

Tree Biology Cyber Infrastructure

Group Leader: *David Neale (UC Davis)* dbneale@ucdavis.edu

Community group members: *Sally Aitken (U British Columbia), Michael Dietze (U Illinois), Dan Kliebenstein (UC Davis), Sarah Mathews (Harvard U), Ram Oren (Duke U), Jill Wegrzyn (UC Davis), Ross Whetten (North Carolina State U)*

The NSF iPlant Plant Sciences Opportunity Team (PSOT) identified Tree Biology as a potential new area for cyber infrastructure (CI) development by the iPlant development team since Tree Biology currently has data that cuts across multiple scales and will provide a test bed to attempt integration from DNA to the Globe. The iPlant leadership approved a pre-conference Tree Biology CI planning meeting to be held in advance of the iPlant annual meeting in Las Vegas, NV on May 24-25, 2010. The Tree Biology CI meeting was organized by David Neale and this proposal grew from discussion held at that meeting.

1. The biological challenge the seed CI is aimed at addressing.

The sub-disciplines of Tree Biology are data rich, however, database resources for these data are not uniformly well-developed across sub-disciplines (Table 1). In Genetics, there are databases for genomic and other "omic" data, but are poorly developed for the vast common garden and population genetic diversity data resources. In Physiology, there are virtually no database resources even though the discipline is data-rich. In Ecology, there are also some important database resources but these do not cover all data types of this community. Systematics and herbaria/arboreta database resources are well-developed and should be quite easy to incorporate into a Tree Biology CI.

One of the most pressing data needs for scaling forests from the individual to the globe is information on ecological traits and physiological rates by species, geo-referenced, and linked to ancillary data (e.g., GIS layers, remote sensing, downscaled past, current and predicted future climate data) that allow us any understanding of the spatial variability in rates and processes at multiple scales from the stand to landscape to biome to globe. This is a key connection to lower levels in the biological hierarchy (population, physiology, etc) and one that cannot be stressed enough. There is a similar need for ecological data, such as tree demographics and ecosystem rates, to be linked, geo-referenced, etc. This would assist with direct analysis, model validation, and the estimation of the current state of ecological systems. This later goal is essential for model initialization and research on ecological data assimilation is a rapidly emerging area of research but one with few established tools that permit the lay user to access. Beyond global change such virtual cyber-networks would provide invaluable information for the biodiversity crisis. Right now we simply do not know enough about enough species. Potential tools would allow connections to the IPTOL. To move forward we do not need information on every species as they are not phylogenetically or ecologically independent, but current knowledge at the lower scales of biological organization is extremely biased and has massive gaps.

2. The societal significance of the challenge.

Global climate change is affecting the health, composition and distribution of forests around the globe. Policy makers and resource managers will need the highest quality science to inform decision-making processes. Tree Biology is a mature discipline that is well positioned to provide the science. Tree Biology, however, spans multiple scales from the genome (DNA) to global ecosystems.

3. A detailed description of the functionalities of the seed CI.

To frame the discussion on what type of tree physiology research we wish to promote, what sort of data we wish to archive, what information we wish to obtain from the data, and what types of analytical tools we wish to apply to the data, we propose concepts outlined by John Passioura

(1979). In this paper, Passioura discussed that physiological research tended to be focused on the behavior of molecules and whole plants and communities, neglecting what's between. He recommended that we assess our research activities using a framework of hierarchically organized systems, measuring understanding by our ability to link adjacent layers. If we cannot link upscale, the phenomenon we study may be trivial; if we cannot link downscale the observation is merely descriptive. DNA to the Globe in Tree Biology offers a rare opportunity to employ Passioura's approach without imposing undue burden on individual. Data collected by tree biologists must address large spatiotemporal heterogeneity based on methods capable of obtaining data slowly. The solution is to (1) standardize data collections to the greatest degree possible while relying as little as possible on individual investigators, and (2) somehow impose data archiving in a manner leading to analyses and syntheses generating the needed parameters. **One objective of this initiative is to suggest a standard measurement protocol and a base data set.**

Many temperate tree species have large geographic ranges that encompass a wide range of climatic conditions. The adaptation of populations to local climates has resulted in considerable population differentiation, and managing this has been a major challenge in forestry. There is a history of long-term field-based provenance trials to determine what populations to use for reforestation and breeding, and how far seed can be transferred from provenance to planting site without maladaptation. These reciprocal transplant studies usually contain samples from many geo-referenced populations planted on many sites in different geographic areas that vary in climatic conditions. Results from these experiments have become invaluable for separating genetic from environmental sources of variation, and for developing models predicting the responses of populations to climate change. Growth-chamber based common garden studies have also generated a wealth of data on the local adaptation and population differentiation of tree populations. These experiments also allow for integration with physiology through assessing population differences in physiological traits including abiotic stress tolerance, disease or insect resistance, growth phenology, gas exchange, and effects of enriched carbon dioxide. Data from these experiments is often limited to height, diameter, survival, and damage scores from biotic or abiotic stresses. **Many of these experiments are decades old, and there are no databases established for archiving and sharing results outside of agencies, universities, or companies who established them.**

Genetic diversity of forest trees has been studied extensively using selectively neutral genetic markers. Many allozyme studies were completed from the 1970's through 1990's, and estimates of population genetic diversity and divergence parameters have been summarized in several reviews by J.L. Hamrick. In recent years a wide variety of marker types has been utilized. Two types of genetic resources would be useful for the tree biology community from this research (1) **a database of available markers by genus and species, particularly primer sequences for markers;** and (2) **a database of genotypic data from geo-referenced populations included in these studies.**

Much of the research in forest ecology is still limited to a collection of idiosyncratic case studies. While this approach to research has been fruitful, it has limited the spatial scales across which ecologists are able to make inference or to anticipate large-scale changes due to land-use, invasive species, pollution, and climate change. These challenges require the ability to synthesize across scales both within ecology and between ecology and the other domains of biology, whether that be by direct statistical analysis, meta-analysis, or by mechanistic modeling. There is an urgent need in forest ecology for cross-site and cross-scale synthesis. Many of the problems with research in forest ecology could be alleviated by an improvement in cyber-infrastructure. **Such an effort should focus on standardized data formats and tools to operate on standard formats.** Key questions could focus on determining both the scaling of different state variables as well as formalizing our understanding of the relationships among

different variables. This could build on existing ecoinformatics work and tools developed at the National Center for Ecological Analysis and Synthesis (NCEAS). **Key needs are the ability to search across existing and emerging ecological databases, to assimilate past data, for example by automated literature searches and data extraction, and to establish repositories for raw data whether it be new data collected as ecology moves forward or it be old data that needs to be captured before it is lost to history.**

OVERVIEW OF CURRENT RESOURCES

Database	Taxon	Data Types
TreeGenes/Dendrome	Conifers and other forest trees	ESTs, Comparative Maps, Protein, Resequencing data, Genotypes, Phenotypes
ConiferGDB	Pinaceae	ESTs and Next-gen sequencing
GDR	Rosaceae	ESTs, Comparative Maps, Proteins, Genotypes, Phenotypes, Genome Annotations
Fagaceae Genomics Web	Fagaceae	ESTs, SNPs, Next-gen sequencing
AspenDB	Populus	ESTs and Microarray
PoplarDB	Populus	ESTs
PopulusDB	Populus	Genome Annotation and Microarray
EucalyptusDB	Eucalyptus	Genome Annotation
EVOLTREE	Conifers, Populus, and Quercus	ESTs, Comparative Maps, Genotypes, Phenotypes

4. Design, development and implementation time line.

The timeline to design, develop and implement the Tree Biology CI prototype is expected to be one year.

5. Management plan.

This prototype project will be co-managed by David Neale and the iPlant development team. Neale and the IT staff person supported by iPlant to the Neale Lab will develop the initial schema and identify all data resources. The iPlant development team will do the programming and interface development to develop the CI.

6. Brief vision of the future more comprehensive CI needs of the discipline, and how the seed CI being developed will be integrated and help facilitate the larger CI

The prototype CI will be developed around a small number of tree species with existing information resources. Such species include *Pinus taeda*, *Populus trichocarpa* and *Eucalyptus grandis*. These prototypes will develop the CI linkages across scales and subdisciplines of forest biology. The long term is to develop a Forest Biology CI that integrates not only across the 1000s of species of forest trees but also across the complex array of associated forest organisms.