PToL Year 4 Roadmap - Q3 Status

iPToL Year 4 Roadmap

Revision 2.3 – Q3

The June-September period has been largely occupied by outreach efforts, with iPToL presence at the following conferences:

- Evolution (oral presentation: NM)
- iEvoBio (2 oral presentations: NM, JE)
- Botany (workshop: NM, SM; poster: MN)
- Plant biology (workshop: NM, poster: MN)
- UseR (poster: NM)

In certain area, the progress of the iPToL has been affected by changes in the discovery environment development plan. This affected in particular data management, visualization and integration into the discovery environment. As a result some original objectives have been deprecated or postponed to a later date to be determined according to the overall development. In some cases, the working groups have taken over additional projects.

Data Assembly

General Data Assembly

After some initial difficulties, the different components of the Data Assembly are now either on-track for delivery or already completed. The data upload and management capabilities have been greatly improved through a refactoring of the DE backend. The work on the ingestion pipelines is also moving forward. Collaborative tools are planned as part of the future DE development.
Deliverables:

- An industrial strength pipeline for data assembly for very large data sets;
- Collaboration and Analysis Tools (contribute one's own data, download data, analyze data, keep data/results private) through the DE.

Strategy:

- Engage iPlant faculty (Nirav Merchant, Sudha Ram, Eric Lyons) for domain expertise in data infrastructure, meta-data management and scientific workflows.

Tasks:

- Robust data upload capability (IRODS; Rion Dooley, Nirav Merchant), late 11Q1?
    - **COMPLETED**
- Meta-data management (Rion Dooley, Nirav Merchant, Sudha Ram), early 11Q2;
    - **POSTPONED/IN PROGRESS**
- Robust data storage and retrieval, collaboration tools in DE (iPlant-wide requirement; core software), 11Q2;
    - **DATA STORAGE: COMPLETED**
    - **COLLABORATION TOOLS: POSTPONED/IN PROGRESS**
- Advanced collaboration tools (iPlant-wide requirement; core software), 11Q3;
    - **POSTPONED/IN PROGRESS**
- ~~Input data validation (core software in collaboration with data integration ETAs);~~
    - **DEPRECATED:** The original database/metadata driven model for the discovery environment has been replaced by a file-based storage. This requires a redesign of the role metadata will have in the new context.
- Multiple sequence alignment generation:
    - PHLAWD (John Cazes, Stephen Smith);
        - **ONGOING**
    - Gordon Burleigh's pipeline (John Cazes, Eric Lyons, Stephen Smith);
        - **STALLED – BEING RESOLVED**
    - Muscle, other alignment strategies (John Cazes, Eric Lyons).
        - **COMPLETED**
- Sequence database(s) bad GIs for PHLAWD, etc (Sheldon McKay and delegates), 11Q2.
    - **NCBI'S GENBANK ADDED**

My-Plant/My-Crop

The backend has been successfully redesigned, which will allow the further development of My-Crop
Deliverables:

- My-Plant: Robust, widely used scientific collaboration network for the plant sciences based on a phylogeny metaphor;
- My-Crop: In support of integrated breeding platform, a scientific interaction site as well as data landing pad.

Strategy:

- Build a phylogenetically-structured social networking website for information sharing and collaboration;
- Generalize and extend to other display/networking paradigms.

Milestones:
1. Official launch 10Q3.
Status:

- 239 users, 57 clades (Feb 3, 2010).

Tasks:

- Refactor back-end for generic 'clade' structure to facilitate other display/organization paradigms (Matt Hanlon), late 11Q1:
    - Node based;
    - Drupal core taxonomy.
        - **COMPLETED**

- Implement Drupal module to ingest and provide basic search functionality for relevant literature citations from the user community and public databases (Steve Mock and delegates), 11Q2;
- Integration with Facebook and other similar sites. Initially with passive linkouts, possible later with Facebook page or app (Steve Mock and delegates), 11Q3;
- Integration (as consumer) of TNRS and iPlant tree viewer (Steve Mock; Matt Hanlon);
- My-Crop (different paradigm for display – also data repository):
    - scoping, early 11Q2;
    - early implementation, late 11Q2;
    - full implementation, early 11Q3.

Trait Evolution

Initial work on the integration of tree stretching (evolutionary) models identified major problems with the optimization routines used by the underlying R package. Because the problems are severe and affect other packages and functions (optim) that are used across the spectrum of biological sciences, the group decided to further investigate the problem, with specific application to phylogenetic questions. A statistics graduate student has been integrated into the group to work on a project that should elucidate the conditions under which the performance of the optimization routines yield unreliable results and formulate suggestions on which routines are more appropriate.
Moreover, members of the group have integrated 4 additional tools into the discovery environment. A new tool written by members of the working group will be integrated and linked from the publication describing it.
Deliverables:

- An infrastructure for trait analysis and ancestral characters estimation.

Strategy:

- Integration of limited number of trait analysis tools with special focus on R scripts (since Sept 2010).

Milestones:
1. Identified set of 5 components for integration:

- Phylogenetically Independent Contrast (PIC);
- Discrete Ancestral Character Estimation (DACE);
- Continuous Ancestral Character Estimation (CACE);
- Tree stretching models;
- Discrete traits correlations (Pagel 94).

2. Released 1st component (PIC) 10Q2.

1. DACE and CACE included in the 3rd release of DE 11Q1;
2. PL, lopper, OUCH, Picante, DTT included in the 0.4 release of the DE 11Q2;
3. Code improvements in 2 existing R programs (ape, geiger) pushed back to community, 11Q
4. Publications:

- Beaulieu J.M., Jhwueng D.C., and O'Meara B.C. Modeling stabilizing selection: Relaxing the Ornstein-Uhlenbeck model of adaptive evolution. Evolution. **(Submitted)**
- O'Meara, B.C. and B. Banbury. Approximate Bayesian computation for comparative methods. (**In preparation**)

Status:

- Tree stretching and Pagel94 to be included in 6th DE release 11Q3.
- Postdoc and grad student in the group are able to independently integrate tools into the DE.

Tasks:

- Tool integration in DE-R scripts and command line tools (Naim Matasci and delegates), early 11Q2;
  - **ONGOING**

- Data uptake- files and external web based data from TreeBase
  - **POSTPONED**: The development of a generic interface for web services will provide a simple, unified solution to remote data access.

- Tree viz integration:
  - New visualization needs (Kris Urie), early 11Q2;
    - **IN PROGRESS**
  - ~~Call backs for new analyses (Adam Kubach, Sonya Lowry), mid 11Q2.~~
    - **DEPRECATED**: A new unified strategy for graphical interactions will be developed for the DE.

- DE integration:
  - Integration (~~Sonya Lowry~~ Nira Merchant and delegates), late 11Q2;
    - **COMPLETED**
  - ~~Analysis and viewer integration (Sonya Lowry and delegates), late 11Q2.~~
    - **DEPRECATED**: A new unified strategy for graphical interactions will be developed for the DE. The metadata mapper will be part of the standalone release.

- Metadata mapping (Kris Urie, Naim Matasci), late 11Q3 (**Q3 new**);
  - **IN PROGRESS**

- Optimization review (Naim Matasci, Kurt Michels), 11Q4/12Q1 (**Q3 new**)
  - **IN PROGRESS**

- Code and documentation release (Naim Matasci, Matthew Helmke), 11Q2.
  - **ONGOING**

Tree Reconciliation and onekp

The sequencing of the 1000 transcriptomes has been completed and the group is now moving into the data analysis. The data for the "deep green" publication is currently being processed and the analysis should be completed by Nov 2011. The group has however identify issues with the current SOAP assembley and is looking into alternatives, including running Trinity Assembler on TACC resources.
Deliverables:

- Applications to perform, visualize and analyze the evolution of gene families from the onekp project with gene-species tree reconciliations.

Strategy:

- Development of an analytical pipeline, a database schema and a visualization tool;
- Populate with data from onekp project.

Milestones:
1.Bioinformatic pipeline for gene-species tree reconciliation completed and database populated with the reconciled trees, 10Q4;
2.Developer preview released with the following features, 11Q1:

- Database containing reconciliations for over 2500 gene families in six examplar species (poplar, grape, cucumber, papaya, soybean and Arabidopsis thaliana), 11Q1;
- GUI with the ability to search and view reconciled trees and to download data;
- Display of species trees and gene trees side-to-side, using the Tree Visualizer developed by the Tree Viz Working Group;
- Interactive mapping of duplication and speciation events between gene and species tree and vice versa;
- Markups for speciation and duplication events on the gene tree nodes and of duplication events on the species tree branches;
- Ability to add additional markups;
- Contextual menus;
- Advanced search functionality, including:
  - BLAST,
  - GO terms and IDs,
  - Gene IDs;
- GO tag clouds for gene families;
- Retrieval of underlying data (sequences and reconciliations).

3. Publications:

- Estill J et al. The TRON ontology (provisional). **(In prep)**.

4. Completed serial pipeline for Bayesian gene trees

Status:

- Ready to receive onekp data 11Q1;
- All the current onekp assemblies (36,998,590 assemblies from 398 total species/tissues) mirrored at TACC;
- Have run blastx searches on each of the onekp assemblies (complete);
- Blast server at TACC for onekp consortium members to search against the assemblies using either single sequences or batches (complete).
- Reconciliation ontology described
- Basic XML integration framework in place

TR Tasks:

- Data modeling and database schema design (Sheldon McKay, Jamie Estille), 10Q3, update 11Q2;
  - **COMPLETED**
- Port tree reconciliation analysis pipeline to TACC HPC resources (Sheldon McKay), early 11Q2;
  - **COMPLETED**
- Adapt tree reconciliation analysis pipeline to use Bayesian tree building method (PrIME-GSR, Sheldon McKay), mid 11Q2;
  - **COMPLETED**
- Update stand-alone web application for onekp/TR data (Naim Matasci), ~~early 11Q2~~ 11Q3;
  - **IN PROGRESS**
- ~~DE integration (Sonya Lowry), 11Q2;~~
  - **DEPRECATED**: A new unified strategy for graphical interactions will be developed for the DE.
- Establish/negotiate onekp data release policy (Sheldon McKay,Jim Leebens-Mack and Gane Wong), 11Q1;
  - **IN PROGRESS**
- Expose services through DE and API (TBD), 11Q2;
  - **NO PROGRESS: Not considered a priority by the WG**
- Release code and documentation (Matthew Helmke), 11Q1;
  - **COMPLETED**
- Adjustments to viz as required (~~Adam Kubach~~, Kris Urie).
  - **ONGOING**

Onekp Tasks:

- Sequencing and transcript assemblies (onekp consortium; external);
  - **COMPLETED**
- Gene cluster identification and alignment (Norm Wickett), ongoing;
  - **COMPLETED**
- Continue to manage onekp data intake and tool implementation at TACC (Michael Gonzales, Sheldon McKay, Chris Jordan).
  - **STALLED: Ineffective management is severely hindering progress in this group.**

TNRS

At this point, the only task remaining is to provide support for multiple authorities and multiple taxonomic codes.
Deliverables:

- A Taxonomic Name Resolution Service that:
  - will query taxonomic data from Tropics and other data services using GNI architecture and global names index allowing for different nomenclatures;
  - recover validated names names using exact and fuzzy matching algorithms;
  - inspect taxonomic status of validated names and convert synonyms where applicable.

Strategy:

- Development of a tool based on TaxaMatch (Tony Rees, CSIRO Marine and Atmospheric Research) and GNI Parser (Dmitry Mozzherin from the Encyclopedia of Life).

Milestones:

1. First release of the tool 10Q4;
2. Completion of Phase 1/Scoping of Phase 2 11Q1;
3. Support for Family and infraspecific epithets 11Q1.
4. Support of synonyms 11Q2.
5. Improvements to GNI parser and TaxaMatch codes pushed back to community, 11Q1
6. Publications:

- Boyle B et al. The taxonomic name resolution service: an online tool for automated standardization of plant names. (**In prep**).

Status:

- Support for multiple data sources, 11Q2;
- Have big tree for plant observation database.

Tasks:

- UI redesign (Nicole Hopkins), 11Q1;
    - **COMPLETED**
- Algorithm to handle synonyms (Jerry Lu), early 11Q2.
    - **COMPLETED**
- Support for multiple authorities, 11Q3.
    - **IN PROGRESS**

Big Trees

Both tools have been ported to a HPC environment and are available at TACC so that component of the work of the Big Tree group can be considered completed. However, continuous improvements to the codebase are undergoing. The tools can now be used to generate large phylogenies as part of the phylogenetic workflow and the perpetually updating big tree. A 55K tree has been published and is available through the DE and on the Tree Viewer. Furthermore, improvements to RAxML have been presented at Evolution11.
Deliverables:

- Computational infrastructure to build ToL.

NINJA/WINDJAMMER.

Strategy:

- Optimization of NINJA (neighbor joining implementation) for HPC.

Milestones:

1. Software rewritten from Java to C with an MPI;
2. On-board distance matrix calculation added (K2P and Jukes Cantor for DNA; Blossum 42 for protein);
3. Six day run time reduced 32-fold to 4.5 hours for 220K species data set;
4. Two/three day run time reduced 1,800-fold to 2 minutes for distance matrix calculation on 220K set.

Status:

- Completed, minor tweaking of MPI.

RAxML

Tasks:

- Implement RAxMl-lite on Ranger, benchmark with various data sets (John Cazes), 11Q1;
    - **COMPLETED (BENCHMARKING?)**
- Implement web interface (relies on foundational API; Steve Mock and delegates), 11Q2.
    - **IN PROGRESS (CIPRES)**

Phylogenetics Workflow and Perpetually Updating Tree

Workflow
A Nascent workflow has been added to the DNA subway as an education tool. This can serve as a model for integrating phylogenetic analysis tools to the DE.

- **IN PROGRESS**

Perpetually updated TOL This is predicated on the completion of the infrastructure for data matrix assembly, RAxML-lite tree building, tree visualization etc. and is being scoped by the iPlant scientific project Management team (Eric Lyons, Sheldon McKay, Matt Vaughn, Nicole Hopkins). Advice will be sought from the iPToL faculty regarding further requirements.

- The basic strategy is an automated workflow that will synch with GenBank or other data repository, build or iterate on a character matrix,

re-run the tree building and update the Discovery Environment.
    - **IN PROGRESS**

Tree Visualization

The shift in development priorities for the Discovery Environment resulted in the postponement of visualization related projects. Therefore the group is now concentrating in further developing the tree viewer as a standalone application and plans to submit a publication describing it, probably early next year.
Deliverables:

- An interactive tree viewer that:
    - Makes possible to view large trees as a stand alone tool;
    - Makes the green plant ToL and sub-trees available in the iPlant DE;
    - Meets the visualization needs of Trait Evolution and Tree Reconciliation and of other applications in the Discovery Environment.

Strategy:

- Development of a *de novo* tree viewer using GWT.

Milestones:

- Tree able to display 500K taxa with semantic zooming;
- Search capabilities;
- Metadata driven node interactions;
- Visual annotations (node and branch colors, thickness, size, etc.);
- Designed API for generalized use;
- User interactions support;
- User interface.

Status:

- Completed development for TR functionality and moved into maintenance mode (pending changes needed to accommodate large datasets).
- Working towards publications.

Tasks:

- Continuing development towards inclusion of TE functionality ~~in the 4th DE release~~ (~~Adam Kubach~~, Kris Urie), ~~11Q2~~ **11Q3**;
    - **IN PROGRESS**
- Integration into the DE (Sonya Lowry), 11Q2;
    - **COMPLETED**
- Standalone release (Karen Cranston), 11Q3
    - **IN PROGRESS**
- Code and documentation release (Karen Cranston, Matthew Helmke), 11Q2.
    - **ONGOING**