Temple University
Department of Biology

-Final Doctoral Thesis Defense-

TITLE

"Big Data Phylogenomics: Methods and Applications"

Sudip Sharma

TIME AND PLACE-

Monday, August 07, 2023
2:00 PM in room 604 located in SERC – Science Technology, Research Center
& also via Zoom
Any questions, please contact the Biology Department @ 215-204-8854

Dissertation Committee
Dr. Sudhir Kumar, Advisory Chair, Department of Biology, Temple University
Dr. S. Blair Hedges, Examining Committee Chair, Department of Biology, Temple University
Dr. Sergei Pond, Examining Committee Member, Department of Biology, Temple University
Dr. Xinghua Shi, External Member, Department of Computer and Information Science, Temple University

Title: Big Data Phylogenomics: Methods and Applications

Abstract: Phylogenomics, the study of genome-scale data containing many genes and species, has advanced our understanding of patterns of evolutionary relationships and processes throughout the Tree of Life. Recent research studies frequently use such large-scale datasets with the expectation of recovering historical species relationships with high statistical confidence. At the same time, the computational complexity and resource requirements for analyzing such large-scale data increase with the number of genes and sites. Therefore, different crucial steps of phylogenomic studies, like model selection and estimating bootstrap confidence limits on inferred phylogenetic trees, are often not feasible on regular desktop computers and generally time-consuming on high-performance computing systems. Moreover, increasing the number of genes in the data increases the chance of including genes that may cause biased and fragile species relationships that spuriously receive high statistical support. Such data errors in phylogenomic datasets are major impediments to building a robust tree of life. Contemporary approaches to detect such data error require alternative tree hypotheses for the fragile clades, which may be unavailable a priori or too numerous to evaluate. In addition, finding causal genomic loci under these contemporary statistical frameworks is also computationally expensive and increases with the number of alternatives to be compared. In my Ph.D. dissertation, I have pursued three major research projects: (1) Introduction and advancement of the bag of little bootstraps approach for placing the confidence limits on species relationships from genome-scale phylogenetic trees. (2) Development of a novel site-subsampling approach to select the best-fit substitution model for genome-scale phylogenomic datasets. Both of these approaches analyze data subsamples containing a small fraction of sites from the full phylogenomic alignment. Before analysis, sites in a subsample are repeatedly chosen randomly to build a new alignment that contains as many sites as the original dataset, which is shown to retain the statistical properties of the full dataset. Analyses of simulated and empirical datasets exhibited that these approaches are fast and require a minuscule amount of computer memory while retaining similar accuracy as that achieved by full dataset analysis. (3) Development of a supervised machine learning approach based on the Evolutionary Sparse Learning framework for detecting fragile clades and associated gene-species combinations. This approach first builds a genetic model for a monophyletic clade of interest, clade probability for the clade, and gene-species concordance scores. The clade model and these novel matrices expose fragile clades and highly influential as well as disruptive gene-species candidates underlying the fragile clades. The efficiency and usefulness of this approach are demonstrated by analyzing a set of simulated and empirical datasets and comparing their performance with the state-of-the-art approaches. Furthermore, I have actively contributed to research projects exploring applications of these newly developed approaches to a variety of research projects.