Summary of proposed work

Open Tree of Life has just released a first draft version of a comprehensive tree of life based on synthesis of published phylogenies and taxonomic data. This tree exists within a framework for community contributions. Our open-source software stack includes components for merging taxonomies, curating input trees, synthesizing trees and taxonomies, browsing the tree and providing feedback. All of the data - input trees, reference taxonomy and synthetic tree - are publicly available for download and through application programming interfaces (APIs). Our methods for synthesis using novel graph database technologies scale to tens of million of nodes and thousands of input trees.

This supplemental funds request will allow OpenTree to update the synthetic tree and reference taxonomy, and continue hosting and improvements to the Open Tree platform. The primary goals of the request are:

1. *Improve the synthetic tree*: Continuously update the synthetic tree, primarily through the incorporation of new input phylogenies but also through better weighting / filtering of input trees using metadata and by enabling users to deploy the synthesis platform for specific clades
2. *Discovering conflict*: Improve communication of the existence of conflicting phylogenetic hypothesis, both in the tree browser ("How much conflict underlies this clade of the synthetic tree?") and also the curation application ("How does my tree compare with other published trees?")
3. *Linking data*: Develop mechanisms to more easily link / share data between Open Tree of Life and other biodiversity resources such as Arbor, Next-gen Phenomics, EOL, ToLWeb and GBIF
4. *Sustainability*: Ensure sustainability through hardening and testing of the software components; responding to feedback from user and developer communities; organizing workshops to grow the curator community

As part of the sustainability efforts, work on all four of these goals will occur across the three institutions to ensure that multiple developers are familiar with the code, that documentation is sufficiently detailed and that infrastructure components work together efficiently. The three curation / synthesis workshops will be primarily organized by PI Cranston, but drawing on connections that the other PIs have to specific target audiences, for example natural history museums at Kansas and Michigan. The following sections detail the specific work that will happen at Duke (PI Cranston), Kansas (PI Holder) and Michigan (PI Smith).

Duke (PI Cranston)

Improving the synthetic tree through community contribution

Improvements to the synthetic tree will largely come through deposition of newly-published phylogenetic trees. Scalability of this community curation depends upon software that incentivizes data deposition, on being able to efficiently incorporate many more input trees, and on being able to communicate to users how their trees are improving the synthetic tree. Curator ranking of input trees currently play a key role in resolving conflict between input trees. Cranston will assess what data our curators use to construct these rankings and will test which metadata best predicts existing rankings. Along with software architect and web application developer, we will motivate curation of these data through visual incentives in the curator application and implement their incorporation into the synthesis methods. By using programming interfaces to existing reference manager tools (e.g., our Mendeley group), we will allow the community to propose phylogenies for inclusion in OpenTree and track the status of data from nominated publications in the curator application. Finally, we will develop interfaces to easily upload data from the local synthesis instances being developed at Michigan into the primary OpenTree data stores.

Visualization of conflict and coverage

Feedback so far on the project indicates that communication of phylogenetic coverage and conflict is a priority. Users want to know why a particular relationship is not in the synthetic tree, the extent of conflict at nodes in the tree and how their input tree affects the synthetic tree. We will incorporate visual and statistical metrics for conflict being developed at Kansas and Michigan into the OpenTree web application, both in the synthetic tree browser and also in the curation application.

Software sustainability

Long-term sustainability of research software depends on a developer community larger than the funded project. Developers are more likely to contribute to a project that is well-documented, tested, and generally follows software engineering best practices. We will increase test coverage of the code base and ensure documentation is in place for all components. In September 2014, OpenTree and Arbor are running a joint hackathon, and this supplement will allow us to collaborate with hackathon participants to bring community-identified projects to completion.

Project management

Personnel at Duke will continue to be responsible for overall project management and community engagement. Cranston will coordinate effort across the three institutions and the advisory board, and organize three curation workshops to engage the systematics community. The software architect will manage the software development efforts, ensuring that the project follows best practices in software engineering and data linking on the web.

Michigan (PI Smith)

Developing methods for understanding conflict

Conflict is inevitable when comparing any set of trees. This conflict may arise as a result of lack of information, horizontal gene transfer, gene tree discordance, or other phenomena. While generating a synthetic tree is important for many questions and downstream analyses, many researchers are interested in the conflict underlying the synthetic analyses. Our tools and database structure allows for the underlying conflict to be present but there is still development to be done to expose this conflict. Specifically, software that we use to load trees for synthesis, treemachine, maps compatible nodes between sets of trees. These nodes can then be queried in our graph databases. Although our focus had been synthesis to produce trees, this design allows for analysis of conflict and congruence. We will develop the tools necessary to better analyze and understand the underlying conflict between the sets of source trees. At Michigan the development will be centered on the backend analyses and these analyses will be incorporated in the web application primarily at Duke University.

Facilitating synthesis by individual researchers

While synthesis on the entire tree of life is important for a number of use cases, researchers also want to be able to conduct synthesis on more targeted clades. For example, a researcher that studies Mammals might use the synthetic tree created by the Open Tree of Life but might also want to conduct their own synthesis using a different set of studies. Researchers may want to compare synthetic trees in their focal clade using different synthesis methods and different source trees. Our current software stack can be deployed on a local system where researchers can conduct synthesis. However, there are many steps involved and it can require quite a bit of software development knowledge. We will develop easier ways for individual researchers without this expertise to conduct local synthesis on sets of trees.

Connectivity to molecular resources

Successful connectivity to a multitude of resources is important for any project of this type. By connecting resources, the usability, power, and exposure of each resource increases. One of the most important biological resources available is NCBI's GenBank. Because it houses essentially of the generated and published molecular data, it is used by almost all practicing biologists generating or using any molecular data. We will develop a connection from the nodes in the Open Tree of Life graph to the data within GenBank. Instead of emphasizing link outs, we will develop simple procedures for extracting molecular data for nodes in the graph. This will not only connect one important resource of biologists, GenBank, but also provide more opportunities for Open Tree of Life to facilitate phylogenetic reconstruction.

Kansas (PI Holder)

Using reticulate trees to express uncertainty and conflict caused by genealogical phenomena.

Currently the summary synthetic tree is a diverging tree with no reticulation events. Reticulate trees can be used to depict horizontal gene transfer (HGT), introgression, or the presence of hybrid taxa. Reticulate trees can also be used to express phylogenetic uncertainty. Open Tree of Life's comprehensive tree has a large proportion of tips that are placed solely on the basis of their taxonomy. In the current user interface, branches leading to these taxa are visually different from other branches, but it is difficult for a biologist to see the entire set of relationships that are equally plausible based on our input trees. In this supplement, Holder and a postdoc will develop approaches for constructing reticulate trees for the specialized case of one input (our reference taxonomy) that is highly unresolved and which contributes most of the tips to the tree. These networks will let us more clearly display the uncertainty in our synthetic tree, and the visualization tools required to support these networks will also allow the project to annotate and display cases of HGT or other biological processes that lead to non-tree like species relationships. Uncertainty that can be captured in reticulate trees will also allow for more precise statements of taxonomic uncertainty than can be achieved by labelling taxa as *incertae sedis*.

Improving access to Open Tree of Life via other cyber infrastructure in evolutionary biology

One key to the success of the Open Tree of Life project as a nexus for phylogenetic information will be the ability to work well with existing platforms. By improving our "link outs" to other resources, the project will clarify to users that the project complements resources that provide information about particular species (e.g. the Encyclopedia of Life) or detailed discussion of the systematics of different clades (e.g the Tree of Life web project). Our reference taxonomy stores the alignment of its IDs to the identifiers in source taxonomy, so we can already provide many links to other resources. However, linking to resources that are not part of our set of taxonomic inputs will also be crucial. Because the same taxonomic name can be used for different meanings, the creation of these links in a reliable way requires some sophisticated analyses. Fortunately, the tools that the project has already built for combining taxonomies will make this task more feasible. With the support of this supplement, we will be able to provide link-outs to iDigBio, EOL, and Tree of Life Web Project pages relevant to a taxon along with warnings in the cases in which the external website uses alternative meanings of the same taxonomic name. This will help biologists realize which connections warrant extra caution. The Open Tree of Life.Arbor/Nescent hackathon will gather input from the community of developers of bioinformatics tools about what services would be most useful. While many of the minor issues raised in these hackathons will be corrected in the third year of Open Tree, the supplemental funding will allow us to design and implement a "version 2" of the Open Tree of Life Application Programming Interfaces (APIs) to help the project fit the needs of iDigBio, ToLWeb, EOL, Arbor, Next-gen phenomics, and GoLife projects. To encourage the use of Open Tree web interfaces by other programmers, we will build adapters between our tools and widely used toolkits such as BioPython and DendroPy.