Using Trees to Capture Reticulate Evolution

Lateral Gene Transfers and Cancer Progression

ALI TOFIGH

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the on Friday the 12th June 2009, at 10:00 a.m. in Svedbergssalen, Albanova, Roslagstullsbacken 21, Stockholm.
Abstract

The historic relationship of species and genes are traditionally depicted using trees. However, not all evolutionary histories are adequately captured by bifurcating processes and an increasing amount of research is devoted towards using networks or network-like structures to capture evolutionary history. Lateral gene transfer (LGT) is a previously controversial mechanism responsible for non tree-like evolutionary histories, and is today accepted as a major force of evolution, particularly in the prokaryotic domain.

In this thesis, we present models of gene evolution incorporating both LGTs and duplications, together with efficient computational methods for various inference problems. Specifically, we define a biologically sound combinatorial model for reconciliation of species and gene trees that facilitates simultaneous consideration of duplications and LGTs. We prove that finding most parsimonious reconciliations is NP-hard, but that the problem can be solved efficiently if reconciliations are not required to be acyclic—a condition that is satisfied when analyzing most real-world datasets. We also provide a polynomial-time algorithm for parametric tree reconciliation, a problem analogous to parametric sequence alignment, that enables us to study the entire space of optimal reconciliations under all possible cost schemes.

Going beyond combinatorial models, we define the first probabilistic model of gene evolution incorporating a birth-death process generating duplications, LGTs, and losses, together with a relaxed molecular clock model of sequence evolution. Algorithms based on Markov chain Monte Carlo (MCMC) techniques, methods from numerical analysis, and dynamic programming are presented for various probability and parameter inference problems.

Finally, we develop methods for analysis of cancer progression, a biological process with many similarities to the process of evolution. Cancer progresses by accumulation of harmful genetic aberrations whose patterns of emergence are graph-like. We develop a model of cancer progression based on trees, and mixtures thereof, that admits an efficient structural EM algorithm for finding Maximum Likelihood (ML) solutions from available cross-sectional data.

Keywords
Lateral Gene Transfer, Horizontal Gene Transfer, Cancer Progression