

Assembling the Tree of Life for the Plant Sciences (iPTOL)

Principal leaders of the proposed collaborative

Main contact

Michael Sanderson, Department of Ecology and Evolutionary Biology, University of Arizona.
Email: sanderm@email.arizona.edu. Phone: 520-626-6848. Research/EOT interests: Computational phylogenetics, plant systematics; quantitative literacy in biology.

Plant Science Community Leaders

Michael Donoghue, Department of Ecology and Evolutionary Biology, Yale University. Email: michael.donoghue@yale.edu. Research interests: Diversity and evolution of flowering plants, using phylogenetic trees to understand patterns of diversification, character evolution, biogeography and ecology. As Director of Yale's Peabody Museum of Natural History he was directly involved in K-12 and family education and outreach activities, including the production of a museum exhibition entitled "Travels in the Great Tree of Life."

Pamela Soltis, Florida Museum of Natural History. University of Florida. Email: psoltis@flmnh.ufl.edu. Research/EOT interests: Angiosperm phylogeny, polyploidy (both ancient and recent), and the origin and evolution of the flower; student mentoring and public outreach through teacher education and museum exhibits and programs.

Douglas Soltis, Department of Botany, University of Florida. Email: dsoltis@botany.ufl.edu. Research/EOT interests: Angiosperm phylogeny, genetic and genomic consequences of genome doubling (both ancient and recent), phylogeography, conservation genetics, and the origin and subsequent diversification of the flower.

Computational Science Community Leaders

Val Tannen, Department of Computer and Information Science, University of Pennsylvania. Email: val@cis.upenn.edu. Research/EOT interests: Databases and bioinformatics; systems for data integration and sharing between collaborating scientists, on data provenance, on phylogenetic data modeling and on the integration of AToL data resources.

Alexandros Stamatakis, Department of Computer Science, Technische Universität München. Email: stamatak@cs.tum.edu. Research/EOT interests: design of algorithmic and HPC solutions for large-scale phylogenetic inference; fostering communication and collaboration between computer scientists and biologists.

Todd Vision, Department of Biology, University of North Carolina. Email: tjv@bio.unc.edu. Research/EOT interests: Computational genomics and genome evolution. He teaches courses in computational and evolutionary genetics at UNC Chapel Hill, and has been an organizer since 2007 of the Phyloinformatics Summer Courses and Phyloinformatics Summer of Code through the National Evolutionary Synthesis Center. Vision has worked with the Destiny Science Bus program to bring inquiry-driven bioinformatics, plant biology and evolutionary biology educational opportunities to underserved secondary students in North Carolina.

Project Summary

Knowledge of evolutionary relationships is fundamental to much of biology, yielding new insights across the plant sciences, from comparative genomics and molecular evolution, to plant development, to the study of adaptation, speciation, community assembly, and ecosystem functioning. Although our knowledge of the phylogeny of green plants has expanded dramatically over the past two decades, the task of assembling a comprehensive “tree of life” for the half million species of green plants remains daunting. The rapid accumulation of data relevant to reconstructing the tree of life has far exceeded all expectations. Sequence databases are growing faster than computer processing power, outpacing even “Moore’s Law” (Goldman and Yang 2008). We now find ourselves awash in unconsolidated information that could shed light on the history of plants. Major barriers confront us in managing and synthesizing the relevant data for assembling the tree, undertaking analyses based on it, visualizing it as it grows ever larger, and developing tools to put this growing knowledge-base to work. This presents a Grand Challenge, the solution to which requires dramatic expansion of a developing intersection between phylogenetic biology and the computer sciences, broadly conceived, with an eye toward integrating data and facilitating research and education across the plant sciences.

We propose to bring plant and computer scientists together to focus on the grand challenge of assembling all knowledge of the phylogeny of green plants to build a comprehensive phylogeny fully 100 times larger than the largest existing trees, to build a cyberinfrastructure for the dissemination of data associated with trees, and to implement scalable “post-tree” analysis tools to enable integration of the plant tree of life with the rest of the botanical sciences. The undertaking to unravel the evolutionary relationships among all living things, and to express this in the form of a phylogenetic tree of life, is one of the most profound scientific challenges ever undertaken, and represents a true “moonshot” for the life sciences. Construction of a cyberinfrastructure for green plant phylogeny will undoubtedly enable key facets of this broader effort in the life sciences.

This is an appropriate time for this proposed grand challenge effort. Most of the information on phylogeny is dispersed among disparate data sets and databases, and thus is not readily accessible to the research and education communities. Furthermore, the necessary cyberinfrastructure improvements are so broad and cross-disciplinary that they are not likely to materialize through normal NSF funding sources. A concerted iPlant effort in this area is not only necessary, but also highly likely to yield enormous successes. Importantly, the plant phylogenetics community has an established history of broad-scale collaboration and has consistently been willing to adopt new cyber-solutions. Equally, this area is also rich in challenging problems of direct interest to computer scientists and bioinformaticians. Early success in addressing the plant phylogeny problem would enable developments throughout the plant sciences, and would be especially useful in connection with other Grand Challenge Projects supported through the iPlant Collaborative that involve comparisons between genes, genomes, or species. In the process, a project focusing on phylogeny would promote the integration of plant evolutionary biologists throughout the iPlant community, and, in doing so, would insure the broad impact of the project as a whole. Finally, the plant tree of life provides exciting opportunities for training and outreach at all levels. Since Darwin, the tree of life has proven to be a very accessible visual metaphor for nonscientists, providing an elegant opening for communicating results in the plant sciences and evolutionary biology to people with diverse backgrounds.

I. THE GRAND CHALLENGE

A. Introduction to the Problem

The mission to unravel the evolutionary relationships among all living things, and to express this in the form of a phylogenetic tree of life, is one of the most profound scientific challenges ever undertaken. With more than 1.8 million known species, and many more millions of undiscovered and extinct species, the size of the tree is almost beyond comprehension (Blackmore 2002; Pennisi 2003; Wheeler 2004). Yet, despite the magnitude of the problem, new methods, new sources of data, and new computational resources have combined to yield unprecedented progress in reconstructing the history of life (e.g., Cracraft and Donoghue 2004; Bininda-Emonds et al. 2007; Moore et al. 2007; Jansen et al. 2007; Dunn et al. 2008; Smith and Donoghue 2008; Hackett et al. 2008). And, as the tree comes more clearly into view, we have increasingly come to appreciate the value of this knowledge. As a telescope into the past, the tree provides a powerful new lens through which to interpret the patterns and processes of evolution, and also the ability to predict the responses of life in the face of rapid environmental change. Just as the sequencing of an entire human genome provided countless, largely unanticipated new biological insights, reconstructing the entire tree of life will undoubtedly fuel fundamental research and the development of practical tools to sustain biological diversity and enhance human well being.

Phylogenetic trees help us understand the tempo and mode of evolution of species and clades, of genomes, of developmental systems, and of the distribution and interactions of organisms in communities and ecosystems. Phylogenetic methods are being put to use in identifying and combating diseases, (in humans and crops; Davies and Pedersen 2008; Bardel et al. 2005), conserving biodiversity (Purvis and Gittleman 2005), and predicting responses to contemporary environmental concerns such as global climate change (Willis et al. 2008) and biological invasions (Proches et al. 2008). In short, knowledge of the tree of life creates a new, exciting, and complex discovery environment that will change how we teach biology at all levels and that will generate new opportunities for interaction with the world at large. The realization of this ideal depends on the completeness of the tree and its accuracy, as well as on its linkages to other information, such as data on genome organization, gene expression, or geographic occurrences and environmental parameters.

Phylogenetic data have accumulated so rapidly that we are now awash in mostly unconsolidated information that could shed light on the history of life. Worldwide efforts to survey biodiversity, as well as organized programs, such as the NSF's "Assembling the Tree of Life" (ATOL) program, have greatly accelerated the rate of accumulation of both molecular and morphological data, and have fostered new analytical methods (e.g., NSF's CIPRES project: "Cyberinfrastructure for Phylogenetic Research"), but we are still largely unprepared to scale this activity up to the levels that are now needed. To appreciate the problem, consider that GenBank already contains sequence data on over 70,000 green plant species, but these data cannot readily be combined into a single tree (Sanderson 2008). Moreover, ATOL and CIPRES efforts have not focused specifically on connecting phylogenetic trees with data beyond the ATOL community to provide new mechanisms for the visualization, navigation, and integration of these data to enable exciting (and largely unpredictable) insights and discoveries.

Plant scientists have been leaders in the global effort to resolve the tree of life, and are widely regarded as being at the vanguard in organizing large collaborative efforts and in pushing the limits of analysis, synthesis, and the use of phylogenetic information (reviewed in Soltis et al. 2005). Through years of coordinated effort, plant phylogeneticists have assembled and analyzed massive datasets, and we are now on the verge of obtaining whole plastid genome data from hundreds – soon thousands – of plant species. The plant phylogenetics community has also been at the forefront of adopting cyberinfrastructure and in developing new scientific and educational uses for phylogenetic trees. Hard problems with special significance in plants, such as reticulating phylogenies and the history of whole genome duplications, have especially excited mathematicians and computer scientists (Nakleh et al 2005; Huber et al. 2006; Sankoff et al. 2007). Recently, a three-month-long workshop devoted to phylogenetics was held at the Isaac Newton Institute for Mathematical Sciences in Cambridge, UK (<http://www.newton.ac.uk/programmes/PLG/index.html>), attended by several members of our team. The

construction of large phylogenies was identified as a core emerging challenge. A concerted effort through the iPlant Collaborative to develop the cyberinfrastructure to support this would surely provide a model for the other biological sciences, and at the same time would foster fundamental research in computer science, bioinformatics and mathematics that will yield solutions of very general interest.

We therefore propose the establishment of a mechanism – the iPlant Tree of Life (iPTOL) project – to address the grand challenge of constructing the plant tree of life to understand the diversification of green plants over the last billion years, and to build a cyberinfrastructure to connect this tree to the rest of the plant sciences and beyond.

B. Impact on the Broader Field

Unraveling the evolutionary relationships among all living things, even among all green plants, and expressing this in the form of a phylogenetic tree of life, is one of the most profound scientific challenges ever undertaken, and represents a true “moonshot” for the life sciences. The successful implementation of this project will have a dramatic impact far beyond phylogeny reconstruction per se and “just” the tree of life. The outcome will be a comprehensive phylogenetic framework with which we can understand how plant genomes and developmental systems have evolved and influenced the course of adaptive evolution and diversification. This framework will revolutionize our ability to infer the movement of plants around the globe, the assembly of ecological communities, and the development and maintenance of biodiversity and functioning ecosystems. Phylogenies of plants find uses in numerous applied areas of plant science as well, such as conserving biodiversity and predicting biogeographic or ecological responses to global climate change and biological invasions. As brief examples, consider plant genomics and plant ecology. Plant genomes have undergone explosive diversification in parallel with the diversification of plants themselves (Vandepoele and Van de Peer 2005). Due to evolutionary processes such as gene and genome duplication, recombination, and horizontal transfer, the phylogenetic histories of individual genes are distinct from one another and the phylogeny of the organisms in whose genomes they reside. In particular, there has been a high frequency of gene and genome duplication in plant evolution (e.g., Vision et al. 2000; Cui et al. 2006; Soltis et al. 2009), with the consequence that simple orthology fails to be a useful concept for plant comparative genomics except between very closely related species. Furthermore, each separate gene lineage has undergone functional evolution in response to its unique organismal and genomic context. This presents a formidable cyberinfrastructure challenge: to elucidate how evolving biochemical pathways and regulatory networks have contributed to developmental and phenotypic innovations in different plant lineages, molecular biologists require tools that relate how the relevant gene families have functionally evolved within the context of the larger organismal phylogeny.

The nodes and edges of the green plant tree of life provide a common coordinate system for plant functional data in the same way that latitude and longitude coordinates enable users to mash-up different georeferenced data sources using Google Earth. We foresee users wanting to explore the plant tree of life by annotating it with functionally significant “phylo-referenced” events such as the duplication and loss of genes, changes in transcriptional or translational regulation, changes in protein domain composition or structure, bouts of positive selection at the amino acid level, and so on, for sets of gene families relevant to particular biological questions.

Integration of these data within the context of gene trees and species trees would allow researchers to address numerous questions. What factors influence the relative importance of subfunctionalization, neofunctionalization, and the retention of ancestral function in duplicated genes? What role has polyploidy played in the origin of innovations such as the flower, the seed, and vascular tissue? Have plants been functionally enriched by horizontal gene transfer from fungal, bacterial, and metazoan symbionts? How have gene networks been co-opted in whole or part at different times for different purposes, e.g. vegetative versus seed desiccation-tolerance (Fisher 2008), or mycorrhizal vs. rhizobial symbiosis (Gherbi et al. 2008)? What genomic differences lead some plant lineages, and not others, to be predisposed to evolve complex phenotypes such as C4 photosynthesis (Sage 2003)? What is the molecular basis for the alternation of diploid and haploid generations in plants? An open platform for

comparative genomic analysis as described here will, in ways limited only by the imagination of future scientists, contribute to our understanding of basic plant biology, inform the translation of basic discovery to agriculture, and contribute to efforts to engineer biosynthetic pathways for human welfare.

A comprehensive green plant tree of life will also provide insights into the processes underlying species diversification and the formation of ecological assemblages. The fastest-growing user community of plant phylogenies is ecologists (e.g., Silvertown et al. 1997; Pugnaire and Valladares 2007), who are exploring how global patterns of species diversity relate to patterns of phylogenetic diversity, trait or characteristic diversity, ecological gradients, and biogeography and plant migration, both historical and present-day. Access to a dynamic and continually updated plant tree of life would let ecologists address numerous specific questions: How deep in the tree can ancestral states of trait values be reconstructed accurately, and do changes in trait values match changes in historical climates and biogeography? Are hotspots of species diversity also hotspots of phylogenetic diversity? Are individuals/taxa from a particular locality phylogenetically clustered? If an individual/taxon is sampled from a particular locality, are related species in geographically and/or ecologically similar places? What are the limitations on community assembly and the relative importance of dispersal vs. phylogenetic constraints (Donoghue 2008)? What environmental factors are most important as selective agents in evolution?

Besides the many basic science questions that can be addressed using the plant tree of life, there are also many uses in applied disciplines, including conservation and restoration, agriculture, the biology of invasive and rare species, and climate change. Of particular importance today is the potential for increased understanding of historical responses of plants to environmental factors (e.g., evolutionary changes in trait values, migration) to help predict shifts in species traits and community assemblage under future climate regimes (Willis et al. 2008; Davies et al. 2008). Better predictions of future communities can then be used to understand potential changes in ecosystem functions, such as carbon turnover rates.

The interaction of genetic and environmental variables is critical in shaping plant morphology (the evolution of developmental programs), adaptive life history strategies (adaptation), and the ecology and evolution of plant assemblages/communities. These genetic and environmental interactions coupled with population biology factors (e.g., population size, natural selection, rates of mutation) play an important role in structuring genome architecture, including the evolution of co-evolved gene complexes. iPTOL is uniquely positioned to build a phylogenetically based infrastructure placing genes, including genes from large gene families, underlying physiological and developmental processes into an ecological and biodiversity context.

In summary, the success of this project will not only greatly expand connectivity within the plant phylogenetics community, but will link this community with the other plant sciences through the development of computational tools that will significantly enhance productivity and understanding. To accomplish our goals we will leverage our past experience in collaborative research, forming a steering committee and working groups to envision and develop the necessary cyberinfrastructure. Additionally, iPTOL will provide fertile ground for collaborations that will advance computer science, bioinformatics, mathematics and the development of cyberinfrastructure more generally.

C. Broad Activities Needed to Make Progress toward Solution of the Grand Challenge

We propose to develop a “discovery environment” to integrate and disseminate phylogenetic information, reconstruct and visualize the green plant tree of life, apply it to known and unanticipated problems in the plant sciences, and serve as a platform for outreach. Key objectives will be: (1) to develop of new tools to streamline the mining and assembly of relevant datasets from disparate databases and data resources to carry out sophisticated, and in many cases, massive phylogenetic analyses; (2) to develop new scalable tools that use the tree of life as a gateway for the analysis of a diverse array of evolutionary, genetic, genomic, developmental, and physiological processes, (3) to build scalable tools to visualize, annotate and communicate very large phylogenetic trees for research and outreach; and (4) to construct incrementally larger and more complete phylogenies, with an ultimate target of the 500,000 described green plant species.

Goals 1 and 4 have specific data requirements, some of which are well understood (e.g., the existence already of a diverse collection of databases of phylogenetic trees that can form the basis of some of our work), whereas others will require development of strategies for data assembly (such as optimal mining of sequence databases to build large trees), as well as an adaptive response to burgeoning new sources of data from projects such as DNA barcode surveys. All four goals require interdisciplinary collaboration with computer scientists, especially in the areas of algorithm development and high performance computing—over and above collaboration with implementation experts at iPlant.

D. Barriers in Performing these Activities

The rapid accumulation of data relevant to reconstructing the tree of life has far exceeded all expectations. GenBank contains 31 million nucleotide sequences for plants. The world's herbaria serve as repositories for more than 100 million plant specimens, each acting as a set of observations on morphology, geographical location, and ecological context and serving as a potential source of genetic resources. These and other potential data sets offer different degrees of accessibility, a variety of interface formats, and much semantic heterogeneity. The task of integrating them faces both technical and management difficulties. Providing the owners of this information with easy-to-use tools and methodologies is crucial for securing their cooperation. Conquering the semantic heterogeneity of these sources requires mapping the structure of their information onto the concepts to which the plant tree of life can offer direct and intuitive access, such as trait evolution.

In addition to this wealth of data, almost all of the computational problems (e.g., sequence alignment; tree optimization; most supertree construction methods; ancestral gene order estimation) in this domain are formally “NP-complete”, meaning they require clever heuristic methods and high-performance implementations to grapple with their fundamental intractability and return reasonable solutions (e.g., Steel 1992; Roch 2006). In fact, this project represents an important opportunity to promote new thinking in computer science.

On the hardware level, the current multi-core revolution in computer architectures also poses new challenges to the development of novel algorithms for phylogenetic analysis. Because parallel computing has become mainstream now even at the desktop level and several of the current top 500 supercomputer systems are based on multi-core or accelerator architectures, it will be necessary to simultaneously develop algorithms and respective parallelization strategies that will fit well onto emerging parallel architectures. Hence, our computational needs are immense; we are now confronting major barriers in managing and synthesizing the relevant data, in producing viable algorithms, in visualization, and in the development of hardware and software tools to put this exploding knowledge-base to work. The enormous wealth of largely unconsolidated data presents a Grand Challenge whose solution clearly requires dramatic expansion at the intersection between phylogenetic biology and the computer sciences.

E. Specific Proposed Computational/Cyberinfrastructure Activities

Large-scale phylogenetic analyses, involving hundreds to thousands of species, have become increasingly common (e.g., Ley et al. 2005). The largest plant phylogenies constructed algorithmically have 2200–4600 species (e.g., Kallersjö et al. 1999; McMahon and Sanderson 2006; Smith and Donoghue 2008), barely 1% of the ~500,000 described plant species. Larger trees have only been assembled with semi-formal methods (Moles et al. 2005). The challenges involved in scaling existing computational capabilities up by two orders of magnitude to obtain trees of 100,000–500,000 species are enormous. Our workshop participants compartmentalized the solution into four goals, three of which provide either the cyberinfrastructure foundation or the tools for dissemination of the results obtained in the final push toward the fourth goal of actually constructing a comprehensive tree of all green plants. These goals are described below in the order in which we believe substantive progress and products can be delivered to the community. Following this detailed description, we illustrate how a user outside of the phylogenetics community might interact with the developing iPTOL cyberinfrastructure.

Goal 1. Database integration and data assembly. The discovery environment we envision will facilitate the use of the tree of life as a way of organizing, visualizing, and using all the data of plant biology in a phylogenetic context. At the same time, large and heterogeneous datasets will need to be assembled, periodically or “on the fly”, in order to enable the large-scale phylogenetic computations necessary to infer a comprehensive plant Tree of Life.

Databases of phylogenetic trees already archive hundreds of thousands of relatively small trees that can form the starting point for synthetic analyses, and several other data sources (e.g., sequences, alignments) will be crucial for the comprehensive analyses. Tree databases range from those designed explicitly for phylogenetic biology (TreeBASE, PhyLoTA Browser, ToL Web Project), to databases aimed at the gene and protein evolution communities (PFAM, Phytome, Phylofacts, PlantTribes). These databases differ in syntax and semantics, and vary tremendously in the sophistication of their application program interfaces. Database integration is a task well-known for its difficulty, and efforts to integrate “everything” have consistently failed. Therefore, we believe that taking an *incremental* approach, especially to on-the-fly integration, would be most productive. We envision first focusing on resources that are critical for subsequent steps and that are managed by iPTOL participants. This will both focus our efforts and demonstrate the benefit of the environment. Subsequently, we should actively invite and enable other parties to integrate their resources into the discovery environment through activities such as hackathons.

A crucial first step in integration of these data will be a specification of persistent and unambiguous identifiers for the *nodes, leaves and branches* of phylogenetic trees. The adoption of such identifiers would permit the attachment of a diversity of properties (gene function, biogeographic distribution, etc) for downstream applications. When applied to a species tree, such identifiers can be used to reconcile the sometimes conflicting systems of Linnean and phylogenetic classification and to facilitate computation over tree topologies. Emerging phylogenetic data standards exist that can be adopted for data exchange and semantics, and will not need to be invented de novo.

The workshop concluded that it was important to provide, within the discovery environment, “one stop shopping” for plant biologists to obtain best current estimates of the phylogeny of sets of genes and species of interest early on, well before we achieve the goals described below of computing more comprehensive plant phylogenies with hundreds of thousands of taxa. Therefore, we envision tree-discovery, tree-synthesis and tree-analysis modes for the iPToL discovery environment. In the first two modes, users choose one or more of species or genes of interest and the tool brings in the best current trees that are relevant, with the option for synthesis of those trees. In the third mode, the users investigate post-tree analyses on the trees built in the first mode, as described next.

Goal 2. Post-tree analysis. Most plant biologists want to use phylogenetic trees for comparative study *after* they have been constructed (by someone else), and numerous software tools are available for such analyses (e.g., Mesquite: Maddison and Maddison 2008). However, most are not designed from the ground up to be scalable to very large trees. Workshop participants identified two specific post-tree tools as highest priority: reconciling gene trees with species trees and inferring ancestral character states. We envision developing analysis tools early that will be useful with existing small to moderate-sized trees *and* will scale to the very large trees we plan to build.

(i) *Tree reconciliation* uses an estimate of the species tree to infer the history of gene duplication and loss, lineage sorting, lateral transfer, and other events in a gene family’s history (Page 2002). It thus has wide applicability in genomics and molecular biology, but has been used relatively infrequently, not because of lack of theory but of implementations. Recently, substantial progress has been made on both algorithms and software development (Durand et al. 2006; Bansal et al. 2007), but important problems remain, including scaling implementations to the size of the largest known gene families and species trees to be estimated, and handling uncertainty in the reconstruction of both gene and species trees.

(ii) Algorithms for *reconstruction of ancestral character states* have been used for decades (Fitch 1971; Pagel 1999), but are now finding increasingly broad utility from ancestral genome reconstruction (Blanchette et al. 2004) to inferences about past climates (Yesson and Culham 2006). Not only do these

inform evolutionary studies, but some have fostered experimental research programs testing functions of ancestral proteins in hypothesized ancient environments (Yokoyama et al. 2008). Some of these algorithms are easily scaled to large trees, but no such software implementation yet exists; others do not scale as well and require algorithmic or engineering improvements.

The kinds of data that end users might expect to bring to these analyses are diverse, especially for (ii). Traits can include anything from sequences, to gene expression levels, to presence/absence of a secondary compound, to climate variables (Evans et al. 2009). Since it is not possible to anticipate every input that users might use, nor every external data source, one challenge is to design the discovery interface such that it provides relatively flexible and simple methods for external data to be imported or retrieved.

Goal 3. Tree visualization. Visualization is a fundamental challenge when trees get larger than 50-100 species. The phylogeny of all plants, if printed out on a giant piece of paper with species labels printed at 10-point font size, would be 2116 meters long, which is almost five times the height of the Empire State Building. Imagine visualization and analytical software that can zoom out to see the landscape of five Empire State Buildings end to end, yet still be capable of zooming in to read the text on newspapers in the hands of people sitting by their office windows. Then imagine this vast and gigantic tree used as a giant computational template to infer the historical patterns for any known biological data -- for example, detecting orthologous, paralogous, and xenologous genes, tracing the historical sequence of genome reorganizations, correlating genomic changes with patterns of plant innovation and adaptation, and examining patterns of adaptation with geographic vicariance, climatic change, or coevolution. These annotations of trees are important for communicating many of the results provided by our work under Goal 2. Moreover, improving visualization is important not only for surmounting this emerging obstacle to interpreting research results, but also for directly reaching out to a much broader audience, because the metaphor of the tree of life is widely understood by the general public.

Several models for scaling tree visualizations have been explored (Munzner et al. 2003; Hughes et al. 2004; Sanderson 2006; Jordan and Piel 2008), but one of the great obstacles to progress in tree visualization is almost more esthetic than technical. What kinds of information ought visualizations convey? We believe it essential to organize a workshop to bring together users, designers, and computer scientists to ask very fundamental questions about the goals of tree visualization to begin a process that will lead to a scalable and informative visualization tool.

Goal 4: Reconstructing green plant phylogenetic trees with upwards of 500,000 taxa. A factor of 100 separates the largest plant phylogenies currently being constructed and the tree of approximately 500,000 extant species of green plants we propose to build. We will approach the daunting scalability and data availability problems by combining incremental increases in analysis size with flexible views about the shape of the final product. With respect to scalability, there is reason for optimism that data sources are available and methods can be scaled up at least one order of magnitude to ~50,000 species. Trees this size can be built using existing algorithms, and we know sequence data are already available in some form for that many plant species. The grand challenge will involve the next order of magnitude. With respect to the shape and scope of the final tree, a conservative strategy would be to accentuate methodological rigor and data homogeneity and build a “forest” (Mossel 2007) of disjointed well-supported large phylogenies from the largest collection of sequence data possible at the time. Anticipating the growth of GenBank and proposals to rapidly augment biodiversity sampling via DNA barcoding projects (Kane and Cronk 2008), this could amount to over 100,000 species in the next five years. The “tree” would not contain that number of species but the sum over a collection of large trees could. Alternatively, a more complete tree could be built by incorporating much more heterogeneous lines of evidence, such as taxonomic “trees” based on classifications, which are far more complete at the species level than are molecular or morphological sequence databases. A model for this strategy is the Phylomatic system (Webb and Donoghue 2005), which incorporates expert knowledge from the APG classification of all angiosperms (Bremer et al. 2003) with inferred phylogenetic trees of smaller clades. We envision a discovery environment that would convey both ends of this spectrum of strategies to the user.

Two separate technical subproblems are central. The first arises when using a single locus (or several

such loci together) for a very large number of taxa. For example, in plants, there are ~30,000 sequences of plastid *rbcL*, which has long been a workhorse gene for plant phylogenetics (Chase et al. 1993). New plant barcoding efforts promise to deliver several such loci for large numbers of plants. The second arises in assembling data from many loci, each covering only a subset of all taxa so that some data are missing for some species. The information in these supertrees or supermatrices is generally fragmented in ways that pose significant problems to tree construction (Sanderson et al. 2008).

Single large alignments. Very good methods are already available for estimating phylogenies from aligned data, including methods for maximum likelihood (RAxML, GARLI, and Phyml, PAUP*), for maximum parsimony (TNT and PAUP*), and Bayesian MCMC (e.g., MrBayes). Of these methods, RAxML and TNT may be the two most scalable programs for ML and MP respectively. Current testing of these methods suggests that analyses of perhaps 50,000 sequences are feasible, given perhaps a week (or longer) and sufficient hardware, as memory requirements can be substantial. RAxML also has a version with very fast bootstrapping to assess confidence limits, and a new (not yet publicly available) version enabled with a stopping criterion for the bootstrapping that makes it possible to obtain highly accurate support estimates without extreme effort. Several subprojects are clearly necessary:

(i) Extending these methods to be able to analyze datasets of this size with less time, taking advantage of specialized hardware where needed. Alexandros Stamatakis (Munich), the designer of RAxML, is one of the principal leaders of our proposal.

(ii) Extending these methods to be able to analyze datasets of 100,000 to 500,000 taxa. To achieve this goal, it will be essential to relax the general requirement of finding a good local optimum. Instead, the goal will be to identify the regions within the tree that cannot be completely resolved (due to lack of data), and to output a tree that is the best approximation to the true tree, given the amount of data and the amount of computational resources. Factors that would lead to such lack of resolution include branches being too short for the amount of available sequence data, rogue taxa that can attach to many places within the tree, or (more generally) branches being so long that the subtrees at the ends of the branch look random with respect to each other. The first scenario is very common in phylogenetic analyses, especially for very large datasets, and is easily represented by polytomies in the output; the challenge, however, is to modify the search heuristic so that these unresolvable nodes are identified early in the search, so that effort is not wasted trying to resolve those nodes. The second and third scenarios call for a different treatment, with the output being potentially a forest of trees (Mossel 2007), rather than a single tree, and hence also call for modification to the search heuristics.

(iii) Developing new methods for estimating confidence levels in large trees. Estimation of accuracy of inferred trees is both computationally intense and statistically complex (Sanderson and Wojciechowski 2000; Alfaro and Holder 2006), but it is essential to interpretation of phylogenies—without support estimates, it is not clear which clades are reliable and what downstream post-tree analyses can be supported. Bootstrapping is a standard approach for estimating support (Felsenstein 1985), but may not be scalable for very large datasets (with more than 50,000 sequences). Alternative methods for estimating support will need to be investigated.

Supertree and supermatrix assembly. Supertrees are trees built from smaller trees constructed from separate loci. Supermatrices are assemblies of the raw sequence alignments into a larger alignment prior to conventional tree construction. Of the various methods that have been developed for supertree assembly (Bininda-Emonds 2004), Matrix Representation with Parsimony (MRP) and its variants have been used most widely. However, there is no production-quality implementation of MRP. Worse yet, even MRP fails to recover highly accurate trees under fairly common conditions such as low taxon overlap between trees (Bininda-Emonds and Sanderson 2001). Therefore, we propose the following:

(i) Develop a production-quality implementation of MRP, so that users can provide just the set of trees on the taxa, and receive the supertree back as output.

(ii) Experiment with other approaches for supertree construction, with the objective of developing a new supertree method with substantially improved accuracy.

Assembly of supermatrices also requires substantial work. Some of the components have been

designed and built (e.g., the *Mor* project, Hibbett et al. 2005; *PhyLoTA* project, Sanderson et al. 2008; *Phylogeny.fr* project, Dereeper et al. 2008), and the basic challenge is to engineer a workflow that puts these pieces together to guide assembly from the data mining stage through sequence clustering, alignment, data set combination, and finally tree building (Ciccarelli et al. 2006). Importantly, there must be feedback from the downstream elements of this pipeline back to the earliest phases, because experience demonstrates that the quality of the final supermatrix assemblies is highly sensitive to early decisions. Improvements made to the analysis of single large alignments, described above, will also apply to phylogenetic inference from supermatrices.

Ancillary algorithmic research: Workshop participants discussed several additional important problems that are either highly relevant to plant phylogeny reconstruction or would generally improve the efficiency or accuracy of tree reconstruction in general. Although none is critical to the success of our grand challenge project, each is sufficiently important to merit attention. Several of the actors in these areas work outside the US, and consequently we propose an international workshop on algorithms in these areas (see below).

(i) *Reticulate phylogeny, polyploidy.* Hybridization is common in many groups of plants, and a large fraction of plant species have hybridization and/or polyploidy somewhere in their evolutionary histories. In these cases, phylogenetic networks rather than trees are appropriate graphical representations. The challenges of this problem go far beyond tree inference and have excited the attention of computer scientists (Nakleh et al. 2005; Huber and Moulton 2006). Considerable work is ongoing, some in collaboration with botanists (Huber et al. 2006; PADRE software), but much remains to be done if new methods are to be applicable to a large phylogeny.

(ii) *Whole genome phylogeny.* Unequivocal evidence emerged from plant whole-genome sequencing for repeated rounds of whole-genome duplication (Vision et al. 2000; Bowers et al. 2003; Paterson et al. 2004; Tuskan et al. 2006; Jaillon et al. 2007; Ming et al. 2008). More generally, segmental duplications, rearrangements, loss of genomic regions, and chromosomal fusion and fission, have occurred in plants. Although fundamental advances have been made in using such data for reconstructing phylogeny (e.g., Blanchette et al. 1999; Boore 2006; Wang et al. 2006), these are still quite limited in the kinds of genomic events they model simultaneously. We will work to extend breakpoint methods and event-based methods to increase the kinds of genome-scale changes they can handle.

(iii) *Multiple sequence alignment.* Although highly accurate estimates of evolutionary histories are possible for many datasets, these estimates require that the true alignment be known, or very accurately estimated. Unfortunately, despite the plethora of alignment methods, none of these methods produce highly accurate alignments on large datasets that have evolved with many insertions and deletions (“indels”) and substitutions. Furthermore, there is increasing evidence that much of the problem in resolving evolutionary histories lies in the alignment process itself -- with poor alignments leading to poor phylogenies (Wong et al. 2008). One new avenue is the estimation of trees from *unaligned* data, but most of these methods are not scalable beyond about 50 sequences (Fleissner et al. 2005; Redelings and Suchard 2005). Two methods, POY and POY* (Liu et al. 2009a), can analyze 100 sequences, given several days, but they do not provide improved estimates of phylogenies. A promising new method, SATe (Liu et al. 2009b), being developed at the University of Texas by Warnow, Linder, and students, can analyze datasets with 1000 sequences in just 48 hours.

Imagining the user experience. How might a plant biologist take advantage of what iPTOL produces as it develops? Consider two examples. First, early in the project, iPTOL will provide non-phylogenetic researchers with the ability to integrate existing phylogenetic trees with valuable tools in order to address non-phylogenetic questions. A plant molecular biologist interested in the patterns of orthology among the genes related to a certain developmental pathway might use the discovery environment to take existing gene trees for different pathway members and reconcile them against a conservatively resolved species tree, thus revealing gene duplications, gene losses, orthology and paralogy. Ultimately, the user will want to visualize the duplications superimposed on multiple trees, each containing hundreds of genes/species. Though a relatively simple use case, it already demonstrates the

outcomes from goal 1 (integration of a species tree and gene tree from two different data sources), goal 2 (embedded reconciliation tools for large unresolved trees) and goal 3 (visualization of large trees).

We can imagine a second use case that would be enabled at a later stage in the project. A biologist wishes to identify whether some recurrent plant innovation (e.g., C_4 photosynthesis, or woodiness) has occurred more frequently at older or younger age intervals within the plant tree of life. This calls for as comprehensive a tree as possible, so that there is even sampling of both early branches and more recent ones. Such a tree would include hundreds of thousands of taxa, a scale that cannot currently be computed, but that is the explicit aim of goal 4, and is predicated on the successful outcome of goal 1 (data integration and assembly). Within the discovery environment, the user would infer the ancestral states (goal 2, post-tree analysis) and would naturally wish to visualize the results (goal 3) as well as summarize them quantitatively. Thus, late in the project, as a broader picture of the plant Tree of Life emerges, broad comparative hypotheses can be analyzed in a phylogenetic framework with the aid of scalable post-tree analysis tools and visualizations.

As the pace of genome sequencing continues to accelerate and the diversity of taxa surveyed blossoms, the data on which these scenarios depend will no longer be the limiting factor, but rather the computational resources that the end-user has at his/her command. In sum, these four goals, while ambitious, could not be more timely and, as a set, comprise a coherent, important and intellectually exciting Grand Challenge.

F. Estimated Collaboration Needs with CISE Domain and Available Expertise

Fortunately, the phylogenetics community has a long track record of collaboration with computer scientists, especially those working on algorithm theory. Our team includes computer scientists, such as Alexandros Stamatakis, Tandy Warnow, Katharina Huber, Val Tannen, and Tamara Munzner, experts in high performance computing, phylogenetic algorithms, databases, and scientific visualization, respectively. In addition, we have identified a large network of CS workers in these areas, many of whom participated in our workshop, with whom we will work throughout the project. Some of our goals involve application of high-performance computing and assessment of emerging parallel architectures to phylogenetic problems. Alexandros Stamatakis is a leader in this young field and has agreed to participate in these efforts. Other key issues relate to database integration. Val Tannen will play a key role managing these efforts, but there is also considerable expertise among our broad community participants in phylogeny-oriented databases, data assembly, and data exchange (e.g., team members William Piel, Michael Sanderson, Todd Vision). Thus, we are in a good position with experienced team members and collaborators to work with iPlant CISE personnel to design the iPTOL discovery environment.

G. Projected Cyberinfrastructure Needs

Database integration and data assembly. Much of our Goal 1 involves integrations of existing tree databases to permit data mining and assembly of data sets for tree building and post-tree analyses. We will need considerable support from database software designers and experts in the area of integration, including implementation of technologies for data exchange. We also require iPTOL to provide the expertise to design and develop the discovery environment (including user interface development for integration of post-tree analysis and visualization tools below), and host a server configuration that provides sufficient processor and memory capacity to enable the execution of multiple simultaneous compute-intensive operations from remote users.

Algorithm engineering for post-tree analyses. Goal 2 involves implementing algorithms that perform quantitative analyses on a previously computed phylogenetic tree. Many of these algorithms are already well understood and implemented, but are simply not scalable or have not been integrated into the kind of broad-based discovery environment we envision. iPlant personnel will be able to leverage a considerable knowledge base in our field on the performance and implementation of these algorithms, required data structures, etc., but new ideas will be needed to engineer scalability.

Visualization. Access to graphic design expertise and technical expertise in computer graphics, such as programmers familiar with OpenGL or other 3D toolkits will be essential for goal 3.

High Performance Computing Requirements. Our proposal's goal 4 envisions scaling phylogenetic tree construction up by between one and two orders of magnitude over the current largest data sets. This is undoubtedly the most computationally demanding of our goals. We will need access to a cluster with Multi-Core nodes, that have a sufficient amount of memory per node (64 or 128GB) and a fast infiniband interconnect to be able to exploit fine-grained and coarse-grained parallelism across nodes and fine-grained parallelism within nodes. Each node should have 16 or 32 cores to allow for rapid parallelization and prototyping with OpenMP or Pthreads and respective production runs. We have, for example developed a yet unreleased OpenMP-based version of MrBayes (Pratas et al. submitted), which currently represents the only viable solution to handle very large data sets using a Bayesian approach. For our maximum likelihood implementation, RAxML, we have shown that fine-grained loop level parallelism can be exploited within multi-core nodes as well as on clusters of multi-core nodes with a fast infiniband interconnect (Ott et al. 2007; Stamatakis and Ott 2008a,b). Therefore, RAxML is currently being considered as application for the SPEC-MPI benchmark. We are in direct contact with Intel, which is currently benchmarking RAxML for us on the new Nehalem multi-core architecture with QuickPath interconnect. The initial results are promising and the SMT (Simultaneous Multi-Threading) option seems to yield additional speedups. The significantly higher memory bandwidth compared to current AMD architectures will become important due to the rapidly growing size of biological datasets. Finally, it would be of great value to allocate funds to a stepwise acquisition of accelerator architectures like high-end FPGAs, GPUs, or the Intel Larrabee to allow for assessment of emerging parallel architectures for phylogenetic inference as they appear on the market.

H. What Falls within the Scope of the Collaborative? What does not?

The grand challenge of building the plant tree of life has broad scope, resting on a diverse array of data and methods, and enabling an equally diverse set of analyses during its construction. A feasible implementation of this project will therefore have to *limit* itself to (i) integration of data types that are both informative for phylogeny reconstruction and accessible in existing databases; (ii) development of algorithms for tree inference and post-tree processing that have potential for scalability to very large trees; and (iii) heuristic solutions to almost all the hard problems in the data analysis pipeline leading from sequence data to phylogenies at the proposed scale, which will necessitate new ways to validate the performance of such an ensemble of heuristics.

It will also have to *exclude* (i) a comprehensive treatment of taxonomic nomenclatural issues (a subject partly treated by the BIEN grand challenge proposal), (ii) integration or assembly of data that are tangential to the goals for the discovery environment or a comprehensive phylogenetic analyses, (iii) work on many types of post-tree analyses that are difficult to standardize for non-expert practitioners and/or will not have broad applicability, and (iv) any duplication of effort that would devalue existing resources or ongoing projects, such as TreeBASE

I. Education, Outreach, and Training (EOT) Objectives

iPlant offers exciting and novel approaches to education, outreach, and training at multiple levels, from K12 to the citizen naturalist to the scientifically literate layperson to the fledgling scientist in training. We envision creative ways to use cyberinfrastructure to teach about plant biology and new opportunities to train teachers and students in the use of cyberinfrastructure. We propose cross-training in biology and computer science for students of all ages and teacher workshops for training in the use and implementation of cyberinfrastructure for teaching plant biology. In fact, our workshop's working group on EOT was so enthusiastic that its recommendations would likely take at least 10 years to implement. We therefore look forward to working with iPlant to select and develop the most innovative and effective EOT activities using iPlant's extensive cyberinfrastructural resources. Furthermore, we propose to collaborate with the other Grand Challenge projects to develop synthetic and integrative teaching tools

and outreach products for plant biology. Finally, we will look to iPlant’s staff for guidance in the development and implementation of assessments that will indicate whether or not our activities have been successful.

Our basic, general goals for K12 education and public outreach at all levels are:

- 1) To develop cyberinfrastructure for application to K12 education and provide training to teachers to integrate the resulting tools into curricula.
- 2) Facilitate access to effective educational materials for a broad public audience (e.g., through websites, YouTube, and new cyberinfrastructure developed through this project).
- 3) Facilitate access to journals, data, and other information for students and post-docs.

We propose to meet these goals through collaboration with personnel from iPlant and from the other Grand Challenge projects, as well as other cyberinfrastructure-based EOT programs, such as PlantingScience from the Botanical Society of America. However, we propose the following specific projects as examples of EOT activities that are relevant to iPTOL.

- 1) *Tree visualization* for use in displaying phylogeny, traits, etc. Using the improved tree visualization methods developed as part of iPTOL, we propose to develop teaching/outreach tools to convey the evolutionary history of plants. We envision the development of multiple trees with varying levels of detail designed for different audiences: K12, citizen scientists, undergraduate biology students, professional scientists. For EOT, trees of varying complexity would show plant phylogeny and have applications to permit overlaying traits on clades, distributions on maps (perhaps using Google Earth?), functional/ecological attributes, etc. Such a holistic view of plant evolutionary history would provide an excellent base from which to develop teaching units on evolution.
- 2) *“DNA-to-tree” module* to permit students to explore the relationship between molecular biology (DNA variation) and biodiversity through their common link – phylogeny. Designed as a pedagogical tool rather than a research tool, this module would contain some elements of MacClade and software for sequence visualization and manipulation for alignment, tree construction, and tree manipulation, thus making use of research methods and illustrating the link between biology and computer science.
- 3) *Social networking opportunities* for students to interact with other students elsewhere or with scientists. This approach and the resulting tools could also be developed for multiple audiences: K12, undergraduates, graduate students. Faculty could use these tools to share course materials and course content and develop novel methods of instruction. This proposal resonated very positively with the teachers who attended the EOT breakout session at our workshop.
- 4) *Develop video clips* on plant evolutionary history for public outreach (for dissemination via YouTube, for example). These might be created *de novo*, or from existing programs such as Nova or Discovery. Surprisingly, existing entries under searches for “tree of life” or “phylogeny” show virtually nothing about phylogenetic relationships among organisms but instead portray the “tree of life” clips from religious or creationist contexts. Interviews with scientists could provide an entree into the realm of interactions with scientists: these same scientists could be virtual mentors through the social networking system described in (3).
- 5) *Teacher workshops* for training in the use of CI; for example, how to implement the tools proposed above and how to integrate them into curricula. We propose to recruit and support both biology and computer science teachers for summer sessions to work with iPTOL and iPlant personnel to develop curricula and to lead the workshops for other teachers. Varying state curricular standards and requirements will be kept in mind to make the teaching materials as broadly useful as possible.
- 6) *Cross-training of students* in biology and computer science at both undergraduate and graduate levels. We propose to develop courses and other approaches to integrating biology and computer science, at both our home institutions and centralized through iPlant.

II. GRAND CHALLENGE TEAM

A. Team Composition (in alphabetical order)

J. Gordon Burleigh, Department of Botany, University of Florida
 Steve Cannon, USDA/ARS, Iowa State University
 Karen Cranston, University of Arizona/Field Museum
 Michael J. Donoghue, Department of Ecology and Evolutionary Biology and Peabody Museum of Natural History, Yale University
 Katharina Huber, School of Computing Sciences, University of East Anglia, UK
 Robert Jansen, Section of Integrative Biology, University of Texas
 Jim Leebens-Mack, Department of Plant Biology, University of Georgia
 Tamara Munzner, Department of Computer Science, University of British Columbia
 William H. Piel, Peabody Museum, Yale University
 Michael J. Sanderson, Department of Ecology and Evolutionary Biology, University of Arizona
 Stephen Smith, NESCent, Durham, NC
 Douglas E. Soltis, Department of Botany, University of Florida
 Pamela S. Soltis, Florida Museum of Natural History, University of Florida;
 Alexandros Stamatakis, Department of Computer Science, Technische Universität München
 Val Tannen, Department of Computer and Information Science, University of Pennsylvania
 Todd Vision, Department of Biology, University of North Carolina
 Tandy Warnow, Department of Computer Science, University of Texas
 Cam Webb, Arnold Arboretum, Harvard University
 Michael Zanis, Department Botany and Plant Pathology, Purdue University
 Amy Zanne, Department of Biology, University of Missouri, St. Louis

B. Expertise Available and Needed

Collectively our group has considerable expertise in phylogeny reconstruction, evolutionary plant biology, algorithm theory, and database and software development. Hence, we have enormous capacity to assist and provide advice in this endeavor. We have only few participants with experience in database integration, user interface development, production quality software development, and high performance computing. In many cases, existing tools could be re-engineered to be scalable and more robust without fundamental rethinking of algorithms. In other cases, new software and technology will be necessary. Undoubtedly, we will need iPlant IT help designing hardware environments and developing software tools for the massive computations needed to meet goal 4, building a tree with a half million species.

C. Projected Personnel Resource Needs

The success of this project in large part will hinge on the direct involvement (“buy in”) of the community. Such involvement will be enhanced by the availability of tangible resources and intellectual opportunities. Several immediate funding needs include:

- Summer salary for the chairs of the five working groups
- Salary and travel support for superuser postdocs, one per working group, who would spend some time working with the IT people at UA and the rest of the time with a faculty member who is part of one of the working groups
- Travel support for graduate student exchange; graduate student stipends
- Summer salary for the project leader
- Workshops (see below)

Workshops: One of the best opportunities for engaging and leveraging the expertise in the external research community is through workshops. A number of workshops will be necessary to achieve both the individual goals and coordination among them. What follows is a noncomprehensive list of exemplars.

(1) Tree visualization. We would invite the developers of the major software programs for visualizing trees and alignments, and the biological researchers who see a need for further development.

(2) Novel algorithm development. There are diverse and substantial research questions associated with how best to scale up phylogenetic algorithms to very large datasets. To address these questions early on, we propose a workshop that is designed to engage the computer science and statistics research communities, and that will need to be particularly inclusive of the international research community. Possible organizers for international workshops include Katharine Huber (U East Anglia) and Alexandros Stamatakis (Munich). It may be desirable to pursue this in partnership with DIMACS (the NSF Center for Discrete Mathematics and Computer Science).

(3) Data assembly and integration. At least one early workshop will be needed to bring together experts in large-scale phylogenetic analysis, data providers, data integration specialists, and iPlant IT staff to determine what data are to be integrated and how that is to be achieved. Some of the phylogenetic data integration issues may begin to be addressed by an upcoming NESCent hackathon on Phylogenetic Database Interoperability in which iPlant staff are scheduled to participate. Principal leader Todd Vision is an organizer of the hackathon and Grand Challenge team member Karen Cranston is a participant.

(4) Mid-course strategic workshop. In Year 2, we propose to organize a large workshop re-uniting the Grand Challenge team with the original workshop invitees (and new participants). The purpose would be to discuss progress and solicit advice on specific sampling strategies for the large tree construction project. For this, it will especially important to invite empirical phylogeneticists in the plant Tree of Life community.

D. Projected Expertise Needed to Assess and Ensure Usability of the Developed Cyberinfrastructure

Prior to our workshop, we identified a group of “super users”, postdocs and junior faculty, typically with expertise in both phylogenetics and programming, and solicited their input on cyberinfrastructure needs. The positive contributions of this group spurred us to propose a postdoc fellowship program (see below), so that we will have a similar group continually on hand to provide high-quality feedback on both the design and usability of the developing cyberinfrastructure. However, we will also need to enlist the aid of more naive end-users for usability testing at multiple stages, from initial design through the multiple releases of the working product. We will require expertise from iPlant in helping us to design and implement a rigorous usability testing plan.

III. ORGANIZATION AND MANAGEMENT

A. Organization of iPTOL personnel

Steering Committee. We propose a Steering Committee consisting of a chair (our current iPTOL director), four other members of our team (not necessarily principal leaders) and four iPlant members. Two of the iPTOL members will be principal leaders; two will rotate in from the team as a whole on two-year rotations. The Steering Committee will meet once every other month via computer conference, once per year face-to-face. Responsibilities of the committee are to: ensure that project culture provides rewards and career development for junior investigators; encourage data sharing and appropriate credit for contributions; facilitate timely publication and internal collaboration; elicit collaboration with members of the national and international research communities; identify common goals; identify key personnel and cyberinfrastructure needs; raise and resolve intellectual property issues; identify milestones and timelines for each major project; keep projects to timelines, providing exhortation (or added resources) when needed; and refine project objectives and strategies as new capabilities come on line.

Working Groups. The five **Working Groups** (each with chair and 4-6 individuals, consisting of relevant faculty, programmers, post-docs, students) correspond to the four major goals of the project plus an algorithms group whose activities would be cross-cutting: thus one each for database integration, post-

tree analyses, visualization, large tree construction, and algorithms.

Meeting frequency for each working group should be at least once per month via computer conference, once per year face-to-face. Chairs will be compensated by summer support where requested. Responsibilities include: identify key programming and cyberinfrastructure needs, pass to steering committee; provide direct, active mentorship; encourage team cohesion in working toward goals; ensure that the working group provides rewards and career development for participants; encourage data sharing and appropriate credit for contributions; facilitate timely publication and internal collaboration; elicit collaboration with members of the national and international research communities; identify common goals; promote cross-training of group members, to enable appreciation of problems in terms of both biology and computer science; develop, test, and promulgate software and hardware solutions for focal problems; keep projects to milestones and timelines; refine group objectives and strategies as new capabilities come on line. Working group leaders will be selected by the Steering Committee.

Postdoc-superusers. We have developed a good model for junior-level superusers who are trained in both phylogenetics and programming (postdoc/junior faculty superusers to date: W. Piel; K. Cranston; A. Stamatakis; G. Burleigh; S. Shiu; S. Smith; R. Vos). We request funding for a fellowship program to support 5 superuser postdocs at any one time (one per working group). Any member of the plant science or computational biology community could apply for funding to host such a postdoc (or postdoc candidates could initiate the process). Postdoc responsibilities would include participating in one of the working groups, spending one month per year at UA working with the cyberinfrastructure team and acting as liaisons between the working groups, iPlant developers, and the postdoc's faculty host. We see this program as a key mechanism to provide incentives for participation from a wide pool of investigators in all the relevant disciplines. Postdocs will be selected via an ad hoc selection committee appointed by the Steering Committee.

Students. In our experience, graduate students can play an important role in exchanges between institutions and benefit from the interdisciplinary training that iPTOL must entail. Much of the best programming in computational phylogenetics is implemented by graduate students, but we view them also as ambassadors between disciplines, between labs, and in our highly international field, between countries. Thus, we will request travel funds to support student exchanges, including training at UA with iPlant team members. Finally, graduate stipend support is requested for two students each semester. These students would also work in association with working groups, mentored by an appropriate investigator in the community, and would be selected by an ad hoc committee appointed by the steering committee.

B. Milestones and Timeline of Project (DE=Discovery Environment)

	<i>Year 1</i>	<i>Year 2</i>	<i>Year 3</i>	<i>Year 4</i>
<i>Database integration</i>	-Technical specification for data integration and user interface -Prototype CI for tree database integration using 1-2 exemplar databases	-Integration of additional data sources -Intensive user interface development and testing		-Integration of databases with overall DE
<i>Post-tree analysis tools</i>	-Design of scalable algorithm for ancestral state reconstruction (ASR)	-Implementation of ASR algorithm in the context of developing CI -Scaling of tree reconciliation algorithms	-Implementation of tree reconciliation	-Integration of tools with the overall DE
<i>Tree Visualization</i>	-Integration of selected <i>existing</i> tree viz tools into CI	-Completion of design specification for tree viz	-Implementation of tree viz	-Integration of tools with the overall DE
<i>Large tree construction</i>	-Scaling of ML and/or NJ methods to 50,000 taxa; benchmarking on real data sets -Design taxon sampling specification	-Construction of data assembly pipelines (incl. sequence acquisition; clustering, alignment, multigene assembly; supermatrix and supertree assembly)	-HPC analyses of data for large tree construction -Exploration of scaling to 500,000 taxa	-Release of large tree(s) results as part of DE
<i>EOT and other activities</i>	-Scalable phylogenetic algorithms workshop -Database integration and assembly workshop	-Tree visualization workshop -Workshop re-uniting original workshop participants to advise on large tree sampling design -Development of teaching materials	-Teacher-training workshop -Production of video clips for EOT; additional teaching materials	-Official releases of tools and DE; talks at meetings -Evaluation and incorporation of community feedback -Teacher-training workshop

C. Progress Review and Monitoring

We recommend that iPlant appoint an external advisory board, with members drawn from outside of the iPTOL team (although perhaps invitees to the workshop would be eligible to serve as we cast a broad net for that meeting). The steering committee would be charged with submitting a report to the board once per year outlining progress on milestones. Members of the advisory board would have open invitations to all workshops and working group meetings.

References

- Alfaro, M. E., and M. T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution and Systematics* 37:19-42.
- Bansal, M., J. G. Burleigh, O. Eulenstein, and A. Wehe. 2007. Heuristics for the gene-duplication problem: An O(N) speed-up for the local search. RECOMB 2007.
- Bardel, C., V. Danjean, J.-P. Hugot, P. Darlu, and E. Génin. 2005. On the use of haplotype phylogeny to detect disease susceptibility loci. *BMC Genetics* 6:24.
- Bininda-Emonds, O. R. P. (ed) 2004. Phylogenetic Supertrees. Kluwer, Boston.
- Bininda-Emonds, O. R. P., and M. J. Sanderson. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565-579.
- Bininda-Emonds, O. R. P., M. Cardillo, K. E. Jones, R. D. E. MacPhee, et al. 2007. The delayed rise of present-day mammals. *Nature* 446: 507-512.
- Blackmore, S. 2002. Biodiversity update: progress in taxonomy. *Science* 298: 365.
- Blanchette, M., E. D. Green, W. Miller, and D. Haussler. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Research* 14:2412-2423.
- Bowers, J. E., B. A. Chapman, J. K. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433-438.
- Bremer, B., K. Bremer, M. Chase, J. Reveal, D. Soltis, P. Soltis, P. Stevens, A. Anderberg, M. Fay, P. Goldblatt, W. Judd, M. Kallersjö, J. Kårehed, K. Kron, J. Lundberg, D. Nickrent, R. Olmstead, B. Oxelman, J. Pires, J. Rodman, P. Rudall, V. Savolainen, K. Sytsma, M. van der Bank, K. Wurdack, J. Xiang, and S. Zmarzty. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141: 399-436.
- Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, and Y.-L. Qiu. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcl*. *Annals of the Missouri Botanical Garden* 80: 528-580.
- Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283-1287.
- Cracraft, J., and M. J. Donoghue (eds). 2004. *Assembling the Tree of Life*. Oxford University Press, New York.
- Cui, L., P. K. Wall, J. H. Leebens-Mack, and 13 co-authors. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738-749.
- Davies, T. J., and A. B. Pedersen. 2008. Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proceedings of the Royal Society of London B* 275: 1695-701.
- Davies, T. J., S. A. Fritz, R. Grenyer, C. D. L. Orme, J. Bielby, J. L. Gittleman, and G. M. Mace. 2008. Phylogenetic trees and the future of mammalian biodiversity. *Proceedings of the National Academy of Sciences, USA* 105: 11556-11563.
- Donoghue, M. J. 2008. A phylogenetic perspective on the distribution of plant diversity. *Proceedings of the National Academy of Sciences, USA* 105:11549 -11555.
- Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. Mobjerg Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745-749
- Durand, D., B. V. Halldorsson, and B. V. Vernot. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology* 13: 320-335.
- Evans, M. E. K., S. A. Smith, R. S. Flynn, and M. J. Donoghue. 2009. Climate, niche evolution, and diversification of the “bird-cage” evening primroses (*Oenothera*, Sections *Anogra* and *Kleinia*). *American Naturalist* 173: 225-240.

- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783-791.
- Fisher, K. M. 2008. Bayesian reconstruction of ancestral expression of the LEA gene families reveals propagule-derived desiccation tolerance in resurrection plants. *American Journal of Botany* 95: 506-515.
- Fitch, W. M. 1971. Toward defining the course of evolution - minimum change for a specific tree topology. *Systematic Zoology* 20: 406-416.
- Fleissner, R., D. Metzler, and A. von Haeseler. 2005. Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology* 54: 548-61.
- Forest, F., R. Grenyer, M. Rouget, T. J. Davies, et al. 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* 445: 757-760.
- Gherbi, H., et al. 2008. SymRK defines a common genetic basis for plant root endosymbioses with AM fungi, rhizobia and *Frankia* bacteria. *Proceedings of the National Academy of Sciences, USA* 105: 4928-4932.
- Goldman, N., and Z. Yang. 2008. Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. *Philosophical Transactions of the Royal Society B-Biological Sciences* 363:3889-3892.
- Gopalan, V., W. G. Qiu, M. Z. Chen, and A. Stoltzfus. 2006. Nexplorer: phylogeny-based exploration of sequence family data. *Bioinformatics* 22: 120-121.
- Hackett S. J., R. T. Kimball, S. Reddy, R. C. K. Bowie, E. L. Braun, et al. 2008. A phylogenomic study of birds reveals their evolutionary history. *Science* 320: 1763-1768.
- Hibbett, D., R. Nilsson, M. Snyder, M. Fonseca, J. Costanzo, and M. Shonfeld. 2005. Automated phylogenetic taxonomy: An example in the homobasidiomycetes (mushroom-forming fungi). *Systematic Biology* 54: 660-668.
- Huber, K. T., and V. Moulton. 2006. Phylogenetic networks from multi-labelled trees. *Journal of Mathematical Biology* 52: 613-632.
- Huber, K. T., B. Oxelman, M. Lott, and V. Moulton. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Molecular Biology and Evolution* 23: 1784-1791.
- Hughes, T., Y. Hyun, and D. A. Liberles. 2004. Visualising very large phylogenetic trees in three-dimensional hyperbolic space. *BMC Bioinformatics* 5: 48.
- Ilic, K., E. Kellogg, P. Jaiswal, F. Zapata, P. F. Stevens, L. P. Vincent, S. Avraham, L. Reiser, A. Pujar, et al. 2007. Plant structure ontology: Unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiology* 143: 587-599.
- Jaillon, O., J. M. Aury, B. Noel, and 53 coauthors. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463-467.
- Jansen, R. K., et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences, USA* 104: 19369-19374.
- Jordan, G. E., and W. H. Piel. 2008. PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics* 24: 1641-1642.
- Kallersjo, M., J. S. Farris, M. W. Chase, B. Bremer, M. F. Fay, C. J. Humphries, G. Petersen, O. Seberg, and K. Bremer. 1998. Simultaneous parsimony jackknife analysis of 2538 rbcL DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Systematics and Evolution* 213: 259-287.
- Kane, N. C., and Q. Cronk. 2008. Botany without borders: barcoding in focus. *Molecular Ecology* 17: 5175-5176.
- Ley, R. E., F. Backhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight, and J. I. Gordon. 2005. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences, USA* 102: 11070-11075.
- Liu, K., S. Nelesen, S. Raghavan, C.R. Linder, and T. Warnow. 2009a. Barking up the wrong treelength:

- The impact of gap penalty on alignment and tree accuracy. To appear, *IEEE Transactions on Computational Biology and Bioinformatics*.
- Liu, K., S. Nelesen, S. Raghavan, C.R. Linder, and T. Warnow. 2009b. Rapid and accurate large-scale co-estimation of sequence alignments and phylogenetic trees. Submitted.
- Mabee, P.M., M. Ashburner, Q. Cronk, G. V. Gkoutos, M. Haendel, E. Segerdell, C. Mungall, and M. Westerfield. 2007. Phenotype ontologies: the bridge between genomics and evolution. *Trends in Ecology and Evolution* 22: 345-50.
- Maddison, W. P., and D.R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5 <http://mesquiteproject.org>
- McMahon, M. M., and M. J. Sanderson. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Systematic Biology* 55: 818-836.
- Ming, R., S. Hou, Y. Feng, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452: 991-996.
- Moles, A., D. Ackerly, C. Webb, J. Tweddle, J. Dickie, and M. Westoby. 2005. A brief history of seed size. *Science* 307: 576-580.
- Moore, M. J., C. D. Bell, P. S. Soltis, and D. E. Soltis. 2007. Using plastid genome scale-data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* 104: 19363-19368.
- Mossel, E. 2007. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4:108-116.
- Munzner, T., F. Guimbretiere, S. Tasiran, L. Zhang, and Y. H. Zhou. 2003. TreeJuxtaposer: Scalable tree comparison using Focus+Context with guaranteed visibility. *ACM Transactions on Graphics* 22: 453-462.
- Nakhleh, L., T. Warnow, C. R. Linder, and K. St John. 2005. Reconstructing reticulate evolution in species - Theory and practice. *Journal of Computational Biology* 12: 796-811.
- Ott, M., J. Zola, S. Aluru, and A. Stamatakis. 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. Proceedings of IEEE/ACM Supercomputing (SC2007) conference, Reno, Nevada, November 2007.
- Page, R. D. M. (ed) 2002. Tangled Trees. University of Chicago Press, Chicago.
- Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* 48: 612-622.
- Paterson, A. H., J. E. Bowers, and B. A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences, USA* 101: 9903-9908.
- Pennisi, E. 2003. Modernizing the tree of Life. *Science* 300: 1692.
- Pratas, F., P. Trancoso, A. Stamatakis, L. Sousa. Fine-grain parallelism for the phylogenetic likelihood functions on multi-cores, Cell/BE, and GPUs. submitted.
- Proches, S., J. R. U. Wilson, D. M. Richardson, and M. Rejmanek. 2008. Searching for phylogenetic pattern in biological invasions. *Global Ecology and Biogeography* 117: 5-10.
- Pugnaire, F. I., and F. Valladares (eds). 2007. Functional Plant Ecology, 2nd edition. CRC Press, New York.
- Purvis, A., and J. L. Gittleman (eds). 2005. Phylogeny and Conservation. Cambridge Univ. Press., London.
- Redelings, B., and M. Suchard. 2005. Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology* 3: 401-418.
- Roch, S. 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 3: 92-94.
- Sage, R. F. 2003. The evolution of C4 photosynthesis. *New Phytologist* 161: 341-370
- Sanderson, M. J. 2006. Paloverde: an OpenGL 3D phylogeny browser. *Bioinformatics* 22: 1004-1006.
- Sanderson, M. J. 2007. Construction and annotation of large phylogenetic trees. *Australian Systematic*

- Botany* 20:287-301.
- Sanderson, M. J. 2008. Phylogenetic signal in the eukaryotic tree of life. *Science* 321: 121-123.
- Sanderson, M. J., and M. F. Wojciechowski. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Systematic Biology* 49:671-685.
- Sanderson, M. J., C. Ane, O. Eulenstein, D. Fernandez-Baca, J. Kim, M. M. McMahon, and R. Piaggio-Talice. 2007. Fragmentation of large data sets in phylogenetic analysis *in* Reconstructing Evolution: New Mathematical and Computational Advances (O. Gascuel, and M. Steel, eds.). Oxford University Press, Oxford.
- Sanderson, M. J., D. Boss, D. Chen, K. A. Cranston, and A. Wehe. 2008. The PhyLoTA Browser: Processing GenBank for Molecular Phylogenetics Research. *Systematic Biology*: In press.
- Sankoff, D., C. Zheng, and Q. Zhu. 2007. Polyploids, genome halving and phylogeny. *Bioinformatics* 23: I433--I439.
- Sennblad, B. et al. 2007. Primetv: a viewer for reconciled trees. *BMC Bioinformatics* 8: 148.
- Silvertown, J. W., M. Franco, and J. L. Harper (eds). 1997. Plant Life Histories: Ecology, Phylogeny and Evolution. Cambridge Univ. Press, London.
- Smith, S. A., and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322: 86-89.
- Soltis, D. E., P. S. Soltis, M. W. Chase, and P. Endress. 2005. Phylogeny and Evolution of Angiosperms. Sinauer Associates, Sunderland, MA.
- Soltis, D. E., C. D. Bell, S. Kim, and P. S. Soltis. 2008. Origin and early evolution of angiosperms. *Annals of the New York Academy of Sciences* 1133: 3-25.
- Soltis, D. E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. Paterson, C. Zheng, D. Sankoff, P. Kerr Wall, and P. S. Soltis. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336-348.
- Stamatakis, A., and M. Ott. 2008a. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. *Philosophical Transactions of the Royal Society B*, 363: 3977-3984.
- Stamatakis, A., and Ott, M. 2008b. Exploiting fine-grained parallelism in the phylogenetic likelihood function with MPI, Pthreads, and OpenMP: A performance study. Pattern Recognition in Bioinformatics (PRIB 2008). Volume 5265 of Springer Lecture Notes in Computer Science, pp. 424-435.
- Steel, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9: 91-116.
- Tuskan, G. A., S. Difazio, S. Jansson, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596-1604.
- Vandepoele, K., and Y. Van de Peer. 2005. Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiology* 137: 31-42.
- Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290: 2114-2117.
- Wang, L. S., T. Warnow, B. M. E. Moret, R. K. Jansen, and L. A. Raubeson. 2006. Distance-based genome rearrangement phylogeny. *Journal of Molecular Evolution* 63:473-483.
- Webb, C. O., and M. J. Donoghue. 2005. Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes* 5: 181-183.
- Wheeler, Q. D. 2004. Taxonomic triage and the poverty of phylogeny. *Philosophical Transactions of the Royal Society of London B*. 359: 571.
- Willis, C. G., B. Ruhfel, R. B. Primack, A. J. Miller-Rushing, and C. C. Davis. 2009. Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change. *Proceedings of the National Academy of Sciences, USA* 105: 17029-17033.
- Wong, K. M., M. A. Suchard, and J. P. Huelsenbeck. 2008. Alignment uncertainty and genomic analysis. *Science* 319: 473-476.

- Yesson, C., and A. Culham. 2006. A phyloclimatic study of *Cyclamen*. *BMC Evolutionary Biology* 6.
- Yokoyama, S., T. Tada, H. Zhang, and L. Britt. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences, USA* 105: 13480-13485.

BIOGRAPHICAL SKETCH

Michael J. Sanderson

Professor, Department of Ecology and Evolutionary Biology,
University of Arizona, Tucson, AZ 85721;
E-mail: sanderm@email.arizona.edu

Professional Preparation

B. S., 1982, (cum laude, Physics) University of Arizona.
Ph. D., 1989, (Ecology and Evolutionary Biology) University of Arizona.

Appointments

2006 – Professor, Dept. of Ecology and Evolutionary Biology, University of Arizona
2001 – 2006 Professor, Section of Evolution and Ecology, University of California, Davis
1997 – 2001 Associate Professor, Section of Evolution and Ecology, University of California, Davis.
1995 – 1997 Assistant Professor, Section of Evolution and Ecology, University of California, Davis.
1992 – 1995 Assistant Professor, Department of Biology, University of Nevada, Reno.
1992 - 1992 (one semester) Adjunct Assistant Professor, Department of Ecology and Evolutionary Biology, University of Arizona.
1989 - 1991 Alfred P. Sloan Foundation Postdoctoral Fellow, L. H. Bailey Hortorium, Cornell University.

Five Publications Relevant to Proposed Research

Sanderson, M. J. 2008. Phylogenetic signal in the eukaryotic tree of life. *Science*, 321:121-123.
Sanderson, M. J., D. Boss, D. Chen, K. A. Cranston, and A. Wehe. 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. *Syst. Biol.* 57:335-346.
Sanderson, M. J., and M. M. McMahon. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol. Biol.*, 7 (suppl. 1): S3.
McMahon, M. M., and M. J. Sanderson. 2006. Phylogenetic Supermatrix Analysis of GenBank Sequences from 2228 Papilionoid Legumes. *Syst. Biol.*, 55:818-836.
Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, B. O'Meara, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. *Science*, 306:1172-1174.

Five Other Publications

Sanderson, M. J. Construction and annotation of large phylogenetic trees. 2007. *Australian Syst. Bot.* 4:287-301.
Sanderson, M. J. 2006. Paloverde: an OpenGL 3-D phylogeny browser. *Bioinformatics*, 22:1004-1006.
Burleigh, J. G., A. C. Driskell, and M. J. Sanderson. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Syst. Biol.* 55:426-440.
Scherson, R. A., R. Vidal, and M. J. Sanderson. 2008. Phylogeny, biogeography and rates of diversification of New World *Astragalus* (Leguminosae) with an emphasis on the South American radiations. *Amer. J. Bot.* 95:1030-1039.
Scotland, R., and M. J. Sanderson. 2004. The significance of few versus many in the tree of life. *Science*, 303:643.

Synergistic Activities

PhyLoTA Browser database: <http://loco.biosci.arizona.edu/pb>.

University of Arizona Biodiversity Informatics Initiative: <http://loco.biosci.arizona.edu/bdii>.
'r8s', vers. 1.71. Software for the analysis of molecular rates of evolution and the reconstruction of divergence times: <http://loco.biosci.arizona.edu/r8s>

TreeBASE, an electronic database of phylogenetic knowledge: <http://www.treebase.org>.

“Workshop in Applied Phylogenetics”. Bodega Marine Lab, Bodega Bay California, Taught every year since 2000, and in 2004 with an appended workshop on supertree methodology.

Organized an NSF-sponsored workshop: “Workshop on Phyloinformatics; October 20-21, 2000.

Collaborators and other Affiliations

Graduate advisor; Michael J. Donoghue, Yale University;

Postdoctoral advisor: Jeff J. Doyle, Cornell University

Graduate advisees and current addresses (7):

Michael Plotkin, UC Davis

Jer-Ming Hu, Assistant Professor, National Taiwan University

Shelah Morita, Postdoc, NC State University

Rosita Scherson, Postdoc, Universidad de Chile

Brian O'Meara, Postdoc, Duke U (NESCent)

Travis Wheeler, U of Arizona

Ed Gilbert, U of Arizona

Postdoctoral Advisees and current addresses (12):

Chris Henze, NASA Advanced Supercomputer Division, Ames Research Lab

Olaf Bininda-Emonds, Technical University of Munich

Susana Magallon, Ass't Prof., UNAM, Mexico City

Charles Nunn, Max Planck Institute, Leipzig

Amy Driskell, Smithsonian Institution

Mike Alfaro, Ass't Prof., Washington State

Rick Ree, Ass't Curator., Field Museum of Natural History

Gordon Burleigh, NESCENT Fellow, Duke

Campbell Webb, Arnold Arboretum, Harvard

Cecile Ane, Ass't Prof., University of Wisconsin

Justen Whittall, Ass't Prof., Santa Clara University

Karen Cranston, U of Arizona

David Hearn, U of Arizona

Brad Boyle, U of Arizona

Collaborators not listed above or in publications (no co-editors in last 24 months):

Boss, Darren, U of Arizona, Chen, D., Iowa State U, Eulenstein, O., Iowa State U,

Fernandez-Baca, D., Iowa State U, Goloboff, P., INSUE, CONICET, Instituto Miguel Lillo,

Argentina, Kim, J., U Pennsylvania, Wehe, A., Iowa State U, Beaman, R., Yale U, grant co-

PI, Cellinese, N., Yale U., grant co-PI, Davis, C., Harvard U, grant co-PI, Hilu, K., Virginia

Tech, grant co-PI, Judd, W., U Florida, grant co-PI, Manchester, S. U Florida, grant co-PI,

Olmstead, R. U Washington, grant co-PI, Qiu, Y.-L., U Michigan, grant co-PI, Sierwald, P.,

Field Museum, grant co-PI, Soltis, D.S., U Florida, grant co-PI, Soltis, P.S., U Florida, grant

co-PI, Sytsma, K., U Wisconsin, grant co-PI, Wurdack, K., Smithsonian Institution, grant co-

PI

MICHAEL J. DONOGHUE

Department of Ecology and Evolutionary Biology
& Peabody Museum of Natural History
Yale University, New Haven, CT 06520-8105
Phone: 203-432-2074; michael.donoghue@yale.edu

PROFESSIONAL PREPARATION:

Michigan State University, East Lansing, MI	Botany	B.S. (honors), 1976
Harvard University, Cambridge, MA	Biology	Ph.D., 1982

APPOINTMENTS:

2000-present G. E. Hutchinson Professor of Ecology and Evolutionary Biology, with appointments in the Department of Geology and Geophysics, in the School of Forestry and Environmental Studies, and as Curator of Botany in the Peabody Museum. Chair, Department of Ecology and Evolutionary Biology (2001-02); Director, Peabody Museum of Natural History (2003-2008); Vice President for West Campus Planning and Program Development (2008-).
1993-2000 Professor of Organismic and Evolutionary Biology, Harvard University, and Director of the Harvard University Herbaria (1995-99).
1998-1999 Visiting Professor, Department of Biology, Stanford University.
1985-1993 Assistant Professor (1985-88), Associate Professor (1988-90), and Professor (1990-92) of Ecology and Evolutionary Biology, University of Arizona.
1982-1985 Assistant Professor of Biology, San Diego State University.

FIVE RELEVANT PUBLICATIONS (of 191):

Donoghue, M. J. 2004. Immeasurable Progress on the Tree of Life. Pp. 548-552 in Cracraft, J. and M. J. Donoghue (eds.), *Assembling the Tree of Life*. Oxford University Press, New York.
Donoghue, M. J. 2005. Key innovations, convergence, and success: macroevolutionary lessons from plant phylogeny. *Paleobiology* 31(2): 77-93.
Moore, B. R., S. A. Smith, and M. J. Donoghue. 2006. Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Syst. Biol.* 55: 662-676.
Howarth, D. G. and M. J. Donoghue. 2006. Phylogenetic analyses of the "ECE" (CYC/TB1) clade reveal duplications that predate the core eudicots. *Proc. Nat. Acad. Sci. USA* 103: 9101-9106.
Smith, S. A. and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322: 86-89.

FIVE ADDITIONAL RELEVANT PUBLICATIONS:

Piel, W. H., M. J. Sanderson, and M. J. Donoghue. 2003. The small-world dynamics of tree networks and data mining in phyloinformatics. *Bioinformatics* 19: 1162-1168.
Webb, C. O. and M. J. Donoghue. 2004. Phylomatic: tree retrieval for applied phylogenetics. *Molecular Ecology Notes*.
Kim, S-T., S. E. Sultan, and M. J. Donoghue. 2008. Allopolyploid speciation in *Persicaria* (Polygonaceae): Insights from a low-copy nuclear marker. *Proc. Nat. Acad. Sci. USA* 105: 12370-12375.
Cantino, P. D., J. A. Doyle, S. W. Graham, W. S. Judd, R. G. Olmstead, P. S. Soltis, D. E. Soltis, and M. J. Donoghue. 2007. Towards a phylogenetic nomenclature of *Tracheophyta*. *Taxon* 56: 822-846.
Edwards, E. J., C. J. Still, and M. J. Donoghue. 2007. The relevance of phylogeny to studies of global change. *Trends Ecol. Evol.* 22: 243-249.

RECENT SYNERGISTIC ACTIVITIES:

Selected Service, 1995-03: President, Society of Systematic Biologists (1994-95); Steering Committee, "Evolution, Science, and Society" (1995-97); DIVERSITAS Steering Committee (2001-), Vice Chair (2002-), Chair, bioGENESIS (2006); NAS Committee, Int. Union Biol. Sci. (1999-04); NCEAS Working Groups: Biogeography (2000-02), Adaptive Radiations (2001-02); Phylogeny and Ecology (2002); Executive Committee, Discovering Life in America (2001-); Visiting Committees: Princeton Univ. (1995, 2002), SUNY Stony Brook (1997), UC Berkeley (1998, 2001), Cornell (1999), Arnold Arboretum (2001, 2003), Brown Univ. (2002); Board of Directors, Natural Science Collections Alliance (2004-).

Selected Honors, 1997-02: Fellow, AAAS (1997); Glaser Distinguished Visiting Prof., Florida Int. Univ. (1998); Eminent Biologist Lecturer, Carnegie Museum (1999), Perspectives in Biology Lecturer, Wake Forest Univ. (1999); Distinguished Alumni Award, Michigan State Univ. (2005); Elected Member, US National Academy of Sciences (2005); Elected Member, American Academy of Arts and Sciences (2008).

Selected Invited Presentation, 2000-05: "Frontiers in phylogenetic biology," Bot. Soc. America (2000); "Transference of function," FASEB, Vermont (2000, with D. Baum); "Phylogenies and global change," Amsterdam (2001); "Historical biogeography," European Soc. Evol. Biol., Denmark (2001); "Plant evolution," Cold Spring Harbor (2001); "Dipsacales flower evolution," Zurich (2002); "Evolution of Biomes," Royal Society, (2004); "Homology," Botanical Society of America (2004); "Biodiversity," Paris (2005).

Additional Activities: *Databases* -- TreeBASE: a database of phylogenetic knowledge (with W. Piel, M. Sanderson,); <http://www.treebase.org>. *Selected Symposia/Meetings Organized* -- "Northern Hemisphere Biogeography" (with J. Wen), International Botanical Congress, Vienna, Austria (2005); "Phytogeography of the Northern Hemisphere" (with P. Manos), NESCent Working Group (2006-08); "Phylogenies and Biodiversity Science," DIVERSITAS Open Science Conference, Oaxaca, Mexico (2005)

COLLABORATORS AND OTHER AFFILIATIONS:

Co-authors and co-editors (past 5 years, excluding students and postdocs – see below): G. Burleigh (NESCent), N. Arens (Smith Coll.), C. Campbell (U. Maine), P. Cantino (Ohio U.), L. Clark (Iowa State), J. Cracraft (Am. Mus. Nat. Hist.), C. David (Harvard U.), C. Delwiche (U. Maryland), J. Doyle (UC Davis), T. Eriksson (Stockholm) T. Field (U. Tennessee), J. Gauthier (Yale), G. Gilbert (UC Santa Cruz), S. Graham (U. British Columbia), N. Havill (Yale U.), B. Jacobs (Leiden), C. Jaramillo (STRI), W. Judd (U. Florida), E. Kellogg (U. Missouri), B. LePage (U. Pennsylvania), M. Loreau (McGill U.), J. Lundberg (Stockholm), P. Manos (Duke U.), D. Miranker (U. Texas), S. Mathews (Arnold Arb.), L. Nakhleh (U. Texas), R. Olmstead (U. Washington), E. Smets (Leiden), M. Smith (Paris), D. Soltis (U. Florida), P. Soltis (U. Florida), P. Stevens (U. Missouri), N. Theis (U. Massachusetts), J. Wiens (Stony Brook), K. Wurdack (Smithsonian Inst.), T. Yahara (Kyushu U.).

Graduate Students (21 total): J. Dice (CalTrans), M. Sanderson (U. Arizona), J. Malusa (Discov. Mag.), L. Abbott, J. M. Porter (Rancho Santa Ana Gard.), G. Bharathan (SUNY Stony Brook), D. Ferguson (Louisiana State U.), C. Maley (Wistar Inst.), G. Weiblen (U. Minnesota), R. Spangler (U. Minnesota), R. Ree (Field Mus.), C. Davis (Harvard U.), C. Bell (U. New Orleans), E. Edwards (Brown), S-T. Kim (Tubingen, Germany), B. Moore (UC Berkeley), R. Novick (Purdue), S. Smith (NESCent). Current: S. Carlson, J. Beaulieu, A. Greenberg.

Recent Postdoctoral Associates (24 total): C. Webb (Arnold Arboretum), R. Winkworth (Fiji), D. Howarth (St. Johns U.), M. Evans (U. Arizona), D. Tank (U. Idaho). Current: E. Lo, W. Clement.

Ph.D. advisor: Carroll E. Wood, Jr., Harvard University

BIOGRAPHICAL SKETCH: PAMELA S. SOLTIS

Address: Florida Museum of Natural History and the Genetics Institute, University of Florida,
Gainesville, FL 32611. e-mail: psoltis@flmnh.ufl.edu

Birthdate: November 13, 1957 Nelsonville, OH

Education: B.A. (Biology), Central College, Pella, IA, 1980, Summa Cum Laude
M. Phil. (Botany), University of Kansas, 1984, with Honors
Ph.D. (Botany), University of Kansas, 1986

Appointments:

Curator, Florida Museum of Natural History, and Professor, Genetics Institute, University of Florida,
October, 2000 – present
Distinguished Professor, University of Florida, 2007 – present
University of Florida Research Foundation Research Professor, 2006-09
Fulbright Distinguished Scholar, Royal Botanic Gardens, Kew, England, and Imperial College,
Silwood Park, England, 2000 – 2001
Assistant, Associate, and Full Professor, Department of Botany, Washington State University, 1986 –
2000
Mellon Senior Fellow, Smithsonian Institution, 1994-95

Five publications most relevant to this proposal:

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes: A research tool for comparative biology. *Nature* 402: 402-404.
Cui, L., P. K. Wall, J. Leebens-Mack, B. G. Lindsay, D. Soltis, J. J. Doyle, P. Soltis, J. Carlson, A. Arumuganathan, A. Barakat, V. Albert, H. Ma, and C. W. dePamphilis. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738-749.
Soltis, D. E., P. S. Soltis, P. K. Endress, and M. W. Chase. 2005. *Phylogeny and Evolution of Angiosperms*. Sinauer Associates, Sunderland, MA.
Bell, C. D., D. E. Soltis, and P. S. Soltis. 2005. The age of the angiosperms: A molecular time-scale without a clock. *Evolution* 59: 1245-1258.
Soltis, P. S. and D. E. Soltis. 2004. The origin and diversification of angiosperms. *American Journal of Botany* 91: 1614-1626.

Five additional relevant publications:

Chanderbali, A. S., V. A. Albert, J. Leebens-Mack, N. S. Altman, D. E. Soltis, and P. S. Soltis. 2009. Transcriptional signatures of ancient floral developmental genetics in avocado (*Persea americana*; Lauraceae). *Proceedings of the National Academy of Sciences, USA*: in press.
Soltis, P. S., S. F. Brockington, M.-J. Yoo, A. Piedrahita, M. Latvis, M. J. Moore, A. S. Chanderbali, and D. E. Soltis. 2009. Floral variation and floral genetics in basal angiosperms. *American Journal of Botany* 96: 110-128.
Soltis, D. E., V. A. Albert, J. Leebens-Mack, J. D. Palmer, R. A. Wing, C. W. dePamphilis, H. Ma, J. E. Carlson, N. Altman, S. Kim, P. K. Wall, A. Zuccolo, and P. S. Soltis. 2008. The *Amborella* genome: An evolutionary reference for plant biology. *Genome Biology* 99: 402 (doi:10.1186/gb-2008-9-3-402).
Soltis, P. S., D. E. Soltis, S. Kim, A. Chanderbali, and M. Buzgo. 2006. Expression of floral regulators in basal angiosperms and the origin and evolution of the ABC model. In D. E. Soltis,

J. Leebens-Mack, and P. S. Soltis (eds.), *Developmental Genetics of the Flower. Advances in Botanical Research* 44:483-506.

Davies, T. J., T. G. Barraclough, M. W. Chase, P. S. Soltis, D. E. Soltis, and V. Savolainen. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences, USA* 101: 1904-1909.

Synergistic Activities (recent representative activities):

Society Service: President, Society of Systematic Biologists, 2004-06
President, Botanical Society of America, 2006-09

Editorial Service: Associate Editor, *Systematic Biology*, 2001-07
Associate Editor, *Evolution*, 2003-07

NSF Workshops: iPlant Grand Challenge Workshop, Biosphere2, AZ (co-organizer)
Species Diversity on Earth, Washington, DC, 2005 (co-organizer)
Databases in Plant Systematics, Gainesville, FL, 2003 (co-organizer)

Research Collaborations: International Polyploidy Conference, London, 2003 (co-organizer)
Deep Time Research Coordination Network, 2001-present
Floral Genome Project, 2001-present
Ancestral Angiosperm Genome Project, 2006-present

Advisory Board Chair: NSF Plant Genome Polyploidy Project (L. Comai, PI)

Conflicts of Interest:

(i) Collaborators & Co-authors: V. Albert (U. Oslo), N. Altman (Penn State U.), M. Bennett (Royal Bot. Gard., Kew), T. Borsch (U. Bonn), P. Cantino (Ohio U.), J. Carlson (Penn State U.), M. Chase (Royal Bot. Gard., Kew), J. Chen (U. Texas), L. Cui (Penn State U.), C. Davis (Harvard), C. dePamphilis (Penn State U.), M. Donoghue (Yale), J. Doyle (Cornell), P. Endress (U. Zurich), S. Farris (Stockholm U.), M. Frohlich (Natural History Museum, London), P. Herendeen (George Washington U.), K. Hilu (Virginia Tech U.), R. Huck (U. Florida), L. Hufford (Washington State U.), H. Kong (Chinese Acad. Sci.), A. Kovarik (Czech Acad. Sci.), J. Leebens-Mack (U. Georgia), A. Leitch (U. London), I. Leitch (Royal Bot. Gard., Kew), H. Ma (Penn State U.), P. Manos (Duke), R. Matyasek (Czech Acad. Sci.), M. Moody (U. Connecticut), L. Mueller (Cornell), L. Oliveira (U. Vicosa), R. Olmstead (U. Washington), Y.-L. Qiu (U. Michigan), L. Ronse DeCraene (Royal Bot. Gard., Edinburgh), M. Sanderson (U. Arizona), V. Savolainen (Imperial College, Silwood), S. Schlarbaum (U. of Tennessee), K. Sytsma (U. Wisconsin), S. Tanksley (Cornell), G. Theissen (U. Jena), L. Zahn (Penn State U.)

(ii) Graduate Advisor: W. L. Bloom

(iii) Thesis Advisor/Postgrad Sponsor: Students: P. G. Wolf (PhD 1990; Utah State U.); R. E. B. Kirkpatrick (MS 1988; UC-Berkeley); T. S. Richter (MS 1990); M. S. Mayer (PhD 1993; U. San Diego); J. L. Schultz (PhD 1996; Lewis-Clark State College); T. M. Hardig (PhD 1998; U. Montevallo); L. M. Cook (PhD 1998; Washington State U.); D. Albach (MS 1998; U. Mainz); J. A. Koontz (PhD 2000; Augustana College); M. A. Gitzendanner (PhD 2000; U. Florida); P. Speranza (PhD 2005; U. de la República); A. Morris (PhD 2006; U. S. Alabama); H. Loring (MS 2006); C. Edwards (PhD 2007; U. Wyoming); R. Vergara (2002 -); S. Brockington (2002-); S. Servick (2006-); M. Latvis (2007 -); C. Segovia (2007 -); N. Miles (2007 -); M. Heaney (2008 -); **Post-docs:** T. Ranker (U. Colorado); S. Novak (Boise State U.); J. C. Pires (U. Missouri); S. Kim (Korean Inst. Biol. Res.); M. Buzgo (Louisiana State U.-Shreveport); H. Wang (U. Florida); J. Tate (Massey U.); C. Bell (U. New Orleans); A. Powell (U. Evansville); V. Symonds (Massey U.); M. Moore (Oberlin); S. Jian (Chinese Acad. Sci.); H. Wang (Chinese Acad. Sci.); A. Doust (Oklahoma State U.); E. Mavrodiev; A. Chanderbali; R. Buggs; L. Zhang

Biographical Sketch: Douglas E. Soltis

Address: Department of Botany & the Genetics Institute, University of Florida, Gainesville, FL 32611; email: dsoltis@botany.ufl.edu

Birthdate: 28 October, 1953. Sewickley, Pennsylvania

Education:

Groveton High School, Alexandria, Virginia, 1971
B.S. (Biology) College of William and Mary, 1975
M.A. (Biology) Indiana University, 1977
Ph.D. (Biology) Indiana University, 1980
Postdoctoral experience, University of British Columbia, summer 1981

Appointments:

Distinguished Professor, University of Florida, 2008-present
Chair, Department of Botany, University of Florida, 2006-present
Professor, University of Florida, 2000-present
Professor, Washington State University, 1990-2000
Acting Director of the Ownbey Herbarium, 1990-91
Associate Professor, Washington State University, 1986-1990
Assistant Professor, Washington State University, 1983-1986
Assistant Professor, The University of North Carolina at Greensboro, 1980-1983
Postdoctoral experience, University of British Columbia, summer 1981
William R. Ogg Fellowship, Indiana University 1979-1980
Associate Instructor, Indiana University, 1975-1979

Five Publications Relevant to this Proposal:

Soltis, D. E., V. A. Albert, J. Leebens-Mack, J. D. Palmer, R. A. Wing, C. W. dePamphilis, H. Ma, J. E. Carlson, N. Altman, S. Kim, P. K. Wall, A. Zuccolo, and P. S. Soltis. 2008. The *Amborella* genome: An evolutionary reference for plant biology. *Genome Biology* 99: 402 (doi:10.1186/gb-2008-9-3-402).

Soltis, D. E., H. Ma, M. Frohlich, P. Soltis, V. Albert, D. Oppenheimer, N. Altman, C. dePamphilis, and J. Leebens-Mack. 2007. The floral genome: an evolutionary history of polyploidy and shifting patterns of gene expression. *Trends in Plant Science* 12: 358-367.

Moore, M. M., C. D. Bell, P. S. Soltis, and D. E. Soltis. 2007. Using plastid genomic-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci., USA* 104: 19363–19368.

Soltis, D. E., M. Gitzendanner, and P. S. Soltis. 2007. A Bayesian analysis of the three-gene, 567-taxon data set for angiosperms. *International Journal of Plant Sciences* 168: 137-157.

Soltis, D. E. and Soltis, P. S. 2004. *Amborella* not a basal angiosperm? Not so fast. *American Journal of Botany* 91: 997-1001.

Five Additional Publications:

Soltis, D.E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson, et al. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* 96: 336-348.

Wang, H-C., M. M. Moore, P. S. Soltis, C. D. Bell, S. R. Manchester, and D. E. Soltis. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences, USA*, in press.

Jian, S., P. S. Soltis, M. Gitzendanner, M. Moore, R. Li, T. Hendry, Y. Qiu, A. Dhingra, C. Bell, and D. E. Soltis. 2007. Resolving an ancient, rapid radiation in Saxifragales. *Systematic Biology* 57: 38-57.

Kim, S., J. Koh, M-J Yoo, H. Kong, Y. Hu, H. Ma, P. S. Soltis, and D. E. Soltis. 2005. Expression of floral MADS-box genes in basal angiosperms: Implications for the evolution of floral regulators. *Plant Journal* 43:724-744.

Soltis, D. E., P. S. Soltis, M. W. Chase, and P. Endress. 2005. *Phylogeny and Evolution of Angiosperms*. Sinauer Associates, Sunderland, MA.

Synergistic Activities (recent and representative):

Associate Editor, *American Journal of Botany*, 2006-present

Co-Organizer of PAG polyploidy symposia, 2005-2009

Co-Organizer of multiple Deep Time symposia and workshops, 2002-2007

Co-Organizer, NSF-DFG Biodiversity Conference, November, 2005

Head, Advisory Committee, Polyploidy—Plant Genome Grant, 2001-2005

Conflicts of Interest:

(i) Collaborators (last 48 months):

D. Albach, U. Mainz; V. Albert, U. Oslo; G. A. Allen, U. Victoria; T. J. Barraclough, Imperial College, Silwood; M. Bennett, Royal Botanic Gardens, Kew; B. Bremer, Royal Swedish Acad. Sci.; K. Bremer, Stockholm U.; M. W. Chase, Royal Botanic Gardens, Kew; P. Crane, U. Chicago; C. dePamphilis, Penn State Univ.; M. J. Donoghue, Yale U.; C-Z. Fan, North Carolina State U.; K. Hilu, Virginia Polytechnic Inst.; L. Hufford, Washington State U.; J. Leebens-Mack, U. Georgia; A. Leitch, U. London; I. Leitch, Royal Botanic Gardens, Kew; H. Ma, Penn State U.; P. Manos, Duke U.; M. Moody, U. Connecticut; D. Nickrent, Southern Illinois U.; R. G. Olmstead, U. Washington; Y-L. Qiu, U. Michigan; J. L. Reveal, U. Maryland; V. Savolainen, Imperial College, Silwood; P. F. Stevens, U. Missouri-St. Louis; E. A. Zimmer, Smithsonian Institution

(ii) Advisors: M.A./Ph.D. Advisor: G. J. Gastony; Postdoctoral Advisor: B. A. Bohm

(iii) Major Thesis Advisor: F. A. Bryan (1983); L. H. Rieseberg (1987; U. Brit. Col.); B. Ness (1989, Anguin College); R. D. Noyes (1989; Univ. Colorado); S. J. Brunfeldt (1990, U. Idaho, deceased); G. M. Plunkett (1994; Virginia Commonwealth); Q-Y. Xiang (1995, North Carolina State U.); L. A. Johnson (1997; Brigham Young U.), D. Strenge (1998, Battelle Research), A. Rabe (1999, Nature Conservancy), R. K. Kuzoff (1998; U. Wisc.-Stevens Pt.), M. E. Mort (1999; U. Kansas), M. Zanis (2002; Purdue U.); Stacie A. Kageyama (2001; U. Oregon); C. Notis (2004; U. Michigan—Flint); C. Edwards (2001-07, U. Wyoming); M. Arakaki (2002-); M-J. Yoo (2002-); S. Brockington (2002-); J. Clayton (2003-); J. Koh (2005-); S. Servick (2006-); N. Miles (2007-); L. Majure (2007-); M. Latvis (2007-); C. Segovia (2007-); M. Heaney (2008-).

(iv) Postdoctoral-Scholar Sponsor: T. Ranker (1987-88; U. Colorado); E. Conti (1994-95; U. of Zurich); M. Fishbein (1998-2000; Portland State U.); J. C. Pires (2000-2001; U. Missouri); A. Scheen (2001-2002; U. Oslo); E. Melendez-Ackerman (2002-2003; U. Puerto Rico); S. Kim (2001-07; Korean Inst. Biol. Res.); M. Buzgo (2002-07; Louisiana State U.-Shreveport); H. Wang (2002-03; U. Florida); J. Tate (2002-2006; Massey U.); C. Bell (2003-04; New Orleans U.); V. Symonds (2005-06; Massey U.); A. Powell (2005-06; U. Evansville); M. Moore (2005-07; Oberlin); S. Jian (2005-06; Chinese Acad. Sci.); H. Wang (2006-07; Chinese Acad. Sci.); A. Doust (2006-07; Oklahoma State U.); E. Mavrodiev (2002-); A. Chanderbali (2003-); R. Buggs (2007-); L. Zhang (2008-).

Val Tannen

Computer and Information Science Department
University of Pennsylvania
Levine Hall, 3330 Walnut St., Philadelphia, PA 19104
val@cis.upenn.edu, (215)898-2665, fax -0587

PROFESSIONAL PREPARATION

Polytechnic Institute of Bucharest, Computer Engineering, BSE/MSE 1977.

Massachusetts Institute of Technology, Applied Mathematics/Computer Science, PhD, 1987.

MIT Laboratory for Computer Science, Postdoc in Computer Science, 1987.

APPOINTMENTS

1999–present, Professor, Department of Computer and Information Science, University of Pennsylvania.

2002–2003, Visiting Professor, MPLA, University of Athens, and Visiting Researcher, FORTH, Crete, Greece (sabbatical leave)

1993–1999, Associate Professor, Department of CIS, UPenn.

1994–1995, Visiting Professor, Université de Paris XI, Orsay, and Visiting Researcher, INRIA, Rocquencourt, France (sabbatical leave).

1987–1993, Assistant Professor, Department of CIS, UPenn.

1985, 1986, Researcher, IBM Thomas J. Watson Research Center.

SELECTED RECENT PUBLICATIONS

“Annotated XML: Queries and Provenance” J.N. Foster, T.J. Green, and V. Tannen. Proceedings PODS 2008.

“Update Exchange with Mappings and Provenance” T.J. Green, G. Karvounarakis, Z. Ives, and V. Tannen. Proceedings VLDB 2007: 675-686

“ORCHESTRA: Facilitating Collaborative Data Sharing” T.J. Green, G. Karvounarakis, N. Taylor, O. Biton, Z. Ives, and V. Tannen. Demo, Proceedings SIGMOD 2007: 1131-1133

“Provenance semirings” T.J. Green, G. Karvounarakis, and V. Tannen. Proceedings PODS 2007, pp.31-40.

“Data Integration in the Life Sciences” S. Cohen Boulakia and V. Tannen (eds.) Proceedings 4th International Workshop, DILS 2007, Philadelphia, Springer LNCS 4544 (LNBI), 2007

“Models for Incomplete and Probabilistic Information” T.J. Green and V. Tannen, Proceedings EDBT 2006, IIDB Workshop, also in Data Engineering Bull., Vol. 29, Mar.2006, pp.17-24.

“Query Reformulation with Constraints” A. Deutsch, L. Popa, and V. Tannen. SIGMOD Record 35(1): 65-73 (2006)

“XML Queries and Constraints, Containment and Reformulation”, A. Deutsch and V. Tannen. Theoretical Computer Science, Vol.336, May 2005, pp.57-87.

“Semantic Web and Databases” C. Bussler, I. Fundulaki, and V. Tannen (eds.) Revised Selected Papers from SWDB, Toronto, 2004. Springer LNCS 3372, 2005.

“MARS: A System for Publishing XML from Mixed and Redundant Storage”, A. Deutsch and V. Tannen. Proceedings VLDB 2003.

“Viewing the Semantic Web through RVL Lenses.” A. Magkanaraki, V. Tannen, V. Christophides, and D. Plexousakis. Proceedings of 2nd Int’l Semantic Web Conference (best paper award) 2003. Full version in J. Web Sem. 1(4): 359-375 (2004)

SYNERGISTIC ACTIVITIES

NSF Presidential Young Investigator 1990–1995, Editor, “Electronic Journal of Discrete Mathematics and Theoretical Comp. Sci.”, Area Editor, “Encyclopedia of Databases”, Springer-Verlag 2008.

Program Committees (recent): 2009 Mendelzon Wksh’p on Foundations of Data Management (AMW), 2009 Int’l Conf. on Extended Database Technology (EDBT), 2009 Int’l Conf. of Data Engineering (ICDE), 2008 Principles of Database Systems Symp. (PODS), 2008 Int’l Conf. on Extended Database Technology (EDBT), 2008 Int’l Conf. on Very Large Databases (Demos) (VLDB), Co-Chair, 2007 Data Integration in Life Sciences (DILS), Vice Chair, 2007 Semantic Web Conf. (ISWC), 2007 Int’l Conf. on Scalable Uncertainty Management (SUM), Area Chair, Scientific and Biological Databases and Bioinformatics for 2006 Int’l Conf. of Data Engineering (ICDE), 2006 Int’l Conf. on Extended Database Technology (EDBT), 2006 Semantic Web Conf. (ISWC), 2005 Principles of Database Systems Symp. (PODS), Co-Chair, 2004 Semantic Web and Databases Wksh’p (SWDB), 2004 Web and Databases Wksh’p (WebDB), 2004 XML Database Symp. (XSym), 2004 Semantic Web Conf. (ISWC), 2003 Int’l Wksh’p on Database Prog. Lang. (DBPL), 2003 Int’l Conf. on Very Large Databases (VLDB), 2002 Int’l World Wide Web Conf. (WWW11), 2001 Int’l World Wide Web Conf. (WWW10), 2001 Int’l Conf. on Database Theory (ICDT),

Keynote and Conference Invited Talks (recent): Int’l Provenance and Annotation Wksh’p, Salt Lake City, June 2008; Symposium on Provenance in Databases, e-Science Institute Edinburgh, May 2008; Workshop on Principles of Provenance, Edinburgh, Nov.2007; Mendelzon Workshop on Foundation of Databases and the Web, Chile, Nov.2006; Int’l Workshop on Exchange and Integration of Data, Brixen-Bressanone, Italy, June 2006; University of Washington / Microsoft Research Summer Institute on Infrastructure for Managing Imprecise Information in Relational Database Systems”, Salish Lodge, Aug.2005, Grids and Applied Language Theory: NeSC Workshop, Edinburgh, Oct.2003; Foundations of XML: Dagstuhl Workshop, Sept.2001.

Recent Invited Lectures at SUNY Buffalo, UC San Diego, Microsoft Research, TU Vienna, University of Washington, Edinburgh University, Indiana University, National Technical University of Athens, INRIA Rocquencourt, Bell Labs, ETH Zurich, AT&T Labs, Ecole Polytechnique Paris, TU Eindhoven, University of Antwerpen, University of Torino, M.I.T., Ecole Normale Supérieure Paris, Imperial College London, University of Southern California, University of Paris VII, Carnegie Mellon University, Hewlett-Packard Laboratories, Stanford University, University of Pisa.

SOME FORMER PhD STUDENTS Atsushi Otori, Tohoku University, Dan Suci Univ. of Washington, Lucian Popa, IBM Almaden, Alin Deutsch, UC San Diego, Arnaud Sahuguet, Google.

Collaborators (not PhD students) in the last 4 years: J.N. Foster, N. Taylor, UPenn, O. Biton, S. Cohen-Boulakia, Univ. Paris 11, W. Piel, Yale, Z. Ives, UPenn.

PhD Students and PostDocs in the last 5 years: T. J. Green, G. Karvounarakis. Total number of PhD students: 9, Total number of postdocs: 3.

PhD Thesis and PostDoc Supervisor: A. Meyer, MIT.

BIOGRAPHICAL SKETCH

Alexandros Stamatakis

Junior Research Group Leader (equiv. to Assist. Prof.), Dept. of Computer Science
Technical University of Munich, D-85748 Garching b. München
E-mail: stamatak@cs.tum.edu

Professional Preparation

Diploma (equiv. to MSc), 2001, (Computer Science) Technical University of Munich.
Dr. rer. nat, 2004, (Computer Science) Technical University of Munich.

Appointments

Oct 2008 – present Junior Research Group Leader, Technical University of Munich,
Germany
Feb 2008 – Sept 2008 Junior Research Group Leader, Ludwig-Maximilians University of
Munich, Germany
July 2006 – Jan 2008 Postdoctoral Fellow, School of Computer and Communication
Sciences, Swiss Federal Institute of Technology, Lausanne, Switzerland
Jan 2005 – June 2006 Postdoctoral Fellow (funded by German Academic Exchange Service)
Foundation for Research and Technology Hellas, Institute of Computer Science,
Heraklion, Greece

Five Publications Relevant to Proposed Research (of 45)

Pattengale, N.D., M. Alipour, O.R.P. Bininda-Emonds, B.M.E. Moret, and A. Stamatakis.
2009. How Many Bootstrap Replicates are Necessary?. *Proceedings of RECOMB 2009*,
Tucson, Arizona, to be published.
Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A Fast Bootstrapping Algorithm for the
RAxML Web-Servers. *Syst. Biol.*, 57(5): 758-771.
Ott, M., J. Zola, S. Aluru, and A. Stamatakis. 2007. Large-scale Maximum Likelihood-based
Phylogenetic Analysis on the IBM BlueGene/L. *Proceedings of IEEE/ACM
Supercomputing (SC2007) conference*, Reno, Nevada, November 2007. **Best Paper
Award Finalist**
Stamatakis, A. 2006. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses
with Thousands of Taxa and Mixed Model. *Bioinformatics* 22(21):2688-2690.
Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A Fast Program for Maximum
Likelihood-based Inference of Large Phylogenetic Trees". *Bioinformatics* 21(4):456-463.

Five Other Publications

Alaxiotis, N., E. Sotiriades, A. Dollas, and A. Stamatakis. 2009. Exploring FPGAs for
accelerating the Phylogenetic Likelihood Function. *Proceedings of HICOMB 2009 (in
conjunction with IPDPS 2009)*, Rome, Italy, to be published.
Stamatakis, A. and M. Ott. 2008. Exploiting Fine-Grained Parallelism in the Phylogenetic
Likelihood Function with MPI, Pthreads, and OpenMP: A Performance Study.
Proceedings of PRIB 2008, volume 5265 of *Springer Lecture Notes in Computer Science*
pp 424-435. **Best Paper Award Finalist**
Stamatakis, A., and M. Ott. 2008. Efficient Computation of the Phylogenetic Likelihood
Function on Multi-Gene Alignments and Multi-Core Architectures. *Philosophical
Transactions of the Royal Society B*, 363: 3977-3984.
Blagojevic, F., D.S. Nikolopoulos, A. Stamatakis, and C.D. Antonopoulos. 2007. Dynamic
Multigrain Parallelization on the Cell Broadband Engine. *Proceedings of ACM SIGPLAN*

Symposium on Principles and Practice of Parallel Programming 2007 (PPoPP 2007), 90-100, San Jose, California. **Best Paper Award**
Bininda-Emonds, O.R.P., A. Stamatakis. 2007. Taxon Sampling versus Computational Complexity and their Impact on obtaining the Tree of Life. In Trevor Hodkinson, John Parnell, and Steve Waldren, editors, *Towards the Tree of Life: taxonomy and systematics of large and species rich clades*, pp 77-95, Volume 72, Special Volume for the Systematics Association, CRC Press.

Synergistic Activities

RAxML web-servers at Vital-IT unit of Swiss Institute of Bioinformatics and CIPRES project portal at San Diego Supercomputer Center
Maintenance & Development of RAxML open-source code for phylogenetic inference
Maintenance & Development of AxParafit & AxPcoords open-source code for co-phylogenetic analysis
Symposium organization (with Susanne Renner) on “Advances in Tree Reconstruction from Complex Data Matrices” at 2009 Evolution Meeting, Moscow, Idaho.
Summer school courses and tutorials on large-scale phylogenetic inference and parallel computing in Bioinformatics: Bernhard-Rensch Summer School at Bremen (2008), BGRS Summer Schools at Novosibirsk (2004, 2006, 2008) German Conference on Bioinformatics at Bielefeld (2004) IEEE Supercomputing Conference at Seattle (2005), Summer School at Munich (2009) Summer School at Cambridge (2009)

Collaborators and other Affiliations

Graduate advisor; Arndt Bode, Technical University of Munich;
Postdoctoral advisors: Bernard Moret, Swiss Federal Institute of Technology, Lausanne
Graduate advisees and current addresses (3):

Michael Ott, TU Munich
Simon Berger, TU Munich
Nikos Alachiotis, TU Munich

Collaborators not listed above or in publications (no co-editors in last 24 months):

Renner, S.S. (LMU Munich), Gottschling, M. (LMU Munich), Grimm, G. (Univ. Tübingen), Auch, A. (Univ Tübingen) Bader, D. (Georgia Tech), Roshan, U. (NJIT), Meier-Kolthoff, J. (Univ. Tübingen), Stockinger, H. (Swiss Institute of Bioinformatics), Johnson, A.D. (NIH), Janies, D. (Ohio State) Curtis-Maury, M. (Virginia Tech) Nindl I. (DKFZ Heidelberg), Stockfleth, E. (Charite Hospital Berlin), Alonso, A. (DKFZ Heidelberg), Gissmann, L. (Charite Hospital Berlin), Bravo, I.G. (Univ. Münster), Hemleben, V. (Univ. Tübingen), Arvelakis, A. (ICS-FORTH Heraklion), Reczko, M. (ICS-FORTH Heraklion), Symeonidis, A. (ICS-FORTH Heraklion) Tollis, I.G. (ICS-FORTH Heraklion), Charalambous. M. (Univ. Cyprus), Trancoso, P. (Univ. Cyprus)

Todd J. Vision

Department of Biology, University of North Carolina at Chapel Hill
Campus Box 3280, Chapel Hill, NC 27599
Phone: 919.843.4507, Fax: 919.962.1625, email: tjv@bio.unc.edu

Professional Preparation

<u>Institution</u>	<u>Area</u>	<u>Degree/Position</u>	<u>Year(s)</u>
University of Chicago	Biological Sciences	B.A.	1992
Princeton University	Evolutionary Genetics	M.S.	1995
Princeton University	Evolutionary Genetics	Ph.D.	1998

Appointments

2007-present	Associate Professor, Department of Biology, UNC Chapel Hill Affiliations: Curriculum in Bioinformatics and Computational Biology, Curriculum in Genetics and Molecular Biology, Carolina Genome Sciences Center
2006-present	Associate Director for Informatics, National Evolutionary Synthesis Center (NESCent), Durham, NC
2001-2007	Assistant Professor, Department of Biology, UNC Chapel Hill
1999-2001	Postdoctoral Associate, Center for Agricultural Bioinformatics, USDA-ARS, Ithaca, NY
1998-1999	Postdoctoral Associate, Dept. of Plant Breeding, Cornell University, Ithaca NY

Most relevant recent publications (10 out of 29)

- Bouck A, Vision TJ (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* 16, 907-924.
- Cheng F, Hartmann S, Gupta M, Ibrahim JG, Vision TJ (2008) A hierarchical model for incomplete alignments in phylogenetic inference. *Bioinformatics*. Epub ahead of print 01/15/2008.
- Ganko EW, Meyers BC, Vision TJ (2007) Divergence in expression between duplicated genes in Arabidopsis. *Molecular Biology and Evolution* 24, 2298-2309.
- Gaulton KJ, Mohlke KL, Vision TJ (2007) A computational system to select candidate genes for complex human traits. *Bioinformatics* 23, 1132-1140
- Hartmann S, Lu D, Phillips J, Vision TJ (2006) Phytome: a platform for plant comparative genomics. *Nucleic Acids Research* 34, D724-730.
- Hartmann S, Vision TJ (2008) Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evolutionary Biology* 8, 95.
- Hemminger BM, Saelim B, Sullivan PF, Vision TJ (2007) Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. *Journal of the American Society for Information Science and Technology (JASIST)* 58, 2341-2352.
- Knies JL, Dank KK, Vision TJ, Hoffman N, Swanstrom RI, Burch CL (2008) Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. *Molecular Biology and Evolution* 25, 1778-1787.
- Lapp H, and 24 others (2007) The 2006 NESCent Phyloinformatics Hackathon: a field report. *Evolutionary Bioinformatics* 3, 357-366.
- Leebens-Mack J, and 26 others (2006) Taking the first steps towards a standard for reporting in phylogenies: Minimal Information About a Phylogenetic Analysis (MIAPA). *OMICS* 10, 231-237.

Synergistic Activities

- Development and support of software programs and databases for genomics (MapPop, FISH, Phytome, mimulusevolution.org).
- Development of “The Power Within”, a high school curriculum module on phylogenomics and the origin of the eukaryotic cell, for the Destiny Science Bus, which delivers inquiry-based science learning activities to underserved minority K-12 students and teachers in NC.
- Co-organizer of NESCent Phyloinformatics Summer Course (2007, 2008); NESCent Summer of Code (2007, 2008); NESCent Hackathons (2007, 2008, 2009).
- Co-organizer: iPlant Green Tree of Life Grand Challenge Workshop, November 2008
- Senior personnel: Dryad digital data archive
- Senior personnel: DataNetOne

Collaborators and other Affiliations

Coauthors on papers & grants (last 5 yrs, excluding students, postdocs & UNC faculty): T. Bradshaw (U Washington), J. Bowers (U Georgia), E. Brenner (NYBG), G. Burleigh (U. Florida), S. Cannon (USDA), F. Chang (Illinois State), M. Clement (Brigham Young), J. Comstock (unaff.), M. Crayton (Xavier U), C. Cunningham (Duke U), C. dePamphilis (Penn. State), R. deSalle (AMNH), J. Doyle (Cornell), J. Eisen (UC Berkeley), Oliver Eulenstein (Iowa State), D. Fernandez-Baca (U. Iowa), X. Gu (Iowa State), M. Gupta (Boston University), J. Harshman (unaff), M. Holder (U Kansas), R. Holland (EMBL), I. Holmes (UC Berkeley), R. Jansen (UT Austin), T. Katayama (Osaka U), E. Kellogg (U Missouri), S. Knapp (U. Georgia), J. Knies (Brown), E. Koonin (NCBI), H. Lapp (Duke U.), J. Leebens-Mack (U Georgia), P. Lewis (U Connecticut), P. Mabee (U South Dakota), A. Mackey (GSK), B. Martin (Oklahoma State), B. Meyers (U of Delaware), W. Michener (U New Mexico), B. Mishler (UC Berkeley), B. Osborne (BioTeam), H. Philippe (U Montreal), W. Piel (Yale), S. Kosakovsky Pond (UC San Diego), W-G. Qiu (CUNY), D. Schemske (Michigan State), C. Pires (U Missouri), Y-L. Qiu (U Michigan), S. Rhee (Stanford U), S.-H. Shiu (Michigan State), K Sjölander (UC Berkeley), D. Soltis (U Florida), P. Soltis (U Florida), J. Stajich (UC Berkeley), L. Stein (CSHL), D. Stevenson (NYBG), A. Stoltzfus (NIST), T. Thierer (Biomatters), A. Vilella (EMBL), L. Szekely (U South Carolina), K. Smith (NESCent), J. Tang (U South Carolina), C. Tauer (Oklahoma State U), J. Tomkins (Clemson U), R. Vos (U British Columbia) T. Warnow (UT Austin), M. Westerfield (U Oregon), J. Willis (Duke U), C. Zmasek (Burnham Inst.)

Graduate students: 2 current: Toby Clarke, Suja Thomas; 2 former: Sugata Chakravarty (MS 2004, UNC Charlotte), Ruchir Shah (PhD 2006, Constella Group).

Postdocs: 10 former: Amy Bouck (Pioneer Hi-Bred), Gordon Burleigh (U Florida), María Chacón (U. Nacional de Colombia), Jixin Deng (Baylor U.), Kirsten Fisher (Cal State, LA), Eric Ganko (Syngenta), Stefanie Hartmann (U Potsdam), Maria Tsompana (SUNY Buffalo), Rajkumar Rathinavelu (Indian Tobacco Co.), Zongli Xu (NIEHS)

Ph.D. advisors: H. Hollocher (U Notre Dame), D. Stratton (U. Vermont)

Postdoc advisors: S. Cartinhour (USDA), S. Tanksley (Cornell U.)