

## Assembling the tree of life to enable the plant sciences (iPTOL): A proposal for an iPlant Grand Challenge Workshop

### Project Personnel:

Primary Lead: Michael Donoghue (Yale University), [michael.donoghue@yale.edu](mailto:michael.donoghue@yale.edu); Co-leads: Michael Sanderson (University of Arizona), Douglas Soltis (University of Florida), Pamela Soltis (University of Florida), Val Tannen (University of Pennsylvania), Todd Vision (University of North Carolina, Chapel Hill)

### Project Summary:

Ever since Darwin heralded “the great Tree of Life,” biologists have attempted to infer the precise order and timing of the branching events that link all species that have ever existed. Owing to the sheer magnitude of this problem, and to its fundamental importance, reconstructing the tree of life is one of the most profound scientific endeavors ever undertaken. Tackling this grand challenge will require the integration of data from multiple sources, including the morphology of living and extinct organisms, and, increasingly, vast quantities of genomic data. It will also require a quantum leap in the capabilities of algorithms to infer phylogenies from these data at the scale imagined. While unprecedented progress has been made recently, the grand synthesis of phylogenetic knowledge that we seek requires the development of a supporting cyberinfrastructure far beyond anything now available. Likewise, tools are needed to harness the power of phylogenetic knowledge – to provide fresh perspectives on problems at all levels of biological organization, from genes, cells, and organisms, to species and ecosystems. It has become clear, for instance, that phylogenetic trees provide a rigorous framework for testing comparative hypotheses in structural, functional, and developmental biology, and can help elucidate processes ranging from adaptation to climate change. However, to be truly useful in these domains, phylogenetic trees and the associated data must be readily accessible, easily tracked and managed, and productively merged with other information.

The rate of accumulation of data relevant to plant phylogeny has far exceeded all expectations, and we now find ourselves awash in largely unconsolidated, haphazardly accumulating information. We face major barriers in managing and synthesizing data, in visualization, and in the development of tools to use this exploding knowledge-base. This area is ripe for a coordinated effort involving phylogeneticists, computer scientists, bioinformaticists, and the down-stream users of phylogenies. It is also a domain in which plant scientists are especially well prepared to lead the way in the development of a new model that can be applied across the entire tree of life. Here we envision a sea-change in the way that researchers and educators can access, integrate, and use phylogenetic information. And, in turn, we anticipate entirely new insights into fundamental problems, such as the origin of green plants, the occupation of land, the evolution of key adaptations such as multicellularity, seeds, and flowers, and the role of polyploidy in the functional diversification of gene and protein families.

We propose to bring together a team of plant and computer scientists to conceive a Grand Challenge Project focused on assembling all knowledge of the phylogeny of plants, to make this knowledge readily accessible, and to integrate it throughout the botanical sciences. Specifically, we will envision “discovery environments” to enable the mining and synthesis of all relevant literature and underlying data, and within which this information could be visualized, disseminated, and utilized in novel ways. Key objectives are the development of new tools to streamline the assembly and analysis of massive datasets, to properly track the provenance of this information, and to integrate phylogenetic knowledge into botanical studies and into education and outreach at all levels.

The necessary developments are so extensive and cross-disciplinary that they are unlikely to materialize through normal NSF funding. A concerted plant phylogeny effort is not only necessary, but is also highly likely to yield successes, as the plant phylogenetics community has an established history of collaboration and has consistently been willing to adopt new cyber-solutions. This area is also rich in problems of direct interest and importance to computer scientists and bioinformaticists. Furthermore, a variety of these problems have begun to be addressed through other major efforts, including the NSF’s CIPRES and ATOL programs. Early solutions to the plant phylogeny problem would enable developments throughout the plant sciences, particularly in connection with other Grand Challenge Projects supported through the iPlant Collaborative.



Acknowledgements: The iPlant Collaborative is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#EF-0735191).