Software School on New Methods for Phylogenomic Analysis

Organizers: Jim Leebens-Mack (Georgia) and Tandy Warnow (Illinois)

Contact emails: [jleebensmack@uga.edu](mailto:jleebensmack@uga.edu) and [warnow@illinois.edu](mailto:warnow@illinois.edu)

This software school will teach new software tools for multiple sequence alignment, gene tree estimation, species tree estimation, phylogenetic network estimation, and metagenomic data analysis.

All software tools are publicly available, most in open-source form on Github, and some have Google Users Groups. The software training will be sufficient to enable the participants to use the tools on their datasets, and follow-up support (through Google Users Groups, or elsewhere) will also be available. All participants should bring their own laptops, and should have installed the required software in advance of the tutorial. Mac and Linux users will have full access to all tutorials; some of the programs also support Windows. Participants with Windows machines should contact the tutorial instructor in advance for additional instructions. No fee will be charged for attending the workshop.

**Note: A full breakfast will served from 7:30-8:30 AM at MR 7 (Level 3). Please wear your badge.**

Brief overview:

A. Phylogenetic network estimation (addressing ILS, HGT, hybridization)

* Phylonet (Luay Nakhleh and Yun Yu), MR 4, 11:15-12:15
* PhyloNetworks/TICR/SNaQ (Cécile Ané and Claudia Solis-Lemus), MR 4, 8:30-9:45 (Part 1) and 10:00-11:00 (Part 2)

B. Species tree estimation in the presence of ILS

* ASTRAL (Siavash Mirarab), MR 7, offered twice:
  + 1:45-3:15 (Last name A-L) and
  + 3:45-5:15 (Last name M-Z)
* ASTRID (Pranjal Vachaspati), MR 7, 11:15-12:15
* SVDquartets (David Swofford), MR 7, 8:30-9:45

C. Multiple sequence alignment

* SATé/PASTA (Nam Nguyen and Mike Nute), MR 4, offered twice:
  + 1:45-3:15 (Last name M-Z)
  + 3:45-5:15 (Last name A-L)

D. Metagenomic data analysis

* TIPP (Nam Nguyen), MR 7, 10-11 AM

**ASTRAL**

Location: MR 7

Times: 1:45-3:15 (Last name M-Z) and 3:45-5:15 (Last name A-L)

Instructor: Siavash Mirarab ([smirarab@ucsd.edu](mailto:smirarab@ucsd.edu))

Description: ASTRAL is a method to estimate a species tree from a collection of input gene trees that is statistically consistent under the multi-species coalescent (MSC) model (i.e., in the presence of gene tree heterogeneity due to incomplete lineage sorting). ASTRAL uses a quartet-based method to estimate the species tree, and does not require rooted gene trees as input. In simulation studies, ASTRAL has been shown to be highly accurate, even on large datasets with tens of thousands of gene trees and thousands of taxa.

The tutorial will cover analysis of a biological dataset using ASTRAL, including estimations of species trees, calculations of branch support and tree length, and visualization. Parameter choices for different datasets properties, including incomplete gene trees, poorly resolved gene trees, and dataset size will be discussed as well. If time permits, advanced features of ASTRAL such as handling of multiple individuals per species will be introduced.

Participants should download and install ASTRAL onto their laptops in advance of the tutorial, and bring a dataset of estimated gene trees for analysis during the tutorial. Most datasets can be analyzed very quickly.

* Mirarab, Siavash, and Tandy Warnow. "ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes."Bioinformatics 31.12 (2015): i44-i52.
* Mirarab, Siavash, Rezwana Reaz, Md. Shamsuzzoha Bayzid, Theo Zimmermann, M Shel Swenson, and Tandy Warnow. “ASTRAL: Genome-Scale Coalescent-Based Species Tree.” Bioinformatics 30.17 (2014): i541–i548.
* Software at [https://github.com/](https://github.com/crsl4/PhyloNetworks)smirarab/ASTRAL , and is implemented in Java

**ASTRID**

Location: MR 7

Time: 11:15-12:15

Instructor: Pranjal Vachaspati ([pr@nj.al](mailto:pr@nj.al))

Description: ASTRID is a method to estimate a species tree from a collection of input gene trees that is statistically consistent under the multi-species coalescent (MSC) model (i.e., in the presence of gene tree heterogeneity due to incomplete lineage sorting). ASTRID is related to the NJst method (Liu and Yu, 2011), and uses a distance-based method to estimate the species tree. Like ASTRAL, ASTRID does not require rooted gene trees as input. In simulation studies, ASTRID has been shown to be highly accurate, even on large datasets with tens of thousands of gene trees and thousands of taxa. ASTRID is generally very fast, and faster in practice for very large datasets than ASTRAL. ASTRID computes the species tree topology, and can be followed by ASTRAL to compute branch support and branch lengths. The tutorial will cover analysis of a biological dataset, including estimations of species trees, calculations of branch support and tree length, and visualization.

Participants should download and install ASTRID onto their laptops in advance of the tutorial, and bring a dataset of estimated gene trees for analysis during the tutorial. Most datasets can be analyzed very quickly.

* Publication: Pranjal Vachaspati and Tandy Warnow. "ASTRID: Accurate Species TRees from Internode Distances." BMC Genomics 16. Suppl 10 (2015): S3.
* Tutorial: <https://github.com/pranjalv123/ASTRID/blob/master/tutorial.md>
* Software: <https://github.com/pranjalv123/ASTRID>

**Multiple Sequence Alignment**

Location: MR 4

Times: 1:45-3:15 (last name A-L) and 3:45-5:15 (last name M-Z)

Instructors: Nam-phuong Nguyen ([namphuon@illinois.edu](mailto:namphuon@illinois.edu)) and Mike Nute ([nute2@illinois.edu](mailto:nute2@illinois.edu))

Description: This tutorial will present SATé and PASTA (Practical Alignment using SATé and Transitivity). PASTA and SATé are methods for co-estimating alignments and maximum likelihood trees from unaligned sequences, but PASTA is faster, more accurate, and can analyze larger datasets than SATé.

This tutorial will cover how to use PASTA, which has been shown to align very large datasets (even one million sequences). PASTA runs on a GUI with nearly the same interface as SATé. The command line version provides some additional features, but the GUI version is easy to learn, and runs with little effort on MACs.

This tutorial will use the GUI version of PASTA. Participants should download and install PASTA GUI onto their laptops in advance of the tutorial.

PASTA

* Mirarab, S., Nguyen, N. Guo, S., Wang, L., Kim, J. and Warnow, T. “PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences.” Journal of Computational Biology 22(5):337-386, (2015). For an downloadable version of the paper, see <http://tandy.cs.illinois.edu/pasta-download.pdf>
* PASTA Software (implemented in Python) can be downloaded from <https://github.com/smirarab/pasta>
* Tutorial at <https://github.com/smirarab/pasta/blob/master/pasta-doc/pasta-tutorial.md>

**PhyloNet (phylogenetic network estimation from gene trees)**

Location: MR 4

Time: 11:15-12:15

Instructors: Yun Yu ([yy9@rice.edu](mailto:yy9@rice.edu)) and Luay Nakhleh ([Nakhleh@cs.rice.edu](mailto:Nakhleh@cs.rice.edu))

Description: PhyloNet is a software package for phylogenetic network inference and analysis. For inference, the software package has several utilities that take as input gene trees from multiple loci and infer phylogenetic networks. The input gene trees can be single point estimates for the loci, or collections of trees (e.g., from a bootstrap analysis or a sample of the posterior). The networks can be inferred using parsimony or statistical techniques (maximum likelihood or Bayesian). The inferred

networks are represented in the Rich Newick format, which can be readily visualized using Dendroscope. PhyloNet also has utilities for comparing phylogenetic network topologies, computing gene tree probabilities on phylogenetic

networks, and many more.

Participants should download and install PhyloNet on their laptops in advance of

the tutorial

PhyloNet:

* Y. Yu, R.M. Barnett, and L. Nakhleh (2013) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Systematic Biology 62(5): 738-751.
* Y. Yu, J.H. Degnan, and L. Nakhleh (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genetics 8(4): e1002660.
* Y. Yu, J. Dong, K.J. Liu, and L. Nakhleh (2014) Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences 111(46): 16448-16453.
* Y. Yu and L. Nakhleh (2015) A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16(Suppl 10): S10.
* Software can be downloaded from <http://bioinfo.cs.rice.edu/PhyloNet>

**PhyloNetworks (TICR/SNaQ)**

Location: MR 4

Times: 8:30-9:45 AM (Part 1), 10:00-11:00 (Part 2)

Instructors: Claudia Solis-Lemus ([solislemus@wisc.edu](mailto:solislemus@wisc.edu)) and Cécile Ané ([cecile.ane@wisc.edu](mailto:cecile.ane@wisc.edu))

PhyloNetworks Part I: overview of TICR pipeline; introduction to Julia and unrooted hybridization networks; network visualization and estimation from multi-locus data with SNaQ.

PhyloNetworks Part II: network uncertainty: number of reticulations, bootstrap; TICR test: adequacy of a tree with ILS only.

Description: Phylogenetic networks are necessary to represent the tree of life expanded by edges to represent events such as hybridizations, horizontal gene transfers or gene flow. This tutorial will present two software tools: TICR (Tree Incongruence Checking in R) and SNaQ (Species Networks applying Quartets). TICR is a nonparametric statistical test to determine whether a species tree (or a population tree) adequately explains the data, or whether reticulation needs to be invoked. SNaQ is a method for constructing phylogenetic networks (modeling both ILS and HGT) given a set of sequence alignments, and is part of the PhyloNetworks software package. The input to both SNaQ and TICR is a set of unrooted quartet frequencies, which can be obtained from a set of estimated unrooted gene trees, or from BUCKy to account for gene tree error. These methods can analyze datasets with many gene trees (e.g., several thousand) and up to (approximately) 30 species.

Participants should download and install R and Julia onto their laptops in advance of the tutorial. Please see [https://github.com/crsl4/PhyloNetworks.jl/wiki](https://webmail.illinois.edu/owa/redir.aspx?SURL=ITtHGRkWc_UTbPDnTZx6KzRlel3RB8hAWURZA_.&URL=https%3a%2f%2furldefense.proofpoint.com%2fv2%2furl%3fu%3dhttps-3A__github.com_crsl4_PhyloNetworks.jl_wiki%26d%3dCwMF-g%26c%3d8hUWFZcy2Z-Za5rBPlktOQ%26r%3dg6LIYhia8j0snAxVdlx6t9X8kB4ZOvadPvn8VWo9_oQ%26m%3d31ol54fAkA1fxqouJGkN12KejtBpuhcrZL7LQyc1O4o%26s%3d34hIU5vzzxfu-0UbCRtLLeXh4INGManc8QE2NrCsmrg%26e%3d" \t "_blank).

SNaQ

* C. Solis-Lemus C. and C. Ané 2015. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. arXiv:1509.06075 pp 1-32 and PLoS Genetics 12(3):e1005896

TICR

* N. Stenz, B. Larget, D.A. Baum and C. Ané 2015. Exploring tree-like and non-tree-like patterns using genome sequences: an example using the inbreeding plant species Arabidopsis thaliana (L.) Heynh. Systematic Biology 64:809-823.

**SVDquartets**

Location: MR 7

Time: 8:30 AM – 9:45 AM

Instructor: David Swofford ([david.swofford@duke.edu](mailto:david.swofford@duke.edu))

Description: SVDquartets is a program to compute a score based on singular value decomposition of a matrix of site pattern frequencies corresponding to a split on a phylogenetic tree. These quartet scores can be used to select the best-supported topology for quartets of taxa, which in turn can be used to infer the species phylogeny using quartet methods.

The input to SVDquartets is a set of multiple sequence alignments, one per locus. The estimation of the species tree is not based on gene trees, but only on the site patterns.

The SVDquartets program relies on [GSL](http://www.gnu.org/software/gsl/), which must be installed on your system prior to running SVDquartets. You may have to change the Makefile to give the correct path to the GSL libraries.

* Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent, *Bioinformatics*, 30(23): 3317-3324.
* J. Chifman and L. Kubatko, 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. J. Theoretical Biology, Vol. 374, 7 June 2015, pp. 35-47.
* Software can be downloaded from <http://www.stat.osu.edu/~lkubatko/software/SVDquartets/>

**TIPP (Metagenomic taxon identification and abundance profiling)**

Location: MR 7

Time: 10:00-11:00

Instructor: Nam-phuong Nguyen ([namphuon@eng.ucsd.edu](mailto:namphuon@eng.ucsd.edu))

Description: Abundance profiling (also called ‘phylogenetic profiling’) is a crucial step in understanding the diversity of a metagenomic sample, and one of the basic techniques used for this is taxonomic identification of metagenomic reads.

TIPP is a new marker-based taxon identification and abundance profiling method. TIPP combines SATé**-**enabled phylogenetic placement with statistical techniques to control the classification precision and recall, and results in improved abundance profiles. TIPP is highly accurate even in the presence of high indel errors and novel genomes, and matches or improves on previous approaches.

Please download and install TIPP before attending the tutorial; the README is available at <https://github.com/smirarab/sepp/blob/master/README.TIPP.md>

Publication:

* Nguyen, Nam , Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. “TIPP: Taxonomic identification and phylogenetic profiling.” Bioinformatics (2014). [doi:10.1093/bioinformatics/btu721](http://bioinformatics.oxfordjournals.org/content/30/24/3548.full.pdf).

Software:

* TIPP is part of the SEPP repository at <https://github.com/smirarab/sepp>