Algorithmic Advances and Implementation Challenges: Developing Practical Tools for Phylogenetic Inference
Poster Session Abstracts
Tuesday, November 19, 2024

Branch Length Transforms using Optimal Tree Metric Matching
Shayesteh Arasti, University of California San Diego

The abundant discordance between evolutionary relationships across the genome has rekindled interest in ways of comparing and averaging trees on a shared leaf set. However, most attempts at reconciling trees have focused on tree topology, producing metrics for comparing topologies and methods for computing median tree topologies. Using branch lengths, however, has been more elusive, due to several challenges. Species tree branch lengths can be measured in many units, often different from gene trees. Moreover, rates of evolution change across the genome, the species tree, and specific branches of gene trees. These factors compound the stochasticity of coalescence times. Thus, branch lengths are highly heterogeneous across both the genome and the tree. For many downstream applications in phylogenomic analyses, branch lengths are as important as the topology, and yet, existing tools to compare and combine weighted trees are limited. In this project, we make progress on the question of mapping one tree to another, incorporating both topology and branch length. We define a series of computational problems to formalize finding the best transformation of one tree to another while maintaining its topology and other constraints. We show that all these problems can be solved in quadratic time and memory using a linear algebraic formulation coupled with dynamic programming preprocessing. Our formulations lead to convex optimization problems, with efficient and theoretically optimal solutions.

Estimating Phylogenetic Models Using Incomplete U-Statistics
Max Bacharach, University of California Riverside

This poster will present a new method of using phylogenetic invariants to perform inference. In particular, we consider a novel method (introduced in 2024 by Sturma, Drton, and Leung) which utilizes incomplete U-statistics for performing hypothesis testing of models defined by algebraic and semi-algebraic constraints. In this poster, I will present our results and some novel insights gained from implementing and applying this method to a variety of models from phylogenetics, including both coalescent-based models and group-based substitution models. This is based on joint work with David Barnhill, Marina Garrote-López, Elizabeth Gross, and Bryson Kagy, John Rhodes, and Joy Zhang.

Larch: Efficient manipulation of collections of trees and exploration of tree space
Mary Barker, Fred Hutchinson Cancer Center, Howard Hughes Medical Institute

In the past decades, increased data availability and sequencing as well as the global pandemic have contributed to a dramatic increase in the amount of data available for viral infections. This has corresponded to an increased need for fast and efficient tools for phylogenetic inference, particularly in densely sampled regimes where runtime issues, weak signal, and low bootstrap support prove obstacles for Bayesian or likelihood methods. We developed a graph-based framework that represents large families of optimal trees for densely sampled data. The C++ implementation called Larch interfaces with an SPR-based search software that uses the distribution of known trees to propel the search further.

Instability in Phylogenetic Trees after Taxon Addition
Lena Collienne, Fred Hutch Cancer Center

Online phylogenetic inference methods add sequentially arriving sequences to an inferred phylogeny without the need to recompute the entire tree from scratch. Some online method implementations exist already, but there remains concern that additional sequences may change the topological relationship among the original set of taxa. We call such a change in tree topology a lack of stability for the inferred tree.
We analyze the stability of single taxon addition in a Maximum Likelihood framework across 1,000 empirical datasets. We find that instability occurs in almost 90% of our examples, although observed topological differences do not always reach significance under the AU-test. Changes in tree topology after addition of a taxon rarely occur close to its attachment location, and are more frequently observed in more distant tree locations carrying low bootstrap support. To investigate whether instability is predictable, we hypothesize sources of instability and design summary statistics addressing these hypotheses. Using these summary statistics as input features for machine learning under random forests, we are able to predict instability and can identify the most influential features. In summary, it does not appear that a strict insertion-only online inference method will deliver globally optimal trees, although relaxing insertion strictness by allowing for a small number of final tree rearrangements or accepting slightly suboptimal solutions appears feasible.

State-dependent diversification with fossils: a case study with Canidae
Bruno do Rosario Petrucci, Iowa State University

Questions regarding the effects of traits on diversification rates are ubiquitous in evolutionary biology, with the most popular model category for such analyses being state-dependent speciation and extinction (SSE) models. Much work has been done on their strengths and limitations, but very few simulation and empirical studies have included fossils in their analyses. Simulations have showed that fossils can improve the extinction estimate accuracy of such models, but not solve issues regarding their false positive rates. For this presentation, we will explore a case study of state-dependent diversification using Canidae as a model group. We will use a total-evidence analyses with a fossilized birth-death range (FBDR) model to obtain an updated complete tree of fossil and extant canids. We will then apply BiSSE and HiSSE models to that tree to investigate the effect of diet on canid diversification. We can then explore these results with the help of past work exploring the limitations of SSE models. This project will not only allow us to gain a deeper insight on canid diversification dynamics, but also increase our understanding of the limitations of applying SSE models to paleontological data.

Biological causes and impacts of the widespread lack of tree convergence in phylodynamic inference
Jiansi Gao, Fred Hutchinson Cancer Center

Phylodynamic analysis of genomic datasets has been instrumental in elucidating the evolutionary and transmission dynamics of pathogens. Such analyses often involve averaging over a distribution of phylogenetic trees (inferred previously or simultaneously)---a computationally demanding step---while comprehensive characterization of the tree space and assessment of its impacts on phylodynamic inferences remain limited. Here, by carefully re-running and analyzing 15 classic large phylodynamic analyses, we show that:1) the tree space is complex, and a lack of topological convergence is widespread; 2) difficulties in tree space exploration may frequently stem from a small subset of viral sequences that might have been subject to biological processes---such as recombination, hypermutation, or sequencing error---that violate common assumptions in phylogenetic modeling; 3) tree-wise phylodynamic inferences appear to be minimally affected by poor exploration of tree space, whereas impacts on clade-specific estimates, such as the origin time and introduction history, are more pronounced, and; 4) the sequential and joint inference approaches systematically lead to distinct trees, likely due to the serially and densely sampled viral sequences with limited genetic diversity. We introduce new MCMC diagnostics to uncover details underlying the identified convergence issues. Our findings highlight potential new directions for developing tree-rearrangement mechanisms. Outputs from our comprehensive analyses (over one trillion MCMC samples) will facilitate the further development of scalable phylodynamic methods, providing a robust training and benchmarking dataset for tree-search algorithms.

Improved robustness to gene tree incompleteness, estimation errors, and systematic homology errors with weighted TREE-QMC
Yunheng Han, University of Maryland, College Park

Summary methods are widely used to reconstruct species trees from gene trees while accounting for incomplete lineage sorting; however, it is increasingly recognized that their accuracy can be negatively impacted by incomplete and/or error-ridden gene trees. To address the latter, Zhang and Mirarab (2022) leverage gene tree branch lengths and support values to weight quartets within the popular summary method ASTRAL. Although these quartet weighting schemes improved the robustness of ASTRAL to gene tree estimation error, implementing the weighting schemes presented computational challenges, resulting in the authors abandoning ASTRAL's original search algorithm (i.e., computing an exact solution within a constrained search space) in favor of search heuristics (i.e., hill climbing with nearest neighbor interchange moves from a starting tree constructed via randomized taxon addition). Here, we show that these quartet weighting schemes can be leveraged within the Quartet Max Cut framework of Snir and Rao (2010), with only a small increase in time complexity compared to the unweighted algorithm, which behaves more like a constant factor in our simulation study. Moreover, our new algorithm, implemented within the TREE-QMC software, was highly competitive with weighted ASTRAL, even outperforming it in terms of species tree accuracy on some challenging model conditions, such as large numbers of taxa. In comparing unweighted and weighted summary methods on two avian data sets, we found that weighting quartets by gene tree branch lengths improves their robustness to systematic homology errors and is as effective as removing the impacted taxa from individual gene trees or removing the impacted gene trees entirely. Lastly, our study revealed that TREE-QMC is highly robust to high rates of missing data and is promising as a supertree method.

Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments
Niels Holtgrefe, Delft University of Technology

Phylogenetic networks model the evolutionary history of taxa while allowing for reticulate events such as hybridization and horizontal gene transfer. As is the case for phylogenetic trees, it is often not possible to infer the root location of such a network directly from biological data for several evolutionary models. Hence, we consider semi-directed (phylogenetic) networks: partially directed graphs without a root in which the directed edges represent reticulate evolutionary events. By specifying a known outgroup, the rooted topology can be recovered from such networks. We introduce the algorithm Squirrel (Semi-directed Quarnet-based I nference to Reconstruct Level-1 Networks) which constructs a semi-directed level-1 network from a full set of quarnets (four-leaf semi-directed networks). Our method also includes a heuristic to construct such a quarnet set directly from sequence alignments. To build a network from quarnets, Squirrel first builds a tree, after which it repeatedly solves the Travelling Salesman Problem (TSP) to replace each high-degree vertex by a cycle. We demonstrate Squirrel's performance on randomly generated networks and on real sequence data sets, the largest of which contains 29 aligned sequences close to 1.7 Mpb long. The resulting networks are obtained on a standard laptop within a few minutes. Lastly, we prove that Squirrel is

combinatorially consistent: given a full set of quarnets coming from a triangle-free semi-directed level-1 network, it is guaranteed to reconstruct the original network. Squirrel is implemented in Python, has an easy-to-use graphical user-interface that takes sequence alignments or quarnets as input, and is freely available at https://github.com/nholtgrefe/squirrel.

Tree metrics, tree likelihood, and effective sample size
Bradley Jones, Simon Fraser University

Tree metrics can be used to compare the similarity between two tree topologies. Many different tree metrics have been devised, some based on tree topology directly, others based on applying tree moves and some based on mapping trees to vectors. Besides being able to compare two tree topologies, what other properties do tree metrics have? Do trees at low distances have similar likelihoods? Can tree variation be used to estimate convergence in Markov chain Monte Carlo (MCMC)? We investigated 11 tree metrics. We compared distances between pairs of trees to the differences in their likelihoods in samples of trees generated with MCMC. We also used Lanfear's approximate and pseudo effective sample size (ESS), with different tree metrics, on real data and on toy examples where the true ESS could be determined. We found that closely related trees have similar likelihoods at the beginning of an MCMC chain for most metrics, but this relationship deteriorates further along the chain. Approximate and pseudo ESS estimates were quite reliable for toy data sets, but there was considerable variability when estimating ESS of real data. While each tree metrics has its strengths, there is still the potential for new useful tree metrics to be developed.

reconcILS: A gene tree-species tree reconciliation algorithm that allows for incomplete lineage sorting
Sarthak Mishra, Indiana University Bloomington

Reconciliation algorithms provide an accounting of the evolutionary history of individual gene trees given a species tree. Many reconciliation algorithms consider only duplication and loss events (and sometimes horizontal transfer), ignoring effects of the coalescent process, including incomplete lineage sorting (ILS). Here, we present a new heuristic algorithm for carrying out reconciliation that accurately accounts for ILS by treating it as a series of nearest neighbor interchange (NNI) events. For discordant branches of the gene tree identified by last common ancestor (LCA) mapping, our algorithm recursively chooses the optimal history by comparing the cost of duplication and loss to the cost of NNI and loss. We demonstrate the accuracy of our new method, which we call reconcILS, using a new simulation engine (dupcoal) that can accurately generate gene trees produced by the interaction of duplication, loss, and ILS. Despite being a heuristic method, we show that reconcILS is much more accurate than models that ignore ILS, and at least as accurate or better than leading methods that can model ILS, while also being able to handle much larger datasets. We demonstrate the use of reconcILS by applying it to a dataset of 23 primate genomes, highlighting its accuracy compared to standard methods in the presence of large amounts of ILS.

A novel pipeline for calling transposable elements presence/absence in repetitive regions from whole genome alignments

Rachel Parsons, University of Maryland College Park

Retrotransposons are a type of transposable element (TE) that "jump" around the genome through a copy-and-paste mechanism. An estimated 30-50% of vertebrate genomes are derived from retrotransposons, making them a major mechanism of genome expansion and innovation. Moreover, they are conjectured to be robust evolutionary markers because their mutational process is widely thought to be nearly homoplasy free. This assumption enables TE presence/absence at orthologous loci to be used in conjunction with species tree estimation approaches that account for incomplete lineage sorting (ILS). These prior studies have been limited by the use of short reads and/or genome assembly quality. However, emerging telomere-to-telomere (T2T) projects are producing high quality assemblies of TE rich regions within a variety of species, along with whole genome alignments.
Here, we present a new pipeline for calling TE presence/absence in repetitive regions from whole genome alignments (WGAs). To our knowledge, our pipeline is the first approach that does not require a reference genome, thus minimizing reference-bias and broadening comparisons across non-human primates. Additionally, our method can be applied to both standard WGAs (e.g., those produced by CACTUS) as well as implicit WGAs (e.g., those represented by collections of pairwise alignments). The application of our method to T2T primates consortium data demonstrates the ability of our approach to find examples of ILS and highlights potential challenges calling TE presence/absence across more divergent species and with lower quality assemblies.

Explicit Modular Decomposition

Guillaume Scholz, Universität Leipzig

Cographs are among the best-studied graph classes. Among other characterizations, they are precisely those graphs G that can be represented by a unique rooted tree (T,t), called cotree, whose leaf set is V(G) and whose non-leaf vertices v are assigned a label t(v) in {0,1}.

Recent advances in mathematical biology have shown that cographs are intimately linked to pairwise relationships between genes. For example, the orthology relation collects all pairs of genes whose last common ancestor in their evolutionary history was a speciation event (or, equivalently, has label 1) and thus forms a cograph. In many applications, however, graphs G obtained from biological data usually violate the property of being a cograph and thus cannot be represented by a tree.

This poster presents the novel concept of Explicit Modular Decomposition, that aims at representing graphs and other apparented objects with suitable {0,1}-labelled rooted structures.

This is joint work with Marc Hellmuth and Anna Lindeberg.

A Diffusion-Based Approach for Simulating Forward-in-Time State-Dependent Speciation and Extinction Dynamics

Albert Soewongsono, Washington University in St. Louis

Although diffusion approximations are widely used in population genetics, they remain underused for modelling phylogenetic diversification dynamics. We bridge this gap by establishing a general diffusion-based framework to study a wide and highly influential class of phylogenetics birth-death models, known as cladogenetic state-dependent-speciation-extinction (ClaSSE) models. We validate our diffusion-based framework is reliable under a variety of diversification scenarios. We also derive relationships between model rates and their stationary state frequencies. In summary, our work helps to formalize relationships between evolutionary state patterns, process rates, and mixing times for ClaSSE-type models.

On the effects of selection and mutation on species tree inference

Matthew Wascher, Case Western Reserve University

The effect of selection acting on regions of the genome on the accuracy of species-level phylogenetic inference using methods that do not explicitly model selection is an open question that is relevant to most, if not all, phylogenomic studies. To address this, we derive a mathematical approximation to the Wright-Fisher model with mutation and selection in the limit as the population size becomes large. In contrast to previous approximations based on diffusion processes, our approximation can be used to study the distribution of coalescent times for an arbitrary number of lineages, allowing calculation of the probability distribution of gene genealogies under the coalescent model. We use these calculations to show that direct selection at strengths typically encountered in practice has only a small effect on the distribution of coalescent times, and hence on the distribution of gene trees. This implies that many coalescent-based methods for estimating the species tree topology will be robust to the presence of selection in a subset of the underlying genes. Selection will, however, bias the estimation of speciation times, causing them to underestimate the true speciation times. Our model captures the effects of selection on the genealogies that generate the observed sequence data, but does not model selective pressures that act only on the subsequent sequences or that negatively impact gene tree estimation.