

**PHYL-O'DATA (POD) FROM TREE OF LIFE: INTEGRATION CHALLENGES FROM YELLOW  
SLIMY THINGS TO BLACK CRUNCHY STUFF**

Junhyong Kim  
Department of Biology  
Penn Center for Bioinformatics  
Penn Genomics Institute

**Background:**

The AToL (Assembling the Tree of Life) is a large-scale collaborative research effort sponsored by the National Science Foundation to reconstruct the evolutionary origins of all living things. Currently 31 projects involving 150+ PIs are underway generating novel data including studies of bacteria, microbial eukaryotes, vertebrates, flowering plants and many more. Modern large-scale data collection efforts require fundamental infrastructure support for archiving data, organizing data into structured information (e.g., data models and ontologies), and disseminating data to the broader community. Furthermore, distributed data collection efforts require coordination and integration of the heterogeneous data resources. In this talk, I first introduce the general background of the phylogenetic estimation problem followed by an introduction to the associated data modeling, data integration, and workflow challenges.

***Phylogeny Estimation and Its Utility***

While ideas about genealogical reconstruction have been around since Darwin, quantitative algorithmic approaches to the problem have been developed only in the last 50 years. The basic structure of the problem involves considering all possible tree graph structures compatible with an organismal genealogy and measuring their fits to observed data by various objective functions. There are now many algorithms based on various inferential principles including maximum information, maximum likelihood, Bayesian posterior, etc. Many flavors of the phylogeny reconstruction problem have been shown to be NP-hard and there is a considerable body of literature on associated computational and mathematical problems. Phylogenetic methods provide the temporal history of biological diversity and have been used in many applications. For example: to track the history of infectious diseases; to reconstruct ancestral molecules; to reveal functional patterns in comparative genomics; and even in criminal cases, to infer the relatedness of biological criminal evidence. But, a grand challenge for phylogenetic research is to reconstruct the history of all extent organismal life—the so-called Tree of Life.

***Problems and Challenges***

In modern times, much of phylogenetic estimation is carried out using molecular sequences, some explicitly gathered for phylogenetic research; others, systematically collected and deposited in public databases. There are many data management problems associated with using molecular sequence data even from public databases—which are not discussed here. But, for the frontline researcher the problem starts at the stage of actually collecting the biological material for experimentation. That is, the animal must

be captured, preserved, measured, and recorded along with associated metadata (e.g., capture location). Such specimens must be physically archived (called voucher specimen) and identified if possible and given a name. All of these activities are sometimes called *alpha-taxonomy*.

Once a specimen has been obtained and named then it (or presumed identical specimens) must be measured for relevant traits including extracting molecular sequences. This action of obtaining “relevant measurements” involve gathering characteristics that will be broadly comparable amongst different varieties of organisms—thus require a prior data model of what is or is not relevant. Once the relevant measurements are recorded, the next important step is deriving an equivalence relationship between measurements on different organisms such that the measurements are considered to be evolutionarily comparable to each other. This activity is called “establishing homology relationships” and is a critical prelude to further analysis. An example of such homologizing activity is the alignment of molecular sequences whereby equivalent relations of individual sequence letters are established. This activity of defining relevant characters and homologizing their assembly is called *Systematics* and the final product of this activity is the “data matrix” that encapsulates the data model of relevant measurements and the relational maps of sets of such measurements. Biologists widely disagree on details of such matrices—for example, whether a particular measurement should be described as present or absent; thus, these matrices are best seen as a “data view” of the primary objects. It is common in the literature and in public databases to have available only the fixed data matrix. Given the complicated and uncertain ontology of such matrices, a **critical challenge is to endow the phylogenetic matrices with their provenance information as well as to provide a facility to change the “views” as biologists’ assumptions change.**

Notwithstanding the fundamental problems described above, there are continuing activities to collect empirical measurements and generate data matrices and place them into a structured information source. For example, there are current database efforts within the AToL projects where all projects have sub-aims targeting data storage, access, and sharing; and, a small number of projects have been funded to develop domain specific data models and analysis tools such as databases for 3D morphological data, web-based information storage and dissemination, and mining molecular databases for phylogenetic information. As a simple fact the magnitude of the AToL efforts is insufficient to meet the real-world needs of the AToL projects that will become critical as each empirical project matures. More importantly, there is little or no coordination between these efforts and there is a critical need to enable the integration of distributed, heterogeneous, and changing data sources; provide for reliable data archiving and maintenance of data provenance; and, help manage the complex data collection and analysis processes. Many of the projects are already very mature and domain-specific problems, cultural problems, and legacy problems make it difficult to develop a single solution to the problems. Therefore, another **critical challenge is to provide ways to post-hoc integrate the extremely dispersed and heterogeneous phylogenetic data sources in a scalable manner.**

The ultimate end product of phylogenetic reconstruction is the tree graph depicting the genealogical history and associated data. The associated data is usually mapped to substructures within the tree graph—say the nodes of the graph, usually from a secondary analysis (so-called post-analysis). For example, once the phylogeny is known, there are algorithms available for reconstructing the measurements of putative ancestors; thus, we may assign data matrices to interior nodes of the tree. In actual practice, the researcher may try many different algorithms to reconstruct the tree, each algorithm may generate multiple trees (e.g., because of equivalent optimality score), and given some preliminary tree estimate one may want to modify the data matrix (try a different view) and re-estimate the tree. Furthermore, there is often a large battery of post-analysis routines that involve other calculations including calculations on substructures of estimated tree. As is typical in complicated data analysis, the total analysis may involve a large number of steps, some steps recursive, cyclic, or branching. Thus, a **final challenge is to develop a workflow for phylogenetic analysis that automatically tracks analysis flow and helps manage the complexity in such a way that is useful to the primary researcher and helps other researchers recapitulate analyses carried out by third parties.**

Phyl-O’Data (POD) is an NSF-funded collaborative project (<http://www.phylodata.org>) involving Penn, UC Davis, and Yale that will address these challenges and provide software tools for data integration and analysis workflows to be used by the AToL community and by related efforts in evolutionary genomics.