Models and estimation for phylogenetic trees

Faisal Ababneh* & John Robinson

University of Sydney

The DNA consists of a sequences of bases A,C,G, and T, we will label the four bases by 1,2,3 and 4 respectively. Let $(X_l(t))$ denote the bases that occurs at the l^{th} position $(1 \leq l \leq n)$ at time t. The substitution process $\{X(t), t \geq 0\}$ can be described by the transition matrix of a Markov process $P_{ij}^X(t) = p(X(t) = j|X(0) = i)$. Consider the case when we have two sequences, suppose that we have two independent sequences of nucleotide having the same base length n, and derived from a common ancestor by independent mutation for each site, each described by a Markov process; that is $(X(t), Y(t))$ are two independent Markov process coming from a common ancestor $X(0) = Y(0)$. Define $f_{ij}(t) = p(X(t) = i, Y(t) = j|X(0) = Y(0))$, where f_{ij} is the probability that for a given site, the first and second sequences having the position i and j at that site. We can write F_t depending on the transition matrices as $F_t = P_X^T(t)F_0P_Y(t)$, where F_0 is the probability of the initial frequency for the bases. This can generalized to several sequences. The sequences (taxa) are then arranged in an evolutionary tree (or plylogenetic tree) depicting how taxa diverge from a common ancestor (root). The evolutionary tree is a directed graph showing the relationship between a group of taxa and their hypothetical common ancestors. The root of the tree is a common ancestor of all the taxa, the other nodes are either the contemporary taxa at the tips of the tree or speciation events (internal nodes) from which two new taxa bifurcate. The simplest plylogenetic tree consist of two taxa, the length of each leaf represent the evolutionary time t. Estimation in such models involves many parameters. Early attempts of estimation used extremely simplified models which did not fit the data. We propose tests of homogeneity in some parts of the tree, depending on the results of these tests we propose models appropriate for the data and for which, estimation of the parameters is feasible. We discuss the case of estimation of a five edge tree first on simulated data and we apply the method to real data to estimate the real parameters.

References

Cascual, O. (1994). A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances *Mol Biol Evol*, **11**, 961-963.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, **22**, 240-240.

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

Felsenstein, J. (1996) Inferring Phylogenies from Protein Sequence by Parsimony Distance, and Likelihood Methods. *Methods in Enzymology*, **24**, 418-427.

Goldman, N. (1990). Maximum likelihood inference of pylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analysis . *Syst. Zool.*, **39**, 345-361.

Kimura, M. (1980). A simple method for estimating evolutionary rate in a finite population due to mutational production of neutral and nearly neutral base substitution through comparative studies of nucleotide sequences. *J. Molec. Biol.*, **16**, 111-120.

Lake, J.A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA.* **91**, 1455-1459.

Sokal, R.R. and Michener, C.D. (1958). A Statistical Method for Evaluating Systematic Relationship. *University of Kansas Scientific Bulletin*, **28**, 1409-1438.

Waterman, M.S., Smith, T.F., Singh, M. and W.A.Beyer (1977). Additive evolutionary tree, *Journal Theoretical Biol.*, **64**, 199-213.