2018 Phylogenomics Software Symposium
Institut des Sciences de l'Evolution - Montpellier
August 17, 2018
Abstracts

Dominic J. Bennett

Title: supersmartR: Towards a modular pipeline for phylogenetic tree construction in the R language

Abstract: When constructing time-calibrated phylogenetic trees, researchers are confronted with ever-growing biological databases of various forms, and computational programs that are forever developing. To make use of these data and programs, the phylogeneticist is forced to develop their own workflows, often switching between different programming languages and often requiring user input at multiple stages. In addition, because biological data differs in quality and type, it can be difficult to develop a single, universal process that is able to reliably generate phylogenetic trees for different taxa and from different data sources. SUPERSMART (Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationships of Taxa) is one such platform that can construct phylogenetic trees from various data sources by first constructing higher-level taxon trees followed by lower-level taxon trees which are then combined. To make this pipeline more accessible, we here present supersmartR, a port of SUPERSMART in the widespread and popular R language. The SUPERSMART pipeline has been broken down into a series of interconnected, independent R packages with defined inputs and outputs, allowing users to either recreate the SUPERSMART pipeline entirely or adapt it to their own phylogenetic needs. In addition, several improvements to SUPERSMART are introduced, such as the automated retrieval of the most current sequence data. We present supersmartR as the basis for a more formal, portable, reproducible and modular approach to construction of future phylogenetic workflows that can be readily integrated with the latest software and data sources. (See our GitHub project for latest developments: https://github.com/AntonelliLab/supersmartR) This is unpublished work and is joint with R. Vos, H. Hettling, D. Silvestro and A. Antonelli.

Nicola DeMaio

Title: BADTRIP: reconstructing transmission from within-host SNPs

Abstract: I developed a Bayesian MCMC approach to infer species trees based on the model PoMo. This BEAST2 package allows inference of speciation events from genomic data without the need to specify genes or infer gene trees, but accounting for ILS. This package allows the use of an-

cient genomes, and accounts for sequencing errors. I showcase the applicability of this method by using it to infer transmission trees (a particular case of timed species trees) within the 2014 Ebola outbreak. Part of this work is under minor revision for PLOS Comp Biol, but is also posted on arXiv https://www.biorxiv.org/content/early/2017/11/08/213819. This is joint work with Daniel J Wilson, Nicole Stoesser, Colin Worby, Carolin Kosiol, and Dominik Schrempf.

Jan Kim

Title: High Throughput Multigene Phylogeny: Tools for Tackling the Plant and Fungal Trees of Life

Abstract: The Plant and Fungal Trees of Life (PAFTOL) project aims to generate multigene phylogenies representing all genera of the plant and fungal kingdoms. For plants, a set of 353 loci that (1) occur in a wide range of seed plants and (2) have only one single copy in most genomes have been selected. We use high throughput "next generation" sequencing, combined with target capture, to obtain sequencing data enriched for our target loci, and use a process inspired by the HybPiper pipeline to recover the sequences by applying de novo assembly for each locus individually.

PAFTOL data are continuously generated, requiring frequent running of analysis pipelines. In addition to newly generated data, we also include sequences from various existing sources, including GenBank and the Thousand Plants (1KP) project. We have developed a software framework (currently dubbed "paftools") to provide the extensibility, flexibility and reproducibility that enables us to meet these requirements using a small high performance computing (HPC) cluster. This framework is implemented in Python and makes substantial use of BioPython, as well as numerous bioinformatics tools. Some aspects of our framework are of more general interest, as they address requirements that apply to resources for computing multigene phylogenies in general.

PAFTOL targets the plant and fungal kingdoms in their entirety, and therefore divergence between a sequenced locus and its closest known ortholog may be quite considerable. Our framework supports using the BWA read mapper and the tblastn BLAST tool for identifying reads originating from our loci of interest, and parameterising these for identifying diverged sequences. We also noticed that de Bruijn graph based de novo assemblers such as SPAdes are not ideal for assembling individual loci, and therefore explore alternatives to de Bruijn graph based assembly.

This is joint work with V. Barber, A. Barker, L. Botigué, G. Brewer, S. Dodsworth, W.L. Eiserhardt, N. Epitawalage, F. Forest, E. Gaya, I. Leitch, O. Maurin, T. Niskanen, L. Pokorny, and W.J. Baker.

Carolin Kosiol

Title: IQ-TREE-POMO: Polymorphism-aware tree estimation

Abstract: The increased availability of sequenced genomes both from closely related species and from individuals of the same species, offers a great opportunity to study the speciation and evolutionary history of populations, provided we can properly model the process of sequence evolution using inter and intraspecific data together.

We have developed a new method called POlymorphisms-aware phylogenetic MOdel (PoMo). It extends any DNA substitution model and additionally accounts for polymorphisms in the present and in the ancestral population by expanding the state space to include polymorphic states. It is a selection-mutation model which separates the mutation process from the fixation process. PoMo naturally accounts for incomplete lineage sorting because ancestral populations can be in a polymorphic state. Our method can accurately and time-efficiently estimate the parameters describing evolutionary patterns for phylogenetic trees of any shape (species trees, population trees, or any combination of those). Here, we present the new implementation of our PoMo approach within the IQ-TREE software package (versions 1.6 and higher). IQ-TREE-POMO has several new features that allow to find the best-fit mutation model, to integrate rate heterogeneity among sites (e.g., gamma distribution) and the assessment of branch supports using fast bootstrapping.

The new genome-wide data set of seven baboon populations (genus Papio) present a unique opportunity to apply our method to a primate clade that involves more complex processes than those usually assumed by phylogenetic models. The history of Papio includes episodes of introgression or admixture among genetically distinct lineages. We will discuss the effect of this complex history on genome-wide phylogenetic inference with PoMo as well as other approaches. We will also present as new estimates of divergence times and mutation rates.

This is joint work with Dominik Schrempf, Minh Quang Bui, Nicola De Maio and Arndt von Haeseler.

Elise Lauterbur

Title: OrthoCapture: Facilitating Gene Capture Probe Creation for Non-Model Organisms

Abstract: Targeted gene or exon capture has become increasingly important as a method of allowing targeted gene sequencing for evolutionary studies that is cheaper, faster, and higher coverage. This is important for phylogenetic tests of selection and constraint in candidate genes, which are frequently conducted in taxa without a reference genome. However the lack of genome annotations, or even fully assembled genomes, presents difficulties for creating capture probes for non-model organism genomes. Capture probes are designed based on gene sequences from an existing genome, but gene capture efficiency decreases with sequence divergence between the probes and the input DNA. With too much

sequence divergence, capture may be impaired or even fail to provide usable data for evolutionary analysis. Probe design from multiple related references, allowing capture using pooled, divergent baits has been recommended in such cases. Here, I present OrthoCapture, a tool to mine unannotated genomic data from non-model species and/or their close relatives to allow the creation of complementary capture probes for non-model organisms using multiple genomic sources. This tool provides an integrated method of retrieving genic sequences from the unannotated genomic data, verifying their identity, and concatenating and extending them to the desired length. In tests on six annotated genomes, OrthoCapture reliably retrieves 93-100% of genes queried, compared to 42-69% of annotations existing for queried genes. Average percent identity between exons retrieved by OrthoCapture and their annotated versions is 98-100%. Use of OrthoCapture is via command-line interface on Unix systems, and requires the input of a gene sequence from an annotated reference genome such as Mus musculus or Homo sapiens and a fasta database from a target, unannotated genome (whole-genome shotgun contigs, for example). The output, sequence templates from the non-annotated genomic data, allows probe creation by any commercial company providing gene capture services.

This represents a new method to facilitate preparation for phylogenetic analyses of selection in non-model species, and as such is expected to increase gene capture efficiency in data sets comprised of multiple, related species for which annotated reference genomes do not exist. It is currently being applied in a study of the genomics of mammalian cyanide adaptation. This work is unpublished as of submission of this registration.

Uyen Mai

Title: TreeShrink: Fast and Accurate Detection of Outlier Long Branches in Collections of Phylogenetic Trees

Abstract: Sequence data used in reconstructing phylogenetic trees may include various sources of error. Typically errors are detected at the sequence level, but when missed, the erroneous sequences often appear as unexpectedly long branches in the inferred phylogeny. We propose an automatic method to detect such errors. We build a phylogeny including all the data then detect sequences that artificially inflate the tree diameter. We formulate an optimization problem, called the k-shrink problem, that seeks to find k leaves that could be removed to maximally reduce the tree diameter. We present an algorithm to find the exact solution for this problem in polynomial time. We then use several statistical tests to find outlier species that have an unexpectedly high impact on the tree diameter. These tests can use a single tree or a set of related gene trees and can also adjust to species-specific patterns of branch length. The resulting method is called TreeShrink. We test our method on six phylogenomic biological datasets and an HIV dataset and show that the method successfully detects and removes long branches. TreeShrink removes sequences more conservatively than rogue taxon removal and often reduces gene tree discordance more than

rogue taxon removal once the amount of filtering is controlled. This is joint work with Siavash Mirarab.

Siavash Mirarab

Title: Coalescent-based species tree estimation

Abstract: A major difficulty in species tree estimation from genome-wide data is the discordance of the species trees with gene trees. Discordance can have many causes, and a ubiquitous process that leads to gene trees differing from species trees is incomplete lineage sorting (ILS). The multi-species coalescent model, which extends the classic coalescent model to a phylogenetic context, is the main approach used in practice for modeling ILS. In this talk, we will quickly review the methods available to build a species tree despite ILS, and will then focus on a particular scalable approach called ASTRAL. We will show how ASTRAL works, why it is relevant to ILS, how its output should be interpreted, and how its input should be best prepared. ASTRAL is available at https://github.com/smirarab/ASTRAL.

Erin Molloy

Title: Scaling species tree estimation methods to large datasets using NJMerge

Abstract: In this talk, I will present a new divide-and-conquer approach for scaling phylogeny estimation methods to large datasets that does not require supertree estimation. Instead, the approach operates by (1) dividing the species set into disjoint (instead of overlapping) subsets, (2) constructing trees on the subsets, and (3) merging the subset trees using a distance matrix computed on the full set of species. For this merger step, I will present a new method, called NJMerge, which is a extension of the Neighbor Joining algorithm of Saitou and Nei. I will then show the results of an extensive simulation study demonstrating NJMerge's utility in scaling three popular species tree methods: ASTRAL, SVDquartets, and concatenation analysis using RAxML. NJMerge is available in open source form at https://github.com/ekmolloy/njmerge. This is joint work with Tandy Warnow.

Benoit Morel

Title: ParGenes, an integrated tool for model selection and maximum likelihood (ML) based phylogenetic inference on thousands of independent MSAs on clusters and supercomputers

Abstract: Inferring individual phylogenetic trees on a large set of multiple sequence alignments (MSAs) constitutes a recurring task in phylogenomics. For instance, this step is required to obtain per-gene phylogenies on genomic or transcriptomic datasets, which in turn serve as input for gene tree/species tree reconciliation methods. To the best of our knowledge, no easy-to-use efficient

parallel tool for this task exists. Instead, users typically write custom scripts and submit per-MSA jobs one by one to a cluster. This approach is inefficient with respect to man-hours (re-implementing the same pipeline, job queue monitoring), cpu-hours (due to suboptimal parallel efficiency) and time-to-solution (e.g., typical cluster setups only allow for simultaneously executing a limited number of jobs per user).

Here, we present ParGenes, an integrated tool for model selection and maximum likelihood (ML) based phylogenetic inference on thousands of independent MSAs on clusters and supercomputers. ParGenes schedules multiple independent parallel runs of ModelTest-NG (https://github.com/ddarriba/modeltest) and RAxML-NG (https://github.com/amkozlov/raxml-ng/) from within a single parallel job submission that can leverage the computational power of thousands of cores. It employs simple, yet efficient scheduling heuristic (initial job sorting and dynamic scheduling) to attain high parallel efficiency. This is joint work with Alexey M. Kozlov and Alexandros Stamatakis.

Mike Nute

Title: Benchmarking BALi-Phy

Abstract: The estimation of multiple sequence alignments of protein sequences is a basic step in many bioinformatics pipelines, including protein structure prediction, protein family identification, and phylogeny estimation. Statistical co-estimation of alignments and trees under stochastic models of sequence evolution has long been considered the most rigorous technique for estimating alignments and trees, but little is known about the accuracy of such methods on biological benchmarks. We report the results of an extensive study evaluating the most popular protein alignment methods as well as the statistical co-estimation method BAli-Phy on 1192 protein data sets from established benchmarks as well as on 120 simulated data sets. Our study (which used more than 230 CPU years for the BAli-Phy analyses alone) shows that BAli-Phy is dramatically more accurate than the other alignment methods on the simulated data sets, but is among the least accurate on the biological benchmarks. There are several potential causes for this discordance, including model misspecification, errors in the reference alignments, and conflicts between structural alignment and evolutionary alignments; future research is needed to understand the most likely explanation for our observations. This is joint work with Tandy Warnow and Ehsan Saleh.

Sanna Ollson

Title: Demographic history and molecular adaptation of the Pinus halepensis-brutia complex

Abstract:

Comparative genomics is a powerful approach to understand the evolutionary history of species complexes. We investigate how demographic histories

and ecological preferences have affected genetic divergence and adaptation in a species complex composed of the Aleppo pine (*Pinus halepensis*) distributed widely across the Mediterranean Basin, and the Turkish or Calabrian pine (*Pinus brutia*) present primarily in Turkey and far East Greece. Although the two species form natural hybrids in Turkey where they are sympatric, they have diverged quite distinctly for several key adaptive traits. It is yet unclear whether this divergence is due to distinct demographic histories or to differences in ecological behavior, or to both.

In order to gain more insights on the evolutionary history of this complex species, we analyzed and compared two molecular datasets: i) 186 candidate genes sequenced in 47 individuals of *P. halepensis* and 12 individuals of *P. brutia*; ii) seven transcriptomes of *P. halepensis* and four transcriptomes of *P. brutia*. In both cases the sampling design aimed at covering the entire distribution of the two species. By combining demographic models and selection tests our main objectives were the following: i) to infer the speciation events and demographic histories; ii) to identify genes involved in adaptation and speciation.

The two molecular datasets analyzed with different approaches separate well the two species. Isolation with Migration models (IMa2) suggest that the divergence of these two species happened 24.3-50.1 mya during global climate cooling, and that eastern Aleppo pine populations have similar levels of genetic diversity than those in Turkish pine, while genetic diversity seems to have been depleted when this species colonized its western range. We investigated possible ancestral admixture events and gene flow, using SNAPP, smc++ and f_4-statistics. Finally, we used simulations based on the inferred demography to detect adaptive genes or genes affecting speciation.

This is joint work with Zaida Lorenzo, Francesca Bagnoli, Sara Pinosio, Mario Zabal-Aguirre, G. G. Vendramin, Santiago C. González-Martínez, and Delphine Grivet.

Lisa Pokorny

Title: Building the Plant Tree of Life: A Proof of Concept for Flowering Plant Families

Abstract: Evolutionary trees are powerful tools for prediction, species discovery, monitoring and conservation. Through comparative analysis of DNA sequence data, the backbone of the plant tree of life is relatively well understood, and many subcomponents have been studied in great detail. However, DNA data are still lacking for numerous genera and the vast majority of species of plants, preventing their accurate placement within this evolutionary framework and hindering downstream science. To better understand how the world's plants are related to each other and how they have evolved, we have initiated a project at the Royal Botanic Gardens, Kew to complete the Plant and Fungal Trees of Life (PAFTOL). We will utilise our collections and work with our collaborative networks to produce extensive new DNA sequence data (whole plastid genomes, 353 single-to-low-copy nuclear loci) for a representative species from

each genus of plants using high-throughput sequencing technologies. In here we present results at the family level across flowering plants and test their current classification. This comprehensive investigation of phylogenetic relationships will be a rich resource enabling the discovery and study of evolutionary patterns in plants, and will provide a unifying framework for comparative research. The project is an essential step towards the compilation of genomic data for all known species. This is joint work with Steven Dodsworth, Abigail Barker, Laura R. Botigué, Grace Brewer, Wolf L.T. Eiserhardt, Niroshini Epitawalage, Ester Gaya, Jan T. Kim, Ilia J. Leitch, Olivier Maurin, Joe Parker, Felix Forest, and William J. Baker.

Eric Tannier

Title: Treerecs with Seaview: gene tree inference from alignment to reconciliation, with a graphical interface

Abstract: Gene tree species tree reconciliation methods can root gene trees and correct their uncertainties coming from phylogenetic methods based uniquely on multiple sequence alignments. Currently in reconciliation software there is a lack of either performance on certain functions, usability for biologists or integration to standard phylogenetic methods. I will present Treerecs, a phylogenetic software based on reconciliation with duplications and losses. It is simple to install and to use, fast, versatile, with a graphic output, and can be used along with multiple alignment methods like PLL and Seaview.

Tandy Warnow

Title: Scaling statistical multiple sequence alignment

Abstract: Multiple sequence alignment (MSA) is one of the most basic bioinformatics steps, in which a set of molecular sequences (i.e., DNA, RNA, or amino acid sequences) are arranged inside a matrix to identify corresponding positions. MSA calculation is a fundamental first step in many biological analyses. Because of its broad applicability and importance, many MSA methods have been developed and are in wide use today. Unfortunately, many real world biological datasets have features (large size and fragmentary sequences, for example) that make accurate MSA calculation very difficult. Because poorly estimated alignments result in errors in downstream biological analyses, new MSA techniques are needed that can produce accurate alignments on difficult datasets. One of these very promising techniques is BAli-Phy (Redelings and Suchard), a Bayesian statistical method for co-estimation of alignments and trees, yet BAli-Phy is extremely computationally intensive and generally will not converge well on more than about 50 sequences. In this talk I will present SATé, PASTA, and UPP, three methods that have been developed to enable ultra-largescale MSA estimation with high accuracy. These methods use divide-and-conquer strategies that work with selected MSA methods (such as MAFFT, Probcons, PROMALS, or BAli-Phy), and enable these methods to run on large datasets.

I will present results of an extensive simulation study showing that PASTA and UPP achieve high accuracy, even on datasets with 1,000,000 divergent sequences. I will also show that using PASTA and UPP enables BAli-Phy to be used to advantage on ultra-large datasets. PASTA (a direct improvement of SATé by Liu et al., Science 2009) and UPP are available in open source form at GitHub at https://github.com/smirarab/pasta and https://github.com/smirarab/sepp respectively.