Probabilistic Models for Species Tree Inference and Orthology Analysis

IKRAM ULLAH

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 12th June 2015, at 1:00 p.m. in Conference room Air, SciLifeLab, Tomtebodavägen 23A, Solna.
Abstract

A phylogenetic tree is used to model gene evolution and species evolution using molecular sequence data. For artifactual and biological reasons, a gene tree may differ from a species tree, a phenomenon known as gene tree-species tree incongruence. Assuming the presence of one or more evolutionary events, e.g., gene duplication, gene loss, and lateral gene transfer (LGT), the incongruence may be explained using a reconciliation of a gene tree inside a species tree. Such information has biological utilities, e.g., inference of orthologous relationship between genes.

In this thesis, we present probabilistic models and methods for orthology analysis and species tree inference, while accounting for evolutionary factors such as gene duplication, gene loss, and sequence evolution. Furthermore, we use a probabilistic LGT-aware model for inferring gene trees having temporal information for duplication and LGT events.

In the first project, we present a Bayesian method, called DLRSOrthology, for estimating orthology probabilities using the DLRS model: a probabilistic model integrating gene evolution, a relaxed molecular clock for substitution rates, and sequence evolution. We devise a dynamic programming algorithm for efficiently summing orthology probabilities over all reconciliations of a gene tree inside a species tree. Furthermore, we present heuristics based on receiver operating characteristics (ROC) curve to estimate suitable thresholds for deciding orthology events. Our method, as demonstrated by synthetic and biological results, outperforms existing probabilistic approaches in accuracy and is robust to incomplete taxon sampling artifacts.

In the second project, we present a probabilistic method, based on a mixture model, for species tree inference. The method employs a two-phase approach, where in the first phase, a structural expectation maximization algorithm, based on a mixture model, is used to reconstruct a maximum likelihood set of candidate species trees. In the second phase, in order to select the best species tree, each of the candidate species tree is evaluated using PrIME-DLRS: a method based on the DLRS model. The method is accurate, efficient, and scalable when compared to a recent probabilistic species tree inference method called PHYLDOG. We observe that, in most cases, the analysis constituted only by the first phase may also be used for selecting the target species tree, yielding a fast and accurate method for larger datasets.

Finally, we devise a probabilistic method based on the DLTRS model: an extension of the DLRS model to include LGT events, for sampling reconciliations of a gene tree inside a species tree. The method enables us to estimate gene trees having temporal information for duplication and LGT events. To the best of our knowledge, this is the first probabilistic method that takes gene sequence data directly into account for sampling reconciliations that...
contains information about LGT events. Based on the synthetic data analysis, we believe that the method has the potential to identify LGT highways.

**Keywords**

phylogenetics, phylogenomics, gene tree, species tree, expectation maximization, mixture model, dynamic programming, markov chain monte carlo, PrIME, JPrIME