The PhyLoTA Project
AToL: A Phylogenomic Toolbox for Assembling the Tree of Life

Univ. California, Davis
Mike Sanderson
Amy Driskell

Iowa State Univ.
David Fernández-Baca
Oliver Eulenstein

Univ. Pennsylvania
Junhyong Kim

Project Description & Goals

All life on earth is related in a complex genealogy—much of which can be described by a phylogenetic "tree of life." Molecular sequencing technology has fostered vast databases of comparative sequence data which has been widely used to construct small subtrees of this larger tree of life. To date, research has been largely focused on the problem of building a tree from a single data set. Relatively little is known about optimally extracting new phylogenetic data sets en masse from sequence databases and then assembling a synthesis. Novel computational problems as challenging as tree building itself arise naturally in this arena largely because of the scale of the data input. We refer to the collection of gene trees, species trees, and other descriptors of historical relationships among taxa collectively as the *phylota*, by analogy to *biota*. Our AToL efforts focus on new algorithmic and computational problems that arise in reconstructing the large set of phylogenetic trees of sequences encompassed by large sequence databases—in other words, to grapple with the subset of the phylota for which sequence data are relevant.

Our work is both theoretical and empirical. **Theoretical** work is aimed at solving computational problems in the extraction and assembly of data sets and construction of comprehensive phylogenetic trees. **Empirical** work will focus on analysis of three large databases that pose a range of computational challenges. We will also release a **software** "toolbox" that will include implementations of algorithms for extracting optimal phylogenetic data sets from databases and scripts to automate the analysis of large collections of phylogenetic trees.

Current Tools Under Construction

`FarmBlast` – distributes BLAST jobs on a cluster
`blink`, `blinkcomp`, `plink` – perform single-linkage clustering BLAST output; can be distributed on a cluster
`density plot` – visualize sequence cluster space
`biclique exh`, `biclique` – construct concatenated data sets (bicliques)
`grovify` – identify groves in a sequence database
Benchmark data sets – complete sets of all amino acid sequences for "Viridiplantae" from GenBank release 137 and Swiss-Prot version 40.29, assembled into clusters of homologs, classified according to copy number and phylogenetic informativeness, and aligned as standard Nexus-formatted files for phylogenetic analysis.

Research

The phylogenetic information content of large sequence databases – A large portion of our research has been focused on measuring and evaluating the phylogenetic information content of public databases. To date we have examined the entire Swiss-Prot database of 120,000 amino acid sequences for nearly 7500 taxa and a "taxonomically enriched" subset of GenBank,which consisted of 185,000 amino acid sequences for more than 16,000 green plant taxa.

Concatenated data sets (bicliques) – The accuracy of phylogenetic inference is expected to improve through the concatenation of multiple sequences from the same taxon but the enumeration of all possible multi-gene phylogenetic data matrices in a large sequence database is an *NP*-complete problem. However we have implemented exact algorithms to identify all concatenated matrices that are maximal (not contained in larger matrices) and complete (no missing sequences).

Supermatrices – To investigate further the phylogenetic potential of the databases, we assembled supermatrices spanning a small (but broad) taxonomic sample of green plants and metazoans. A "supermatrix" is an *incomplete* concatenated matrix assembled from more than one gene or protein. The resulting green plant supermatrix, for example, contained 69 taxa × 254 genes (2777 seqs and 96,698 characters), an average of 40 genes per taxon. Trees from parsimony analysis of this supermatrix broadly (though not entirely) agree with conventional views on green plant phylogenetic relationships, and many nodes are well supported by bootstrap values.

Supertree Construction – We have been pursuing the improvement and development of methods of supertree and consensus tree reconstruction. Recently we have increased the efficiency of the existing method of "flip supertrees" and have been formulating a new supertree method using additive distances.

Groves – A set of trees (each inferred from a cluster) with enough taxonomic overlap to allow supertree construction is a "grove." We are developing methods for identifying groves in a collection of trees (sets of taxa). The minimum number of groves in a database is a lower bound on the number of supertrees required to encompass all its sequence data.

Tools and Data Sets available on our website **www.phylota.org**