Ancestral inference in molecular biology

Simon Tavaré

Departments of Biological Sciences, Mathematics and Biostatistics, USC

Introduction: This series of lectures describes several statistical problems involving ancestral inference from molecular data. The examples illustrate branching on widely diverse time scales: (a) inferring phylogenetic trees, and reconciling them with fossil estimates of divergence times (b) estimating the age of a mutation from a within-population sample of chromosomes and (c) reconstruction of tumor history within an individual. The common theme is the development of novel computational inference methods for branching processes, in particular for non-standard sampling schemes. Open problems will be highlighted during the lectures.

Lecture outlines:

Lecture 1. Introduction to basic molecular biology. DNA sequence data, mitochondrial DNA and microsatellites. Overview of ancestral inference using molecular data and illustrative examples.

Lecture 2. Mutation models for DNA sequence data. Species trees, gene trees and phylogeny. The basic computations for deterministic trees. A statistical look at rate variation.

Lecture 3. Inference in population genetics: the coalescent. Statistical inference in population genetics: uses of the coalescent. Markov chain Monte Carlo (MCMC) methods.

Lecture 4. The age of a unique event polymorphism.

Lecture 5. The time since loss of mismatch repair in tumors.

Lecture 6. Molecular and fossil estimates of primate divergence times: a reconciliation?

References: The papers below give more details of topics to be discussed in the lectures. Papers marked with a †can be downloaded from the Villars button at `http://www-hto.usc.edu/papers/abstracts/lists/tavare.html` as either postscript or pdf files. Other references will be provided during the lectures.

An overview:

†Tavaré S (1999) Random trees in molecular genetics. *Bull. I.S.I.*, **52**, 269-272.

Coalescent theory:

Hudson RR (1991) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology,* ed. D Futuyma, J Antonovics, **7**, 1-44. Oxford University Press.

Griffiths RC, Tavaré S (1994) Simulating probability distributions in the coalescent. *Theor. Popn. Biol.*, **46**, 131-159.

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. *Statistical Science*, **9**, 307-319.

Donnelly P and Tavaré S (1995) Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.*, **29**, 401-421.

Kuhner MK, Yamato J and Felsenstein J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421-1430.

Kuhner MK, Yamato J and Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429-434.

†Markovtsova L, Marjoram P and Tavaré S (1999) The age of a unique event polymorphism. *Genetics*, submitted.

†Markovtsova L, Marjoram P and Tavaré S (2000) The effects of rate variation on ancestral inference in the coalescent. *Genetics*, submitted.

Tumor history:

†Tsao J-L, Tavaré S, Salovaara R, Jass JR, Altonen LA and Shibata D (1999) Colorectal adenoma and cancer divergence: evidence of multi-lineage progression. *Am. J. Pathol.*, **154**, 1815-1824.

†Tsao J-L, Yatabe Y, Salovaara R, Järvinen HJ, Mecklin J-P, Altonen LA, Tavaré S and Shibata D (2000) Genetic reconstruction of individual colorectal tumor histories. *Proc. Natl. Acad. Sci. USA*, **97**, 1236-1241.

Species trees:

Felsenstein J (1981) Evolutionary trees from DNA sequence data: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368-376.

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *TREE*, **9**, 367–372.

Hillis DM, Moritz C and Mable BK (1996) *Molecular Systematics*, Second Edition, Sinauer Associates, Inc.

Mau R, Newton MA and Larget B (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1-12.