PhD Thesis title: Couplet based Analysis of Phylogenetic Trees

Candidate: Sourya Bhattacharyya
PhD Student, Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Supervisor: Prof. Jayanta Mukhopadhyay

Abstract:

A *phylogenetic tree* represents the evolutionary relationships among a group of taxa. For a particular set of taxa, biomolecular sequences (such as DNA, protein, etc.) of a particular type of gene are sampled from them, and are used to construct a phylogenetic tree (also referred to as the *gene tree*). For different genes sampled from a group of taxa, different gene trees thus can be generated. These gene trees often exhibit conflicting topologies due to independent evolutionary histories of respective genes, differences in gene sequences, variation in the phylogenetic tree construction methods employed, etc. In addition, biological processes like gene duplication, loss, incomplete lineage sorting, etc., contribute to such conflicts in the gene tree topologies. So, a *species tree* which depicts the true evolutionary relationships of the concerned taxa set, needs to be constructed from these conflicting gene tree topologies. Current thesis focuses on two different approaches for species tree construction.

The first approach, termed as *supertree* construction, quests for resolving the topological conflicts among gene trees with the consensus topologies of individual subtrees. Such problem is shown to be NP hard, and various heuristics have been proposed in the existing studies. The most common and popular technique is based on the subtree decomposition. Here, input trees are first decomposed into fixed size subsets like triplets (set of three taxa) or quartets (set of four taxa), which are then synthesized to form a supertree. Time and space complexities of these methods are, however, at least of the order of the subtree size employed. The current thesis proposes two novel supertree construction methods which employ the evolutionary relationships among individual taxa pairs (couplets). These methods are shown to exhibit much lower time and space complexities, and higher performances, with respect to the existing methods.

The second approach considers modeling *incomplete lineage sorting* (ILS) as the reason of discordance between the gene trees. ILS occurs due to the rapid speciation and short branches in respective gene trees, resulting failure of two or more lineages in a population to coalesce. During ILS, consensus gene tree topologies do not agree with the true species tree topology. So, supertree methods are not applicable for resolving the ILS based topological conflicts. Existing studies employing statistical analysis (such as maximum likelihood, or Bayesian techniques) are computationally intensive, and are not applicable for a large number of taxa or gene trees. Summary based methods involving quartet topologies, or divide and conquer based statistical approaches, also incur huge computation. A few of the existing methods employ couplet based topological analysis for species tree inference. Such couplet based methods exhibit the lowest computational complexity. Current thesis proposes two novel approaches which extend these couplet based methods, for species tree inference. These new methods are shown to exhibit improved performances from the existing couplet based approaches.

Overall, the thesis presents four methods for species tree construction. All of these techniques involve couplet based evolutionary analysis, either by following the consensus logic (supertrees), or by modeling ILS, to resolve the topological incongruence among gene trees to generate the final species tree.