methods for storing data for long-term use and urges journals to adopt more stringent data-sharing policies.

"I think what we need is a major change in our mindset about just how important it is to deposit your data - this has to be a standard part of what we do," said co-author Doug Soltis, a distinguished professor at the Florida Museum of Natural History on the UF campus and UF's biology department. "Because if it's not there, it's lost forever. These are really, really important for long-term use, as we're seeing

now in our efforts to build a tree."

Estimates of the amount of missing data were based on 7,539 peer-reviewed studies about animals, fungi, seed plants, bacteria and various microscopic organisms. Soltis said the missing genetic data has required project collaborators to contact hundreds of researchers to request information, or attempt to reproduce the sequence alignments and analyses, which is extremely labor intensive.

"There are ambiguities with the alignments, you have to make certain judgment calls, and so an alignment that I do is not going to be the same as an alignment that somebody else does," said lead author Bryan Drew, a postdoctoral researcher in UF's biology department. "It's hard to assess a publication's validity in a lot of cases if you don't have access to the alignments. To me, that's the biggest problem with all of this."

Challenges include complicated mechanisms for uploading data and inconsistencies between journals – some require or strongly recommend data be stored in an online database and others do not, Drew said. The most widely used, publicly accessible databases include GenBank, TreeBASE and Dryad. Most journals require DNA sequences be deposited in GenBank, but comparatively few require the sequence alignments to be publicly archived. When study co-authors emailed researchers to obtain missing information, a majority did not respond, and the co-authors were rarely successful in retrieving the data.

"A lot of the authors I contacted said their data was in TreeBASE, but they were unaware of the next step needed after acceptance by the journal – the researchers didn't know they had to go back into TreeBASE and actually make the data available to the public," Drew said.

Elizabeth Kellogg, a professor in the department of biology at the University of Missouri-St. Louis who was not involved with the study, said she is not surprised about the large amount of missing information.

"They're absolutely right that when people are publishing papers, you want to document your results as much as you can," Kellogg said. "But many journals aren't requiring that extra step, so some researchers are only submitting the minimum to have their studies published. "There are databases for archiving, but some of their interfaces are somewhat cumbersome, and if you haven't previously done this, it can appear to be a daunting task."

Source URL (retrieved on 12/12/2013 - 3:48pm):

<http://www.scientificcomputing.com/news/2013/09/scientists-encounter-holes-tree-life-push-better-data-storage>