Workshop on Algorithms and Software in Phylogenetics
11–13 September 2023

Abstracts

Talk 1

MAST: Phylogenetic inference with mixtures across sites and trees

Bui Quang Minh, Australian National University, Australia

Abstract

Hundreds or thousands of loci are now routinely used in modern phylogenomic studies. Concatenation approaches to tree inference assume that there is a single topology for the entire dataset, but different loci may have different evolutionary histories due to incomplete lineage sorting, introgression, and/or horizontal gene transfer; even single loci may not be treelike due to recombination. To overcome this shortcoming, we introduce the mixture across sites and trees (MAST) model, which uses a mixture of bifurcating trees to represent multiple histories in a single concatenated alignment. We implemented the MAST model in a maximum-likelihood framework in the popular phylogenetic software, IQ-TREE. Simulations show that we can accurately recover the true model parameters. We applied the MAST model to multiple primate datasets and found that it can recover the signal of incomplete lineage sorting in the Great Apes, as well as the asymmetry in minor trees caused by introgression among several macaque species. These results suggest that the MAST model is able to analyse a concatenated alignment using maximum likelihood, while avoiding some of the biases that come with assuming there is only a single tree. The MAST model can therefore offer unique biological insights when applied to datasets with multiple evolutionary histories.

Joint work with Thomas Wong, Caitlin Cherryh, Allen Rodrigo, Matthew Hahn, and Rob Lanfear

Talk 2

Non-parametric quartet-based estimation species tree inference

Siavash Mir Arabbaygi, University of California at San Diego, USA

Abstract

In this talk, I review recent advances in the ASTRAL family of methods for species tree inference from gene trees. Starting from the traditional setting, maximizing the number of quartets shared between the species tree and the gene tree, I will cover several recent advances. The first focus will be on a new family of methods designed for dealing with gene tree estimation error by weighting quartets according to quantities that can be measured from the gene trees. The second direction is the incorporation of duplication and loss events, including positive theoretical results and new algorithmic ideas. I end by pointing some directions of (near) future research related to quartet-based tree estimation for species tree inference.

Talk 3

Leaping through tree space: continuous inference for phylogenies

Neil Scheidwasser, University of Copenhagen, Denmark

Abstract

Phylogenetic inference, the task of reconstructing the evolutionary history of taxa from input data, is fundamental to the analysis of evolutionary processes in various domains, from linguistics to ecology and epidemiology. Over the past decades, a substantial body of research has been dedicated to developing computationally efficient methods for phylogenetic inference. These traditionally rely on heuristic approaches that use tree rearrangements such as nearest-neighbour interchange, subtree-prune and regraft, or tree bisection and reconnection operations. However, these operations are prone to convergence to poor local minima due to the inability of these heuristic operations to provide sufficient directions of change towards a better tree. To address these challenges, using a continuous representation of phylogenetic tree topology, we propose a new objective function derived from the balanced minimum evolution criterion, which allows for large topological changes using adaptive gradient descent methods. The approach achieves competitive performance with respect to state-of-the-art frameworks with an overall complexity comparable to neighbour-joining.

Talk 4 (Keynote)

Phylogenetic reconstructions based on chromosomal rearrangements as an alternative at the intrinsic complexity of protein evolution

Alessandra Carbone, Sorbonne Université-CNRS, France

Abstract

I will introduce some observations unraveling the “genome-wide” complexity of protein evolution and the interest of using chromosomal rearrangements for the reconstruction of phylogenetic diversity.

Multiple genome comparison, in contrast to pairwise comparison, allows to reconstruct accurate phylogenetic relationships between species from gene order. PhyChro performance is evaluated on two data sets of 13 vertebrates and 21 yeast genomes by using up to 130,000 and 179,000 breakpoints, respectively, a scale of genomic markers that has been out of reach for a long time. PhyChro reconstructs very accurate tree topologies even at known problematic branching positions. PhyChro is very fast, reconstructing the vertebrate and yeast phylogenies in <15 min.

Talk 5

Inferring horizontal gene transfers on a species phylogeny using character traits

Manuel Lafond, Université de Sherbrooke, Canada

Abstract

The reconstruction of phylogenetic trees or networks if often done using gene and protein sequences. However, there are several contexts in which character traits can provide alternate and possibly better phylogenetic signals. This includes, for instance, gene expression profiles or the presence/absence of protein domains, which may hide evolutionary relationships even in divergent sequences. In this talk, I will discuss the applicability of character traits for the prediction of horizontal gene transfers along a species tree. In the absence of convergent evolution, the appearance of a trait on two independent clades of a tree may indicate a transfer between the two clades. This observation leads us to study the problem of explaining incompatible characters on a tree through the insertion of transfer arcs, resulting in the inference of tree-based networks. I will discuss past and novel ways to extend established models on trees to networks, including perfect phylogenies and Dollo parsimony. I will then focus on tree-based network models and discuss several algorithmic results and open problems.

Talk 6

Non-binary Tree Reconciliation with Endosymbiotic Gene Transfer

Mathieu Gascon, Université de Montréal, Canada

Abstract

Reconciling a non-binary gene tree with a binary species tree can be done efficiently in the absence of horizontal gene transfers, but becomes NP-hard in the presence of gene transfers. In this presentation, the focus will be on the special case of *endosymbiotic gene transfers* (EGT), i.e. transfers between the mitochondrial and nuclear genome of the same species. More precisely, given a multifurcated (non-binary) gene tree with leaves labeled 0 or 1 depending on whether the corresponding genes belong to the mitochondrial or nuclear genome of the corresponding species, we investigate the problem of inferring a most parsimonious Duplication, Loss and EGT (DLE) Reconciliation of any binary refinement of the tree. This can be done through a two-steps heuristic: ignoring the 0-1 labeling of leaves, output a binary resolution minimizing the Duplication and Loss (DL) Reconciliation and then, for such resolution, assign a known number of 0s and 1s to the leaves in a way minimizing EGT events. While the first step corresponds to the well studied non-binary DL-Reconciliation problem, the complexity of the label assignment problem corresponding to the second step is unknown. It turns out that it is NP-complete, even when the tree is restricted to a single polytomy and, surprisingly, even if transfers can occur in only one direction. I will present various heuristics and algorithmic results for this problem.

Talk 7

Ploidy profiles

Katharina T. Huber, University of East Anglia, UK

Abstract

Polyploidisation is an evolutionary phenomenon by which an organism acquires multiple copies of its complete set of chromosomes. This number is sometimes called the *ploidy level* of a species. The main strategy for reconstructing the evolutionary past of a dataset that has undergone polyploidization is to first construct a so called multiple-labelled tree from the dataset and to then somehow derive a phylogenetic network from it that explains the dataset's *ploidy profil*e (i.e. the ploidy levels of the species that make up the dataset). The following question therefore arises: How much can be said about that past if such a tree is not readily available? In this talk, we first formalize this question and then report on recent results that shed some light into it.

Talk 8

SplitsTree - phylogenetic trees, rooted networks and unrooted networks all in one

Daniel Huson, University of Tübingen, Germany

Abstract

The original purpose of the SplitsTree program (H.,1998, H. and Bryant, 2006) was to make splits-based methods available to biologists, in particular first the split-decomposition method (Bandelt and Dress, 1992) and then later the neighbor-net method (Bryant and Moulton, 2004). We have developed a new release SplitsTree CE that aims at providing a much wider range of algorithms for the computation, analysis and visualization of unrooted and rooted trees and networks. A key design idea is to treat trees and rooted networks alike in the program. We are currently working on pushing this idea further so as to treat trees, rooted networks and unrooted (split) networks uniformly. This involves further extending the extended Newick format, "it is time for a standard representation of unrooted phylogenetic networks" (not just rooted ones).

Talk 9

Reconstructing semi-directed network topology under Markov models

Mark Jones, TU Delft, Netherlands

Abstract

Markov models for DNA sequence evolution have recently been introduced for phylogenetic networks. Under these models, (generic) identifiability of the semi-directed network topology has been shown for certain network classes, using techniques from algebraic geometry. Such results indicate, roughly speaking, that the network topology is characterized by the distributions of characters one may (reasonably) expect to see in aligned DNA sequence data.

In this presentation, I will discuss recent developments in combining algebraic and combinatorial techniques, to extend identifiability results to larger network classes. I will also talk about the challenges involved in translating generic identifiability results into constructive algorithms.

Talk 10

A Scalable Method for Inferring Phylogenetic Networks from Trees

Louxin Zhang, National University of Singapore, Singapore

Abstract

The reconstruction of phylogenetic networks is an important but challenging problem in phylogenetics and genome evolution, as the space of phylogenetic networks is vast and cannot be sampled well. One approach to the problem is to solve the minimum phylogenetic network problem, in which phylogenetic trees are first inferred, then the smallest phylogenetic network that displays all the trees is computed. The approach takes advantage of the fact that the theory of phylogenetic trees is mature and there are excellent tools available for inferring phylogenetic trees
from a large number of biomolecular sequences. A tree-child network is a phylogenetic network satisfying the condition that every non-leaf node has at least one child that is of indegree one. Here, we develop a new method that infers the minimum tree-child network by aligning lineage taxon strings in the phylogenetic trees. This algorithmic innovation enables us to get around the limitations of the existing programs for phylogenetic network inference. Our new program, named ALTS, is fast enough to infer a tree-child network with a large number of reticulations for a set of up to 50 phylogenetic trees with 50 taxa that have only trivial common clusters in about a quarter of an hour on average.

Talk 11 (keynote)

Beyond Simple Trees: Trees of Clusters (Clonal Trees) and Clusters of Trees (Phylogenetic Networks)

Luay Nakhleh, Rice University, USA

Abstract

In this talk, I will describe the multispecies network coalescent (MSNC) model, which extends the MSC model so that it operates within the branches of a phylogenetic network. This extended model naturally allows for modeling vertical and horizontal evolutionary processes acting within and across species boundaries. In particular, it simultaneously accounts for gene tree incongruence across loci due to both hybridization and incomplete lineage sorting. I will then briefly describe a host of methods that we have developed for phylogenetic network inference under the MSNC. The methods differ by the mathematical criterion they employ, the data they take as input, as well as the information they infer. I will also discuss practical issues facing phylogenetic network inference in practice, including the challenges of inferring networks in the presence of allopolyploidy.

Talk 12

Transfer Bootstrap, Robustness of Branch Supports with Respect to Taxon Sampling

Olivier Gascuel, Institut de Systématique, Evolution, Biodiversité (ISYEB UMR7205 – CNRS, Muséum National d'Histoire Naturelle, SU, EPHE, UA), France

Abstract

The bootstrap method is based on resampling alignments and reestimating trees. Felsenstein's bootstrap proportions (FBP; Felsenstein 1985) is the most common approach to assess the reliability and robustness of sequence-based phylogenies. However, when increasing taxon-sampling (i.e., the number of sequences) to hundreds or thousands of taxa, FBP will tend to return low supports for deep branches. The Transfer Bootstrap Expectation (TBE; Lemoine et al. 2018) has been recently suggested as an alternative to FBP. TBE is measured using a continuous transfer index in [0,1] for each bootstrap tree, instead of the {0,1} index used in FBP to measure the presence/absence of the branch of interest. TBE has been shown to yield higher and more informative supports, without inducing falsely supported branches. Nonetheless, it has been argued that TBE must be used with care due to sampling issues, especially in datasets with high number of closely related taxa. In this study, we conduct multiple experiments by varying taxon sampling and comparing FBP and TBE support values on different phylogenetic depth, using both simulated and empirical datasets. Our results show that the main critic of TBE stands in extreme cases, but that TBE is still very robust to taxon sampling in most simulated and empirical cases, while FBP is inescapably negatively impacted by high taxon sampling. We suggest guidelines and good practices in TBE computing and interpretation.
Joint work with: Paul Zaharias and Frédéric Lemoine.

Talk13

Pre-process before computing distances between phylogenetic trees!

Simone Linz, School of Computer Science, University of Auckland, New Zealand

Abstract

In phylogenetic analyses, it is not uncommon to obtain different phylogenetic trees for the same data set. For example this can be due to the use of different tree inference methods or noise in a date set. In both cases, the resulting tree incongruences motivate the use of distance measures to quantify the dissimilarities between two trees. Popular distances between phylogenetic trees are, for example, the tree bisection and reconnection (TBR) distance between unrooted trees and the rooted subtree prune and regraft (rSPR) distance between rooted trees. Although these distances are NP-hard to compute, they are also fixed-parameter tractable. In this talk, we describe a series of results on the size of the TBR and SPR kernel, i.e. the size of two phylogenetic trees after pre-processing that can be applied prior to any exact or heuristic calculation of the TBR and rSPR distance.

This is joint work with Steven Kelk and Ruben Meuwese (Maastricht University).

Talk 14

Computing a consensus for 1-nested phylogenetic networks

Vincent Moulton, University of East Anglia, UK

Abstract

An important and well-studied problem in phylogenetics is to compute a consensus tree for a collection of rooted phylogenetic trees, all whose leaf-sets are some set *X* of species. Recently, it has become of interest to develop consensus approaches that can also be applied to collections of phylogenetic networks. In this talk, we present an algorithm for computing a consensus for a collection of so-called 1-nested phylogenetic networks. Our approach builds on a previous result by Roselló et al. that describes an encoding for any 1-nested phylogenetic network with leaf-set *X* in terms of a collection of ordered pairs of subsets of *X*. More specifically, we characterize those collections of ordered pairs that arise as the encoding of some 1-nested phylogenetic network, and then use this characterization to help compute a consensus network for a collection of *t* 1-nested networks in $O(t|X|^2+|X|^3)$ time. Applying our algorithm to a collection of phylogenetic trees yields the well-known majority rule consensus tree. Our approach leads to some new directions for future work which we shall also briefly mention.

Talk 15

The structure theorem for rooted binary phylogenetic networks: theory, applications, and challenges

Momoko Hayamizu, Waseda University, Japan

Abstract

Phylogenetic networks offer a versatile model for representing complex evolutionary histories and non-tree-like data. Tree-based networks, a prominent subclass of phylogenetic networks, encompass various well-known subclasses and possess both high descriptive power and mathematical tractability. In this talk, I will describe the structure theorem for rooted binary phylogenetic networks and explain how it furnishes linear-time or linear-delay algorithms for a range of computational problems, including counting, listing, optimising, and ranking spanning trees called subdivision trees (a.k.a. support trees) as well as quantifying the deviation from being tree-based networks. Furthermore, I will discuss the challenges and future research directions in translating these theoretical advances into practical software development, emphasising the potential for collaboration and innovation within the field. (Part of this work, specifically the algorithm for ranking spanning trees, is joint work with Kazuhisa Makino.)

Talk 16

Computing a Consensus Phylogeny via Leaf Removal

Lusheng Wang, City University of Hong Kong, China

Given a set of phylogenetic trees with the same leaf-label set *X*, we wish to remove some leaves from the trees so that there is a tree *T* with leaf-label set *X* displaying all the resulting trees. Note that the labels of leaves removed from one input tree may be different from those of leaves removed from another input tree. One objective is to minimize the total number of leaves removed from the trees, whereas the other is to minimize the maximum number of leaves removed from an input tree. Chauve et al. refer to the problem with the first (respectively, second) objective as *AST-LR* (respectively, *AST-LR-d*), and they show that both problems are NP-hard, where NP is the class of problems solvable in non-deterministic polynomial time. They further present algorithms for the parameterized versions of both problems. In this article, we point out that their algorithm for the parameterized version of AST-LR is flawed and present a new algorithm. Since neither Chauve et al.'s algorithm for AST-LR-*d* nor our new algorithm for AST-LR looks practical, we further design integer-linear programming (ILP for short) models for AST-LR and AST-LR-*d*, and we discuss speedup issues when using popular ILP solvers (say, GUROBI or CPLEX) to solve the models. Our experimental results show that our ILP approach is quite efficient.

Talk 17

The k-Robinson-Foulds Measures for Labeled Trees

Elahe Khayatian, National University of Singapore, Singapore

Abstract

Investigating the mutational history of tumor cells is important for understanding the underlying mechanisms of cancer and its evolution. Now that the evolution of tumor cells is modeled using labeled trees, researchers are motivated to propose different measures for the comparison of mutation trees and other labeled trees. While the Robinson-Foulds distance is widely used for the comparison of phylogenetic trees, it has weaknesses when it is applied to labeled trees. There are of course some measures for labeled trees; however, most of them only allow for the comparison of each pair of trees in which nodes are labeled by disjoint sets. Here, *k*-Robinson-Foulds dissimilarity measures are introduced for labeled trees comparison. These measures are able to compare trees in which two different nodes are labeled by not necessarily disjoint sets or even multisets.

This is a joint work with Gabriel Valiente and Zhang Louxin.

Talk 18

Algorithms for Inferring the Ancestry of an Admixed Individual

Yufeng Wu, University of Connecticut, USA

Abstract

Suppose a person collected his/her own genome and wants to learn something about his/her ancestry. One popular question about ancestry concerns population admixture. The genome of an individual from an admixed population consists of segments originated from different ancestral populations. Most existing ancestry inference approaches can provide an estimate of the percentage (called admixture proportion) of the genome that originated from a specific ancestral population. We further suppose someone wants to estimate admixture proportions of recent ancestors from his/her own genome. For example, let us say an individual has 50% European and 50% Asian ancestry. This individual’s parents can also be 50-50 European/Asian ancestry, but it is also possible one is 100% European and the other is 100% Asian. More generally, can one learn more about the ancestry of recent ancestors from his/her own genome?

In this talk, I will present research results on ancestry inference for estimating admixture proportions of recent ancestors from an extant genome. I will first briefly describe PedMix, one of the first methods for such inference. A major downside of PedMix is that it can only estimate admixture proportions for parents and grandparents, and cannot scale for more distantly related ancestors. I will then focus on PedMix2, a general approach that can estimate admixture proportions of all recent ancestors of this individual. To the best of our knowledge, there is no other method that can practically infer ancestors beyond grandparents from an extant individual’s genome. I will present results on both simulated and real data to show that PedMix2 performs well in ancestry inference.

Talk 19

Planar Phylogenetic Networks

Taoyang Wu, University of East Anglia, UK

Abstract

A rooted phylogenetic network is a directed acyclic graph with a single root, whose sinks correspond to a set of species. As such networks are useful for representing the evolution of species that have undergone reticulate evolution, there has been great interest in developing the theory behind and algorithms for constructing them. However, unlike evolutionary trees, these networks can be highly non-planar, which can make them difficult to visualise and interpret. In this talk I will discuss properties of planar rooted phylogenetic networks and algorithms for deciding whether or not rooted networks have certain special planarity properties, including three natural subclasses of planar rooted phylogenetic networks: namely outer, terminal, and upward planar networks. This is based on joint work with Vincent Moulton.

Talk 20

Inferring phylogenetic networks via minimizing deep coalescence cost

Paweł Górecki, University of Warsaw, Institute of Informatics

Abstract

Despite the computational challenges, our recent research demonstrates that computing the similarity between a phylogenetic network and a gene tree can be efficiently achieved in practice. This is accomplished by minimizing the deep coalescence cost between the gene tree and a tree displayed by the network, using a new approach that resolves conflicts in reticulation scenarios. In this presentation, I will provide a detailed explanation of how this approach can be employed to solve a classical problem - inferring a phylogenetic network from a set of gene trees by minimizing the deep coalescence cost. To illustrate the effectiveness of this method, I will present both simulated and empirical examples, including optimizations that significantly improve network inference speed. Finally, I will highlight the difficulties associated with interpreting the resulting networks.

Based on joint work with Natalia Rutecka, Agnieszka Mykowiecka and Dawid Dąbkowski