UNIVERSITY OF SYDNEY

SCHOOL OF MATHEMATICS AND STATISTICS

Statistics Seminar

Friday, 14 September 2007, 2.00pm

Carslaw 373

General Markov Models for Nucleotide Sequence Evolution

Vivek Jayaswal
(University of Sydney)

Abstract

The aim of many molecular phylogenetics studies is to infer the most probable sequence of gene evolution. The order in which the genes evolve gives rise to a branching pattern referred to as the tree topology. In addition to the tree topology, phylogeneticists are interested in estimating the rate of evolution along the individual branches and these are modeled as Markov processes. Under the assumption that the nucleotide sites within a gene are independent and identically distributed (iid), the most general model is the one proposed by Barry and Hartigan (Stat Sci., 1987:191-210). Since the iid assumption is often violated by real data sets, we generalize the Barry and Hartigan model by relaxing the assumption of identical distribution. We achieve this by allowing a site to be either variable or invariant (BH+I model) and by allowing the variable sites to evolve at k different rates (BHk+I model). We use the maximum-likelihood method to estimate the parameters for the new models and apply these models to real and simulated data sets. We show that these models satisfy the constraint of internal consistency; a necessary condition for analyzing evolutionary trees where the last common ancestor is unknown. We use the BH+I model to analyze a bacterial data set where most of the existing models (including those that allow non-identical distribution of sites) fail due to lack of stationarity and homogeneity. We use parametric bootstrap to (a) show that the data are consistent with the BH+I model and (b) determine the tree topology that best explains the observed data. Finally, we briefly discuss the $BH_k + I$ model.

Enquiries about the Statistics Seminar should be directed to
Rafał Kulik (rkuli@maths.usyd.edu.au)