

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: February 01, 2025 at 3:30 PM

Speaker Information

Navid Bin Hasan (Std No. 0422052035) is a part time M.Sc. student in the department of CSE, BUET. He completed his undergraduate studies from Bangladesh University of Engineering and Technology (BUET) in 2022. His research interest lies in the field(s) of Bioinformatics, Machine Learning, and Software Engineering. He is currently doing his postgraduate thesis under the supervision of Dr. Md. Shamsuzzoha Bayzid. He will be speaking about his ongoing research in this talk.



Leveraging and correcting weighted quartet distributions for enhanced species tree inference from genome-wide data

Species tree estimation from genes sampled from throughout the whole genome is challenging in the presence of gene tree discordance, often caused by incomplete lineage sorting (ILS), where alleles can coexist in populations for periods that may span several speciation events. Quartet-based summary methods for estimating species trees from a collection of gene trees are becoming popular due to their high accuracy and theoretical guarantees of robustness to arbitrarily high amounts of ILS. ASTRAL, the most widely used quartet-based method, aims to infer species trees by maximizing the number of quartets in the gene trees that are consistent with the species tree. An alternative approach (as in wQFM) is to infer quartets for all subsets of four species and amalgamate them into a coherent species tree. While summary methods can be highly sensitive to gene tree estimation errors (GTEE)--especially when gene trees are derived from short alignments--quartet amalgamation offers an advantage by potentially bypassing the need for gene tree estimation. However, greatly understudied is the choice of weighted quartet inference method and downstream effects on species tree estimations under realistic model conditions.

In this study, we investigated a broad range of methods for generating weighted quartets and critically assessed their impact on species tree inference. Our results on a collection of simulated and empirical datasets suggest that amalgamating quartets weighted based on gene tree frequencies (GTF) typically produces more accurate trees than leading quartet-based methods like ASTRAL and SVDquartets. Further enhancements in GTF-based weighted quartet estimation were achieved by accounting for gene tree uncertainty through the utilization of a distribution of trees for each gene (instead of a single tree) by employing traditional nonparametric bootstrapping methods or Bayesian MCMC sampling. Our study provides evidence that the careful generation and amalgamation of weighted quartets, as implemented in methods like wQFM, can lead to significantly more accurate trees compared to widely employed methods like ASTRAL, especially in the face of gene tree estimation errors.

On the other hand, gene tree estimation error (GTEE) arising from a combination of reasons (ranging from analytical factors to more biological causes, as in short gene sequences) can potentially impact the accuracy of phylogenomic inference. We, for the first time, introduce the problem of correcting the quartet distribution induced by a set of estimated gene trees, which involves updating the weights of the quartets to

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: February 01, 2025 at 3:30 PM

better reflect their relative importance within the gene tree distribution. We present QT-WEAVER, the first method of its kind, which learns the conflicts within the quartet distribution induced by a given set of gene trees and generates an updated quartet distribution by adjusting the weights accordingly. QT-WEAVER is a general-purpose technique needing no explicit modeling of the subject system or reasons for GTEE or gene tree heterogeneity. Experimental studies on a collection of simulated and empirical data sets suggest that QT-WEAVER can effectively account for GTEE, which results in a substantial improvement in the species tree accuracy. Additionally, the concept of quartet conflicts and related algorithmic and combinatorial innovations introduced in this study will benefit various quartet-based computations. Therefore, QT-WEAVER advances the state-of-the-art in species tree estimation from gene trees in the face of GTEE.