

Advanced Models of Substitutions

Objective: To give a presentation of about 60 minutes at the end of the week covering the key aspects of the advanced models of substitutions.

This project is devoted to advanced models of substitution. Advanced will in this context mean models that cannot be described by a simple rate matrix on a single position or will use advanced techniques to analyze question arising from simple models. Example of advanced models could include: context-dependent models, models involving several nucleotides at the same time, non-markovian models, models with a large state space (codons for instance), models of signals (k-mers, where k could be 8-12), lumpability in Markov Models, models described by complex equilibrium distributions, how to approximate an empirically derived rate matrix by a rate matrix obeying some combinatorial constraints.

The Big Questions Are:

What are the key characteristics of the true process of evolution as known from empirical studies of genomes from different phyla?

What are the key classes of simple models and their history?

Can realistic sequence models be interpreted in terms of biochemical events?

Maximal Contents of Presentation

Principles of Sequence Models (continuous time markov chain theory)

Key simple Models

Advanced Models and their use

Context Dependent Models

Multinucleotide Models

Non-Markovian Models

Lumping States

Complex equilibrium distributions

Incorporation of selection in models

Inference in these models

Their Application in data analysis

Ancestral Analysis

Biochemical Interpretations of events and models

What remains to be done?

Starting pointers

http://en.wikipedia.org/wiki/Substitution_model

Choi, S.C., Redelings, B.D., and Thorne, J.L. (2008) Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. *Phil. Trans. R. Soc. B*.

Evans and Speed (1993) Invariants of some probability models used in phylogenetic inference *Ann. Statist.* 21, 355-377

Felsenstein, J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach *J Mol Evol.* 1981;17(6):368-76.

Hobolth, A. and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in Genetics and Molecular Biology*, 4, 18

Hobolth, A. and Stone, A.C. (2009). Efficient simulation from finite-state, continuous-time Markov chains with incomplete observations. *Annals of Applied Statistics*

Hobolth, A. (2008). A Markov Chain Monte Carlo Expectation Maximization algorithm for statistical analysis of DNA sequence evolution with neighbour-dependent substitution rates. *Journal of Computational and Graphical Statistics*, 17, 138-164

Goldman, N. (1993) Simple diagnostic statistical tests of models for DNA substitution. *Journal of Molecular Evolution* 37:650-661.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* 11:725-736.

Goldman, N. and Whelan, S. (2002) A novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution* 19: 1821-1831.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120

Kirby, Muse, Stephan W. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A.* 1995 Sep 26;92(20):9047-9051

Hobolth, A. (2008). A Markov Chain Monte Carlo Expectation Maximization algorithm for statistical analysis of DNA sequence evolution with neighbour-dependent substitution rates. *Journal of Computational and Graphical Statistics*, 17, 138-164

Hobolth and Stone (2009) EFFICIENT SIMULATION FROM FINITE-STATE, CONTINUOUS-TIME MARKOV CHAINS WITH INCOMPLETE OBSERVATIONS

Hobolth, A. and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in Genetics and Molecular Biology*, 4, 18

O'Brien, Minin, and Suchard (2009) Learning to Count: Robust Estimates for Labelled Distances between Molecular Sequences *Mol.Biol.Evol.* 26(4):801-814.

Minin and Suchard (2008) Fast, accurate and simulation-free stochastic mapping 3995-363 *Phil. Trans. R. Soc. B*

Minin and Suchard (2008) Counting labeled transitions in continuous-time Markov models of evolution *J. Math. Biol.* 56:391-412

Jukes, TH and Cantor, CR. 1969. Evolution of protein molecules. Pp. 21-123 in H. N. Munro, ed. *Mammalian protein metabolism*. Academic Press, New York

Pedersen and Jensen A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames *Molecular Biology and Evolution* 18:763-776 (2001)

Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17, 57-86, 1986. American Mathematical Society. Thorne

Yang (2006) *Computational Molecular Evolution* OUP

Yap (2009) Similar States in Continuous Time Markov Chains *J.Appl. Prob.* 46:497-506.