PhD Dissertation Presentation

November 29, 2017

3:00 pm - 5:00 pm

3105 Engineering Building

CSE Conference Room

SCALABLE PHYLOGENETIC ANALYSIS AND FUNCTIONAL INTERPRETATION OF GENOMES WITH COMPLEX EVOLUTIONARY HISTORIES

By

Hussein El Abbass Hejase

Advisor: Dr. Kevin Liu

Phylogenomics involves the inference of a genome-scale phylogeny. A phylogeny is typically inferred using sequences from loci of multiple organisms by reconstructing gene trees and then reconciling them into a species phylogeny. Many studies have shown that evolutionary processes such as gene flow, incomplete lineage sorting, recombination, selection, gene duplication and loss, have shaped our genomes and played a major role in the evolution of a diverse array of metazoans, including humans and ancient hominins, mice, bacteria, and butterflies. The aforementioned evolutionary processes are primary causes of gene tree discordance, which introduce different loci in a genome that exhibit local genealogical variation.

In this dissertation, we develop a method for fast and accurate inference of phylogenetic networks using large-scale sequence data. The advent of high-throughput sequencing technologies has brought about two main scalability challenges: (1) dataset size in terms of the number of taxa and (2) the evolutionary divergence of the taxa in a study. We explore the impact of both dimensions of scale on phylogenetic network inference and then introduce a new phylogenetic divide-and-conquer method which we call FastNet. We show using synthetic and empirical data spanning a range of evolutionary scenarios
that FastNet outperforms the state-of-the-art in terms of accuracy and computational requirements.

Furthermore, we develop methods that use better and more accurate phylogenies to functionally interpret genomes. One way to study and understand the biological function of genomes is through association mapping, which pinpoints statistical associations between genotypic and phenotypic characters while modeling the relatedness between samples. Many methods have been proposed to perform association mapping while accounting for sample relatedness. However, the state of the art predominantly utilizes the simplifying assumption that sample relatedness is effectively fixed across the genome. Recent studies have shown that sample relatedness can vary greatly across different loci within a genome where gene trees could differ from each other and the species phylogeny. Thus, there is an imminent need for methods to model this phenomenon, which is referred to as local genealogical variation. We address this methodological gap by introducing two methods, Coal-Map and Coal-Miner, which account for sample relatedness locally within loci and globally across the entire genome. We show through simulated and empirical datasets that these newly introduced methods offer comparable or typically better statistical power and type I error control compared to the state-of-the-art.

**Publications:**


