

ASSEMBLING THE TREE OF LIFE TO ENABLE THE PLANT SCIENCES (iPTOL): A PROPOSAL FOR AN iPLANT GRAND CHALLENGE WORKSHOP

Organizers:

Michael J. Donoghue, Department of Ecology and Evolutionary Biology and Peabody Museum of Natural History, Yale University; 203-432-2074 (phone), 203-432-7907 (fax), michael.donoghue@yale.edu (primary contact for this proposal): Donoghue's research focuses on the uses of phylogenies in understanding morphological and molecular evolution, ecology, and biogeography. As Director of the Peabody Museum he has managed numerous K-12 outreach programs, including a museum exhibition entitled *Travels in the Great Tree of Life*.

Michael J. Sanderson, Department of Ecology and Evolutionary Biology, University of Arizona; 520-626-6848 (phone), 520-621-9190 (fax), sanderm@email.arizona.edu: Sanderson's research emphasizes method and algorithm development for data mining and large scale phylogenetic tree construction.

Douglas E. Soltis, Department of Botany, University of Florida; 352-273-1963 (phone), 352-846-2154 (fax), dsoltis@botany.ufl.edu: Douglas Soltis's research focuses on angiosperm phylogeny, especially the Saxifragales, and the genetic and genomic consequences of polyploidy.

Pamela S. Soltis, Florida Museum of Natural History, University of Florida; 352-273-1964 (phone), 352-846-2154 (fax), psoltis@flmnh.ufl.edu: Pamela Soltis's research interests are in angiosperm phylogeny, the origin and evolution of the floral genetic program, and the evolution by polyploidy.

Val Tannen, Department of Computer and Information Science, University of Pennsylvania; 215-898-2665 (phone), 215-898-0587 (fax), val@cis.upenn.edu: Tannen's research is on foundations of data management with applications to phyloinformatics; he leads the NSF AToL project "Processing Phylodata" (pPOD) as well as the database focus for the NSF CIPRES project.

Todd J. Vision, Department of Biology, University of North Carolina at Chapel Hill; 919-843-4507 (phone), 919-962-1625 (fax), tjv@bio.unc.edu: Vision's research focuses on computational and informatic applications to evolutionary genetics and genome evolution, with an emphasis on the macroevolution of gene order and gene content in the angiosperms.

SUMMARY OF THE iPLANT PHYLOGENY PROJECT

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.

GRAND CHALLENGE WORKSHOP DESCRIPTION

Statement of the Scientific Problem

Amazing progress has been made over the past few decades in resolving the Tree of Life. Plant scientists have been leaders in this global effort, and are widely regarded as being in the vanguard both in organizing large collaborative efforts and in pushing the limits in the analysis, synthesis, and use of phylogenetic information. While there is much to be proud of, the magnitude of the task that still lies ahead is daunting, and we find ourselves struggling to keep up with the massive datasets that are accumulating through our own efforts and those of our colleagues throughout the plant sciences. Data-mining pipelines are underdeveloped, as are tools for the assembly, analysis, and visualization of larger and larger trees. We still do not have proper tools to integrate and display data from multiple sources (e.g., molecular sequences, genomic data, expression data, morphological and developmental data, fossil evidence), and the results of most phylogenetic studies have not been assembled and made readily accessible, and, therefore, remain effectively unavailable to the wide variety of potential user communities in research and in education.

All of this emphasizes the clear need to greatly expand the intersection between phylogenetic biology and computer science, computational biology, and bioinformatics. Some important attempts have been made recently to bridge this gap, such as the NSF-funded CIPRES project. This effort has highlighted both the enormous needs that exist, but has also led to the realization that solutions to these problems require efforts

that focus on the needs of a particular research community. The iPlant Collaborative presents a unique opportunity to connect plant phylogenetic biologists with colleagues in the relevant computer sciences, to craft real world solutions, not only to deliver better phylogenetic inferences, but to render this information truly useful and to enable research and education across the plant sciences. In turn, this effort would provide a model that could extend well beyond the plant sciences.

The purpose of the Grand Challenge Workshop that we propose is to begin this process; that is, to bring together a set of relevant plant phylogeneticists with computer scientists and informatics experts to visualize a cyberinfrastructure that would propel us all forward. Importantly, the beneficiaries of such a development would not just be phylogenetic biology and the plant sciences, but the computer sciences as well, as many of the problems in this realm are of general theoretical and practical interest. Likewise, our ultimate aim is to enable new, synthetic research and education across the entire scientific community. Approaches and tools developed through the iPlant Collaborative will be useful to those studying other branches of the Tree of Life and other problems – plant structure, function, and gene family evolution, to name just a few – that require phylogenetic analysis of large datasets.

It is important to stress that the likelihood of success is especially high in the plant sciences, where the plant Tree of Life community has already demonstrated a willingness to collaborate on a grand scale. Beginning in the early 1990's, a series of major efforts have been made to compile and analyze large datasets of gene sequences from all three plant genomes (nuclear, plastid, mitochondrial). These team efforts involved many of the investigators associated with the present proposal, who have helped to build and maintain an international network of expertise and data sources. This has spawned several NSF Research Coordination Networks (RCN's), including the Deep Green, Deep Time, and Deep Gene projects. As a result of these efforts, angiosperms became the first major group of organisms to be classified based on molecular phylogenetic analyses (the so-called APG system), and recently we have provided the outline of a truly phylogenetic classification of all vascular plants. These advances mark a turning point not only in terms of our knowledge, but in the way that we operate as a community.

Currently, there are six NSF-funded Assembling the Tree of Life (AToL) projects focused on green plants (on basal green plants, liverworts, seed plants, angiosperms, and monocots, and on the use of whole plastid genome sequences), each of which coordinates the activities of a community of plant scientists, involving morphologists, paleobotanists, and molecular phylogeneticists, with direct connections to ongoing genomics and evo-devo efforts. Excellent communication has been established among these efforts, which is especially evident in their development of common informatics solutions (e.g., TOLKIN).

As an example of a specific effort, the angiosperm AToL group has resolved 12 of the most problematical deep-level nodes, providing a firm foundation for future studies. We have explored challenges in the analysis of large datasets, including supermatrix approaches with large amounts of missing data. In campanulids, we have constructed a supertree of ca. 5,000 (of ~30,000) species from >200 phylogenetic studies. We are currently building a 16-gene dataset from all three plant genomes for hundreds of species to provide a new, comprehensive phylogenetic backbone (replacing our widely used 3-gene, 567-taxon tree). We have also assembled matrices of over 100 complete plastid genome sequences. Finally, it is noteworthy that plant phylogeneticists were among the first to promote global databases of phylogenetic data matrices and trees (e.g., TreeBASE), and have been leaders in the development of key data-mining and analytical tools (e.g., PhyLoTa, Phytome), as well as resources to increase the accessibility of phylogenetic information (e.g., APWeb, Phylomatic).

Why the problem requires cutting-edge computer science, bioinformatics, and computational biology tools, rather than off the shelf solutions

Building an effective cyberinfrastructure for phylogenetic biology will require solutions to a host of practical and theoretical problems in computer science, bioinformatics and phylogenetics itself. Some are well characterized but computationally daunting, such as developing good heuristic solutions for several NP-

complete problems for which very large data inputs are now at hand (multiple sequence alignment; construction of phylogenetic trees; assembly of synthetic "supertrees"). Others are not so well characterized but are emerging as problems unanticipated before the availability of such large quantities of sequences and other data (phylogenetic incongruence between different regions of the genome; the complexities of gene family diversification). Still other problems arise because of the breadth and heterogeneous nature of data than can be brought to bear on building phylogenetic trees: morphology, development, gene expression and other postgenomic data, etc. The canonical informatics problem of data integration is exemplified in our domain by the vast array of diverse datasets that retain footprints of evolutionary history.

Few of these problems currently have off-the-shelf solutions. There is, however, substantial prior experience in the phylogenetics community for building a next-generation phylogenetics cyberinfrastructure. On the one hand, the diversity of mathematical and computational methodologies for solving specific problems in phylogenetics has never been higher (e.g., recent work on building reticulate histories in hybridizing plants). This has been enabled in large part by substantial buy-in from the math and computer science communities. On the informatics side, the NSF-funded Cyberinfrastructure for Phylogenetic Research (CIPRES) project, which will soon be ending, tackled several key issues, including an upgrade of TreeBASE, the community repository of phylogenetic trees. Through its AToL and other programs, NSF has supported several smaller infrastructure-related projects, including TOLWeb, TOLKIN, PhyLoTA, pPod, and others. Moreover, many bioinformatics resources have added phylogenetic components in recent years (e.g., the phylogenetic trees in PFAM and GenBank's BLAST server). However, an overarching infrastructure fostering true high-level integration is still largely missing. These efforts need to be focused and coordinated, and much more attention needs to be given to synthesizing knowledge and making it truly useful for research and education.

Description of the datasets currently available or that will soon be available

Phylogenetic biology in many respects is not data limited. GenBank alone contains at least one DNA sequence for about 10% of all species known to science: this includes some 26 million sequences across 67,000 species of green plants. Phenotypic data have also been databased for protein structure (PDB), gene expression patterns (NCBI GEO), pathway networks (KEGG), and external morphology (MorphBank, etc.), to name just a few. These tend to be smaller repositories, but their information content is high owing to the inherent complexity of their traits compared to simple nucleotide sequences. Other kinds of databases are also relevant: paleontological databases archive important data on fossil taxa (Paleobiology Database); biodiversity collections databases (e.g., GBIF) can tie species to georeferenced collection sites, bringing geography, ecology and climate into the mix; and nomenclatural-taxonomic databases, which help to solve important problems in disambiguating the identities of species. In the next few years, the greatest explosion of growth in data for phylogenetics will no doubt arise from next-generation sequencing technologies that promise to deliver whole genome sequences for a substantially larger sample of the plant tree of life than we have seen up to now.

Thus we need not worry about stimulating the growth of the relevant databases. If anything, the problem is largely the opposite – the phylogenetics community is wallowing in so much data that it is getting increasingly difficult to navigate it, summarize it, render it into a usable representations, or, importantly, to make grand syntheses based upon it.

Goals and outcomes of this GCW

The primary goal of the workshop is to bring experts on plant phylogeny and several related botanical disciplines together with computer scientists, computational biologists, and bioinformaticians, to envision a cyberinfrastructure to enable phylogenetic research and to make the results of such research maximally useful across the plant sciences. Among those present will be leaders of a number of phyloinformatics projects that are actively trying to bridge this gap. The meeting will be designed to accommodate the exchange of information and the identification of current barriers, and then to map these barriers to specific cyber-solutions.

We expect this GCW to lead to the development of a Grand Challenge Project within the iPlant Collaborative, and that in this context “discovery environments” would be developed to advance the cause of integration, synthesis, and training. In the meantime, a white paper will be produced from the GCW to summarize the current state of affairs and to project a vision for the future. The GCW will also serve the immediate purpose of stimulating connections among the participants and the many perspectives and projects that they represent.

Special attention will be paid to the needs of the wide variety of communities who might benefit by the use of phylogenetic knowledge. This includes research communities spanning the plant sciences, from genomics to ecosystem ecology. We also will address challenges and opportunities in education and outreach, including the K-12 and college levels, informal science education, and coordination with ongoing efforts, for example at National Evolutionary Synthesis Center (NESCent).

What other groups are working in this or related areas? Which people from these groups are likely to be involved in the GCW?

A variety of ongoing projects bear on the aims of this GCW, and we intend to bring representatives of these efforts together, along with other plant phylogeneticists and computer scientists. These include the NSF-funded CIPRES project, various ATOL projects, and a variety of databasing efforts and web resources. The leaders of most of these relevant projects will be invited to participate and have already expressed interest in joining us. Finally, developers and users of various pipelines and tools for management of genomic data will be represented.

Format of the meeting and outline of the agenda

We envision a three-day meeting of ca 35-40 participants, including representatives from the iPlant Collaborative and the Board of Directors. The meeting would be held during the winter months in Oracle, AZ. Owing to various planning and timing constraints that we have already identified, the meeting could take place in late November, 2008, or in late January, 2009. We imagine the workshop taking place on a Thursday-Saturday. Participants might arrive early on Wednesday for a hike in nearby Catalina State Park, and for a reception that evening at the Oracle Conference Center. The first day of the meeting would be devoted to short presentations by participants to highlight progress and barriers in plant phylogeny and to review a variety of relevant computational tools. During the second day a variety of break-out groups will circumscribe the key barriers and their possible cyber-solutions, identifying common themes, overlaps, and possible synergies. The morning of the third day would center on education and outreach opportunities, on mobilizing the relevant communities, and on a time-table for future activities. The final afternoon would focus on outlining a Grand Challenge Project and on specific writing tasks and other commitments. On the Sunday following the meeting, the organizers and selected others will remain in Oracle to begin to write up the results of the GCW, and to plan the next steps.

Day 1 (Thursday)

AM – Session 1, 8:00 am – 12:30 pm, D. Soltis, moderator

8:00 Welcome, introductions, and opening comments by GCW PI Michael Donoghue and iPlant Representatives (Rich Jorgensen, others) – Overview of the iPlant Collaborative; “Grand Challenge Projects”; goals for this workshop, i.e., to identify cyberinfrastructure needs for the assembly and use of the Plant Tree of Life; development of the core of a proposal for an iPlant Grand Challenge Project

8:30-12:30 *Phylogenetic knowledge across green plants: progress, issues, and barriers*

PM – Session 2, 1:15 – 5:30 pm, P. Soltis, moderator

1:15-6:00 *Analytical and informatics challenges*

Day 2 (Friday)

AM – Session 3, 8:00 am – 12:30 pm, M. Sanderson, moderator

8:00 Take stock of the presentations made on Day 1, and summarize major conclusions. With this background, return to the issue of a “Grand Challenge Problem.” What are the major issues and major needs? What issues require special attention from the computer science, cyberinfrastructure, and informatics communities?

9:45 Coffee

10:00 Break-out groups to connect needs to possible cyber-solutions (5 groups of 7-8 people)

11:30 Summaries and identification of topics for further discussion

PM – Session 4, 1:30 – 5:30 pm, V. Tannen, moderator

1:30 Break-out groups on topics identified in the AM sessions, each to focus on refining a problem and its possible solutions (the exact make-up and sizes of the break-out groups will depend on the issues identified)

3:00 Coffee

3:30 Group presentations and general discussion, with the aim of identifying common themes and possible synergies

Day 3 (Saturday)

AM – Session 5, 8:00 am – 12:30 pm, T. Vision, moderator

Organizing the community and forming our iPlant team.

8:00 Break-out groups to discuss organization and governance, sociological barriers, mechanisms for communication, integration across sub-teams working on major sub-problems, communication with other iPlant initiatives, and possible education, training, and outreach opportunities.

10:00 Coffee

10:30 Reconvene to share suggestions and develop consensus

11:00 Development of an outline for a Grand Challenge Project proposal

PM – Session 6, 1:00 – 5:30 pm, M. Donoghue, moderator

1:00 Break-out groups to draft key sections of the proposal

3:00 Coffee

3:30 Continue drafting

4:00 Reconvene to discuss progress, remaining tasks, coordinate writing and other assignments

How can the iPlant Collaborative be useful to this GCW effort?

We trust that the iPlant Collaborative team in Tucson will provide logistical support for the meeting at Biosphere2, especially re. transportation issues, lodging, and meals, and will also coordinate with Conference Center staff on details of the meeting itself, including a reception, coffee breaks, and meals. Beyond this base-line support, we look to the Collaborative especially for advice and assistance in coordinating our GCW efforts with other such iPlant efforts. We would be delighted to entertain suggestions about plant and computer scientists whom we could include from other projects and other areas of expertise. Specifically, although a number of excellent computer scientists are among our prospective participants, we would welcome additional suggestions, and we would like to include iPlant computer staff who are already in place. Similarly, although a variety of plant science disciplines are represented by our prospective participants, we would benefit from additional suggestions, especially to connect with model-organism genome projects and their associated informatics projects. Likewise, member of our group may be appropriate in forming links to other iPlant projects.